UNIVERSIDADE FEDERAL DE MINAS GERAIS

INSTITUTO DE CIÊNCIAS BIOLÓGICAS

LABORATÓRIO DE GENÉTICA CELULAR E MOLECULAR

PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

Tese de doutorado

Ômicas: *Corynebacterium pseudotuberculosis o nosso cavalo de batalha.*

BELO HORIZONTE

2017

Vasco Ariston de Carvalho Azevedo

Ômicas: *Corynebacterium pseudotuberculosis o nosso cavalo de batalha.*

Tese apresentada ao programa de Pós-graduação interunidades em Bioinformática da Universidade Federal de Minas Gerais

Orientador: Prof. Dr. José Miguel Ortega

BELO HORIZONTE

2017

Eu dedico este trabalho `a meu Pai, minha mãe, irmãos, esposa, filhos, sobrinhos e a meus amigos verdadeiros.

ÍNDICE

AGRADECIMENTOS

“Quem tem amigo verdadeiro pode dizer que tem duas almas.”
Arturo Graf

Aos meus pais, por me darem os instrumentos para sonhar,

Aos meus amigos, colegas e alunos, que colaboraram em transformá-los em realidade,

A minha esposa e aos meus filhos, por me manterem sonhando.

RESUMO DO PROJETO

O principal objetivo do projeto é dar continuidade às pesquisas desenvolvidas pelo Professor Vasco Ariston de Carvalho Azevedo na área de genética de microrganismos, as quais visam desenvolver ferramentas para erradicar a doença denominada Linfadenite Caseosa (LC). Tal patologia acomete principalmente caprinos e ovinos, causando grandes perdas econômicas na caprinovinocultura. Nesse contexto, faz-se necessário um estudo genético aprofundado e amplo do microrganismo *Corynebacterium pseudotuberculosis* para um melhor entendimento de suas bases moleculares, principalmente aquelas relacionadas ao desenvolvimento da doença no hospedeiro. O trabalho, que já está sendo realizado no Laboratório de Genética Celular e Molecular (LGCM) do Departamento de Biologia Geral do Instituto de Ciências Biológicas (ICB), da Universidade Federal de Minas Gerais, tem contado com a participação de aproximadamente 30 estudantes e pesquisadores que empregam técnicas de Biologia Molecular e Imunologia Aplicada em estudos de genômica e proteômica de *C. pseudotuberculosis*.

Durante o período de execução deste projeto, os seguintes estudos envolvendo *C. pseudotuberculosis* têm sido priorizados, quais sejam: (i) seqüenciamento do genoma do microrganismo; (ii) caracterização do genoma de várias linhagens de *C. pseudotuberculosis* por meio de análises *in silico* de *Genome Survey Sequences* (GSS) para abordagem pangenômia e de epidemiologia molecular; (iii) avaliação do papel dos fatores sigma alternativos na regulação de genes de virulência de *C. pseudotuberculosis*; (iv) técnicas de microarranjos de DNA para avaliação e estudo do microrganismo; (v) análises de plasticidade genômica; (vi) mapa proteômico da fração de proteínas secretadas; (vii) obtenção de mutantes através de um sistema de transposição TnFuZ, e posterior teste para o desenvolvimento de estratégias alternativas de vacinação; (viii) isolamento, clonagem e a caracterização molecular do gene *hsp60* do microrganismo e avaliação do seu potencial como antígeno vacinal com vistas ao desenvolvimento de vacinas de subunidade protéica e de DNA; (xix)

desenvolvimento de vacinas por meio de vacinologia reversa; (x) desenvolvimento de testes diagnósticos moleculares por PCR; (xi) estudos de soroprevalência da enfermidade causada pelo microrganismo; (xii) caracterização de antígenos do microrganismo para o diagnóstico da linfadenite caseosa subclínica em ovinos e caprinos.

A continuidade do trabalho contribuirá para o conhecimento acadêmico-científico nas áreas de Genômica e Proteômica, com o objetivo de fornecer dados e informações que poderão ajudar a elucidar os mecanismos moleculares e as bases genéticas da virulência deste microrganismo. Nesse sentido, novas perspectivas para o desenvolvimento de vacinas e técnicas diagnósticas para o controle e erradicação da LC em rebanhos caprinos e ovinos poderão ser geradas.

Palavras-chave: *Corynebacterium pseudotuberculosis*, genômica, proteômica, vacinas, diagnóstico, linfadenite caseosa.

I. INTRODUÇÃO GERAL

“Para ser grande, sê inteiro: nada
Teu exagera ou exclui.
Sê todo em cada coisa. Põe quanto és
No mínimo que fazes.
Assim em cada lago a lua toda
Brilha, por que alta vive.”
Fernando Pessoa.

I.I Introdução e Referencial Teórico

Estima-se que o rebanho mundial de caprinos e ovinos é de aproximadamente 900 milhões de cabeças. O Brasil possui cerca de 23 milhões de animais, sendo 37% de caprinos e 63% de ovinos. Dos caprinos, 1,4% encontram-se na Região Norte, 93% no Nordeste, 2,4% no Sudeste, 1,9% no Sul e 1% no Centro-Oeste. Com relação ao rebanho ovino, 2,8% encontram-se na Região Norte, 49% no Nordeste, 2,8% no Sudeste, 4% no Sul e 4,9% no Centro-Oeste (EMBRAPA Semi-Árido). A exportação de peles ovinas acumulada no período de 1992 a 1999 foi de US$ 87,1 milhões e, no período de 2000, esse número foi de US$ 7,1 milhões. Já a exportação de peles caprinas registradas no período de 1992 a 1999 foi de US$ 25,9 milhões e, em 2000, representou cerca de US$ 0,3 milhões. Com a crescente procura por subprodutos da ovinocaprinocultura, há um número maior de empresários dispostos a investir nessas atividades, o que resulta em uma busca maior por tecnificação nas duas culturas e por medidas de saúde preventiva mais eficazes e de pronto acesso (EMBRAPA Semi-Árido). A linfadenite caseosa (LC) é uma doença crônica causada pela bactéria *Corynebacterium pseudotuberculosis*, que acomete principalmente pequenos ruminantes e é responsável por perdas econômicas significativas relacionadas à redução da produtividade e eficiência reprodutiva dos animais infectados (Dorella *et al*. 2006). Essa enfermidade é prevalente em países como Austrália, Nova Zelândia, África do Sul, Estados Unidos e Brasil, onde a atividade de ovino e caprinocultura é intensa (Williamson, 2001; Arsenault *et al.*, 2003; Paton *et al.*, 2003). No Brasil, estimativas demonstram que a maior parte dos animais está infectada e que a prevalência clínica esteja em torno de 30% (Ribeiro *et al.*, 2001). Os estados da Região Nordeste são considerados os mais afetados, porém o estado de Minas Gerais (Região Sudeste), mesmo possuindo um rebanho relativamente reduzido, tem apresentado relatos de problemas como conseqüência dessa patologia, a qual foi registrada por 84,3% dos produtores rurais (Faria *et al.*, 2004). Estudos recentes realizados em aproximadamente 200 propriedades do estado de Minas Gerais reveleram uma soroprevalência de LC de 70% em ovinos e 80% em caprinos, o que corrobora a importância da linha de pesquisa envolvendo o agente causador dessa enfermidade.

A transmissão da LC entre caprinos e ovinos ocorre, principalmente, por meio de ferimentos superficiais na pele, os quais podem ser causados tanto por procedimentos de manejo como tosquia, castração, tratamento do cordão umbilical e agulhas contaminadas quanto por fatores naturais como pequenos acidentes com arbustos e/ou objetos pontiagudos (Alves *et al*., 1997).

A doença desenvolve-se lentamente, sendo a LC externa a forma mais freqüente, a qual é caracterizada pela formação de abscessos em nódulos linfáticos superficiais e em tecidos subcutâneos. Esses abscessos podem se desenvolver em órgãos internos, como pulmões, rins, fígado e baço, caracterizando a LC visceral (Piontkowski e Shivvers, 1998). Em alguns casos, a infecção produz sinais clínicos pouco evidentes no animal, o que leva à impossibilidade de identificá-los até seu abate ou morte, dificultando, assim, não só o controle, mas também a obtenção de dados sobre a prevalência dessa doença (Paton *et al.*, 1994, Arsenault *et al.*, 2003). O tratamento da enfermidade não é eficiente por meio de uso de antibióticos, pois estes não penetram na cápsula dos abscessos. Dessa forma, o controle da LC deve ser realizado com base em medidas profiláticas com a identificação dos animais infectados, impedindo que os mesmos tenham contato com os animais saudáveis (Williamson, 2001, Dorella *et al*., 2006).

O presente trabalho de tese tem como principal objetivo apresentar os principais trabalhos feitos com *C. pseudotuberculosis* pelo doutorando usando estra tégias ômicas com a finalidade de compreender as bases genéticas e mecanismos moleculares envolvidos na patofisiologia da enfermidade e, dessa forma, desenvolver técnicas diagnósticas e medidas profiláticas mais eficazes contra patógeno.

II CAPÍTULO

II.I GENÔMICA

II.I.1 Progression of 'OMICS' methodologies for understanding the pathogenicity of *Corynebacterium pseudotuberculosis*: the Brazilian experience.

Dorella FA, Gala-Garcia A, Pinto AC, Sarrouh B, Antunes CA, Ribeiro D, Aburjaile FF, Fiaux KK, Guimarães LC, Seyffert N, El-Aouar RA, Silva R, Hassan SS, Castro TL, Marques WS, Ramos R, Carneiro A, de Sá P, Miyoshi A, **Azevedo V**, Silva A.
Comput Struct Biotechnol J. 2013 Oct 13;6: e201303013. doi: 10.5936/csbj.201303013. eCollection 2013. Review.

Este é um artigo de revisão que aborda as principais metodologias das ciências “ômicas” utilizadas até 2013 para a compreensão de estudos a respeito da virulência e patogenicidade de *Corynebacterium pseudotuberculosis*. Relata desde o início dos primeiros projetos de sequenciamento desta bactéria, quando se alcançou os primeiros dados genômicos, até a geração de dados mais complexos, como dados transcriptômicos e proteômicos. O objetivo principal desta revisão foi demonstrar a importância da utilização de ferramentas e integração dos dados *in silico* com os dados *in vitro* já disponíveis pelo nosso laboratório, para desvendar candidatos alvos que pudessem ser utilizados no desenvolvimento de uma vacina ou de um diagnóstico eficaz para todos os hospedeiros que são acometidos por *C. pseudotuberculosis*, erradicando assim as doenças.

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

Volume No: 6, Issue: 7, e201303013, http://dx.doi.org/10.5936/csbj.201303013

CSBJ

Progression of 'omics' methodologies for understanding the pathogenicity of *Corynebacterium pseudotuberculosis*: the Brazilian experience

Fernanda A. Dorella [a], Alfonso Gala-Garcia [a], Anne C. Pinto [a], Boutros Sarrouh [a], Camila A. Antunes [a], Dayana Ribeiro [a], Flavia F. Aburjaile [a], Karina K. Fiaux [a], Luis C. Guimarães [a], Núbia Seyffert [a], Rachid A. El-Aouar [a], Renata Silva [a], Syed S. Hassan [a], Thiago L. P. Castro [a], Wanderson S. Marques [a], Rommel Ramos [b], Adriana Carneiro [b], Pablo de Sá [b], Anderson Miyoshi [a], Vasco Azevedo [a,*], Artur Silva [b,*]

Abstract: Since the first successful attempt at sequencing the *Corynebacterium pseudotuberculosis* genome, large amounts of genomic, transcriptomic and proteomic data have been generated. *C. pseudotuberculosis* is an interesting bacterium due to its great zoonotic potential and because it causes considerable economic losses worldwide. Furthermore, different strains of *C. pseudotuberculosis* are capable of causing various diseases in different hosts. Currently, we seek information about the phylogenetic relationships between different strains of *C. pseudotuberculosis* isolates from different hosts across the world and to employ these data to develop tools to diagnose and eradicate the diseases these strains cause. In this review, we present the latest findings on *C. pseudotuberculosis* that have been obtained with the most advanced techniques for sequencing and genomic organization. We also discuss the development of *in silico* tools for processing these data to prompt a better understanding of this pathogen.

4TH Annual World DNA and Genome Day 2013

Background

Corynebacterium pseudotuberculosis is a Gram-positive facultative intracellular pathogen that belongs to the class *Actinobacteria*. This pathogen is aerobic, has mycolic acid in its cell wall, displays pleomorphic forms, does not sporulate or encapsulate, is non-motile and possesses fimbriae. *C. pseudotuberculosis* has great infectious potential and implications for zoonotic transmission, as it affects goats, sheep, horses, buffaloes, camels, cattle and primates, causing different symptoms. Furthermore, it has already been reported as the causative agent of more than 33 cases of infection in humans. *C. pseudotuberculosis* presents two biovars, ovis (nitrate negative reduction) and equi (nitrate positive reduction); the former biovar is mainly associated with the globally-distributed disease Caseous Lymph Adenitis (CLA), which affects the lymph nodes and visceral organs of goats and sheep and causes economic losses by compromising the skin, weight, milk and meat production of the animals as well as causing death and compromising the carcass. Although many vaccines exist, they are mainly intended for use in sheep and goats and provide variable levels of protection [1, 2, 3].

C. pseudotuberculosis has interesting survival mechanisms and is able to utilize different strategies to adapt to its environment. Once it is successfully established within a host and able to replicate inside phagocytic cells, this pathogen will evade the immune system with apparent ease. As a result, chronic infections may last for most, if not all, of an animal's life [3]. In the present review, we report the latest information regarding the genomic, proteomic and transcriptomic features of this interesting microorganism, and we attempt to correlate such information with its virulence and pathogenicity.

Genomics

The beginning of the C. pseudotuberculosis *genome projects*

The first attempt to identify the genomic sequence of *C. pseudotuberculosis* was performed by Dorella and collaborators [4], in which genomic libraries of the 1002 strain of this species were constructed using a bacterial artificial chromosome (BAC) vector. This high-quality genomic library, containing approximately 1,800 clones, harbored inserts ranging from 24.5-121 kbp. Partial characterization of this library through a BAC end-sequencing strategy, namely the identification of genome survey sequences (GSS), generated 215 GSS at relatively low cost; these were deposited on the NCBI website. Using these sequences for *in silico* analysis, it was possible to identify putative genes involved in virulence based on their similarity to other deposited sequences and generate a catalog of genes, such as the putative siderophore-binding protein (GSS number BH740428) that increased our biological knowledge of the microorganism. The high quality, low redundancy and absence of contaminants in the library, together with the large number of clones it contained, permitted this library to serve as a physical map for the characterization of the *C. pseudotuberculosis* genome. Moreover, library characterization also allowed for confirmation of the close phylogenetic relationship between *C. pseudotuberculosis* and *C. diphtheriae*, *C. glutamicum*, *C. efficiens* and *C. jeikeium* [4]. Based on this initiative, a project to sequence the first entire genome of *C. pseudotuberculosis* 1002 strain was started by the Rede Genoma de Minas Gerais (RGMG-Brazil) in 2006 and concluded in 2009. This genome was sequenced using the Sanger di-deoxy method and has now been assembled, annotated and deposited in the NCBI database under accession number CP001809.

[a] Laboratório de Genética Celular e Molecular, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Brazil
[b] Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém-PA, Brazil

* Corresponding author.
E-mail address: asilva@ufpa.br (or) vascoariston@gmail.com

Table 1. Strains of *Corynebacterium pseudotuberculosis* that were deposited in NCBI between 2009 and 2012 and their structural characteristics.

Strain	Biovar	Animal/ Host	Site of Isolation	Country of Isolation	Sequencing Technology	Chromosome	Size (Mb)	GC%	Genes	Proteins
1002	*ovis*	Goat	Abscess	Brazil, Bahia	454, Sanger	1	2.34	52.2	2,203	2,090
C231	*ovis*	Sheep	Abscess	Australia	454	1	2.33	52.2	2,204	2,091
162	*equi*	Camel	Abscess	UK	SOLiD v3	1	2.29	52	2,150	2,002
258	*equi*	Horse	Not specified	Belgium	SOLiD v3	1	2.31	52.1	2,195	2,088
CIP5297	*equi*	Horse	Not specified	Kenya	SOLiD v2	1	2.32	52.1	2,194	2,060
PAT10	*ovis*	Sheep	Abscess	Patagonia	SOLiD v2	1	2.34	52.2	2,200	2,079
I19	*ovis*	Bovine	Not specified	Israel	SOLiD v2	1	2.34	52.2	2,213	2,095
31	*equi*	Buffalo	Not specified	Egypt	Ion Torrent, SOLiD v3	1	2.34	52.2	2,170	2,063
FRC41	*ovis*	Human	Inguinal lymph node	France	454	1	2.34	52.2	2,171	2,110
267	*ovis*	Llama	Submandibular abscess	California	SOLiD v3	1	2.34	52.2	2,249	2,148
316	*equi*	Horse	Subcutaneous abscess	California	Ion Torrent	1	2.31	52.1	2,234	2,106
01/06	*equi*	Horse	Abscess	California	Illumina	1	2.28	52.2	2,127	1,963
3/99-5	*ovis*	Sheep	Abscess	Scotland	Illumina	1	2.34	52.2	2,239	2,142
42/02	*ovis*	Sheep	Abscess	Australia	Illumina	1	2.34	52.2	2,164	2,051
P54B96	*ovis*	Antelope	Liver, lung, mediastinal lymph node	South Africa	Ion Torrent, SOLiD v3	1	2.34	52.2	2,205	2,084

The challenge of next-generation sequencing

The *C. pseudotuberculosis* genome project has expanded its boundaries and today the network includes the Rede Paraense de Genômica e Proteômica, which has worked with all of the versions of SOLiD™ (Life Technologies) since v.2 and now employs the most advanced next-generation sequencing (NGS) platforms: the SOLiD™ 5500 series (Life Technologies) and Ion Torrent PGM (Life Technologies). These NGS platforms can sequence more than one bacterial genome per day, thus demonstrating the feasibility of sequencing *C. pseudotuberculosis* strains. This important partnership has also contributed computational resources to process the huge amount of data generated by these new DNA-sequencing technologies.

To date, fifteen strains of *C. pseudotuberculosis* have been sequenced (Table 1), employing all of the presently available technologies. Based on the data obtained by sequencing, the average G+C content of all 15 strains is 52.2%; each genome has an average of approximately 2,195 genes, and total genome sizes range from 2.28 to 2.34 Mb. The sequencing of several strains of *C. pseudotuberculosis* is paving the way for further studies. In 2011, Barh and colleagues compared four genomes of *C. pseudotuberculosis* (strains FRC41, 1002, C231 and I19) with eight other sequenced genomes of pathogens belonging to a group that includes genera such *Corynebacterium, Mycobacterium, Nocardia* and *Rhodococcus,* which are commonly found in humans, goats, sheep, cattle and horses [5]. As a result of this comparative genomic analysis, potential molecular targets were identified for the production of drugs and vaccines.

The study of the diversity among strains promotes our understanding of gene rearrangement, genomic plasticity as loss and gains and inversions in the genome. In addition, this research provides valuable information regarding molecular epidemiology, microevolution, lineage-specific genes and common genes among the isolates [6], contributing to the development of new therapies that are more effective for the control of caseous lymphadenitis (CLA).

Structural genome

Of the fifteen genomes deposited at NCBI, nine belong to the biovar ovis, and six belong to the biovar equi (Table 1). While the ovis strains have almost no genetic differences, the grouping of the equi strains appears to be asymmetric in relation to biovar ovis. Therefore, it is important to detect the differences between *ovis* and *equi* to develop a common vaccine or diagnostic tool for all of them. Typically, vaccines against *C. pseudotuberculosis* infection designed for sheep do not have equal efficacy in goats, although both species are usually infected by bacteria belonging to the biovar *ovis*. Thus, the vaccines developed for *C. pseudotuberculosis* biovar *ovis* may not have the same efficacy in hosts infected with biovar *equi,* which further complicates treatment of *C. pseudotuberculosis* by different animal breeders [1].

Interestingly, no major differences between the structural characteristics of biovars *equi* and *ovis* have been observed, such as the numbers of CDS, genes or proteins, which are very similar between strains of both biovars (Table 1). The differential pathogenicities of the biovars might be due to the presence of genes that are strain-specific, as each pathogen appears to preferentially infect particular hosts, therefore causing different disease symptoms. Thus, specific genes and other unknown process may underlie host preference and determine the different symptoms of the infection process [1].

Features that are common among all of the strains are GC content and the number of ribosomal clusters. GC content is related to different intrinsic or extrinsic factors, and a high GC content suggests that the genetic material has greater stability, providing a more robust genome that suffers less from the influence of environmental variations [7].

With regard to the number of rDNA operons, all strains present four copies, and each ribosome consists of one 5S, one 16S and one 23S. This fact may possibly be related to the slower replication of *C. pseudotuberculosis* compared to *Escherichia coli*, which has seven copies of the rDNA operon, or *C. glutamicum*, which has six copies, considering that ribosomal operons can perform diverse functions related to the control of protein synthesis [8].

Software and databases for Corynebacterium pseudotuberculosis genome analysis

A rapid increase in the number of complete genomes over the past few decades in the form of large molecular datasets in public databases has provoked researchers to develop numerous computational tools and public or proprietary databases. These holistic approaches have facilitated the rapid study and understanding of the innumerable biological functions that are encoded by genomic DNA. The barrier to unraveling prokaryotic genomes has been eliminated using the next generation of high-throughput sequencing technologies, such as SOLiD, GS FLX, Ion Torrent PGM and Illumina, which have prominent advantages over Sanger sequencing. However, although these technologies significantly reduce the cost and time for genome sequencing, they still pose challenges for various aspects of data processing and analysis, such as the assembly of short reads [9]. A number of user-friendly interfaces and stand-alone computational tools have been developed to evaluate the genomic and transcriptomic data obtained from these high-throughput platforms.

Presently, bioinformaticians have developed and are further revising some useful tools and software packages using different algorithms and in-house scripts. A brief description and application of each software program for the data analysis of *C. pseudotuberculosis* and/or taxonomically related organisms is presented below.

1- Pathogenicity Island-Prediction Software (PIPS):

This software is designed to predict the pathogenicity islands (PAIs) in bacterial genomes, utilizing multiple features in an integrative manner. PAIs are large genomic regions acquired through horizontal gene transfer, which have in common the following: deviations in G+C content and codon usage, the presence of transposase and virulence factors, flanking insertion sequences and/or tRNA genes and their absence in non-pathogenic organisms of the same genus or related species. PIPS uses these multiple features to detect PAIs. For validation purposes, PIPS was utilized with model organisms of the genera *Corynebacterium* and *Escherichia*, and the results showed that PIPS provided better accuracy (85-88%) and superior efficiency compared with the other available software tools. This software is easy to install on a personal computer and provides a user-friendly interface for students and researchers [10].

2- Quality Assessment Software (QA):

This software is used to analyze the quality of sequence reads from next-generation platforms. The software removes the reads, which present average quality below the Phred quality cutoffs. The process of quality filtering reduces miss-assemblies and incorrect mapping against the reference genome that are attributable to low quality sequences from the raw data. The software helps to review graphs that show the distribution of quality values from the sequencing reads, including the average and the accumulated quality for each base. Libraries of fragments from SOLiD sequencing of *C. pseudotuberculosis* (Cp162) and *Exiguobacterium antarcticum* (B7) were used as sample data to test the software. QA is a Java-based program that is available at http://qualevaluato.sourceforge.net [11]. A new version of this software, called Quality Assessment Long Reads [12], was developed to apply the Phred quality filter over Ion Torrent PGM data due to the read length: ≈120 bp for the first release of the platform and ≈400 bp with a recent protocol.

3- Singular Value Decomposition (SVD):

This is a very useful technique for information retrieval that helps to uncover the relationships between elements that are not prima facie related. In turn, this leads to the improved inference of evolutionary relationships between amino acid sequences of different species. SVD produces a revised distance matrix for a set of related elements and provides results resembling the internationally accepted scientific gold standard of Linnaean taxonomy. The SVD-based computations establish non-obvious, relevant relationships among the clustered elements, providing a deterministic method for grouping related species. This approach was initially developed to reduce the time needed for information retrieval and analysis of very large-scale genome and proteome data sets in the complex Internet environment. The results obtained by this technique are in close approximation with results based on Linnaean taxonomy, which indicates that SVD can indicate evolutionary relationships of species and construct better quality clusters and phylogenetic trees [13].

The analysis of prokaryotic genomes can be further aided with new algorithmic methods and tools and advancements in bioinformatics and computational biology. These techniques will provide more opportunities to study in detail the "OMICS" of specific organisms. Unifying current and upcoming computational resources to provide a global and integral picture of biology is important and can be achieved by mutual cooperation among researchers from distinct areas.

4- Core StImulon (CSI)

One methodology for performing RNA sequencing (RNA-seq) analyses is the *de novo* approach, which is commonly used when reference genomes are not available in biological databases. An important feature of this method is that it identifies shared transcripts among stimulons (i.e., the set of expressed genes under a given condition), which can permit the selection of possible candidates for vaccine studies through searches for the specific genes of an organism in addition to permitting the identification of new transcripts that have not been previously annotated. We sequenced the cDNA of *Corynebacterium pseudotuberculosis* strain 1002 using the SOLiD V3 system under the following conditions: osmotic stress (2 M), acidity (low pH), heat shock (50°C) and a control condition. To identify the transcripts that were shared among the stimulons and integrate this information with the BLAST and BLAST2GO results, the software CoreStImulon (CSI) was developed, which allows genes to be characterized in terms of their ontology [14].

5- FunSys

FunSys software, which is a stand-alone tool with a user-friendly interface, was developed to evaluate and correlate the differential expression profiles from RNA-seq and proteomics datasets. FunSys produces charts and reports based on the results of the analysis of differential expression (generated using other software) to aid in the interpretation of the results [15].

Proteomics

Proteomics of C. pseudotuberculosis

Unlike genomic studies, proteomics evaluates the protein profile of a cell, tissue or organism [16, 17]. Proteomic studies can provide valuable information about changes in protein synthesis, post-translational modifications and protein-protein interactions, thereby increasing knowledge of physiological phenomena for a specific condition and helping to establish a fundamental understanding of an organism's cellular physiology and virulence factors [16, 18]. The global expression of bacterial proteins is required for growth, survival or pathogenicity, and cataloguing these proteins in response to a determined condition is a key step toward understanding the physiology of these microorganisms [18, 19]. To identify virulence factors and obtain further information about the biology of pathogenic bacteria, studies have been performed using proteomics to characterize whole cells, cytoplasmic and membrane proteomes and the secretome/exoproteome of these pathogens [18].

The primary proteomic studies involving *C. pseudotuberculosis* were intended to analyze the extracellular protein fraction. This protein fraction is associated with the uptake of nutrients, cell-to-cell communication, proteolysis, hemolysis, detoxification, escape from the immune system and destruction of competing microorganisms in their respective environments. However, during the process of adaptation and survival in hostile environments, pathogenic bacteria need to secrete different molecules for adhesion, invasion, proliferation and survival in the host cell [20, 21]. Thus, the study of extracellular proteins is a useful strategy to identify new virulence factors and target immunogenics [22].

Initially, with the aim of identifying new targets for the development of immunodiagnostics and vaccine targets to combat CLA, various research groups conducted proteomic studies using one-dimensional electrophoresis based on sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) and immunoblotting to characterize the whole cell fraction and extracellular proteins of *C. pseudotuberculosis*. The bacteria in these studies were grown in complex media containing exogenous proteins that would contaminate extractions of extracellular proteins [23, 24]. Studies showed that the use of chemically defined medium (CDM) is an effective strategy to identify bacterial components for therapeutic applications [25]. In this context, a CDM for *C. pseudotuberculosis* growth in macromolecule-free conditions was developed [26]. The evaluation of humoral and cellular immune responses of goats experimentally infected with *C. pseudotuberculosis* showed that interferon-γ (IFN- γ) detection using excreted-secreted antigen after cultivation of this pathogen in CDM provided more specific results compared with the use of whole cell sonicated antigen [27]. This suggested that the bacterial growth in CMD and use of the secreted protein fraction may be an interesting strategy for the study of immunogenic proteins of *C. pseudotuberculosis*.

To optimize the process of obtaining the extracellular fraction, Paule and colleagues [27] established an efficient protocol for extracting the extracellular proteins of *C. pseudotuberculosis* based on the three-phase partitioning (TPP) technique. After analyzing the protein extract by SDS-PAGE and immunoblotting, it was possible to detect proteins that were not detected in previous studies [28]. Notably, all of the results obtained by Paule and colleagues [28] only indicated the molecular weights of the proteins or reactivity of the proteins against the sera of infected animals without protein characterization by mass spectrometry (MS).

The *C. pseudotuberculosis* genome project [29] generated information about the pathogenicity and virulence of this microorganism. From the genomic data, the *in silico* pan-exoproteome of *C. pseudotuberculosis* has been deduced [30]. However, how these gene products interact and what their functions are in physiological processes must be elucidated. To respond to these questions and validate gene annotations, proteomic approaches have been applied to characterize the exoproteome of *C. pseudotuberculosis*.

Comparative proteomics

Studies have demonstrated that comparative proteomics is a powerful strategy to characterize bacterial proteomes, and thus it has been adopted to characterize the proteomes of various pathogenic bacteria [21, 31, 32]. A comparative proteomic study was conducted using the "shotgun proteomics" approach to characterize the exoproteome of two strains, *Cp1002* and *CpC231*, of *C. pseudotuberculosis*, both of which belong to the biovar *ovis* but were isolated from different hosts (goat and sheep, respectively). This study combined the techniques of TPP [27] and gel-free separation using liquid chromatography coupled with mass spectrometry (LC-MS), called TPP-LC/MS^E [33]. The two strains were maintained on BHI agar or in broth and in CDM to study proteome growth. The results obtained from this work showed quantitative and qualitative changes between the exoproteomes of both strains. Furthermore, this strategy permitted the characterization of 93 extracellular proteins of *C. pseudotuberculosis* that were associated with the physiology and virulence of this pathogen [33]. The identified proteins that play a role in virulence include phospholipase D (PLD), the main virulence factor of *C. pseudotuberculosis*, which is associated with the spread of bacteria within the host [34]; iron siderophore binding protein (FagD), a component of an iron uptake system [35]; and serine proteinase (CP40), which showed protective activity against infection by *C. pseudotuberculosis* [36]. However, these proteins were identified only in the extracellular proteome of *CpC231*, suggesting that these proteins may not be secreted by *Cp1002*, which may influence the pathogenesis of this strain [33].

Another approach that has been employed to analyze the exoproteome of *C. pseudotuberculosis* is serological proteome analysis (SERPA), which involves 2-DE immunoblotting and identification of antigenic spots by an MS technique. This strategy has been applied to several pathogenic bacterial species to identify virulence factors, target the development of drugs and vaccines and conduct immunodiagnostics [37, 38]. In this context, Seyffert et al. [39] conducted a preliminary serological secretome analysis of *C. pseudotuberculosis* and evaluated the exoproteome of strain 1002 *ovis*. The use of the SERPA approach enabled the characterization of six immunoreactive proteins against the serum of animals infected with *C. pseudotuberculosis*. These identified proteins represent potential targets for developing vaccine targets and diagnostics to combat CLA.

Currently, with advances in proteomic studies, new techniques have been developed and applied for the study of several pathogens.

Thus, the application of different proteomic approaches is a powerful strategy to characterize the proteome of *C. pseudotuberculosis* and broaden our knowledge of the physiology and pathogenesis of this pathogen.

Transcriptomics

The mechanisms with which pathogenic microorganisms surpass the hostile conditions found in a host are of great importance for successful infection, and the genes related to such adaptations constitute clear targets for the development of new diagnostics and vaccines. The advent of RNA microarrays and high-throughput RNA-seq technologies has allowed not only the comprehensive assessment of differential gene expression in bacteria but also the identification of genetic structures such as operons, transcriptional start sites, non-coding regulatory RNAs and small RNAs [40].

Similar to *M. tuberculosis*, *C. pseudotuberculosis* infects and persists inside macrophages, although it does not prevent fusion between the phagosome and lysosome. Because this bacterium is subjected to different stresses in the phagolysosome, Pinto and colleagues [14] evaluated its transcriptome following *in vitro* exposure to high osmolarity (sodium chloride at a final concentration of 2 M), heat shock (50°C) or acidic pH (5.0) by performing RNA-seq with SOLiD technology. When the sets of genes expressed only under each stress condition, which together compose the core stimulon of *C. pseudotuberculosis*, were examined, most of the targets identified were related to oxidation and reduction events, while cell division and the cell cycle were the second- and third-most upregulated processes, respectively. According to the Gene Ontology database, some of the genes in the core stimulon are directly involved in stress responses; one example involves an encoder of a two-component system response-regulator protein that is also linked to pathogenesis. Other genes highlighted by the authors include *dps* (a gene involved in resistance to oxidative stress) and a gene that encodes for one component of the ABC-type iron-uptake system.

The assessment of global transcriptional profiles in bacteria constitutes a key strategy for unveiling mechanisms that are important for virulence and pathogenicity; thus, any efforts to increase the feasibility of RNA-seq experiments are welcome when studying pathogens such as *C. pseudotuberculosis*. Because large portions of sequencing reads are mapped to ribosomal RNA genes, Castro and colleagues [41] tested a new methodology, based on denaturing high-performance liquid chromatography, to deplete ribosomal transcripts from bacterial total RNA samples using *C. pseudotuberculosis* (biovar *equi*) as a model organism. With the elimination of 78% to 92% of rRNA, which are levels that resemble those obtained with a conventional subtraction kit, this new method offers financial advantages for researchers who have access to a chromatographic system.

The elucidation of which gene products of *C. pseudotuberculosis* are directly involved in survival and adaptation during infection has yet to come. As global gene expression profiling will most likely provide key knowledge for the development of effective prophylactic measures in the future, researchers will certainly take a step ahead by integrating information from both transcriptomic and proteomic approaches.

Future of this field

Studies of prokaryotic genomes, transcriptomes and proteomes have been considerably improved with the development of new experimental methods, algorithms and tools and advances in bioinformatics and computational biology.

The main objective of these studies is to find clues that may be useful in developing a vaccine and a diagnostic approach that is effective for all hosts that suffer from *C. pseudotuberculosis* infection. Another goal is to elucidate the physiology, pathogenicity and virulence mechanisms of this bacterium.

In response to advances in molecular biology in the last few years, much information regarding biological systems has been elucidated via a variety of genome-sequencing projects. However, sequencing reveals little about how the proteins of an organism operate individually or together to perform their functions. The integration of both current and upcoming resources to provide a global and integral biological picture is important and can be achieved by mutual cooperation between researchers from distinct areas.

Acknowledgements

This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Financiadora de Estudos e Projetos (FINEP), Universidade Federal do Pará, Universidade Federal de Minas Gerais, Fundação Amazônia Paraense (FAPESPA) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

Citation

Dorella FA, Gala-Garcia A, Pinto AC, Sarrouh B, Antunes CA, Ribeiro D, Aburjaile FF, Fiaux KK, Guimarães LC, Seyffert N, El-Aouar RA, Silva R, Hassan SS, Castro TLP, Marques WS, Ramos R, Carneiro A, Sá PD, Miyoshi A, Azevedo V, Silva A (2013) Progression of 'OMICS' methodologies for understanding the pathogenicity of *Corynebacterium pseudotuberculosis*: the Brazilian experience. Computational and Structural Biotechnology Journal. 6 (7): e201303013. doi: http://dx.doi.org/10.5936/csbj.201303013

References

1. Dorella, F.A., Pacheco, L. G., Oliveira, S. C., Miyoshi, A., Azevedo, V. (2006) *Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. Vet Res, 37, 201-218.
2. Bastos, B.L., Dias Portela, R.W., Dorella, F.A., Ribeiro, D., Seyffert, N., et al. (2012) *Corynebacterium pseudotuberculosis*: Immunological Responses in Animal Models and Zoonotic Potential. J Clin Cell Immunol S4:005. doi:10.4172/2155-9899.S4-005.
3. Dorella, F.A., Pacheco, L.G.C, Seyffert, N., Portela, R.W., Miyoshi, A., Azevedo, V. (2009) Antigens of Corynebacterium pseudotuberculosis ad prospects for vaccine development. Exp Rev Vaccines, 8, 205-213.
4. Dorella, F.A., Fachin, M.S., Billault, A., Dias Neto, E., Soravito, C., Oliveira, S.C., et al. (2006) Construction and partial characterization of a *Corynebacterium pseudotuberculosis* bacterial artificial chromosome library through genomic survey sequencing. Genet Mol Res, 5,653-63.
5. Barh D, Jain N, Tiwari S, Parida BP, D'Afonseca V, Li L, Ali A, Santos AR, Guimarães LC, de Castro Soares S, Miyoshi A, Bhattacharjee A, Misra AN, Silva A, Kumar A, Azevedo V. (2011) A novel comparative genomics analysis for common drug and vaccine

targets in *Corynebacterium pseudotuberculosis* and other CMN group of human pathogens. Chem Biol Drug Des,78, 73-84.

6. Muzzi, A., Donati, C. (2011) Population genetics and evolution of the pan-genome of *Streptococcus pneumoniae.* Int J Med Microbiol, 301,619-22.
7. Wu, H., Zhang, Z., Hu, S., and Yu, J. (2012) On the molecular mechanism of GC content variation among eubacterial genomes. Biology Direct, 7, 2.
8. Martin, J.F., Barreiro, C., González-Lavado, E., and Barriuso, M. (2003) Ribosomal RNA and ribosomal proteins in corynebacteria. J Biotechnol ,104, 41-53.
9. Cerdeira, L.T., Carneiro, A.R., Ramos, R.T.J., Almeida, S.S., D'Afonseca, V., Schneider. M.P.C., Baumbach, J., Tauch, A., McCulloch, J.A., Azevedo, V., Silva, A. (2011) Rapid hybrid *de novo* assembly of a microbial genome using only short reads: *Corynebacterium pseudotuberculosis* I19 as a case study. J Microbiol Methods, 86, 218–223.
10. Soares, S.C., Abreu, V.A., Ramos, R.T., Cerdeira, L., Silva, A., Baumbach, J., et al. (2012) PIPS: pathogenicity island prediction software, *PloS one* 7: e30848.
11. Ramos, R.T., Carneiro, A.R., Baumbach, J., Azevedo, V., Schneider, M.P., and Silva, A. (2011) Analysis of quality raw data of second generation sequencers with Quality Assessment Software. *BMC Res Notes* 4: 130.
12. Ramos, R.T.J., Carneiro, A.R., Soares, S.C., Santos, A.R., Almeida, S.,Guimarães, L., Figueira, F., Barbosa, E., Tauch, A., Azevedo, V., Silva, A. (2013) Tips and tricks for the assembly of a *Corynebacterium pseudotuberculosis* genome using a semiconductor sequencer. Microbial BiotechnologySpecial Issue: The Corynebacterium Cell Factory, 6, 150–156.
13. Santos, A.R., Santos, M.A, Baumbach, J., McCulloch, J.A., Oliveira, G.C, Silva, A., et al. (2011) A singular value decomposition approach for improved taxonomic classification of biological sequences. BMC Genomics, 12(Suppl 4):S11.
14. Pinto, A.C., Ramos, R.T., Silva, W.M., Rocha, F.S., Barbosa, S., Miyoshi, A., et al. (2012) The core stimulon of Corynebacterium pseudotuberculosis strain 1002 identified using ab initio methodologies. Integr Biol, 4, 789-794.
15. de Sa, P., Pinto, A., Ramos, R.T., Coimbra, N., Barauna, R., Dall'agnol, H., et al. (2012) FunSys: Software for functional analysis of prokaryotic transcriptome and proteome. Bioinformation, 8, 529-531.
16. Wu, H.J., Wang, A.H., and Jennings, M.P. (2008) Discovery of virulence factors of pathogenic bacteria. Curr Opin Chem Biol, 12, 93-101.
17. Parkash, O., and Singh, B.P. (2012) Advances in Proteomics of *Mycobacterium leprae.* Scand J Immunol, 75, 369-378.
18. Curreem, S.O., Watt, R.M., Lau, S.K., and Woo, P.C. (2012) Two-dimensional gel electrophoresis in bacterial proteomics. Protein Cell, 3, 346-63.
19. Osman, K.M., Ali, M.M., Radwan, M.I., Kim, H.K., and Han, J. (2009) Comparative proteomic analysis on *Salmonella Gallinarum* and *Salmonella Enteritidis* exploring proteins that may incorporate host adaptation in poultry. J Proteomics, 21:815-21.
20. Hueck, C.J. (1998) Type III protein secretion systems in bacterial pathogens of animals and plants. Microb Mol Biol Rev, 62, 379–433.
21. Trost, M., Wehmhöner, D., Kärs, U., Dieterich, G., Wehland, J., and Jänsch, L. (2005) Comparative proteome analysis of secretory proteins from pathogenic and nonpathogenic *Listeria* species. Proteomics, 5, 1544-1557.
22. Sibbald, M.J.J.B., Ziebandt, A.K., Engelmann, S., Jong, A., Harmsen, H.J.M., Raangs, G.C., et al. (2006) Mapping the pathways to staphylococcal pathogenesis by comparative secretomics. Microb Mol Biol Rev, 70,755–788.
23. Muckle, C.A., Menzies, P.I., Li, Y., Hwang, Y.T., and van Wesenbeeck, M. (1992) Analysis of the immunodominant antigens of *Corynebacterium pseudotuberculosis.* Vet Microbiol, 30, 47-58.
24. Braithwaite, C.E., Smith, E.E., Songer, J.G., and Reine, A.H. (1993) Characterization of detergent-soluble proteins of *Corynebacterium pseudotuberculosis.* Vet. Microbiol, 38, 59-70.
25. James, B.W., Williams, A., and Marsh, P.D. (2000) The physiology and pathogenicity of *Mycobacterium tuberculosis* grown under controlled conditions in a defined medium. J Appl Microbiol, 88, 669-677.
26. Moura-Costa, L.F., Paule, B.J.A., Freire, S.M., Nascimento, I., Schaer, R., Regis, L.F., et al. (2002) Meio sintético quimicamente definido para o cultivo de *Corynebacterium pseudotuberculosis.* Rev Bras Saúde Prod Na, 3, 1-9.
27. Paule, B.J.A.; Azevedo, V.; Regis, L.F.; Carminati, R.; Bahia, C.R.; Vale, V.L.C. et al. (2003) Experimental *Corynebacterium pseudotuberculosis* primary infection in goats: kinetics of IgG and interferon-g production, IgG avidity and antigen recognition by Western blotting. Vet Immunol Immunopathol, 96, 129–139.
28. Paule, B.J., Meyer, R., Moura-Costa, L.F., Bahia, R.C., Carminati, R., Regis, L.F., et al. (2004) Three-phase partitioning as an efficient method for extraction/concentration of immunoreactive excreted-secreted proteins *of Corynebacterium pseudotuberculosis.* Protein Expr Purif, 34, 311-166.
29. Ruiz, J.C., D'Afonseca, V., Silva, A., Ali, A., Pinto, A.C., Santos A.R., et al. (2011) Evidence for Reductive Genome Evolution and Lateral Acquisition of Virulence Functions in Two *Corynebacterium pseudotuberculosis* Strains. Plos One 6:e18551.
30. Santos, A.R,, Carneiro, A., Gala-García, A., Pinto, A., Barh, D., Barbosa, E., et al. (2012) The *Corynebacterium pseudotuberculosis* in silico predicted pan-exoproteome. BMC Genomics 13: Suppl 5:S6.
31. Sengupta, N., Alam, S.I., Kumar, B., Kumar, R.B., Gautam, V., Kumar, S., and Singh, L. (2010) Comparative proteomic analysis of extracellular proteins o*f Clostridium perfringens* Type A and Type C strains. Infect immun, 78, 3957–3968.
32. Muthukrishnan, G., Quinn, G.A., Lamers, R.P., Diaz, C., Cole, A.L., Chen, S., and Cole AM. (2011) Exoproteome of *Staphylococcus aureus* Reveals Putative Determinants of Nasal Carriage. J Proteome Res, 10, 2064-2078.
33. Pacheco, L.G., Slade, S.E., Seyffert, N., Santos, A.R., Castro, T.L., Silva, W.M., et al. (2011) A combined approach for comparative exoproteome analysis of *Corynebacterium pseudotuberculosis.* BMC Microbiology, 11, 12.
34. McKean, S.C., Davies, J.K., Moore, R.J. (2007) Expression of phospholipase D the major virulence factor of *Corynebacterium pseudotuberculosis,* is regulated by multiple environmental factors and plays a role in macrophage death. Microbiology. 153::2203-2211.
35. Billington, S.J., Esmay, P.A., Songer, J.G., Jost, B.H. (2002) Identification and role in virulence of putative iron acquisition genes from *Corynebacterium pseudotuberculosis.* FEMS Microbiol Lett. 2002 Feb 19;208(1):41-5.
36. Walker J, Jackson HJ, Eggleton DG, Meeusen EN, Wilson MJ, Brandon MR. (1994) Identification of a novel antigen from *Corynebacterium pseudotuberculosis* that protects sheep against caseous lymphadenitis. Infect Immun. 62:2562-2567.
37. Vytvytska, O., Nagy, E., Bluggel, M., Meyer, E.H., Kurzbauer, R., Huber, L.A., and Klade, C.S. (2002) Identification of vaccine

candidate antigens of *Staphylococcus aureus* by serological proteome analysis. Proteomics, 2, 580–590.

38. Cash, P. (2011) Investigating pathogen biology at the level of the proteome. Proteomics, 11, 3190-202.
39. Seyffert, N., Pacheco, L.G.C., Silva, W.M., Castro, L.P.C., Santos, A.V., Santos, A., et al. (2011) Preliminary serological secretome analysis of *Corynebacterium pseudotuberculosis*. Jiomics, 1, 193-197.
40. Sorek, R., and Cossart, P. (2010) Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. Nat Rev Genet, 11, 9-16.
41. Castro, T.L.P., Seyffert, N., Ramos, R.T., Barbosa, S., Carvalho, R.D., Pinto, A.C., Carneiro, A.R., Silva, W.M., Pacheco, L.G., Downson, C., Schneider, M.P., Miyoshi, A., Azevedo, V., Silva, A. (2013) Ion Torrent-based transcriptional assessment of a Corynebacterium pseudotuberculosis *equi* strain reveals denaturing high-performance liquid chromatography a promising rRNA depletion method. Microb Biotechnol, 6, 168-77.

Keywords:
Corynebacterium pseudotuberculosis, SOLiD next generation sequencing, Ion Torrent next generation sequencing, SDS-PAGE, mass spectrometry, RNA-seq

Competing Interests:
The authors have declared that no competing interests exist.

II.I.2 Evidence for reductive genome evolution and lateral acquisition of virulence functions in two *Corynebacterium pseudotuberculosis* strains.
Ruiz JC, D'Afonseca V, Silva A, Ali A, Pinto AC, Santos AR, Rocha AA, Lopes DO, Dorella FA, Pacheco LG, Costa MP, Turk MZ, Seyffert N, Moraes PM, Soares SC, Almeida SS, Castro TL, Abreu VA, Trost E, Baumbach J, Tauch A, Schneider MP, McCulloch J, Cerdeira LT, Ramos RT, Zerlotini A, Dominitini A, Resende DM, Coser EM, Oliveira LM, Pedrosa AL, Vieira CU, Guimarães CT, Bartholomeu DC, Oliveira DM, Santos FR, Rabelo ÉM, Lobo FP, Franco GR, Costa AF, Castro IM, Dias SR, Ferro JA, Ortega JM, Paiva LV, Goulart LR, Almeida JF, Ferro MI, Carneiro NP, Falcão PR, Grynberg P, Teixeira SM, Brommonschenkel S, Oliveira SC, Meyer R, Moore RJ, Miyoshi A, Oliveira GC, **Azevedo V**.
PLoS One. 2011 Apr 18;6(4): e18551. doi: 10.1371/journal.pone.0018551.

Montagem, anotação e análises computacionais dos dois primeiros genomas de *Corynebacterium pseudotuberculosis* realizados integralmente por nosso grupo de pesquisa. Esse trabalho criou as bases de conhecimento e competência técnico-científica para o projeto do pangenoma da *C.pseudotuberculosis* que hoje contabiliza mais de 30 genomas do gênero depositados nos últimos 10 anos e mais de 40 citações. Somam-se a essas conquistas outras dezenas de teses de mestrado e doutorado com professores empossados em diversas universidades no Brasil e no exterior.

Evidence for Reductive Genome Evolution and Lateral Acquisition of Virulence Functions in Two *Corynebacterium pseudotuberculosis* Strains

Jerônimo C. Ruiz[1]¶, Vívian D'Afonseca[2]¶, Artur Silva[3], Amjad Ali[2], Anne C. Pinto[2], Anderson R. Santos[2], Aryanne A. M. C. Rocha[2], Débora O. Lopes[4], Fernanda A. Dorella[2], Luis G. C. Pacheco[2,20], Marcília P. Costa[5], Meritxell Z. Turk[2], Núbia Seyffert[2], Pablo M. R. O. Moraes[2], Siomar C. Soares[2], Sintia S. Almeida[2], Thiago L. P. Castro[2], Vinicius A. C. Abreu[2], Eva Trost[6], Jan Baumbach[7], Andreas Tauch[6], Maria Paula C. Schneider[3], John McCulloch[3], Louise T. Cerdeira[3], Rommel T. J. Ramos[3], Adhemar Zerlotini[1], Anderson Dominitini[1], Daniela M. Resende[1,8], Elisângela M. Coser[1], Luciana M. Oliveira[9], André L. Pedrosa[8,10], Carlos U. Vieira[11], Cláudia T. Guimarães[12], Daniela C. Bartholomeu[13], Diana M. Oliveira[5], Fabrício R. Santos[2], Élida Mara Rabelo[14], Francisco P. Lobo[13], Glória R. Franco[13], Ana Flávia Costa[2], Ieso M. Castro[15], Sílvia Regina Costa Dias[14], Jesus A. Ferro[16], José Miguel Ortega[13], Luciano V. Paiva[17], Luiz R. Goulart[11], Juliana Franco Almeida[11], Maria Inês T. Ferro[16], Newton P. Carneiro[12], Paula R. K. Falcão[18], Priscila Grynberg[13], Santuza M. R. Teixeira[13], Sérgio Brommonschenkel[19], Sérgio C. Oliveira[13], Roberto Meyer[20], Robert J. Moore[21], Anderson Miyoshi[2], Guilherme C. Oliveira[1,22], Vasco Azevedo[2]*¶

1 Research Center René Rachou, Oswaldo Cruz Foundation, Belo Horizonte, Minas Gerais, Brazil, **2** Department of General Biology, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, **3** Department of Genetics, Federal University of Pará, Belém, Pará, Brazil, **4** Health Sciences Center, Federal University of São João Del Rei, Divinópilis, Minas Gerais, Brazil, **5** Department of Veterinary Medicine, State University of Ceará, Fortaleza, Ceará, Brazil, **6** Department of Genetics, University of Bielefeld, CeBiTech, Bielefeld, Nordrhein-Westfale, Germany, **7** Department of Computer Science, Max-Planck-Institut für Informatik, Saarbrücken, Saarlan, Germany, **8** Department of Pharmaceutical Sciences, Federal University of Ouro Preto, Ouro Preto, Minas Gerais, Brazil, **9** Department of Phisics, Federal University of Ouro Preto, Ouro Preto, Minas Gerais, Brazil, **10** Department of Biological Sciences, Federal University of Triangulo Mineiro, Uberaba, Minas Gerais, Brazil, **11** Department of Genetics and Biochemistry, Federal University of Uberlândia, Uberlândia, Minas Gerais, Brazil, **12** Brazilian Agricultural Research Corporation (EMBRAPA), Sete Lagoas, Minas Gerais, Brazil, **13** Department of Biochemistry and Immunology, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, **14** Department of Parasitology, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, **15** Department of Pharmacy, Federal University of Ouro Preto, Ouro Preto, Minas Gerais, Brazil, **16** Department of Technology, State University of São Paulo, Jaboticabal, São Paulo, Brazil, **17** Department of Chemistry, Federal University of Lavras, Lavras, Minas Gerais, Brazil, **18** Brazilian Agricultural Research Corporation (EMBRAPA), Campinas, São Paulo, Brazil, **19** Department of Plant Pathology, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil, **20** Department of Biointeraction Sciences, Federal University of Bahia, Salvador, Bahia, Brazil, **21** CSIRO Livestock Industries, Australia, **22** Center of Excellence in Bioinformatics, National Institute of Science and Technology, Research Center René Rachou, Oswaldo Cruz Foundation, Belo Horizonte, Minas Gerais, Brazil

Abstract

Background: *Corynebacterium pseudotuberculosis*, a Gram-positive, facultative intracellular pathogen, is the etiologic agent of the disease known as caseous lymphadenitis (CL). CL mainly affects small ruminants, such as goats and sheep; it also causes infections in humans, though rarely. This species is distributed worldwide, but it has the most serious economic impact in Oceania, Africa and South America. Although *C. pseudotuberculosis* causes major health and productivity problems for livestock, little is known about the molecular basis of its pathogenicity.

Methodology and Findings: We characterized two *C. pseudotuberculosis* genomes (Cp1002, isolated from goats; and CpC231, isolated from sheep). Analysis of the predicted genomes showed high similarity in genomic architecture, gene content and genetic order. When *C. pseudotuberculosis* was compared with other *Corynebacterium* species, it became evident that this pathogenic species has lost numerous genes, resulting in one of the smallest genomes in the genus. Other differences that could be part of the adaptation to pathogenicity include a lower GC content, of about 52%, and a reduced gene repertoire. The *C. pseudotuberculosis* genome also includes seven putative pathogenicity islands, which contain several classical virulence factors, including genes for fimbrial subunits, adhesion factors, iron uptake and secreted toxins. Additionally, all of the virulence factors in the islands have characteristics that indicate horizontal transfer.

Conclusions: These particular genome characteristics of *C. pseudotuberculosis*, as well as its acquired virulence factors in pathogenicity islands, provide evidence of its lifestyle and of the pathogenicity pathways used by this pathogen in the infection process. All genomes cited in this study are available in the NCBI Genbank database (http://www.ncbi.nlm.nih.gov/genbank/) under accession numbers CP001809 and CP001829.

Citation: Ruiz JC, D'Afonseca V, Silva A, Ali A, Pinto AC, et al. (2011) Evidence for Reductive Genome Evolution and Lateral Acquisition of Virulence Functions in Two *Corynebacterium pseudotuberculosis* Strains. PLoS ONE 6(4): e18551. doi:10.1371/journal.pone.0018551

Editor: Igor Mokrousov, St. Petersburg Pasteur Institute, Russian Federation

Received November 29, 2010; **Accepted** March 11, 2011; **Published** April 18, 2011

Funding: This project received financial support from the following grants: FAPEMIG (Fundação de Amparo à pesquisa do Estado de Minas Gerais) (CBB-1181/0 and REDE-186/08 to Guilherme Oliveira), NIH (National Institutes of Health) - Fogarty (TW007012 to Guilherme Oliveira). Guilherme Oliveira and Vasco Azevedo are CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) fellows. The work also received support from CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: vasco@icb.ufmg.br

◑ These authors contributed equally to this work.

Introduction

Corynebacterium pseudotuberculosis is a facultative intracellular pathogen that mainly infects sheep and goats, causing the disease called caseous lymphadenitis (CL). This bacterium can also cause ulcerative lymphangitis in equines; superficial abscesses in bovines, pigs, deer and laboratory animals; arthritis and bursitis in ovines; pectoral abscesses in equines and, more rarely, in camels, caprines and deer [1-3]. In both disease manifestations, its main characteristic is abscessing of the lymph nodes [4]. Rare cases of human infection have also been reported [5,6].

Despite the broad spectrum of hosts, the high incidence of CL reported from various countries, including Australia, New Zealand, South Africa, the United States of America, Canada and Brazil, mainly refers to small ruminants [7-11]. According to the World Animal Health Organization, among 201 countries that reported their sanitary situations, 64 declared the presence of animals with CL within their borders (OIE, 2009). The highest prevalence of CL has been reported in Brazil [12]. Pinheiro and colleagues (2000) reported 66.9% of animals with clinical signs of CL in the state of Ceará. In Minas Gerais state, a prevalence of 75.8% was reported for sheep [13] and 78.9% for goats [14]. In Australia, 61% of sheep flocks showed signs of infection [15]. In the USA, the prevalence ranges up to 43% [16]. Similar levels have been reported from the Canadian province of Quebec, with a prevalence of 21 to 36% [10]. In the United Kingdom, 45% of the producers that were polled reported abscesses in their sheep [9].

The high prevalence of CL in sheep and goats has made studies on ways to detect *C. pseudotuberculosis* in these hosts increasingly important; an efficient means to accomplish this would be a valuable tool for the control of this disease. Currently, there is no sufficiently sensitive and specific diagnostic test for subclinical CL. Diagnosis is currently achieved only by routine bacterial culture of purulent material collected from animals that have external abscesses, with subsequent biochemical identification of the isolates [17]. A few vaccines against CL are currently available, although they have not been licensed for use in many countries. Not all vaccines that have been developed for sheep are effective in goats. It is usually necessary to adjust vaccination programs to each animal host species [18].

Considering the current unfortunate status of CL prevalence in the world, especially in Brazil and Australia, there is a pressing need for more efficient alternatives for disease control that not only cure sick animals but also minimize or even prevent the onset of disease in herds. One of the major efforts to eradicate this disease involves the identification of genes that are related to the *C. pseudotuberculosis* pathogenicity and lifestyle. As an intracellular facultative pathogen, *C. pseudotuberculosis* exhibits several characteristics in its genome, such as gene loss, low GC content and a reduced genome [19] that differ from those of non-pathogenic *Corynebacterium* species. The finding of seven putative pathogenicity islands containing classical virulence elements, including genes for iron uptake, fimbrial subunits, insertional elements and secreted toxins [20], probably mostly acquired through horizontal transfer, contributes to our understanding of how this species causes disease. Comprehensive knowledge of an organism's genome facilitates an exhaustive search for candidates for virulence genes, vaccine and antimicrobial targets, and components that could be used in diagnostic procedures.

The information retrieved from a single genome is insufficient to provide an understanding of all *C. pseudotuberculosis* strains. Comparative genomics can shed light on the molecular attributes of a strain that affect its virulence, host specificity, dissemination potential and resistance to antimicrobial agents [21,22]. Furthermore, comparison of entire genome sequences of strains belonging to the same species, but from different geographic, epidemiological, chronological and clinical backgrounds, as well as affecting different hosts, would be useful for determining the molecular basis of these differences. As part of an effort to provide means to control CL, we examined the genomes of two strains of *C. pseudotuberculosis* isolated from sheep and goats, respectively, and compared them to each other and to the genomes of two other strains already available in a public database [6,23].

Results

Corynebacterium pseudotuberculosis genome

Overviews of the *C. pseudotuberculosis* genomes can be seen in Figure 1. The genomes are available in the NCBI GenBank database under accession numbers Cp1002:CP001809 and CpC231:CP001829.

The two strains are very similar, with an amino acid similarity of at least 95% between their predicted proteins. In their genomic composition, the isolates were found to have the same mean i) GC content, ii) gene length, iii) operon composition and iv) gene density. However, some significant differences were observed in: i) genome size, ii) number of pseudogenes and iii) lineage-specific genes (Table 1).

Gene order in *C. pseudotuberculosis*

To determine whether synteny was maintained between the two *C. pseudotuberculosis* strains, we made a comparative analysis of global gene order. As expected, the two *C. pseudotuberculosis* strains showed high synteny conservation; approximately 97% of their genes were found to be conserved in the comparison between the two strains. Previous studies provide evidence of a high degree of conservation of gene order in four *Corynebacterium* genomes, *C. diphtheriae*, *C. glutamicum*, *C. efficiens* and *C. jeikeium*, showing only 10

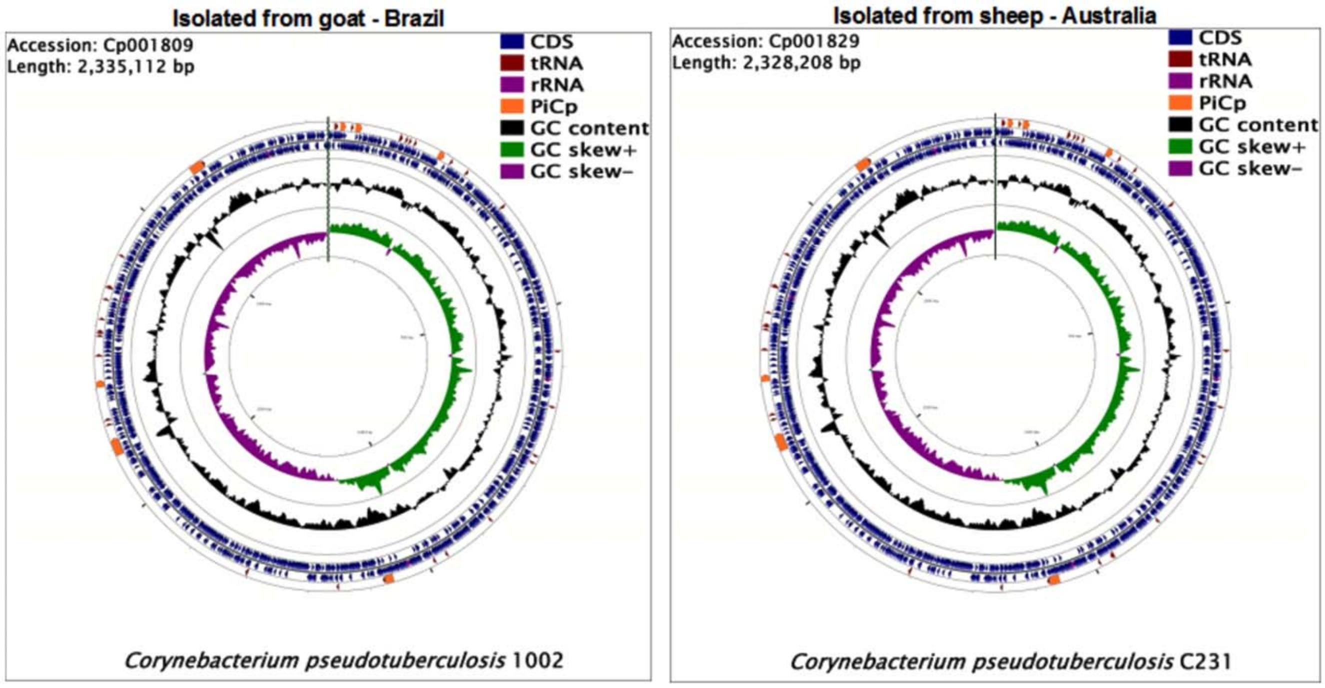


Figure 1. The whole genome of *Corynebacterium pseudotuberculosis.* Cp1002 strain isolated from a goat in Brazil and CpC231 strain isolated from sheep in Australia. Highlighted in yellow are the pathogenicity islands (PiCps) of *C. pseudotubeculosis* and its location in the genomes.
doi:10.1371/journal.pone.0018551.g001

gene-order breakpoints; rearrangement events during evolution in this species appear to be rare [24,25]. We checked the validity of this conclusion by making a comparative analysis of the genomes of the two *C. pseudotuberculosis* strains against *C. diphtheriae*, the *Corynebacterium* species that is most closely related to *C. pseudotuberculosis* [26,27].

Both *C. pseudotuberculosis* genomes showed a high degree of conservation in gene position, when compared to the *C. diphtheriae* genome, with few rearrangement points. This finding supports the hypothesis of a high degree of synteny conservation in this genus [25].

Table 1. General features of the genomes of two *Corynebacterium pseudotuberculosis* strains.

Genome feature	Cp1002	CpC231
Genome size (bp)	2,335,112	2,328,208
Gene number	2111	2103
Operon predicted number	474	468
Pseudogene number	53	50
tRNA number	48	48
rRNA operon	4	4
Gene mean length (bp)	964	968
Gene density (%)	0.88	0.88
Coding percentage	84.9	85.4
GC content (gene) (%)	52.88	52.86
GC content (genome) (%)	52.19	52.19
Lineage-specific genes	52	49

doi:10.1371/journal.pone.0018551.t001

Pathogenicity islands (PAIs)

Pathogenicity islands in bacterial genomes can be characterized by looking for characteristics linked to horizontal gene transfer, such as differences in codon usage, G+C content, dinucleotide frequency, insertion sequences, and tRNA flanking regions, together with transposase coding genes, which are involved in incorporation of DNA by transformation, conjugation or bacteriophage infection [28].

Pathogenicity islands had not been reported for *C. pseudotuberculosis*; to date; we used a multi-pronged approach called PIPS (submitted article) to identify the putative PAIs of *C. pseudotuberculosis*. Seven regions with most or all of the characteristics of horizontally-acquired DNA were found in both strains, Cp1002 and CpC231: i) base composition and/or codon usage deviations, ii) tRNA flanking, and iii) transposase genes. These regions were not found in a non-pathogenic species belonging to the same genus, *C. glutamicum*, and were classified as putative pathogenicity islands in *C. pseudotuberculosis* (PiCp). PiCps encode for proteins involved in the ABC transport system, for glycosil transferase, a two-component system, the *fag* operon and phospholipase D Table 2 provides a list of some genes found in the PAIs, with their respective functions.

Genetic composition of *C. pseudotuberculosis* Pathogenicity Islands

The genetic composition of PAIs can shed light on the lifestyle of pathogenic bacteria, since they include virulence genes that mediate mechanisms of adhesion, invasion, colonization, proliferation into the host and evasion of the immune system [29,30]. In addition, PAIs are characterized as being unstable regions that can be affected by insertions and deletions, influencing bacterial adaptability to new environments and hosts [31]. Here follows descriptions of the most relevant genetic elements found in the *C. pseudotuberculosis* pathogenicity islands. For more information, see

Table 2. Genes and proteins present in pathogenicity islands of the *Corynebacterium pseudotuberculosis* strain genomes.

PAI	Cp1002	CpC231	Protein
	tnp7109-9	tnp7109-9	Transposase for insertion sequence
	pld	pld	Phospholipase D precursor (PLD)
PiCp 1	fag C	fag C	ATP binding cytoplasmic membrane protein - FagC
	fag B	fag B	Iron-enterobactin transporter - FagB
	fag A	fag A	Integral membrane protein - FagA
	fag D	fag D	Iron siderophore binding protein - FagD
	mgtE	mgtE	Mg2+ transporter mgtE
	malL	malL	Oligo-1,6-glucosidase
PiCp 2	tetA	tetA	Putative tetracycline-efflux transporter
	cskE	cskE	Anti-sigma factor
	sigK	sigK	ECF family sigma factor K
	dipZ	dipZ	Integral membrane C-type cytochrome biogenesis protein DipZ
	potG	potG	Putrescine ABC transport system
	afuB	afuB	Putative transport system permease (iron)
PiCp 3	afuA	afuA	Iron (Fe3+) ABC superfamily ATP binding cassette transporter, binding protein
	glpT	glpT	Glycerol-3-phosphate transporter
	phoB	phoB	Two-component regulatory protein
	lcoS	lcoS	Two-component sensor protein, sensor histidine kinase
	ciuA	ciuA	Putative iron transport system binding (secreted) protein
	ciuB	ciuB	Putative iron transport system membrane protein
PiCp 4	ciuC	ciuC	Putative iron transport system membrane protein
	ciuD	ciuD	Putative iron ABC transport system
	ciuE	ciuE	Putative siderophore biosynthesis related protein
	σ70	σ70	Putative RNA polymerase sigma factor 70
	Pseudogene	Pseudogene	Putative chromosome segregation ATPase
PiCp 5	hsdR	hsdR	Putative type III restriction-modification system
	pfoS	pfoS	PfoR superfamily protein
	htaC	htaC	HtaA family protein
	guaB3	guaB3	Inosine 5-monophosphate dehydrogenase
PiCp6	pipA1	pipB	Proline iminopeptidase
	mfsD1	mfsD1	Major facilitator superfamily domain-containing protein 1
	dcd	dcd	Deoxycytidine triphosphate deaminase
	udg	udg	UDP-glucose 6-dehydrogenase
	lysS1	lysS1	Lysyl-tRNA synthetase
	alaT	alaT	Aminotransferase AlaT
	ureA	ureA	Urease gamma subunit
	ureB	ureB	Urease beta subunit
	ureC	ureC	Putative urease subunit alpha
PiCp 7	ureE	ureE	Urease accessory protein
	ureF	ureF	Urease accessory protein
	ureG	ureG	Urease accessory protein
	ureD	ureD	Urease accessory protein
	fepC2	fepC2	ABC superfamily ATP binding cassette transporter
	fecD	fecD1	Iron(III) dicitrate transport system permease fecD
	phuC	phuC	Iron(III) dicitrate transport permease-like protein yusV
	arsR	arsR1	ArsR-family transcription regulator

doi:10.1371/journal.pone.0018551.t002

the list of these orthologous genes in other *Corynebacterium* species in the Table S1 (online supporting information).

PiCp 1. *C. pseudotuberculosis* PiCp 1 harbors key genes involved in virulence and pathogenicity; these include PLD, the major virulence factor of this organism, which plays a role in spreading through the host; the *fag* operon, responsible for extracellular iron acquisition and, consequently, for survival in hostile environments; and a transposase gene, probably responsible for insertion of the island into the *C. pseudotuberculosis* genome. The finding that *C. ulcerans* can produce phospholipase D protein [32] indicates acquisition of PiCp1 by both *C. pseudotuberculosis* and *C. ulcerans*.

PiCp 2. Gene *mgtE* of island 2 has Mg^{2+} influx activity [33]. In prokaryotes, Mg^{2+} has been identified as an important regulatory signal that is essential for virulence, since it is involved in thermal adaptation, protecting bacteria from heat shock caused by fever in warm-blooded mammals [34]. Translation of the *mgtE* gene is regulated by changes in cytosolic Mg^{2+} concentration; loss of MgtE reduces biofilm formation and motility in the pathogenic bacteria *Aeromonas hydrophila* [33].

The protein MalL (*malL*), a maltose-inducible α-glucosidase, hydrolyzes various disaccharides, such as maltose and isomaltose, which can serve as carbon and energy sources [35,36].

The *tetA* gene codes for a tetracycline-efflux transporter protein that extrudes antibiotics from the cell and confers resistance to biofilm cells. The *tetA* gene is often carried by transmissible elements, such as plasmids, transposons, and integrons [37], thus explaining its presence in a PAI.

The *sigK* gene is an extracytoplasmic function sigma factor (sigma ECF) regulated by cskE, an anti-sigma factor. Another sigma ECF, *sigK*, mediates targeted alterations in bacterial transcription via transduction of extracellular signals. In *M. tuberculosis*, *sigK* regulates several genes (*Rv2871*, *mpt83*, dipZ, *mpt70*, *Rv2876*, and *mpt53*). Also, *sigK* mutations produce reduced quantities of the antigens MPT70 and MPT83 in vitro, and only induce strong expression during infection of macrophages [38–40].

PiCp2 also harbors a *dipZ* gene, which is regulated by *sigK* and seems to play a role in macrophage infection by *M. tuberculosis*, although its function is not clearly elucidated. DipZ is found as two separate proteins in most bacteria: CcdA and TlpA-like. Also, a full-length *dipZ* gene, found in the phylum *Actinobacteria*, is present exclusively in pathogenic bacteria (*C. diphtheriae*, *C. jeikeium*, *M. avium*, *M. kansasii*, *M. marinum*, *M. ulcerans* and *M. tuberculosis*) [40].

PiCp 3. *potG* gene, of the *pot*FGHI operon, is a membrane-associated/ATP-binding protein that provides energy for putrescine (polyamine) uptake from the periplasmic space [41]. Although the *potFGHI* operon is a putrescine-specific transport system, *potG* is downregulated by another polyamine (spermine), which is produced only by eukaryotes. Carlson et al. (2009) demonstrated that transcription of the *potG* gene in *Francisella tularensis* decreases with high levels of spermine, while transcription of IS elements ISFtu1 and ISFtu2 increases in response to high levels of spermine in macrophages responding to bacterial infection. Also, many of the upregulated genes of *F. tularensis* (pseudogenes and transposase genes) are located near the IS elements in the chromosome [42].

The gene *glpT* belongs to the organophosphate:phosphate antiporter family of the major facilitator superfamily (MFS); it mediates transport of glycerol 3-phosphate (G3P) across the membrane in bacteria [43].

The PhoPR system regulates expression of various genes involved in metabolic, virulence and resistance processes in several intracellular bacterial pathogens [44]. Based on the information obtained from the complete genome sequence of *C. pseudotuberculosis*, we found that the PhoPR system is constituted of the *phoP* (714 bp) and *phoR* (1506 bp) genes, separated by a small 39-bp sequence, suggesting that these two genes are transcribed by a bicistronic operon. The size and organization of this system in *C. pseudotuberculosis* is similar to those of other Gram-positive bacteria [45]. Live bacteria attenuated via *phoP* inactivation are also promising vaccine candidates against tuberculosis. Several studies have reported the efficacy of attenuated mutant strains of *M. tuberculosis* as vaccines [46,47]. Phylogenetic relationships within the class *Actinobacteria* strongly suggest correlation of the *C. pseudotuberculosis* PhoPR system with virulence mechanisms. The *phoP* gene is an important subject for regulation studies; and is also a probable vaccine candidate against CL.

PiCp4. The operon *ciuABCDE* (*corynebacterium* iron uptake) was described in *C. diphtheriae* as an iron transport and siderophore biosynthesis system. Proteins involved in iron acquisition are recognized as virulence factors, since they help pathogens to obtain iron from a host by using siderophores to strip iron from carrier proteins, such as transferrin, lactoferrin, and hemoglobin-haptoglobin [48,48].

PiCp5. Island 5 harbors a gene (*pfoS*) related to the *pfoR* superfamily. The *pfo*R gene was previously characterized as responsible for positive regulation of production of perfringolysin A (*pfoA*) and other toxins in *Clostridium perfringens* [50]. The virulence factors regulated by *pfoR* have not been totally elucidated. However, it is well known that deactivation of this gene inhibits hemolysis through negative regulation of several *C. perfringens* toxins. *Clostridium perfringens* harbors a phospholipase C gene (*plc*) that serves a function similar to that of phospholipase D [51]. Additionally, PiCp 5 contains a putative sigma 70 factor that is responsible for transporting the transcription machinery to specific promoters. Interestingly, the putative sigma 70 factor presents a nonsense mutation in *C. pseudotuberculosis* strain C231, which could be responsible for differential gene expression.

PiCp6. The *pipA1* gene, which codes for a proline iminopeptidase, may have a role in pathogenesis, since it catalyses the removal of N-terminal proline residues from peptides; it also has a role in energy production [52]. In addition, a PIP-type protein is required for virulence of *Xanthomonas campestris pv. campes*tris [53].

PiCp7. Island 7 harbors a urease operon that is also present in *C. glutamicum*; it is flanked, on both sides, by regions that are absent in the non-pathogenic *C. glutamicum*. This mosaicism is a common feature of pathogenicity islands [54]. The *ure* operon presents a codon usage deviation in *C. glutamicum*, as in *C. pseudotuberculosis*, indicating that this region is a putative genomic island in *C. glutamicum*.

The *ure* operon is responsible for nitrogen acquisition through hydrolysis of urea to carbamate and ammonia. Production of ammonia by uropathogenic and enteropathogenic bacteria causes cellular damage and compromises the action of the host's immune system [55]. Considering this fact, due to the intramacrophagic location of *C. pseudotuberculosis* and the finding of this operon in a non-pathogenic bacterial species, additional studies will be needed to elucidate how *C. pseudotuberculosis* obtains urea from the host and how this operon affects pathogenicity.

PiCp 7 also harbors a lysyl-tRNA synthetase (*lysS*), responsible for lysine incorporation into its respective transfer tRNA. The importance of *lysS* would normally make its location on a PAI inviable, since it is essential for cell metabolism. However, it is the only tRNA synthetase gene that is duplicated in the genome.

Protein classification of *C. pseudotuberculosis* in the biological process

Using the controlled vocabulary of functional terms proposed by the Gene Ontology (GO) Consortium for gene products

Table 3. Subcellular prediction of the protein locations derived from complete genomes of *Corynebacterium* species.

Category/Species	Ce	CgB	CgK	CgR	Cj	Cd	Cu	Cp1002	CpC231	Total
Cytoplasm	2,158	2,11	2,082	2,158	1,49	1,594	1,432	1,399	1,389	15,812
Cytoplasm	504	557	541	561	333	375	332	364	356	3,923
PSE	230	254	249	252	197	204	179	201	201	1,967
Secreted	102	136	121	109	100	99	79	95	107	948
Total	2,994	3,057	2,993	3,08	2,12	2,272	2,02	2,059	2,053	22,648

Ce: *C. efficiens*; CgB: *C. glutamicum B*; CgK: *C. glutamicum K*; CgR: *C. glutamicum R*; Cj: *C. jeikeium;* Cd: *C. diphtheriae;* Cu: *C. urealyticum;* Cp1002: *C. pseudotuberculosis* 1002; CpC231: *C. pseudotuberculosis* C231. PSE: potential surface exposure.
doi:10.1371/journal.pone.0018551.t003

classification [56], the predicted proteomes of the two genomes were analyzed according to the three organizing principles of gene ontology: cellular component, biological process and molecular function. The most abundantly represented categories are linked to metabolic processes in the two strains (cellular metabolic, biosynthetic, primary and macromolecule processes).

The gene products composition characterized using GO terminology suggests that *C. pseudotuberculosis* is a facultative intracellular pathogen. It is commonly found that pathogens specialized for an intracellular lifestyle have a high proportion of proteins linked to the above-mentioned processes. Moreover, the low proportion of proteins linked to the metabolism of secondary metabolites is an indication that *C. pseudotuberculosis* does not possess the metabolic machinery to deal with secondary metabolites, because they are supplied by the host.

Sub-cellular localization of *C. pseudotuberculosis* proteins

Prediction of the sub-cellular localization of *C. pseudotuberculosis* proteins was made by *in silico* analysis, using the SurfG+ tool [57]. Surfg+ is a pipeline for protein sub-cellular prediction, incorporating commonly used software for motif searches, including SignalP, LipoP and TMHMM, along with novel HMMSEARCH profiles to predict protein retention signals. Surfg+ starts by searching for retention signals, lipoproteins, SEC pathway export motifs and transmembrane motifs, roughly in this order. If none of these motifs are found in a protein sequence, then it is characterized as being cytoplasmic. A novel possibility introduced by Surfg+ is the ability to distinguish between integral membrane proteins versus PSE (potentially surface-exposed proteins). This is done by a parameter that determines the expected cell wall thickness, expressed in amino acids. Using published information or electron microscopy, it is possible to estimate cell wall thickness value for procaryotic organisms. *C. pseudotuberculosis* proteins were classified into four different sub-cellular locations: cytoplasmic, membrane, PSE (potentially surface exposed), or secreted. The *C. pseudotuberculosis* genomes were compared to those of other species of the genus, including C. *diphtheriae, C. efficiens, C. glutamicum*, C. *jeikeium* and C. *urealyticum*, also predicted by Surfg+, based on published cell wall thicknesses. Table 3 shows the number of predicted proteins in each sub-cellular location.

Comparison of the frequencies of subcellular occurrence of the *C. pseudotuberculosis* proteins and other *Corynebacterium* proteomes was made with Chi-square tests. The ratio between the four groups (cytoplasmic, membrane anchored, potentially exposed and secreted proteins) was found to be nearly constant among the *Corynebacterium* species. The proportions of the four protein categories cited above were similar to published data [58,59]. Song and colleagues (2009) showed that approximately 30% of proteins secreted in gram-positive bacteria are exported through the Sec pathway. Few proteins (n = 27) were predicted to be secreted by the Tat pathway in Cp1002. About 2% of the proteins predicted to be secreted presented tertiary structures. In terms of proportions of secreted proteins, Cp1002 and CpC231 are at the higher end of the spectrum. They present 4.61 and 5.21%, respectively, predicted secreted proteins (Table 3).

Differences in metabolic pathways in the two strains of *C. pseudotuberculosis*

Automated reconstruction of the *C. pseudotuberculosis* Cp1002 metabolic pathways identified 156 pathways and 744 enzymatic reactions. As expected, quite similar results were encountered for strain CpC231: 154 pathways and 754 reactions (Table 4). Proteins of predicted functions that did not map to pathways, such as transport reactions, enzymes, transporters, and compounds, were also identified. The metabolic pathway database can be accessed online at http://corynecyc.cebio.org. This database enabled us to visualize and compare the metabolism of these two *C. pseudotuberculosis* strains (Figure 2).

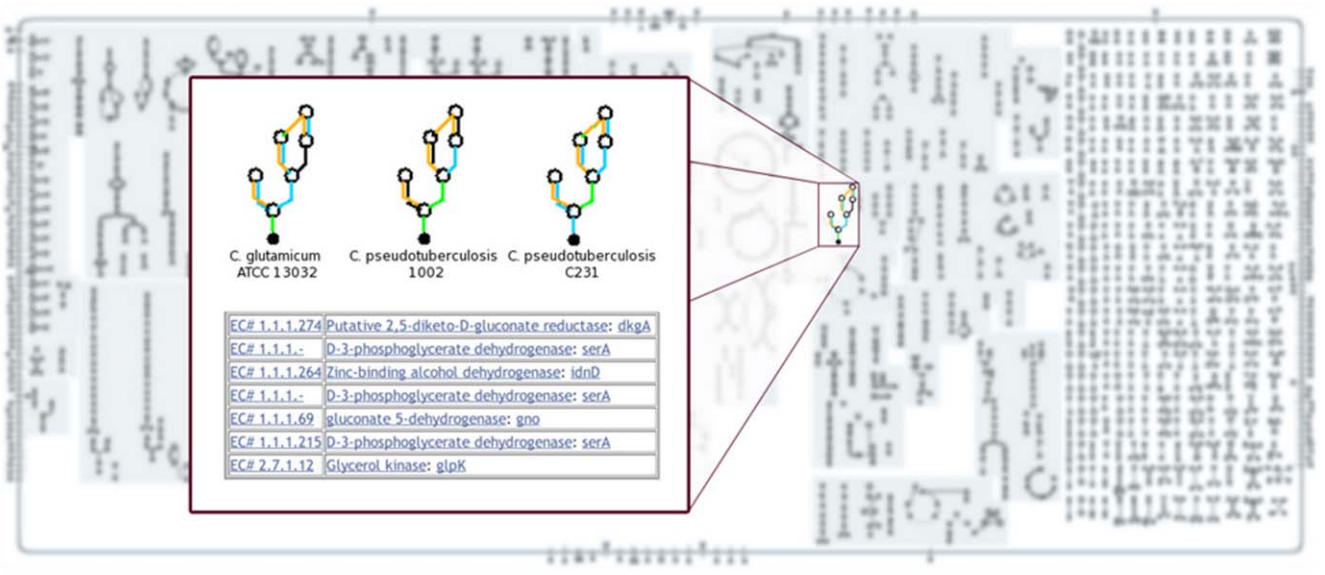
We made a comparative analysis of transport reactions, pathways, compounds and proteins for *C. pseudotuberculosis* strains Cp1002 and CpC231 (Table 5). Despite the high similarity of the metabolic pathways, some differences were observed.

The metabolic pathways in each of the two bacterial strains (Cp1002 and CpC231) were classified into several pathway classes; each pathway class was further broken down to show the distribution of pathways among the next-level subclasses. Analysis of the metabolism database of *C. pseudotuberculosis* strains Cp1002 and CpC231 revealed specific pathway differences between the two strains. Overall, CpC231 had 13 specific metabolic pathways

Table 4. Comparative summary of the *Corynebacterium pseudotuberculosis* strain gene data types.

Data Type	Cp1002	CpC231
Gene products	2,059	2,053
Pathways	156	154
Enzymatic Reactions	744	754
Transport Reactions	8	4
Polypeptides	2,065	2,059
Enzymes	516	506
Transporters	10	10
Compounds	639	651

doi:10.1371/journal.pone.0018551.t004



EC# 1.1.1.274	Putative 2,5-diketo-D-gluconate reductase: dkgA
EC# 1.1.1.-	D-3-phosphoglycerate dehydrogenase: serA
EC# 1.1.1.264	Zinc-binding alcohol dehydrogenase: idnD
EC# 1.1.1.-	D-3-phosphoglycerate dehydrogenase: serA
EC# 1.1.1.69	gluconate 5-dehydrogenase: gno
EC# 1.1.1.215	D-3-phosphoglycerate dehydrogenase: serA
EC# 2.7.1.12	Glycerol kinase: glpK

Figure 2. ***Corynebacterium glutamicum*** **metabolic pathways overview.** *C. glutamicum* reactions are presented in blue and the reactions shared with *C. pseudotuberculosis* C231 and 1002 in red and green, respectively. By clicking on any compound or reaction, a window pops up showing details of each pathway. The fatty acid biosynthesis initiation pathway is the chosen example since computational evidence indicates it is not present only in strain C231.
doi:10.1371/journal.pone.0018551.g002

not found in strain Cp1002, and the latter had 11 metabolic pathways not found in strain CpC231 (Table 6).

Two amine and polyamine biosynthesis pathways, choline degradation I and glycine betaine biosynthesis I (Gram-negative bacteria), were found in strain Cp1002 but not in strain CpC231. Strain CpC231 was found to have an extra amino acid biosynthesis pathway, the citrulline-nitric oxide cycle. Strain Cp1002 was found to have three additional carbohydrate biosynthesis pathways: gluconeogenesis, trehalose biosynthesis II and trehalose biosynthesis III. Strain CpC231 showed three cofactor biosynthesis, prosthetic group and electron carrier pathways, corresponding to adenosylcobalamin biosynthesis from cobyrinate a,c-diamide I, heme biosynthesis from uroporphyrinogen II and siroheme biosynthesis. Strain Cp1002 showed only one unique cofactor biosynthesis pathway, heme biosynthesis from uroporphyrinogen I. Two extra pathways of fatty acid and lipid biosynthesis were found in strain Cp1002, cardiolipin biosynthesis I and fatty acid biosynthesis initiation I. Strain CpC231 showed only the biotin-carboxyl carrier protein. Among metabolic regulator biosynthesis genes, strain CpC231 showed the citrulline-nitric oxide cycle. Strain CpC231 also showed an extra pathway, the canavanine biosynthesis pathway, part of secondary metabolite biosynthesis.

Among degradation/utilization/assimilation pathways, strain Cp1002 showed an extra pathway: glycerol degradation II, for alcohol degradation, as well as choline degradation I for amine and polyamine degradation. Strain CpC231 was found to have two additional pathways, 2-ketoglutarate dehydrogenase complex and citrulline-nitric oxide cycle, for amino acid pathways; strain Cp1002 showed only one extra pathway, valine degradation I. Among carboxylate degradation pathways, involving fatty acid and lipid degradation, strain Cp1002 showed two extra pathways: one corresponding to acetate formation from acetyl-CoA I, and the second linked to triacylglycerol degradation. Two inorganic nutrient metabolism pathways were found in strain CpC231 but not in strain Cp1002: nitrate reduction III (dissimilatory) and nitrate reduction IV (dissimilatory), and a nucleoside and nucleotide degradation and purine deoxyribonucleoside recycling degradation pathway.

Finally, when we analyzed the generation of precursor metabolites and energy, strain CpC231 showed three extra pathways: 2-ketoglutarate dehydrogenase complex, nitrate reduction III (dissimilatory) and nitrate reduction IV (dissimilatory). The differences are presented in Table 6.

Metabolic pathways in *C. pseudotuberculosis* compared to other *Corynebacterium* species

The web interface enabled us to visually compare the metabolic pathways of strains Cp1002 and CpC231 reactions (Figure 2) with those of four other bacteria of the genus *Corynebacterium*: *C. diphtheriae, C. efficiens, C. glutamicum, and C. jeikeium*. Using these diagrams we were able to easily spot reactions present in *C. pseudotuberculosis* and absent in other *Corynebacterium* species.

A comparative analysis of reactions, pathways, compounds and proteins was also done for *C. pseudotuberculosis* and other closely-related bacteria in the same genus. The list of *C. pseudotuberculosis* specific pathways is shown in Table 7.

We found that *C. pseudotuberculosis* has several pathways that are not found in other species of the genus *Corynebacterium*. However, little information is available about these pathways in *Corynebacterium* spp. We found no published information concerning the following pathways: asparagine biosynthesis II, citrulline-nitric oxide cycle (amino acid biosynthesis and degradation), pyrimidine deoxyribonucleotide salvage pathways, methylglyoxal degradation III, reductive monocarboxylic acid cycle, chitobiose degradation, conversion of succinate to propionate, ammonia oxidation I (aerobic), nitrate reduction IV (dissimilatory), D-glucarate degradation, betanidin degradation, D-galactarate degradation, and ammonia oxidation I (aerobic).

Some studies reported five pathways: lysine biosynthesis V, glycerol degradation II, alanine degradation IV, lysine degradation I and phospholipases. However, none of the studies, except for those concerning lysine degradation I and phospholipase

Table 5. Comparative summary of the number of pathways of *Corynebacterium pseudotuberculosis* strains Cp1002 and CpC231.

Pathway Class - Pathway subclass	Cp1002	CpC231
Biosynthesis	105	104
- Amine and Polyamine Biosynthesis	5	3
- Amino acid Biosynthesis	25	26
- Aminoacyl-tRNA Charging	1	1
- Aromatic Compound Biosynthesis	1	1
- Carbohydrate Biosynthesis	10	7
- Cell structure Biosynthesis	4	4
- Cofactor, Prosthetic Group, Electron Carrier Biosynthesis	27	29
- Fatty Acid and Lipid Biosynthesis	8	7
- Metabolic Regulator Biosynthesis	1	2
- Nucleoside and Nucleotide Biosynthesis	12	12
- Other Biosynthesis	1	1
- Secondary Metabolites Biosynthesis	1	2
Degradation/Utilization/Assimilation	53	54
- Alcohol Degradation	2	1
- Aldehyde Degradation	1	1
- Amine and Polyamine Degradation	5	4
- Amino Acid Degradation	11	12
- C1 Compound Utilization and Assimilation	4	4
- Carbohydrate Degradation	7	7
- Carboxylate Degradation	5	4
- Degradation/Utilization/Assimilation - Other	5	5
- Fatty Acid and Lipid Degradation	3	2
- Inorganic Nutrient Metabolism	4	6
- Nucleoside and Nucleotide Degradation and Recycling	2	3
- Secondary Metabolite Degradation	5	5
Generation of precursor metabolites and energy	16	19
Total	163	164

doi:10.1371/journal.pone.0018551.t005

pathways, involved *C. pseudotuberculosis*. Most of these studies were carried out with *C. glutamicum*.

Four papers concerning *C. glutamicum* were found for the lysine degradation I pathway [60–63]. Studies have focused on: acetohydroxyacid synthase, a novel target for improvement of L-lysine production [62], improvement of L-lysine formation by expression of the *Escherichia coli* pntAB genes [61], genetic and functional analysis of soluble oxaloacetate decarboxylase [63], and modeling and experimental design for metabolic flux analysis of lysine-producing *Corynebacteria* by mass spectrometry [64].

Six studies were found concerning the glycerol degradation II pathway, one performed with C. *diphtheria* [65] and four with C. *glutamicum* [66–69]. In the sixth study, made with *C. glutamicum*, we found information on the alanine degradation IV pathway [64].

Approximately 140 studies, of which 107 were made with *C. glutamicum* alone, dealt with the lysine degradation I pathway, in which cadaverine is biosynthesized from L-lysine. Cadaverine is reported to be essential for the integrity of the cell envelope and for normal growth of the organism, as well as for inhibiting porin-mediated outer membrane permeability, thereby protecting cells from acid stress [70,71].

All studies of specific phospholipase pathways were carried out with *C. pseudotuberculosis*. Phospholipases hydrolyze phospholipids and are ubiquitous in all organisms. Several types of phospholipases were reported; phospholipase D is the best studied and has been considered a major virulence factor for *C. pseudotuberculosis* [72,73]. In our analyses, none of the five bacteria of the genus *Corynebacterium* were found to have pathways belonging to the following subclasses: siderophore biosynthesis; chlorinated compound degradation; cofactor, prosthetic group, electron carrier, and hormone degradation. Clearly more biochemical studies are needed. Our current study brings new insight to relevant biochemical pathways that can be further explored experimentally.

We made a comparative summary of the metabolic pathways of *C. pseudotuberculosis* strains Cp1002 and CpC231 and *C. glutamicum* (Table 8). *C. glutamicum* has several metabolic pathways not found in *C. pseudotuberculosis* Cp1002 and/or in *C. pseudotuberculosis* CpC231. Overall, *C. glutamicum* has approximately 40 additional metabolic pathways.

Among biosynthesis pathways, *C. glutamicum* showed around 30 extra pathways when compared to the two strains of *C. pseudotuberculosis*. These involve pathways of amino acid biosynthesis, aminoacyl-tRNA charging, cofactors, prosthetic groups, electron carrier biosynthesis, fatty acid and lipid biosynthesis and secondary metabolite biosynthesis. However, the two strains of *C. pseudotuberculosis* also have specific pathways that were not found in *C. glutamicum*, these being the pathways of amine and polyamine biosynthesis, carbohydrate biosynthesis and nucleoside and nucleotide biosynthesis.

Among the degradation/utilization/assimilation pathways, *C. glutamicum* presented around 20 extra pathways, when compared to *C. pseudotuberculosis* Cp 1002 and *C. pseudotuberculosis* CpC231. These specific pathways of *C. glutamicum* correspond to pathways of amine and polyamine degradation, amino acid degradation, aromatic compound degradation, carbohydrate degradation, carboxylate degradation, chlorinated compound degradation and the metabolism of inorganic nutrients. Again, the two strains of *C. pseudotuberculosis* also had specific pathways involving degradation/utilization/assimilation, fatty acid and lipid degradation and secondary metabolite degradation that were not found in *C. glutamicum*.

We found 25 pathways involving generation of precursor metabolites and energy in *C. glutamicum*, while *C. pseudotuberculosis* Cp1002 had only 16 and *C. pseudotuberculosis* CpC231 had 19.

Discussion

General aspects of the *C. pseudotuberculosis* genome

The *C. pseudotuberculosis* genome has proven to be one of the smallest genomes of the *Corynebacterium* genus sequenced so far, with Cp1002 being the smallest and Cp231 the fourth smallest, larger only than Cp1002, *C. lipophiloflavum* DSM 44291 (2,293,743 bp) and *C. genitalium* ATCC 33030 (2,319,774 bp); the latter two are both human pathogens. *Corynebacterium pseudotuberculosis* has a very small genetic repertoire, with considerable gene loss when compared to non-pathogenic species such as *C. glutamicum* and *C. efficiens*. When predicted proteomes were compared, *C. pseudotuberculosis* showed a loss of approximately 1,220 genes, in comparison with *C. glutamicum*. Classification of these proteins using GO terminology showed that the majority are linked to metabolic processes, such as cellular, primary, biosyn-

Table 6. Table listing the *Corynebacterium pseudotuberculosis* strain-specific pathways.

Pathway Class	**Cp1002**	**CpC231**
Pathway Name		
Biosynthesis - Amines and Polyamines Biosynthesis		
choline degradation I	present	absent
glycine betaine biosynthesis I (Gram-negative bacteria)	present	absent
Biosynthesis - Amino acid Biosynthesis		
citrulline-nitric oxide cycle	absent	present
Carbohydrates Biosynthesis		
gluconeogenesis	present	absent
trehalose biosynthesis II	present	absent
trehalose biosynthesis III	present	absent
Biosynthesis - Cofactor, Prosthetic Group, and Electron Carrier Biosynthesis		
adenosylcobalamin biosynthesis from cobyrinate a,c-diamide I	absent	present
heme biosynthesis from uroporphyrinogen I	present	absent
heme biosynthesis from uroporphyrinogen II	absent	present
siroheme biosynthesis	absent	present
Biosynthesis - Fatty Acid and Lipid Biosynthesis		
biotin-carboxyl carrier protein	absent	present
cardiolipin biosynthesis I	present	absent
fatty acid biosynthesis initiation I	present	absent
Secondary Metabolite Biosynthesis		
canavanine biosynthesis	absent	present
Biosynthesis - Metabolic Regulators Biosynthesis		
citrulline-nitric oxide cycle	absent	present
Degradation - Alcohols Degradation		
glycerol degradation II	present	absent
Degradation - Aldehyde Degradation		
methylglyoxal degradation I	absent	present
methylglyoxal degradation III	present	absent
Degradation - Amine and Polyamine Degradation		
choline degradation I	present	absent
Degradation - Amino Acid Degradation		
2-ketoglutarate dehydrogenase complex	absent	present
citrulline-nitric oxide cycle	absent	present
valine degradation I	present	absent
Degradation - Carboxylate Degradation		
acetate formation from acetyl-CoA I	present	absent
Degradation - Fatty Acid and Lipids Degradation		
triacylglycerol degradation	present	absent
Inorganic Nutrients Metabolism		
nitrate reduction III (dissimilatory)	absent	present
nitrate reduction IV (dissimilatory)	absent	present
Degradation - Nucleoside and Nucleotide Degradation and Recycling		
purine deoxyribonucleoside degradation	absent	present
Generation of precursor metabolites and energy		
2-ketoglutarate dehydrogenase complex	absent	present
nitrate reduction III (dissimilatory)	absent	present
nitrate reduction IV (dissimilatory)	absent	present

doi:10.1371/journal.pone.0018551.t006

Table 7. List of *Corynebacterium pseudotuberculosis* specific metabolic pathways that were compared to those of closely-related bacteria, including *C. diphtheriae*, *C. glutamicum*, *C. efficiens*, and *C. jeikeium*.

Pathway Class
Pathway Name
Biosynthesis - Amino acid Biosynthesis
Asparagine biosynthesis II
Lysine biosynthesis V
Biosynthesis - Metabolic Regulators Biosynthesis
Citrulline-nitric oxide cycle
Biosynthesis - Nucleoside and Nucleotide Biosynthesis
Salvage pathways of pyrimidine deoxyribonucleotides
Degradation - Alcohol Degradation
Glycerol degradation II
Degradation - Aldehyde Degradation
Methylglyoxal degradation III
Degradation - Amino Acid Degradation
Alanine degradation IV
Citrulline-nitric oxide cycle
Lysine degradation I
Degradation - C1 Compound Utilization and Assimilation
Reductive monocarboxylic acid cycle
Degradation - Carbohydrate Degradation
Chitobiose degradation
Degradation - Carboxylate Degradation
Conversion of succinate to propionate
Degradation - Fatty Acid and Lipid Degradation
Phospholipases
Inorganic Nutrients Metabolism
Ammonia oxidation I (aerobic)
Nitrate reduction IV (dissimilatory)
Degradation - Secondary Metabolite Degradation
D-glucarate degradation
Betanidin degradation
D-galactarate degradation
Generation of precursor metabolites and energy
Ammonia oxidation I (aerobic)

doi:10.1371/journal.pone.0018551.t007

thetic, macromolecule, nitrogen compound and oxidation reduction processes.

Other characteristics of the *C. pseudotuberculosis* genome include the lowest GC content in the *Corynebacterium* genus, this being 52% in both the goat and sheep strains, followed by *C. diphtheriae* with a GC content of 53%. This contrasts with *C. urealyticum*, which has a GC content of 64%. Furthermore, *C. pseudotuberculosis* has a higher number of predicted pseudogenes and a lower number of tRNAs, when compared to other species of the *Corynebacterium* genus for which genome sequences are available.

Merjeh et al. (2009) made a comparative analysis of 317 genomes of bacteria with different lifestyles (free-living, facultative intracellular and obligate intracellular). They found evidence that peculiar characteristics in bacterial genomes can drive the organisms to certain lifestyles. All characteristics cited in their work were identified in the *C. pseudotuberculosis* genomes. Lower GC content generally can occur due to gene loss, which is a means to contract the genome in response to a specialized environment. Moreover, presence of a higher number of pseudogenes could be evidence of bacterial mechanisms to generate non-functional genes and subsequent gene loss [19]. In addition, the high proportion of proteins linked to primary metabolism, and the small proportion of proteins related to secondary metabolism, is usually seen in facultative intracellular organisms. Taking these aspects of the genomic architecture of *C. pseudotuberculosis* into account, it can be affirmed that *C. pseudotuberculosis* has a facultative intracellular lifestyle.

High similarity in the genome architecture

Usually, pseudogenes are characterized as genes that have lost their function in the genome, due either to changes in the reading frame (frameshifts) or to a premature stop codon. Pseudogenes are common in prokaryotes; most have been linked to a sudden change in the environment of the pathogen, with simultaneous loss of metabolic and respiratory activities [74].

The high number of pseudogenes in these two strains of *C. pseudotuberculosis* (52 in Cp1002 and 50 pseudogenes in CpC231) suggest an evolutionary process involving a contracting genome in this species. An example of this is also seen in *Mycobacterium leprae*, which has a large number of pseudogenes (around 1,000). When we compare *M. leprae* to *M. tuberculosis*, the latter has both considerably fewer genes and a higher number of pseudogenes that can drive this gene loss.

Virulence factors acquired

Identification of pathogenicity islands (PAIs) in pathogenic bacteria is highly relevant for understanding the reasons behind different responses to vaccines and the biological mechanisms leading to genome plasticity. The biovars *equi* and *ovis* of *C. pseudotuberculosis* cause distinct diseases in their hosts; assessment of virulence genes could help identify genes involved in these host-specific differences.

Virulence genes, which are central to distinguishing pathogenic from non-pathogenic species, are present in PAIs in large numbers. Additionally, the fact that PAIs are a consequence of horizontal transfer events indicates that the virulence factors they contain can help increase the adaptability of strains to different host environments. This increase in adaptability is demonstrated by the finding of genes with functions associated with uptake of iron (*fag* operon), carbon (*malL*) and Mg^{2+} from the host, since this uptake improves survival under stress conditions, such as iron depletion, starvation and heat shock. Furthermore, PAIs of *C. pseudotuberculosis* present genes that respond to a macrophagic environment (*potG*, *sigK* and *dipZ*), which sheds new light on the mechanisms responsible for the intramacrophagic lifestyle of this organism.

Gene Sharing among *C. pseudotuberculosis* strains

Considering the four available genomes of *C. pseudotuberculosis* strains (Cp1002, CpC231, and CpI19 pFRC41), we identified 1,851 whole genes shared among them (Figure 3).

This repertoire of genes is vast for this specie, since, among the four isolates the maximum number of genes is 2,377 (called the pangenome of the species). When we compare the number of genes shared by these four *C. pseudotuberculosis* strains with a study of 17 strains of the bacterium *E. coli* [75], we conclude that *C. pseudotuberculosis* has a greater proportion of shared genes. In isolates of *E. coli*, 2,220 genes constituted the core genome, less

Table 8. Comparative summary of *Corynebacterium pseudotuberculosis* strains Cp1002 and CpC231 and *C. glutamicum* pathways.

Pathway Class	**Cp1002**	**CpC231**	***C. glutamicum***
- Pathway subclass			
Biosynthesis	105	104	131
- Amine and Polyamine Biosynthesis	5	3	3
- Amino acid Biosynthesis	25	26	29
- Aminoacyl-tRNA Charging	1	1	3
- Aromatic Compound Biosynthesis	1	1	1
- Carbohydrate Biosynthesis	10	7	9
- Cell structure Biosynthesis	4	4	4
- Cofactor, Prosthetic Group, Electron Carrier Biosynthesis	27	29	38
- Fatty Acid and Lipids Biosynthesis	8	7	14
- Metabolic Regulator Biosynthesis	1	2	1
- Nucleoside and Nucleotide Biosynthesis	12	12	10
- Other Biosynthesis	1	1	1
- Secondary Metabolite Biosynthesis	1	2	6
Degradation/Utilization/Assimilation	53	54	72
- Alcohols Degradation	2	1	2
- Aldehyde Degradation	1	1	1
- Amine and Polyamine Degradation	5	4	6
- Amino Acid Degradation	11	12	15
- Aromatic Compound Degradation	0	0	9
- C1 Compound Utilization and Assimilation	4	4	2
- Carbohydrate Degradation	7	7	10
- Carboxylate Degradation	5	4	6
- Chlorinated Compound Degradation	0	0	4
- Degradation/Utilization/Assimilation - Other	5	5	2
- Fatty Acid and Lipid Degradation	3	2	2
- Inorganic Nutrient Metabolism	4	6	9
- Nucleoside and Nucleotide Degradation and Recycling	2	3	1
- Secondary Metabolite Degradation	5	5	4
Generation of precursor metabolites and energy	16	19	25
Total	163	164	206

doi:10.1371/journal.pone.0018551.t008

than half of the genes in this species, with a mean of 5,000 genes in each genome [75]. Other significant information that emerges from this data is that the *C. pseudotuberculosis* genomes are extremely similar, since we found no significant change in the composition of the repertoire of genes for this species after adding the two new strains (Figure 3).

Gene Sharing between *C. pseudotuberculosis* and other *Corynebacterium* species

Previous comparative studies of sequences of the rpoB gene of *C. pseudotuberculosis* and *C. diphtheriae* have suggested a close relationship between them [27,76]. In our current study, we confirmed this close relationship with several types of evidence: i) a similar codon bias, ii) high similarity at the amino acid level and iii) conserved synteny. Synteny analysis of the genomes of the two *C. pseudotuberculosis* strains compared to *C. diphtheriae* indicates that these genomes are highly conserved; the gene position is conserved within the species. This observation reinforces the conclusions of previous research claiming conserved synteny in this genus, which indicated that few rearrangement events occurred during evolution [25].

Corynebacterium pseudotuberculosis shares more orthologous genes with *C. glutamicum* (1,345 genes), *C. efficiens* (1,330), *C. diphtheriae* (1,263 genes) and *C. auricumucosum* (1,273 genes); it shares only 1,030 genes with *C. jeikeium* and *C. kroppenstedtii*.

The larger number of genes shared between *C. pseudotuberculosis*, *C. glutamicum* and *C. efficiens* (72%), compared to other species (pathogenic species, 60%), may be a result not only of their close relationships, but also because a comparison is made among species with a larger gene repertoire, such as *C. glutamicum* and *C. efficiens*, which are non-pathogenic microorganisms, thus increasing the possibility of sharing genes.

Lineage-specific genes in *C. pseudotuberculosis*

Most of the lineage-specific genes are involved in processes of virulence, pathogenicity, drug resistance and response to certain types of stress. These factors can increase the adaptability of microorganisms to the niches they inhabit, but they are not

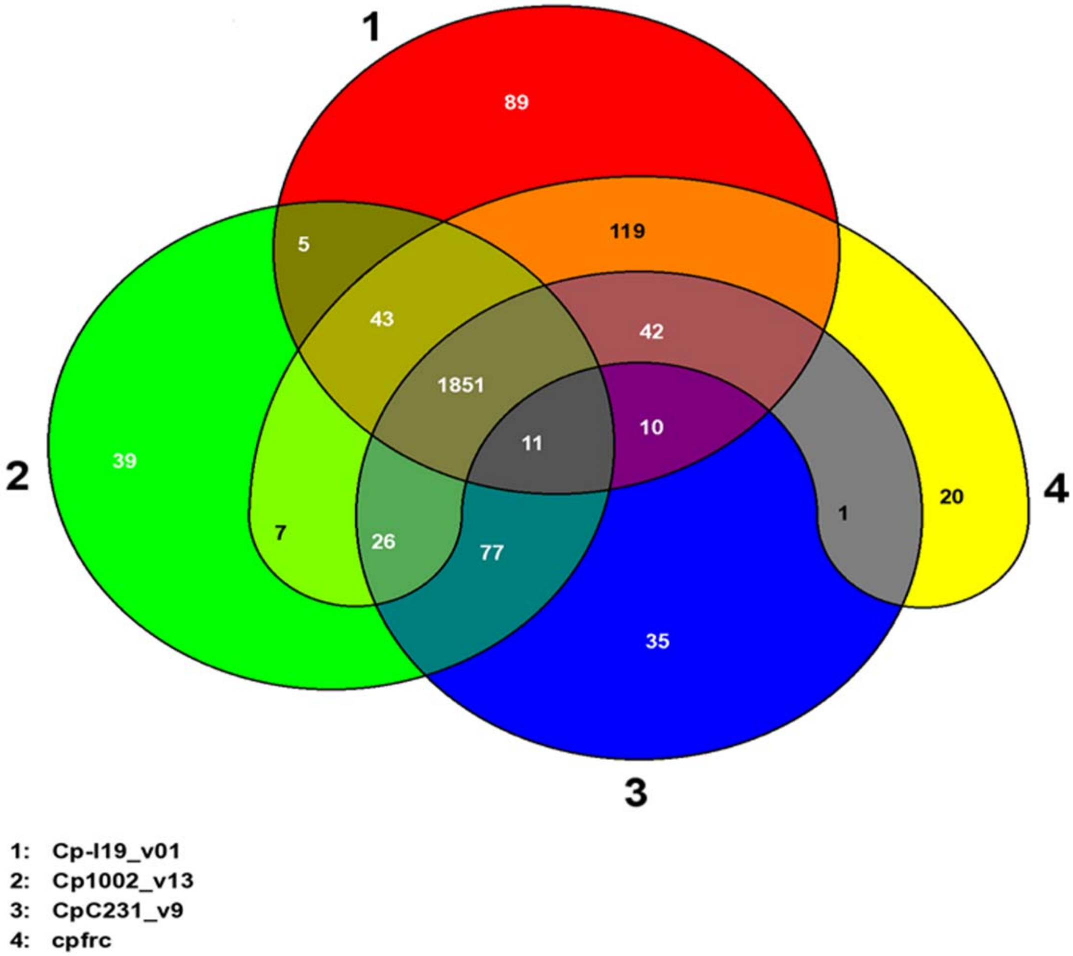


Figure 3. Venn diagram illustrating the three genomic categories of four *Corynebacterium pseudotuberculosis* strains: core, accessory and extended genome. Data obtained from the comparison of the predicted proteomes of four *C. pseudotuberculosis* speices in the EDGAR program (Blom et al., 2009). In red: Cp-I19; green: Cp1002; blue: CpC231 and yellow: CpFRC41. The remaining colors illustrate the shared genes among strains. The numbers within the forms indicate the number of shared genes.
doi:10.1371/journal.pone.0018551.g003

indispensable to the survival of pathogens. Moreover, some copies of these genes can be acquired by horizontal transfer. These genes are not ORFans; they already have been characterized in other species. The terminology 'lineage-specific' portrays only some genes found among the four strains in our study; the same genes may be found in other species.

We found 49 lineage-specific genes in CpC231 and 52 in Cp1002. For most of them, we did not have a descriptive characterization of their products, and they were classified as hypothetical proteins. In addition, many of these identified genes, in both strains, encode membrane and secreted proteins and pseudogenes. On the other hand, some well-characterized proteins were found in the genome. One example is found in CpC231, which has the gene called *pth*A; this gene encodes an effector system of type III secretion and is related to bacterial growth and host cell lesions, as found in *Xanthomonas campestris* [77]. This gene may be a good target for understanding the development of *C. pseudotuberculosis* CpC231 inside the host and the necrosis seen in CL abscesses, where it plays the same role in this pathogen.

In Cp1002, a very interesting gene was found, *tat*A, which encodes a membrane protein translocase, involved in the secretion of proteins in their final conformation, through the inner membrane to the extracellular environment. This gene is interesting because it is independent of the Sec secretion system and is a unique copy among the strains, suggesting that Cp1002 may have other routes for secretion. Regarding the large number of hypothetical proteins found in this strain, it may harbor genes that came from horizontal transfer, including some from phylogenetically-distant organisms, for which genomic molecular characterization has not been made.

Finally, lineage-specific genes may be good tools for understanding the host-pathogen interaction and may be good targets for the development of computational tools for differentiation between these strains, for molecular epidemiology.

Biochemical properties of *C. pseudotuberculosis*

In the latest review of the biochemical properties of *C. pseudotuberculosis* [76], Dorella and colleagues gathered information concerning its metabolism, virulence and pathogenesis. They reported that the peptidoglycan in the cell wall is based on meso-DAP acid, and that arabinose and galactose are major cell-wall sugars. Our analyses predicted all of the reactions of the peptidoglycan biosynthesis II pathway; the meso-DAP acid compound was found as a product/substrate of the reaction catalyzed by UDP-N-acetylmuramyl tripeptide synthase (6.3.2.13). The complete pathway of UDP-galactose biosynthesis was also found; although there was no evidence of biosynthesis of arabinose, we detected a membrane transporter, known as arabinose efflux permease.

We also found short-chain mycolic acids; 10 variations of acids of this type were encountered, including 6-O-cis-keto-mycolyl-trehalose-6-phosphate, and 6-O-mycolyl-trehalose-6-phosphate. The two strains of *C. pseudotuberculosis* showed considerable fermentation ability, with several fermentation pathways, including glycolysis III, mixed acid fermentation and pyruvate fermentation to acetate IV, ethanol I and lactate.

Several sugar degradation pathways were also found in the two strains of *C. pseudotuberculosis*, including galactose, lactose, sucrose and L-and D-arabinose degradation. We confirmed that, as reported by Dorella et al. (2006), all these pathways produce acids and no gasses, generating large amounts of energy.

It was also previously reported that *C. pseudotuberculosis* is phospholipase D and catalase positive. Our analysis showed that both phospholipase D and catalase are involved in important processes. The main molecular functions of phospholipase D are phospholipase D activity, magnesium ion binding, NAPE-specific phospholipase D activity and sphingomyelin phosphodiesterase D activity. Catalase, which is produced by the *cat* gene, is involved in response to oxidative stress and oxidation reduction. Although two enzymes of the denitrification pathway (nitrate reduction I) were found, absence of the remaining enzymes is probably the determining factor for the inability of these strains to reduce nitrate to N^2, as reported by Dorella et al. (2006).

We also detected iron acquisition genes (*fag*) A, B, C and D in both strains of *C. pseudotuberculosis* [78]. Genes *fagA* and *fagB* produce the integral membrane proteins FagA, an iron-enterobactin transporter, and FagBy; both have important roles, including ion, transmembrane, organic acid and protein transport. The ATP binding cytoplasmic membrane protein, FagC, produced by gene *fagC*, has two main molecular functions: ATP binding and ATPase activity. Finally, gene *fagD* produces the iron siderophore binding protein, FeAcquisition gene D, which has a role in iron ion transmembrane transport activity.

Computational reconstruction of the *C. pseudotuberculosis* pathways in our database not only allowed us to better visualize the metabolism of this bacterium, but also to compare it to closely related species. The main purpose of this analysis was to describe *C. pseudotuberculosis* metabolism by computational means, providing a predictive tool for "wet-lab" research.

Methods

Bacterial strains and growth conditions

Corynebacterium pseudotuberculosis 1002 biovar ovis (herein referred to as Cp1002) is a wild strain, isolated from a caprine host in Brazil. *Corynebacterium pseudotuberculosis* C231 biovar ovis (herein referred to as CpC231) is also a wild strain, isolated from an ovine host in Australia. Both strains were confirmed to be *C. pseudotuberculosis* by routine biochemical tests (API CORYNE, Biomerieux, Marcy l'Etoile, France). These strains were maintained in brain-heart-infusion broth (BHI – HiMedia Laboratories Pvt. Ltda, India) at 37°C, under rotation.

Preparation of high molecular weight DNA

Chromosomal DNA extraction was performed as follows: 50 mL of 48–72 h cultures of the two strains were centrifuged at 4°C and 2000 x *g* for 20 min. Cell pellets were re-suspended in 1 mL Tris/EDTA/NaCl [10 mM Tris/HCl (pH 7.0), 10 mM EDTA (pH 8.0), and 300 mM NaCl] and centrifuged again under the same conditions. Supernatants were discarded, and the pellets were re-suspended in 1 mL TE/lysozyme [25 mM Tris/HCl (pH 8.0), 10 mM EDTA (pH 8.0), 10 mM NaCl, and 10 mg lysozyme/mL]. Samples were then incubated at 37°C for 30 min. Thirty milliliters of 30% (w/v) sodium N-lauroyl-sarcosine (Sarcosyl) were added to each sample and the mixtures were incubated for 20 min at 65°C, followed by incubation for 5 min at 4°C. DNA was purified using phenol/chloroform/isoamyl alcohol (25:24:1) and precipitated with ethanol. DNA concentrations were determined spectrophotometrically, and the DNA was visualized in ethidium bromide-stained 0.7% agarose gels.

Construction of *Corynebacterium pseudotuberculosis* genomic libraries and Sanger sequencing

For the shotgun strategy used to sequence *C. pseudotuberculosis* 1002, four small fragment libraries were constructed using the TOPO Shotgun cloning kit and the pCR4 Blunt-TOPO vector (Invitrogen), according to the manufacturer's instructions. Sanger sequencing was carried out using the Minas Gerais Genome Network (http://rgmg.cpqrr.fiocruz.br). A total of 6,144 forward and reverse reads were produced using the DYEnamic Dye Terminator kit and run in a Megabace 1000 automated sequencer (GE Healthcare).

Genome Sequencing

Cp1002 was sequenced using both Sanger and pyrosequencing technologies. Pyrosequencing was carried out using 454 Life Sciences (Branford, CT). A total of 397,147 high quality reads and 86,154,153 high quality bases were obtained, which translates into approximately 31-fold coverage. The average length of the sequences was 253 bases. The sequences were delivered after quality filtering and preassembly with the Newbler assembler (454 Life Sciences).

CpC231 was sequenced with a Roche-454 FLX sequencer at the Australian Animal Health Laboratory, Geelong, Australia. A total of 347,361 reads generated 80,336,550 bases, giving 34-fold coverage of the genome. *De novo* assembly of the filtered sequence data was carried out using the Newbler software. This assembly produced 10 large contigs in four scaffolds. The remaining gaps in the genomic sequence were closed by PCR walking and Sanger sequencing of the resulting fragments.

Treatment and assembly data

The raw Sanger data obtained from sequencing were processed using the Phred-Phrap-Consed package [75]. Possible contaminants (plasmid DNA, sequences with similarity to vectors and other contaminants) were discarded using the Cross_match program (www.phrap.org). The quality value used in the base-calling program was Q = 40 (Probability of incorrect base call 1 in 10,000/base call accuracy 99.99%). An assembly using Phrap parameters (Force Level: 40 and Gap Length: 10,000) was carried out.

The 454 data were processed using the Newbler assembler (454 Life Sciences), and the final genomic consensus sequence was obtained using the Phrap algorithm.

Genome annotation

The annotation procedures involved the use of several algorithms in a multi-step process. Structural annotation was performed using the following software: FgenesB: gene predictor (www.softberry.com); RNAmmer: rRNA predictor [79]; tRNA-scan-SE: tRNA predictor [80]; and Tandem Repeat Finder: repetitive DNA predictor (tandem.bu.edu/trf/trf.html). Functional annotation was performed by similarity analyses, using public databases and InterProScan analysis [81]. Manual annotation was performed using Artemis [82].

Identification and confirmation of putative pseudogenes in the genome was carried out using Consed. Manual analysis was performed based on the Phred quality of each base in the frameshift area. This analysis enabled the identification of erroneous insertions or deletions of bases in the genome information produced by the sequencing process, and it avoided identification of false-positive pseudogenes.

Predictions of the cellular locations of *Corynebacterium* proteins were made using the program SurfG Plus (version 1.0), with a minimum protein size of 73 amino acids. Classification of predicted proteins in functional categories was made using the BLAST2GO program (www.blast2go.org). The cutoff value used was 10−6 (http://www.blast2go.org/).

In silico Identification of Pathogenicity Islands

In order to accurately identify and classify putative Pathogenicity Islands (PAIs) in the corynebacterial genomes, we developed a combined computational approach using several in-house scripts to integrate the prediction of diverse algorithms and databases, namely: Colombo-SIGIHMM [83], Artemis [82], tRNAscan-SE [80]; EMBOSS-geecee [84], ACT: the Artemis Comparison Tool [85], and mVIRdb [86].

In silico metabolic pathway construction

The two main data sources used for reconstructing the *C. pseudotuberculosis* metabolic pathways were the genome sequence file in FASTA format and the genome annotation file in GBK format. Metabolic pathways databases for strains 1002 and C231 were created using the Pathway tools 13 software, developed by SRI International [87]. The Pathway tools software contains algorithms that predict metabolic pathways of an organism from its genome by comparison to a reference pathways database known as MetaCyc [88]. Construction of a metabolic pathways database was done using BioCyc [89], in order to compare the different bacteria, *C. diphtheriae* NCTC 13129, *C. efficiens* YS-314, *C. glutamicum* ATCC 13032, and *C. jeikeium* K411, to the deduced *C. pseudotuberculosis* pathways.

Comparative analysis of *Corynebacterium pseudotuberculosis* strains

Comparative analyses were made for the two *C. pseudotuberculosis* strains. Similarity analyses of the two genomes were made using the BLAST - NCBI [90,91] and InterProScan databases. The Mauve algorithm (gel.ahabs.wisc.edu/mauve) and the ACT tool were used to identify whether blocks had undergone gene rearrangements or remained preserved. The Plotter program of the MUMMer 3.22 package (mummer.sourceforge.net) was used for synteny analysis.

Supporting Information

Table S1 Orthologous genes present inside PAIs regions of *C. pseudotuberculosis* and their counterparts in other Corynebacterium species.
(DOC)

Acknowledgments

The authors acknowledge the scientific support of two genomics networks: Rede Paraense de Genômica e Proteômica and Rede Genoma de Minas Gerais. In addition, the authors thank for valuable contribution of all Public Institutions involved and their respective co-authors during the development of the present work.

Author Contributions

Conceived and designed the experiments: JCR AS MPC RJM AM GCO VA. Performed the experiments: FAD SB MITF GCO AM VA VD EMC LMO MCP SRCD AFC JFA. Analyzed the data: JCR AS RJM GCO AM VA VD ARS FAD LGCP MZT NS TLPC JM AZ SCS SSA VACA DMR. Contributed reagents/materials/analysis tools: VA GCO GRF DOL ALP CUV CTG DCB DMO FRS EMR IMC JMO LVP LRG JAF MITF NPC PRKF SMRT SB SCO. Wrote the paper: JCR AS RJM GCO AM VA VD ARS FAD LGCP MZT NS TLPC JM AZ SCS SSA VACA DMR ET JB AT. Obtained permission for use of cell line: RJM RM AM VA. Bioinformatic support: JCR GCO AD FPL PG.

References

1. Ayers JL (1977) Caseous lymphadenitis in goat and sheep: review of diagnosis, pathogenesis, and immunity. JAVMA 171: 1251–1254.
2. Brown CC, Olander HJ, Alves SF (1987) Synergistic hemolysis-inhibition titers associated with caseous lymphadenitis in a slaughterhouse survey of goats and sheep in northeastern Brazil. Can J Vet Res 51: 46–49.
3. Merchant IA, Packer RA (1967) The genus *Corynebacterium*. In: Merchant IA, Packer RA, eds. Veterinary Bacteriology and Virology. USA: The Iowa State University Press. pp 425–440.
4. Piontkowski MD, Shivvers DW (1998) Evaluation of a commercially available vaccine against *Corynebacterium pseudotuberculosis* for use in sheep. J Am Vet Med Assoc 212: 1765–1768.
5. Join-Lambert OF, Ouache M, Canioni D, Beretti JL, Blanche S, et al. (2006) *Corynebacterium pseudotuberculosis* necrotizing lymphadenitis in a twelve-year old patient. Pediatr Infect Dis J 25(9): 848–851.
6. Trost E, Ott L, Schneider J, Schroder J, Jaenicke S, et al. (2010) The complete genome sequence of *Corynebacterium pseudotuberculosis FRC41* isolated from a 12-year-old girl with necrotizing lymphadenitis reveals insights into gene-regulatory networks contributing to virulence. BMC Genomics 11(1): 728–745.
7. Connor KM, Quirie MM, Baird G, Donachie W (2000) Characterization of united kingdom isolates of *Corynebacterium pseudotuberculosis* using pulsed-field gel electrophoresis. J Clin.Microbiol 38: 2633–2637.
8. Ben Saïd MS, Ben Maitigue H, Benzarti M, Messadi L, Rejeb A, et al. (2002) Epidemiological and clinical studies of ovine caseous lymphadenitis. Arch Inst Pasteur Tunis 79: 51–57.
9. Binns SH, Bailey M, Green LE (2002) Postal survey of ovine caseous lymphadenitis in the United Kingdom between 1990 and 1999. Vet Rec 150: 263–268.
10. Arsenault J, Girard C, Dubreuil P, Daignault D, Galarneau JR, et al. (2003) Prevalence of and carcass condemnation from maedi-visna, paratuberculosis and caseous lymphadenitis in culled sheep from Quebec, Canada. Prev Vet Med 59: 67–81.
11. Paton MW, Walker SB, Rose IR, Watt GF (2003) Prevalence of caseous lymphadenitis and usage of caseous lymphadenitis vaccines in sheep flocks. Aust Vet J 81: 91–95.
12. Pinheiro RR, Gouveia AMG, Alves FSF, Haddad JP (2000) Aspectos epidemiológicos da caprinocultura cearense. Arquivo Brasileiro de Medicina Veterinária e Zootecnia 52: 534–543.
13. Guimarães AS, Seyffert N, Portela RWD, Meyer R, Carmo FB, et al. (2009) Caseous lymphadenitis in sheep flocks of the state of Minas Gerais, Brazil: prevalence and management surveys. Small Ruminants Research 87(1): 86–91.
14. Seyffert N, Guimarães AS, Pacheco LGC, Portela RW, Bastos BL, et al. (2010) High seroprevalence of caseous lymphadenitis in brazilian goat herds revealed by *Corynebacterium pseudotuberculosis* secreted proteins-based ELISA. Res Vet Sci 88: 50–55.
15. Eggleton DG, Middleton HD, Doidge CV, Minty DW (1991) Immunisation against ovine caseous lymphadenitis: comparison of *Corynebacterium pseudotuberculosis* vaccines with and without bacterial cells. Aust Vet J 68: 317–319.
16. Stoops SG, Renshaw HW, Thilsted JP (1984) Ovine caseous lymphadenitis: disease prevalence, lesion distribution, and thoracic manifestations in a population of mature culled sheep from western United States. Am J Vet Res 45(3): 557–61.
17. Ribeiro MG, Júnior JGD, Paes AC, Barbosa PG, Júnior GN, et al. (2001) Punção aspirativa com agulha fina no diagnóstico de *Corynebacterium pseudotuberculosis* na linfadenite caseosa caprina. Arq Inst Biol 68: 23–28.
18. Dorella FA, Estevam EM, Pacheco LGC, Guimarães CT, Lana UGP, et al. (2006) In vivo insertional mutagenesis in *Corynebacterium pseudotuberculosis*: an efficient means to identify DNA sequences encoding exported proteins. Appl Environ Microbiol 72: 7368–7372.

19. Merhej V, Royer-Carenzi M, Pontarotti P, Raoult D (2009) Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. Biol Direct 4: 13–37.
20. Webb SAR, Karleh CM (2008) Bench-to-bedside review: Bacterial virulence and subversion of host defences. Critical Care 12: 234–241.
21. Dobrindt U, Hentschel U, Kaper JB, Hacker J (2002) Genome plasticity in pathogenic and nonpathogenic enterobacteria. Curr Top Microbiol Immunol 264: 157–175.
22. Hall BG, Ehrlich GD, Hu FZ (2010) Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing. Microbiology 156(4): 1060–8.
23. Silva A, Schneider MPC, Cerdeira L, Barbosa MS, Ramos RTJ, et al. (2011) Complete genome sequence of *Corynebacterium pseudotuberculosis* I19, a strain isolated from a cow in Israel with bovine mastitis. J Bacteriol 193(1): 323–324.
24. Nakamura Y, Nishio Y, Ikeo K, Gojobori T (2003) The genome stability in *Corynebacterium* species due to lack of the recombinational repair system. Gene 317: 149–155.
25. Tauch A, Kaiser O, Hain T, Goesmann A, Weisshaar B, et al. (2005) Complete genome sequence and analysis of the multiresistant nosocomial pathogen *Corynebacterium jeikeium* K411, a lipid-requiring bacterium of the human skin flora. J Bacteriol 187: 4671–4682.
26. Khamis A, Raoult D, La Scola B (2004) rpoB gene sequencing for identification of *Corynebacterium* species. J Clin Microbiol 42(9): 3925–31.
27. Khamis A, Raoult D, La Scola B (2005) Comparison between *rpoB* and 16S rRNA gene sequencing for molecular identification of 168 clinical isolates of *Corynebacterium*. J Clin Microbiol 43: 1934–1936.
28. Dobrindt U, Hochhut B, Hentschel U, Hacker J (2004) Genomic islands in pathogenic and environmental microorganisms. Nat Rev Microbiol 2: 414–424.
29. Karaolis DK, Johnson JA, Bailey CC, Boedeker EC, Kaper JB, et al. (1998) A *Vibrio cholerae* pathogenicity island associated with epidemic and pandemic strains. Proc Natl Acad Sci U.S.A 95: 3134–3139.
30. Schumann W (2007) Thermosensors in eubacteria: role and evolution. J Biosci 32: 549–557.
31. Hentschel U, Hacker J (2001) Pathogenicity islands: the tip of the iceberg. Microbes Infect 3: 545–548.
32. McNamara PJ, Cuevas WA, Songer JG (1995) Toxic phospholipases D of *Corynebacterium pseudotuberculosis*, *C. ulcerans* and *Arcanobacterium haemolyticum*: cloning and sequence homology. Gene 156: 113–118.
33. Moomaw AS, Maguire ME (2008) The unique nature of Mg2+ channels. Physiology (Bethesda) 23: 275–285.
34. O'Connor K, Fletcher SA, Csonka LN (2009) Increased expression of Mg(2+) transport proteins enhances the survival of *Salmonella enterica* at high temperature. Proc Natl Acad Sci U.S.A 106: 17522–17527.
35. Schönert S, Buder T, Dahl MK (1998) Identification and enzymatic characterization of the maltose-inducible alpha-glucosidase mall (sucrase-isomaltase-maltase) of *Bacillus subtilis*. J Bacteriol 180: 2574–2578.
36. Yamamoto H, Serizawa M, Thompson J, Sekiguchi J (2001) Regulation of the glv operon in *Bacillus subtilis*: yfia (*glvR*) is a positive regulator of the operon that is repressed through *ccpA* and *cre*. Bacteriol 183: 5110–5121.
37. May T, Ito A, Okabe S (2009) Induction of multidrug resistance mechanism in *Escherichia coli* biofilms by interplay between tetracycline and ampicillin resistance genes. Antimicrob Agents Chemother 53: 4628–4639.
38. Smith I (2003) *Mycobacterium tuberculosis* pathogenesis and molecular determinants of virulence. Clin Microbiol Rev 16: 463–496.
39. Saïd-Salim B, Mostowy S, Kristof AS, Behr MA (2006) Mutations in *Mycobacterium tuberculosis* RV0444c, the gene encoding anti-sigK, explain high level expression of *mpb*70 and *mpb*83 in *Mycobacterium bovis*. Mol Microbiol 62: 1251–1263.
40. Veyrier F, Saïd-Salim B, Behr MA (2008) Evolution of the mycobacterial sigK regulon. J Bacteriol 190: 1891–1899.
41. Vassylyev DG, Tomitori H, Kashiwagi K, Morikawa K, Igarashi K (1998) Crystal structure and mutational analysis of the *Escherichia coli* putrescine receptor: structural basis for substrate specificity. J Biol Chem 273: 17604–17609.
42. Carlson PEJ, Horzempa J, O'Dee DM, Robinson CM, Neophytou P, et al. (2009) Global transcriptional response to spermine, a component of the intramacrophage environment, reveals regulation of *Francisella* gene expression through insertion sequence elements. J Bacteriol 191: 6855–6864.
43. Enkavi G, Tajkhorshid E (2010) Simulation of spontaneous substrate binding revealing the binding pathway and mechanism and initial conformational response of *glp*T. Biochemistry 49: 1105–1114.
44. Pérez E, Samper S, Bordas Y, Guilhot C, Gicquel B, et al. (2001) An essential role for *pho*P in *Mycobacterium tuberculosis* virulence. Mol Microbiol 41(1): 179–87.
45. Soto CY, Menéndez MC, Pérez E, Samper S, Gómez AB, et al. (2004) IS*6110* Mediates Increased Transcription of the *phoP* Virulence Gene in a Multidrug-Resistant Clinical Isolate Responsible for Tuberculosis Outbreaks. J Clin Microb 42(1): 212–219.
46. Aguilar D, Infante E, Martin C, Gormley E, Gicquel G, Pando RH (2006) Immunological responses and protective immunity against tuberculosis conferred by vaccination of Balb/C mice with the attenuated *Mycobacterium tuberculosis* (*phoP*) SO2 strain. Clin Exper Immunol 147: 330–338.
47. Gonzalo-Asensio J, Mostowy S, Harders-Westerveen J, Huygen K, Hernández-Pando R, et al. (2008) PhoP: A Missing Piece in the Intricate Puzzle of *Mycobacterium tuberculosis* Virulence. PLoS ONE 3(10): 1–11.
48. Carson SD, Klebba PE, Newton SM, Sparling PF (1999) Ferric enterobactin binding and utilization by *Neisseria gonorrhoeae*. J Bacteriol 181: 2895–2901.
49. Kunkle CA, Schmitt MP (2005) Analysis of a *dtxR*-regulated iron transport and siderophore biosynthesis gene cluster in *Corynebacterium diphtheriae*. J Bacteriol 187: 422–433.
50. Shimizu T, Okabe A, Minami J, Hayashi H (1991) An upstream regulatory sequence stimulates expression of the perfringolysin o gene of *Clostridium perfringens*. Infect Immun 59: 137–142.
51. Urbina P, Flores-Díaz M, Alape-Girón A, Alonso A, Goni FM (2009) Phospholipase C and sphingomyelinase activities of the *Clostridium perfringens* alpha-toxin. Chem Phys Lipids 159: 51–57.
52. Selby T, Allaker RP, Dymock D (2003) Characterization and expression of adjacent proline iminopeptidase and aspartase genes from *Eikenella corrodens*. Oral Microbiol Immunol 18: 256–259.
53. Zhang L, Jia Y, Wang L, Fang R (2007) A proline iminopeptidase gene upregulated in planta by a *lux*R homologue is essential for pathogenicity of *Xanthomonas campestris* pv. campestris. Mol Microbiol 65: 121–136.
54. Böltner D, MacMahon C, Pembroke JT, Strike P, Osborn AM (2002) R391: a conjugative integrating mosaic comprised of phage, plasmid, and transposon elements. J Bacteriol 184: 5158–5169.
55. Burne RA, Chen YY (2000) Bacterial ureases in infectious diseases. Microbes Infect 2: 533–542.
56. Huntley RP, Binns D, Dimmer E, Barrell D, O'Donavan C, et al. (2009) QuickGO: a user tutorial for the web-based Gene Ontology browser. Database 10: 1–19.
57. Barinov A, Loux V, Hammani A, Nicolas P, Langella P, et al. (2009) Prediction of surface exposed proteins in *Streptococcus pyogenes*, with a potential application to other gram-positive bacteria. Proteomics 9: 61–73.
58. Song C, Kumar A, Saleh M (2009) Bioinformatic comparison of bacterial secretomes. *Genomics Proteomics* Bioinformatics 7: 37–46.
59. Wooldridge L, Lissina A, Cole DK, van den Berg HA, Price DA, Sewell AK (2009) Tricks with tetramers: how to get the most from multimeric peptide-MHC. Immunology 126: 147–164.
60. Wittmann C, Kiefer P, Zelder O (2004) Metabolic Fluxes in *Corynebacterium glutamicum* during Lysine Production with Sucrose as Carbon Source. Applied and Environmental Microbiology 70: 7277–7287.
61. Kabus A (2007) Expression of the *Escherichia coli pntAB* genes encoding a membrane-bound transhydrogenase in *Corynebacterium glutamicum* improves l-lysine formation. Appl Microbiol Biotechnol 75: 47–53.
62. Blombach B, Arndt A, Auchter M, Eikmanns BJ (2009) L-Valine production during growth of pyruvate dehydrogenase complex-deficient *Corynebacterium glutamicum* in the presence of ethanol or by inactivation of the transcriptional regulator SugR. Appl Environ Microbiol 75: 1197–1200.
63. Klaffl S, Eikmanns BJ (2010) Genetic and Functional Analysis of the Soluble Oxaloacetate Decarboxylase from *Corynebacterium glutamicum*. Journal of Bacteriology 192: 2604–2612.
64. Wittmann C, Heinzle E (2001) Modeling and experimental design for metabolic flux analysis of lysine-producing *Corynebacteria* by mass spectrometry. Metab Eng 3(2): 173–91.
65. Parche S, Thomae AW, Schlicht M, Titgemeyer F (2001) *Corynebacterium diphtheriae*: a PTS View to the Genome. J Mol Microbiol Biotechnol 3(3): 415–422.
66. Rübenhagen R, Rönsch H, Jung H, Krämer R, Morbach S (2000) Osmosensor and osmoregulator properties of the betaine carrier *betP* from *Corynebacterium glutamicum* in proteoliposomes. J Biol Chem 275: 735–741.
67. Rittmann D, Schaffer S, Wendisch VF, Sahm H (2003) Fructose-1,6-bisphosphatase from *Corynebacterium glutamicum*: expression and deletion of the *fbp* gene and biochemical characterization of the enzyme. Arch Microbiol 180: 285–292.
68. Kiefer P, Heinzle E, Zelder O, Wittmann Z (2004) Comparative Metabolic Flux Analysis of Lysine-Producing *Corynebacterium glutamicum* Cultured on Glucose or Fructose. Applied and Environmental Microbiology 70: 229–239.
69. Rumbold K, Buijsen HJJ, Overkamp KM, Groenestijn JW, Punt PJ, Werf MJ (2009) Microbial production host selection for converting second-generation feedstocks into bioproducts. Microbial Cell Factories 8: 1–11.
70. Casalino M, Prosseda G, Barbagallo M, Lacobino A, Ceccarini P, et al. (2010) Interference of the *cadC* regulator in the arginine-dependent acid resistance system of *Shigella* and enteroinvasive *E. coli*. Int J Med Microbiol 300(5): 289–95.
71. Alvarez-Ordóñez A, Fernández A, Bernardo A, López M (2010) Arginine and lysine decarboxylases and the acid tolerance response of *Salmonella typhimurium*. Int J Food Microbiol 136: 278–282.
72. Hodgson AL, Carter K, Tachedjian M, Krywult J, Corner LA, et al. (1999) Efficacy of an ovine caseous lymphadenitis vaccine formulated using a genetically inactive form of the *Corynebacterium pseudotuberculosis* Phospholipase D. Vaccine 17: 802–808.
73. D'Afonseca V, Moraes PM, Dorella FA, Pacheco LGC, Meyer R, et al. (2008) A description of genes of *Corynebacterium pseudotuberculosis* useful in diagnostics and vaccine applications. Genet Mol Res 7: 252–260.
74. Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, et al. (2001) Massive gene decay in the leprosy bacillus. Nature 409: 1007–1011.
75. Rasko DA, Rosovitz MJ, Garry SA, Emmanuel FM, Fricke WF, et al. (2008) The pan-genome structure of *Escherichia coli:* comparative genomic analysis of *E. coli* commensal and pathogenic isolates. Journal of Bacteriology 190(20): 6881–6893.

76. Dorella FA, Pacheco LGC, Oliveira SC, Miyoshi A, Azevedo V (2006) *Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. Vet Res 37: 201–218.
77. Shiotani H, Yoshioka T, Yamamoto M, Matsumoto R (2008) Susceptibility to citrus canker caused by *Xanthomonas axonopodis* pv. citri depends on the nuclear genome of the host plant. J Gen Plant Pathol 74: 133–137.
78. Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8: 186–194.
79. Lagesen K, Hallin P, Rødland EA, Staerfeldt H, Rognes T, et al. (2007) Rnammer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35: 3100–3108.
80. Lowe TM, Eddy SR (1997) Trnascan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25: 955–964.
81. Zdobnov EM, Apweiler R (2001) Interproscan--an integration platform for the signature-recognition methods in INTERPRO. Bioinformatics 17: 847–848.
82. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, et al. (2000) Artemis: sequence visualization and annotation. Bioinformatics 16: 944–945.
83. Waack S, Keller O, Asper R, Brodag T, Damm C, et al. (2006) Score-based prediction of genomic islands in prokaryotic genomes using hidden markov models. BMC Bioinformatics 7: 142.
84. Rice P, Longden I, Bleasby A (2000) Emboss: the European molecular biology open software suite. Trends Genet 16: 276–277.
85. Carver TJ, Rutherford KM, Berriman M, Rajandream M, Barrell BG, et al. (2005) ACT: the Artemis Comparison Tool. Bioinformatics 21: 3422–3423.
86. Zhou CE, Smith J, Lam M, Zemla A, Dyer MD, et al. (2007) Mvirdb--a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. Nucleic Acids Res 35: D391–394.
87. Karp PD, Paley S, Romero P (2002) The pathway tools software. Bioinformatics 18(1): S225–32.
88. Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, et al. (2008) The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. Nucleic Acids Res 36: D623–31.
89. Caspi R, Karp PD (2007) Using the metacyc pathway database and the biocyc database collection. Curr Protoc Bioinformatics Chapter 1: Unit1.17.
90. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.
91. Krauthammer M, Rzhetsky A, Morozov P, Friedman C (2000) Using BLAST for identifying gene and protein names in journal articles. Gene 259: 245–252.

II.I. 3 Genome sequence of *Corynebacterium pseudotuberculosis* biovar equi strain 258 and prediction of antigenic targets to improve biotechnological vaccine production.
Soares SC, Trost E, Ramos RT, Carneiro AR, Santos AR, Pinto AC, Barbosa E, Aburjaile F, Ali A, Diniz CA, Hassan SS, Fiaux K, Guimarães LC, Bakhtiar SM, Pereira U, Almeida SS, Abreu VA, Rocha FS, Dorella FA, Miyoshi A, Silva A, **Azevedo V**, Tauch A.
J Biotechnol. 2013 Aug 20;167(2):135-41. doi: 10.1016/j.jbiotec.2012.11.003. Epub 2012 Nov 29.

Em trabalho anterior de nosso grupo, foram descritas 7 Ilhas de Patogenicidade (PAIs) em *C. pseudotuberculosis* 1002 e C231, ambas do biovar *ovis*. No processo de sequenciar os 15 genomas de *C. pseudotuberculosis*, nosso grupo foi confrontado com a necessidade de caracterizar as linhagens do biovar *equi* e encontrar novas PAIs e alvos vacinais que pudessem elicitar uma resposta imune contra ambos os biovares, *ovis* e *equi*. Neste artigo, foi utilizado o software PIPS para predizer 4 PAIs adicionais (PICP 8-11) em *C. pseudotuberculosis* 258, biovar *equi*, que estão de acordo com dados encontrados em um trabalho paralelo com *C. pseudotuberculosis* 316, biovar *equi*. Além disso, foram observados padrões específicos de deleções em PAIs de ambas linhagens do biovar *equi* quando comparadas com *C. pseudotuberculosis* 1002, biovar *ovis*. Finalmente, foi aplicada a estratégia de vacinologia reversa, em uma abordagem de genômica subtrativa, para identificar proteínas conservadas entre os biovares *ovis* e *equi* que possam ser reconhecidas pelo sistema imune.

Accepted Manuscript

Title: Genome sequence of Corynebacterium pseudotuberculosis biovar equi strain 258 and prediction of antigenic targets to improve biotechnological vaccine production

Authors: Siomar C. Soares, Eva Trost, Rommel T.J. Ramos, Adriana R. Carneiro, Anderson R. Santos, Anne C. Pinto, Eudes Barbosa, Flávia Aburjaile, Amjad Ali, Carlos A.A. Diniz, Syed S. Hassan, Karina Fiaux, Luis C. Guimarães, Syeda M. Bakhtiar, Ulisses Pereira, Sintia S. Almeida, Vinícius A.C. Abreu, Flávia S. Rocha, Fernanda A. Dorella, Anderson Miyoshi, Artur Silva, Vasco Azevedo, Andreas Tauch

PII:	S0168-1656(12)00708-0
DOI:	doi:10.1016/j.jbiotec.2012.11.003
Reference:	BIOTEC 6222
To appear in:	*Journal of Biotechnology*
Received date:	4-6-2012
Revised date:	3-11-2012
Accepted date:	8-11-2012

Please cite this article as: Soares, S.C., Trost, E., Ramos, R.T.J., Carneiro, A.R., Santos, A.R., Pinto, A.C., Barbosa, E., Aburjaile, F., Ali, A., Diniz, C.A.A., Hassan, S.S., Fiaux, K., Guimarães, L.C., Bakhtiar, S.M., Pereira, U., Almeida, S.S., Abreu, V.A.C., Rocha, F.S., Dorella, F.A., Miyoshi, A., Silva, A., Azevedo, V., Tauch, A., Genome sequence of Corynebacterium pseudotuberculosis biovar equi strain 258 and prediction of antigenic targets to improve biotechnological vaccine production, *Journal of Biotechnology* (2010), doi:10.1016/j.jbiotec.2012.11.003

Highlights

Corynebacterium pseudotuberculosis strains 258 and CIP 52.97, biovar *equi*, present 11 pathogenicity islands.

Three pathogenicity islands of biovar *equi* strains present large deletions compared with biovar *ovis* strains.

49 proteins were predicted as antigenic in *C. pseudotuberculosis* strains 1002, CIP52.97 and 258.

Putative antigenic proteins are involved in genome rearrangement and survival under stress conditions.

Genome sequence of *Corynebacterium pseudotuberculosis* biovar *equi* strain 258 and prediction of antigenic targets to improve biotechnological vaccine production

Siomar C. Soares [a,b,c,*], Eva Trost [a,b], Rommel T. J. Ramos [d], Adriana R. Carneiro [d], Anderson R. Santos [c], Anne C. Pinto [c], Eudes Barbosa [c], Flávia Aburjaile [c], Amjad Ali [c], Carlos A. A. Diniz [c], Syed S. Hassan [c], Karina Fiaux [c], Luis C. Guimarães [c], Syeda M. Bakhtiar [c], Ulisses Pereira [c], Sintia S. Almeida [c], Vinícius A. C. Abreu [c], Flávia S. Rocha [c], Fernanda A. Dorella [c], Anderson Miyoshi [c], Artur Silva [d], Vasco Azevedo [c,1], Andreas Tauch [b,1]

[a] *CLIB Graduate Cluster Industrial Biotechnology, Centrum für Biotechnologie, Universität Bielefeld, 33615 Bielefeld, Germany*

[b] *Institut für Genomforschung und Systembiologie, Centrum für Biotechnologie, Universität Bielefeld, 33615 Bielefeld, Germany*

[c] *Laboratório de Genética Celular e Molecular, Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Pampulha, Belo Horizonte, MG, Brazil*

[d] *Instituto de Ciências Biológicas, Universidade Federal do Pará, Guamá, Belém, PA, Brazil*

* Corresponding author. Tel.: +49 (521) 106-12253; fax: +49 (521) 106-890415.

E-mail address: siomars@gmail.com

[1] These authors share the senior authorship.

ABSTRACT

Corynebacterium pseudotuberculosis is the causative agent of several veterinary diseases in a broad range of economically important hosts, which can vary from caseous lymphadenitis in sheep and goats (biovar *ovis*) to ulcerative lymphangitis in cattle and horses (biovar *equi*). Existing vaccines against *C. pseudotuberculosis* are mainly intended for small ruminants and, even in these hosts, they still present remarkable limitations. In this study, we present the complete genome sequence of *C. pseudotuberculosis* biovar *equi* strain 258, isolated from a horse with ulcerative lymphangitis. The genome has a total size of 2,314,404 bp and contains 2,088 predicted protein-coding regions. Using *in silico* analysis, eleven pathogenicity islands were detected in the genome sequence of *C. pseudotuberculosis* 258. The application of a reverse vaccinology strategy identified 49 putative antigenic proteins, which can be used as candidate vaccine targets in future works.

Keywords: Corynebacterium pseudotuberculosis; Caseous lymphadenitis; Genome sequence; Pathogenicity island; Reverse vaccinology

1. Introduction

Corynebacterium pseudotuberculosis is a Gram-positive, non-motile, pleomorphic, and facultative anaerobic bacterium of the *Actinomycetales* order (Jones and Collins, 1986). It is a facultative intracellular microorganism that can proliferate inside macrophages (Dorella et al., 2006). The taxonomic identification of *C. pseudotuberculosis* is mainly performed taking into account morphological and biochemical features (Jones and Collins, 1986), and through the use of nitrate reduction tests to classify the species into the biovars *equi* (positive nitrate reduction) and *ovis* (negative nitrate reduction) (Biberstein et al., 1971).

C. pseudotuberculosis biovar *ovis* is the causative agent of caseous lymphadenits (CLA), a disease with high economic importance in respect to goat- and sheep-raising. CLA causes less wool production at shearing and reduced prices at the abattoir due to weight loss and carcass condemnation (Hodgson et al., 1999). The disease has a worldwide incidence and presents a high prevalence in meat-producing countries like Australia, New Zealand, South Africa, United States, Canada, and Brazil (Arsenault et al., 2003; Dorella et al., 2006; Paton et al., 2003). The main reason for the wide spread of CLA is related to the high resistance of the bacteria to low temperatures and humid places and the ability to promptly invade animals through skin lesions (Augustine and Renshaw, 1986; Yeruham et al., 2004). Moreover, the visceral form of the disease is normally detected only in slaughter houses, which contributes to the very low detection rate of CLA (Yeruham et al., 2003). Finally, the wide spread of CLA also results from the inability of antibiotics to reach the bacteria due to the abscess capsule and the intra-macrophagic lifestyle (Williamson, 2001). Infections of horses by *C. pseudotuberculosis* biovar *equi* appear as external abscesses, ulcerative lymphangitis, and in a visceral form affecting internal organs (Aleman et al., 1996; Pratt et al., 2005).

Due to the high veterinary importance of *C. pseudotuberculosis*, and having in mind the inefficiency of antibiotics, several vaccine strategies have already been developed, including the use of attenuated or inactivated bacteria, cell wall fractions, and DNA vaccines (Dorella et al., 2009). Current vaccines are mainly based on formalin-inactivated phospholipase D (PLD), the major protective antigen of *C. pseudotuberculosis* and a virulence factor, which promotes the dissemination of the pathogen by triggering vascular permeabilization, hemolysis, and probably vacuole membrane disruption (Hodgson et al., 1999; Selvy et al., 2011). However, although vaccine strategies exist, vaccinated animals present variable protection levels; not all vaccines available for use in sheep have the same efficiency in goats; they are not licensed in all countries; and they still present side effects (Brogden et al., 1990; Dorella et al., 2009; Eggleton et al., 1991; Ellis, 1991; Holstad, 1989; LeaMaster et al., 1987; Windsor, 2011). Moreover, although many potential targets of *C. pseudotuberculosis* biovar *ovis* have been identified based on reverse vaccinology in literature (Barh et al., 2011), there is still a lack of research targeting diseases caused by *C. pseudotuberculsois* biovar *equi*. Animals infected by *C. pseudotuberculosis* biovar *equi* present cross-immunity to *C. pseudotuberculosis* biovar *ovis* strains, but the opposite has not been observed (Barakat et al., 1984; Biberstein et al., 1971; Steinman et al., 1999). All these factors point to the need for better characterizing virulence factors of *C. pseudotuberculosis* biovar *equi* and performing comprehensive comparisons of virulence factors from both biovars for the development of new vaccine strategies, which are able to protect not only small ruminants, but also horses and cattle.

In this work, we describe the sequencing of *C. pseudotuberculosis* biovar *equi* strain 258, isolated from a horse with ulcerative lymphangitis in Belgium. Furthermore, we compare this strain with *C. pseudotuberculosis* biovar *equi* strain CIP52.97 (Cerdeira et al., 2011b) and *C. pseudotuberculosis* biovar *ovis* strain 1002 (Ruiz et al., 2011), aiming to find new targets, which can be used in vaccine strategies against the different diseases caused by the species.

2. Material and methods

2.1. Genome sequencing of C. pseudotuberculosis *258*

The genome sequence of *C. pseudotuberculosis* 258 was obtained by sequencing a fragment library with the next-generation genome sequencer SOLiD v3. The generated reads were submitted to a quality filter using the software Quality Assessment (Ramos et al., 2011), where reads with a medium quality below phred 20 were discarded. The software SAET was then used to perform error corrections (http://solidsoftwaretools.com/gf/project/saet), thereby selecting reads with high quality scores. These reads were submitted to *de novo* assemblying with the assemblers Velvet (Zerbino and Birney, 2008) and Edena (Hernandez et al., 2008), which perform data processing based on Eurelian path and overlap-layout-consensus methods, respectively. As the resulting contigs contain data from two different methodologies, the software Simplifier (https://sourceforge.net/projects/simplifier/) removed redundant sequences, aiming to facilitate the subsequent manual curation of the genome sequence. Contig orientation and ordering were performed in two steps: first, the contigs were subjected to BLASTN genome comparisons with the reference strain *C. pseudotuberculosis* FRC41 (Trost et al., 2010) as described previously (Cerdeira et al., 2011a); and second, the alignments were uploaded into the software G4ALL (http://sourceforge.net/projects/g4all/) for manual curation and contig extension, resulting in a scaffold sequence. Finally, the software CLC BIO (http://www.clcbio.com/) was used to align short reads (50 bp) with the draft genome in a recursive manner to perform the final gap closure and to generate the complete genome sequence (Tsai et al., 2010).

2.2. Genome annotation and curation

The complete genome sequence of *C. pseudotuberculosis* 258 was functionally annotated using the following softwares: FgenesB (http://linux1.softberry.com/); RNAmmer (Lagesen et al., 2007); tRNAscan-SE (Lowe and Eddy, 1997); InterproScan (Zdobnov and Apweiler, 2001); Artemis and non-redundant proteins database for manual annotation and curation of coding sequences (Rutherford et al., 2000). The genome sequence of *C. pseudotuberculosis* 258 has been deposited in the GenBank database with accession number CP003540.

2.3. Genome plasticity analysis of C. pseudotuberculosis *258*

The identification of pathogenicity islands in the genome of *C. pseudotuberculosis* 258 was performed with PIPS through the detection of regions presenting deviations in genomic signatures and absence in the non-pathogenic organism *C. glutamicum* ATCC 13032 (Soares et al., 2012). The plasticity comparison between strain 258 and *C. pseudotuberculosis* 1002 (CP001809), *C. pseudotuberculosis* CIP52.97 (CP003061), *C. diphtheriae* NCTC 13129 (BX248353), *C. ulcerans* BR-AD22 (CP002791), and *C. glutamicum* ATCC 13032 (BX927147) was performed with the software BRIG (Alikhan et al., 2011). All genome sequences were retrieved from the GenBank database.

2.4. Prediction of putative antigenic targets of C. pseudotuberculosis

The published program Vaxign (He et al., 2010) was used for the prediction of vaccine targets. Vaxign performs a dynamic vaccine target prediction based on input sequences. The utility of this program was demonstrated by predicting vaccine candidates against uropathogenic *Escherichia coli* (UPEC). The identification of genes coding for antigenic proteins was performed using reverse vaccinology and the following rules: (I) The most antigenic proteins are normally those that are somehow exposed to the host and can be promptly recognized by the immune system, like secreted proteins, surface-exposed proteins, and membrane proteins (Rappuoli, 2001); (II) MHC I and II binding properties with adhesion probability greater than 0.51 and absence of similarity to mammalian proteins (He et al., 2010); (III) protein conservation among different genomes, in this case biovar *equi* and *ovis* strains (He et al., 2010); (IV) virulence factors are better targets and are often encoded in pathogenicity islands (Rappuoli, 2001). Therefore, proteins encoded by shared pathogenicity islands are appropriate candidates, but this rule does not exclude the targets from step III.

As for the rule I, the subcellular location of predicted proteins of *C. pseudotuberculosis* strains 1002, CIP 52.97, and 258 was identified by the use of the SurfG+ software, which classifies proteins according to the presence or absence of signal peptides, retention signals, and transmembrane helices (Barinov et al., 2009). A prerequisite for SurfG+ to better differentiate integral membrane proteins from potentially surface exposed proteins is the use of cell wall measures, which were obtained in this study by electron microscopy with an EM10A equipment (Zeiss). Briefly, *C. pseudotuberculosis* strains were grown in 100 mL of Brain Heart Infusion broth for 48h and centrifuged. The resulting pellet (~500 µl) was poured into an Eppendorf tube, fixed in 2.5% gluteraldehyde in 0.1 M sodium cacodylate buffer (pH 7.2) for 6 h at 8ºC and washed 3 times with 0.1 M sodium cacodylate buffer (pH 7.2). After buffer washing, the sample was post-fixed in 1% osmium tetroxide in 0.1 M sodium cacodylate buffer (pH 7.2) + 1.5% potassium ferrocyanide for 90 min, washed with 0.1 M sodium cacodylate buffer (pH 7.2), dehydrated in graded ethanol (50% EtOH, 70% EtOH, 95% EtOH, and 100% EtOH) and embedded in Eponate-Araldite resin. Ultrathin sections were obtained using uranyl acetate and lead citrate and, posteriorly, examined in a Zeiss-EM-10A (Melo et al., 1993). The micrographs were obtained using a CCD Mega view III camera.

The candidate proteins predicted by SurfG+ were analyzed by the software Vaxign (He et al., 2010) in order to apply rule II. As the aim of this work was to search for vaccine candidates common in both biovars (*equi* and *ovis*), the predicted proteomes were screened for proteins that are potentially antigenic in all three strains (rule III). To achieve this goal, we used the Artemis Comparison Tool (Carver et al., 2005) with BLAST alignment comparison files and searched for antigenic proteins that present more than 70% similarity in 70% of their extension in all three strains. Finally, as for the rule IV, we screened for antigenic targets harbored by shared pathogenicity islands in the three strains.

3. Results and discussion

3.1. General features of the C. pseudotuberculosis *258 genome*

The sequencing of genomic DNA of *C. pseudotuberculosis* 258 produced a total of 70,521,987 reads with a size of 50 bp. After quality filtering and error correction, 40,589,132 reads with high quality scores were selected, corresponding to an 868× genome coverage when compared to the 2,3 Mb genome sequence of the reference strain *C. pseudotuberculosis* FRC41 (Trost et al., 2010). The reads were submitted to *de novo* assemblying with Velvet and Edena, generating 8,004 contigs. Redundant sequences were then removed, reducing the number of contigs to 2,289. The reference genome of *C. pseudotuberculosis* FRC41 was used for subsequent contig orientation and ordering, resulting in 655 arranged genomic sequences. After gap closure with CLC BIO, a complete genome sequence of *C. pseudotuberculosis* 258 was generated, consisting in size of 2,314,404 bp with a G+C content of 52.15% (Fig. 1). According to the manual annotation, the genome of *C. pseudotuberculosis* 258 contains 2,088 protein-coding genes, 4 rRNA operons, 49 tRNA genes, and 46 pseudogenes (Table 1). These data are in the range known from the genome analyses of *C. pseudotuberculosis* 1002 and *C. pseudotuberculosis* CIP 52.97 (Table 1).

3.2. Detection of pathogenicity islands in C. pseudotuberculosis *258*

Appropriate candidates for the development of vaccines normally are involved in the virulence mechanisms of the bacterium and, therefore, are expressed during infection. One of the most striking feature of virulence genes is there high abundance within pathogenicity islands; large horizontally acquired genomic regions, which present deviations in genomic signatures and are absent in related non-pathogenic organisms. The PIPS software was used to detect pathogenicity islands in the genome of *C. pseudotuberculosis* 258, as it includes all the above mentioned features to predict genomic islands in an integrative manner (Soares et al., 2012). PIPS found 11 pathogenicity islands in *C. pseudotuberculosis* 258 (Fig. 2), including PICP 1−7 already described in the literature (Ruiz et al., 2011). Furthermore, as per comparison of biovar *equi* and *ovis* strains of *C. pseudotuberculosis* and further analysis of the recently released *C. ulcerans* genomes (Trost et al., 2011), we have assessed 4 additional PAIs identified by PIPS (PICP 8−11), which also showed regions of genomic plasticity, i.e. insertions, deletions, and substitutions, and were classified as new putative pathogenicity islands of *C. pseudotuberculosis* (Fig. 2). Briefly, PICP 9 (CP258_0560−CP258_0575) presents large deletions in *C. pseudotuberculosis* strains 258 and CIP 52.97 when compared to the strain 1002. PICPs 8 (CP258_0171−CP258_0179), 10 (CP258_1622−CP258_1635), and 11 (CP258_2091−CP258_2103) are located in putative hotspots for pathogenicity islands, which present a high degree of plasticity also in the genomes of *C. ulcerans* BR-AD22 and *C. diphtheriae* NCTC 13129 (Fig. 2).

3.3. Prediction of candidate vaccine targets for C. pseudotuberculosis

The subcellular location of predicted proteins of *C. pseudotuberculosis* strains 1002, CIP 52.97, and 258 was identified with the SurfG+ software. As a prerequisite for the use of SurfG+, we have taken electron microscopy images of the three *C. pseudotuberculosis* strains (Fig. 3) and have measured their cell wall sizes, which correspond to 24.54 nm, 19.89 nm, and 24.11 nm, respectively (Table 2). After using the membrane sizes as parameter in SurfG+, we have classified 646−680 gene products as secreted proteins, putative surface-exposed (PSE) proteins or membrane proteins (Table 2; Table 3 [rule I]). The proteins predicted by SurfG+ were further analyzed with the software Vaxign, resulting in the detection of proteins with antigenic properties in the *C. pseudotuberculosis* strains 1002, CIP 52.97, and 258 (Table 3 [rule II]). Further analysis considering only vaccine candidates that are shared by all three strains, and excluding those that were not predicted as antigenic in at least one of the strains, resulted in 49 proteins (Table 4; Table 3 [rule III]).

After searching for antigenic proteins, which are encoded by shared pathogenicity islands (Table 3 [rule IV]), we found one candidate protein, Cp258_1473 (also named Cp1002_1466 and CpCIP5297_1478), which is annotated as uncharacterized protein HtaC and revealed similarity to HtaA superfamily domain proteins. The HtaA superfamily is a well characterized group of membrane-associated and surface-exposed heme receptors, which act in heme sequestration from the host to acquire iron in environments where this component is scarce, thereby playing a critical role in the ability of pathogens to cause disease (Allen and Schmitt, 2009; Anzaldi and Skaar, 2010). *C. pseudotuberculosis* also presents another antigenic protein involved in iron acquisition, represented by the *fhuD* gene (Table 4), which codes for a surface-exposed substrate-binding protein involved in the transport of ferrichrome or other hydroxamate siderophores (Clarke et al., 2000). Besides these two iron acquisition proteins, the potentially antigenic proteins PbpA, PbpB, MalE, RpfA, RuvA, CopC, and NrfC also deserve attention (Table 4).

The *pbpA* and *pbpB* genes code for penicillin-binding proteins (PBPs), a diverse family of secreted proteins, which are the primary targets of β-lactam antibiotics (Georgopapadakou and Liu, 1980). In Gram-negative bacteria, PBPs are related to peptidoglycan polymerization and are essential for bacterial cell elongation, septation, and modulation of cellular morphology. They can play an important role in biofilm formation and, therefore, in pathogenesis evolution (Ghosh et al., 2008).

The *malE* gene product is a carbohydrate-binding protein, which is probably anchored to the cell membrane as it is a putative lipoprotein. The MalE protein was shown to be highly elevated in expression during various phases of host-pathogen interaction, with a putative role in pathogenesis, which is also evidenced by the elicitation of host immune response in humans infected by group A *Streptococcus* (Shelburne et al., 2007).

The *rpfA* gene codes for a resuscitation-promoting factor, a family of proteins distributed through actinobacteria, which plays an important role in bacterial growth and in restoring the culturability of dormant mycobacteria. Moreover, these proteins are essential for viability of *Micrococcus luteus* and mutation of different paralogs in *Mycobacterium tuberculosis* showed differential attenuation in virulence and reduced ability to proliferate in the lungs and spleens of infected mice (Biketov et al., 2007; Tufariello et al., 2006). Finally, studies with *M. tuberculosis* indicated that the resuscitation-promoting factor is a promising candidate for inclusion as an antigen in novel tuberculosis vaccines in terms of its immunogenicity and protective efficacy (Romano et al., 2012).

The *ruvA* gene encodes a Holliday junction branch migrase protein, which plays an important role in immune evasion of several bacteria. The protein is responsible for creating antigenic variation by facilitating the ATP-dependent branch migration of heteroduplex DNA in Holliday junctions, resulting in targeted genome rearrangements (Dresser et al., 2009). Although this protein is very important for bacterial survival, its putative secretion and precise role in virulence, if any, has to be studied and elucidated in *C. pseudotuberculosis*.

The product of the *copC* gene is a copper resistance protein, which acts in copper mobilization and, therefore, has a potential role in bacterial copper homeostasis. Copper plays a dual role in bacteria, as it is a cofactor for the activity of several essential enzymes, but is also toxic in excess. In order to maintain essential biochemical reactions and to prevent toxic levels of copper inside the bacterial cell, microorganisms strictly controls the uptake, distribution, and efflux of this compound, corroborating the importance of *cop* genes in bacterial survival (Djoko et al., 2007; Puig et al., 2002).

The *nrfC* gene encodes the small subunit of the cytochrome c nitrite reductase, which is part of the formate-dependent nitrite reductase complex catalyzing the conversion of nitrite, a toxic compound in high concentrations, to ammonia. This physiological process plays a pivotal role in bacterial growth under anaerobic conditions where nitrite accumulates in the cells due to the use of nitrate as an alternative electron acceptor to oxygen (Cole, 1996).

Finally, the products of the *pld* genes of the three *C. pseudotuberculosis* strains revealed variable results and, therefore, were not included in the data set. The only *pld* gene product predicted to be secreted and antigenic was that of *C. pseudotuberculosis* 1002. The PLD from *C. pseudotuberculosis* CIP 52.97 was predicted to be a cytoplasmic protein and the PLD from *C. pseudotuberculosis* 258, although apparently secreted, presents a low adhesion probability (data not shown). These variations in the prediction of protein features may be related to small differences in the sequence of signal peptides (CIP 52.97) and epitope sites (258). This observation requires further investigations *in vitro*, as PLD is the major virulence factor of both biovars of *C. pseudotuberculosis* and currently used for standard vaccine production.

4. Conclusions

In this work, we present the genome sequence of the *C. pseudotuberculosis* biovar *equ*i strain 258 and compare it to other strains from the same genus in a search for regions of genome plasticity. Moreover, we used reverse vaccinology to predict new antigenic targets, which can be used in the development of new vaccine strategies for hosts of both biovars of *C. pseudotuberculosis* and we gave some insights into the putative functions of the respective proteins. However, additional *in vitro* and *in vivo* experimental analyses and further work on the pan-genome level with *C. pseudotuberculosis* strains isolated from a broad spectrum of animal hosts, including camel and buffalo in the case of biovar *equi*, are still necessary to create effective vaccines against *C. pseudotuberculosis* diseases.

Acknowlegments

SCS was supported by the CLIB Graduate Cluster Industrial Biotechnology and a CAPES-DAAD scholarship. ET acknowledges the receipt of a scholarship of the CLIB Graduate Cluster Industrial Biotechnology. The authors thank Prof. Christopher G. Dowson (University of Warwick, UK) for kindly providing the bacterial strain for genome sequencing.

References

Aleman, M., Spier, S.J., Wilson, W.D., Doherr, M., 1996. *Corynebacterium pseudotuberculosis* infection in horses: 538 cases (1982-1993). J. Am. Vet. Med. Assoc. 209, 804−809.

Alikhan, N., Petty, N.K., Ben Zakour, N.L., Beatson, S.A., 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics 12, 402.

Allen, C.E., Schmitt, M.P., 2009. HtaA is an iron-regulated hemin binding protein involved in the utilization of heme iron in *Corynebacterium diphtheriae*. J. Bacteriol .191, 2638−2648.

Anzaldi, L.L., Skaar, E.P., 2010. Overcoming the heme paradox: heme toxicity and tolerance in bacterial pathogens. Infect. Immun. 78, 4977−4989.

Arsenault, J., Girard, C., Dubreuil, P., Daignault, D., Galarneau, J.R., Boisclair, J., Simard, C., Bélanger, D., 2003. Prevalence of and carcass condemnation from maedi-visna, paratuberculosis and caseous lymphadenitis in culled sheep from Quebec, Canada. Prev. Vet. Med. 59, 67−81.

Augustine, J.L., Renshaw, H.W., 1986. Survival of *Corynebacterium pseudotuberculosis* in axenic purulent exudate on common barnyard fomites. Am. J. Vet. Res. 47, 713−715.

Barakat, A.A., Selim, S.A., Atef, A., Saber, M.S., Nafie, E.K., El-Edeeby, A.A., 1984. Two serotypes of *Corynebacterium pseudotuberculosis* isolated from different animal species. Revue Scientifique et Technique de l'OIE 3, 151−163.

Barh, D., Jain, N., Tiwari, S., Parida, B.P., D'Afonseca, V., Liwei, L., Ali, A., Santos, A.R., Guimarães, L.C., de Castro Soares, S., Miyoshi, A., Bhattacharjee, A., Misra, A.N., Silva, A., Kumar, A., Azevedo, V., 2011. A novel comparative genomics analysis for common drug and vaccine targets in *Corynebacterium pseudotuberculosis* and other CMN group of human pathogens. Chem Biol Drug Des 78, 73-84.

Barinov, A., Loux, V., Hammani, A., Nicolas, P., Langella, P., Ehrlich, D., Maguin, E., van de Guchte, M., 2009. Prediction of surface exposed proteins in *Streptococcus pyogenes*, with a potential application to other Gram-positive bacteria. Proteomics 9, 61−73.

Biberstein, E.L., Knight, H.D., Jang, S., 1971. Two biotypes of *Corynebacterium pseudotuberculosis*. Vet. Rec. 89, 691−692.

Biketov, S., Potapov, V., Ganina, E., Downing, K., Kana, B.D., Kaprelyants, A., 2007. The role of resuscitation promoting factors in pathogenesis and reactivation of *Mycobacterium tuberculosis* during intra-peritoneal infection in mice. BMC Infect. Dis. 7, 146.

Brogden, K.A., Chedid, L., Cutlip, R.C., Lehmkuhl, H.D., Sacks, J., 1990. Effect of muramyl dipeptide on immunogenicity of *Corynebacterium pseudotuberculosis* whole-cell vaccines in mice and lambs. Am. J. Vet. Res. 51, 200−202.

Carver, T.J., Rutherford, K.M., Berriman, M., Rajandream, M.A., Barrell, B.G., Parkhill, J., 2005. ACT: the Artemis Comparison Tool. Bioinformatics 16, 3422−3423.

Cerdeira, L.T., Carneiro, A.R., Ramos, R.T., de Almeida, S.S., D'Afonseca, V., Schneider, M.P., Baumbach, J., Tauch, A., McCulloch, J.A., Azevedo, V.A., Silva, A., 2011a. Rapid hybrid de novo assembly of a microbial genome using only short reads: *Corynebacterium pseudotuberculosis* I19 as a case study. J. Microbiol. Methods 86, 218−23.

Cerdeira, L.T., Schneider, M.P.C., Pinto, A.C., de Almeida, S.S., dos Santos, A.R., Barbosa, E.G.V., Ali, A., Aburjaile, F.F., de Abreu, V.A.C., Guimarães, L.C., Soares, S.D.C., Dorella, F.A., Rocha, F.S., Bol, E., Gomes de Sá, P.H.C., Lopes, T.S., Barbosa, M.S., Carneiro, A.R., Jucá Ramos, R.T., Coimbra, N.A.D.R., Lima, A.R.J., Barh, D., Jain, N., Tiwari, S., Raja, R., Zambare, V., Ghosh, P., Trost, E., Tauch, A., Miyoshi, A., Azevedo, V., Silva, A., 2011b. Complete genome sequence of *Corynebacterium pseudotuberculosis* strain CIP 52.97, isolated from a horse in Kenya. J. Bacteriol. 193, 7025−7026.

Clarke, T.E., Ku, S.Y., Dougan, D.R., Vogel, H.J., Tari, L.W., 2000. The structure of the ferric siderophore binding protein FhuD complexed with gallichrome. Nat. Struct. Biol. 7, 287−291.

Cole, J., 1996. Nitrate reduction to ammonia by enteric bacteria: redundancy, or a strategy for survival during oxygen starvation?. FEMS Microbiol. Lett. 136, 1−11.

Djoko, K.Y., Xiao, Z., Huffman, D.L., Wedd, A.G., 2007. Conserved mechanism of copper binding and transfer. A comparison of the copper-resistance proteins PcoC from *Escherichia coli* and CopC from *Pseudomonas syringae*. Inorg. Chem. 46, 4560−4568.

Dorella, F.A., Pacheco, L.G., Seyffert, N., Portela, R.W., Meyer, R., Miyoshi, A., Azevedo, V., 2009. Antigens of *Corynebacterium pseudotuberculosis* and prospects for vaccine development. Expert Rev. Vaccines 8, 205−213.

Dorella, F.A., Pacheco, L.G.C., Oliveira, S.C., Miyoshi, A., Azevedo, V., 2006. *Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. Vet. Res. 37, 201−218.

Dresser, A.R., Hardy, P., Chaconas, G., 2009. Investigation of the genes involved in antigenic switching at the *vlsE* locus in *Borrelia burgdorferi*: an essential role for the RuvAB branch migrase. PLoS Pathog. 5, e1000680.

Eggleton, D.G., Middleton, H.D., Doidge, C.V., Minty, D.W., 1991. Immunisation against ovine caseous lymphadenitis: comparison of *Corynebacterium pseudotuberculosis* vaccines with and without bacterial cells. Aust. Vet. J. 68, 317−319.

Ellis, J.A., 1991. Antigen specificity of antibody responses to *Corynebacterium pseudotuberculosis* in naturally infected sheep with caseous lymphadenitis. Vet. Immunol. Immunopathol. 28, 289−301.

Georgopapadakou, N.H., Liu, F.Y., 1980. Penicillin-binding proteins in bacteria. Antimicrob. Agents Chemother. 18, 148−157.

Ghosh, A.S., Chowdhury, C., Nelson, D.E., 2008. Physiological functions of D-alanine carboxypeptidases in *Escherichia coli*. Trends Microbiol. 16, 309−317.

He, Y., Xiang, Z., Mobley, H.L., 2010. Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development. J Biomed Biotechnol. 2010, 297505.

Hernandez, D., François, P., Farinelli, L., Osterås, M., Schrenzel, J., 2008. *De novo* bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res. 18, 802−809.

Hodgson, A.L., Carter, K., Tachedjian, M., Krywult, J., Corner, L.A., McColl, M., Cameron, A., 1999. Efficacy of an ovine caseous lymphadenitis vaccine formulated using a genetically inactive form of the *Corynebacterium pseudotuberculosis* phospholipase D. Vaccine 17, 802−808.

Holstad, G., 1989. *Corynebacterium pseudotuberculosis* infection in goats. IX. The effect of vaccination against natural infection. Acta Vet. Scand. 30, 285−293.

Jones, D., Collins, M.D., 1986. Irregular, nonsporing gram-positive rods, section 15. pages 1261–1579 in Bergey's Manual of Systematic Bacteriology. Williams & Wilkins Co., Baltimore, MD.

Lagesen, K., Hallin, P., Rødland, E.A., Staerfeldt, H., Rognes, T., Ussery, D.W., 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35, 3100−3108.

LeaMaster, B.R., Shen, D.T., Gorham, J.R., Leathers, C.W., Wells, H.D., 1987. Efficacy of *Corynebacterium pseudotuberculosis* bacterin for the immunologic protection of sheep against development of caseous lymphadenitis. Am. J. Vet. Res. 48, 869−872.

Lowe, T.M., Eddy, S.R., 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25, 955−964.

Melo, A.L., Machado, C.R.S., Pereira, R.H., 1993. Host cell adhesion to Schistosoma mansoni larvae in the peritoneal cavity of naive mice. Histological and scanning electron microscopic studies. Revista do Instituto de Medicina Tropical de São Paulo 35(1), 17-22.

Paton, M.W., Walker, S.B., Rose, I.R., Watt, G.F., 2003. Prevalence of caseous lymphadenitis and usage of caseous lymphadenitis vaccines in sheep flocks. Aust. Vet. J. 81, 91−95.

Pratt, S.M., Spier, S.J., Carroll, S.P., Vaughan, B., Whitcomb, M.B., Wilson, W.D., 2005. Evaluation of clinical characteristics, diagnostic test results, and outcome in horses with internal infection caused by *Corynebacterium pseudotuberculosis*: 30 cases (1995-2003). J. Am. Vet. Med. Assoc. 227, 441−448.

Puig, S., Rees, E.M., Thiele, D.J., 2002. The ABCDs of periplasmic copper trafficking. Structure 10, 1292−1295.

Ramos, R.T., Carneiro, A.R., Baumbach, J., Azevedo, V., Schneider, M.P., Silva, A., 2011. Analysis of quality raw data of second generation sequencers with Quality Assessment Software. BMC Res. Notes 4, 130.

Rappuoli, R., 2001. Reverse vaccinology, a genome-based approach to vaccine development. Vaccine 19, 2688−2691.

Romano, M., Aryan, E., Korf, H., Bruffaerts, N., Franken, C.L., Ottenhoff, T.H., Huygen, K., 2012. Potential of *Mycobacterium tuberculosis* resuscitation-promoting factors as antigens in novel tuberculosis sub-unit vaccines. Microbes infect 14, 86-95.

Ruiz, J.C., D'Afonseca, V., Silva, A., Ali, A., Pinto, A.C., Santos, A.R., Rocha, A.A.M.C., Lopes, D.O., Dorella, F.A., Pacheco, L.G.C., Costa, M.P., Turk, M.Z., Seyffert, N., Moraes, P.M.R.O., Soares, S.C., Almeida, S.S., Castro, T.L.P., Abreu, V.A.C., Trost, E., Baumbach, J., Tauch, A., Schneider, M.P.C., McCulloch, J., Cerdeira, L.T., Ramos, R.T.J., Zerlotini, A., Dominitini, A., Resende, D.M., Coser, E.M., Oliveira, L.M., Pedrosa, A.L., Vieira, C.U., Guimarães, C.T., Bartholomeu, D.C., Oliveira, D.M., Santos, F.R., Rabelo, É.M., Lobo, F.P., Franco, G.R., Costa, A.F., Castro, I.M., Dias, S.R.C., Ferro, J.A., Ortega, J.M., Paiva, L.V., Goulart, L.R., Almeida, J.F., Ferro, M.I.T., Carneiro, N.P., Falcão, P.R.K., Grynberg, P., Teixeira, S.M.R., Brommonschenkel, S., Oliveira, S.C., Meyer, R., Moore, R.J., Miyoshi, A., Oliveira, G.C., Azevedo, V., 2011. Evidence for reductive genome evolution and lateral acquisition of virulence functions in two *Corynebacterium pseudotuberculosis* strains. PLoS One 6, e18551.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., Barrell, B., 2000. Artemis: sequence visualization and annotation. Bioinformatics 16, 944−945.

Selvy, P.E., Lavieri, R.R., Lindsley, C.W., Brown, H.A., 2011. Phospholipase D: enzymology, functionality, and chemical modulation. Chem. Rev. 111, 6064−6119.

Shelburne, S.A., Fang, H., Okorafor, N., Sumby, P., Sitkiewicz, I., Keith, D., Patel, P., Austin, C., Graviss, E.A., Musser, J.M., Chow, D., 2007. MalE of group A *Streptococcus* participates in the rapid transport of maltotriose and longer maltodextrins. J. Bacteriol. 189, 2610−2617.

Soares, S.C., Abreu, V.A.C., Ramos, R.T.J., Cerdeira, L., Silva, A., Baumbach, J., Trost, E., Tauch, A., Hirata, R.J., Mattos-Guaraldi, A.L., Miyoshi, A., Azevedo, V., 2012. PIPS: pathogenicity island prediction software. PLoS One 7, e30848.

Steinman, A., Elad, D., Shpigel, N., 1999. Ulcerative lymphangitis and coronet lesions in an Israeli dairy herd infected with *Corynebacterium pseudotuberculosis*. Veterinary Record 145, 604−606.

Trost, E., Al-Dilaimi, A., Papavasiliou, P., Schneider, J., Viehoever, P., Burkovski, A., Soares, S.C., Almeida, S.S., Dorella, F.A., Miyoshi, A., Azevedo, V., Schneider, M.P., Silva, A., Santos, C.S., Santos, L.S., Sabbadini, P., Dias, A.A., Hirata, R.J., Mattos-Guaraldi, A.L., Tauch, A., 2011. Comparative analysis of two complete *Corynebacterium ulcerans* genomes and detection of candidate virulence factors. BMC Genomics 12, 383.

Trost, E., Ott, L., Schneider, J., Schröder, J., Jaenicke, S., Goesmann, A., Husemann, P., Stoye, J., Dorella, F.A., Rocha, F.S., Soares, S.D.C., D'Afonseca, V., Miyoshi, A., Ruiz, J., Silva, A., Azevedo, V., Burkovski, A., Guiso, N., Join-Lambert, O.F., Kayal, S., Tauch, A., 2010. The complete genome sequence of *Corynebacterium pseudotuberculosis* FRC41 isolated from a 12-year-old girl with necrotizing lymphadenitis reveals insights into gene-regulatory networks contributing to virulence. BMC Genomics 11, 728.

Tsai, I.J., Otto, T.D., Berriman, M., 2010. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biol. 11, R41.

Tufariello, J.M., Mi, K., Xu, J., Manabe, Y.C., Kesavan, A.K., Drumm, J., Tanaka, K., Jacobs, W.R.J., Chan, J., 2006. Deletion of the *Mycobacterium tuberculosis* resuscitation-promoting factor Rv1009 gene results in delayed reactivation from chronic tuberculosis. Infect. Immun. 74, 2985−2995.

Williamson, L.H., 2001. Caseous lymphadenitis in small ruminants. Vet. Clin. North Am. Food Anim. Pract. 17, 359−371.

Windsor, P.A., 2011. Control of caseous lymphadenitis. Vet. Clin. North Am. Food Anim. Pract. 27, 193−202.

Yeruham, I., Elad, D., Friedman, S., Perl, S., 2003. *Corynebacterium pseudotuberculosis* infection in Israeli dairy cattle. Epidemiol. Infect. 131, 947−955.

Yeruham, I., Friedman, S., Perl, S., Elad, D., Berkovich, Y., Kalgard, Y., 2004. A herd level analysis of a *Corynebacterium pseudotuberculosis* outbreak in a dairy cattle herd. Vet. Dermatol. 15, 315-320.

Zdobnov, E.M., Apweiler, R., 2001. InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17, 847−848.

Zerbino, D.R., Birney, E., 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. 18, 821−829.

Table 1

Genomic features of *C. pseudotuberculosis* (Cp) strains.

Feature	Cp 1002	Cp CIP 52.97	Cp 258
Biovar	*ovis*	*equi*	*equi*
Isolation	goat (Brazil)	horse (Kenya)	horse (Belgium)
Size	2,335,113 bp	2,320,595 bp	2,314,404 bp
G+C content	52.19%	52.14%	52.15%
Proteins	2,090	2,060	2,088
rRNAs	12	12	12
tRNAs	48	47	49
Genes	2,203	2,194	2,195
Pseudogenes	53	75	46

Table 2

Subcellular location of proteins of *C. pseudotuberculosis* (Cp) strains.

Feature	Cp 1002	Cp CIP 52.97	Cp 258
Cell wall size	24.54 nm	19.89 nm	24.11 nm
Cytoplasmic proteins	1,417	1,411	1,428
Membrane proteins	370	368	370
PSE[a] proteins	211	194	201
Secreted proteins	99	84	89

[a] putative surface-exposed

Table 3

Number of *C. pseudotuberculosis* proteins through each step of reverse vaccinology strategy.

Rules	Cp 1002	Cp CIP 52.97	Cp 258
Rule I	680	646	660
Rule II	71	64	63
Rule III			49
Rule IV			1

495 **Table 4**

496 Putative antigenic proteins identified with Vaxign and shared by *C. pseudotuberculosis* (Cp) strains 1002, CIP52.97, and 258.

Cp 1002	Cp CIP 52.97	Cp 258	Gene name	Subcellular location	Gene product
Cp1002_0016	CpCIP5297_0016	Cp258_0017	-	PSE[a]	ABC transporter substrate-binding protein
Cp1002_0035	CpCIP5297_0037	Cp258_0039	*pbpA*	secreted	Penicillin-binding protein A
Cp1002_0079	CpCIP5297_0090	Cp258_0093	-	PSE	Hypothetical protein
Cp1002_0126a	CpCIP5297_0137	Cp258_0139	-	secreted	Hypothetical protein
Cp1002_0192	CpCIP5297_0202	Cp258_0202	-	PSE	Hypothetical protein
Cp1002_0200	CpCIP5297_0207	Cp258_0206	*pbpB*	secreted	Penicillin-binding protein B
Cp1002_0212	CpCIP5297_0219	Cp258_0218	-	membrane	Hypothetical protein
Cp1002_0220	CpCIP5297_0226	Cp258_0225	-	membrane	Hypothetical protein
Cp1002_0315	CpCIP5297_0321	Cp258_0318	-	PSE	Hypothetical protein
Cp1002_0320	CpCIP5297_0326	Cp258_0323	-	PSE	Hypothetical protein
Cp1002_0377	CpCIP5297_0388	Cp258_0385	*malE*	PSE	Maltotriose-binding protein
Cp1002_0388	CpCIP5297_0399	Cp258_0397	-	secreted	L,D-transpeptidase
Cp1002_0415	CpCIP5297_0426	Cp258_0424	-	secreted	Hypothetical protein
Cp1002_0439	CpCIP5297_0452	Cp258_0450	-	PSE	Manganese ABC transporter, substrate-binding protein
Cp1002_0454	CpCIP5297_0467	Cp258_0464	*htaC*	PSE	Hypothetical protein with HtaA family domain
Cp1002_0535	CpCIP5297_0548	Cp258_0542	-	secreted	Secreted hydrolase
Cp1002_0550	CpCIP5297_0563	Cp258_0557	-	PSE	Hypothetical protein
Cp1002_0594	CpCIP5297_0603	Cp258_0599	*rpfA*	secreted	Resuscitation-promoting factor A
Cp1002_0643	CpCIP5297_0654	Cp258_0648	-	PSE	Uncharacterized metalloprotease
Cp1002_0648	CpCIP5297_0659	Cp258_0653	*gluB*	PSE	Glutamate ABC transporter, substrate-binding protein
Cp1002_0686	CpCIP5297_0701	Cp258_0690	-	secreted	Hypothetical protein
Cp1002_0766	CpCIP5297_0782	Cp258_0771	-	secreted	Hypothetical protein
Cp1002_0876	CpCIP5297_0896	Cp258_0884	*fhuD*	PSE	Iron(3+)-hydroxamate-binding protein FhuD
Cp1002_0883	CpCIP5297_0904	Cp258_0892	-	membrane	Hypothetical protein
Cp1002_0979	CpCIP5297_1002	Cp258_0998	-	PSE	Esterase
Cp1002_1000	CpCIP5297_1017	Cp258_1014	-	secreted	Hypothetical protein
Cp1002_1013	CpCIP5297_1030	Cp258_1027	*yceI*	secreted	Hypothetical protein YceI
Cp1002_1083	CpCIP5297_1102	Cp258_1100	-	PSE	Hypothetical protein
Cp1002_1173	CpCIP5297_1194	Cp258_1192	*ruvA*	PSE	Holliday junction ATP-dependent DNA helicase

Cp1002_1189	CpCIP5297_1210	Cp258_1208	*copC*	PSE	Copper resistance protein CopC
Cp1002_1281	CpCIP5297_1304	Cp258_1301	-	PSE	Hypothetical protein
Cp1002_1356	CpCIP5297_1380	Cp258_1380	-	membrane	Hypothetical protein
Cp1002_1362	CpCIP5297_1386	Cp258_1386	-	PSE	Hypothetical protein
Cp1002_1379	CpCIP5297_1404	Cp258_1403	-	PSE	Hypothetical protein
Cp1002_1466	CpCIP5297_1478	Cp258_1473	-	PSE	Hypothetical protein
Cp1002_1503	CpCIP5297_1517	Cp258_1510	*thiX*	PSE	Thiamine biosynthesis protein ThiX
Cp1002_1506	CpCIP5297_1520	Cp258_1514	-	secreted	Guanyl-specific ribonuclease
Cp1002_1540	CpCIP5297_1554	Cp258_1548	-	PSE	Hypothetical protein
Cp1002_1604	CpCIP5297_1617	Cp258_1606	-	PSE	Hypothetical protein
Cp1002_1684	CpCIP5297_1700	Cp258_1697	-	PSE	Hypothetical protein
Cp1002_1763	CpCIP5297_1781	Cp258_1780	*lpqE*	secreted	Lipoprotein LpqE
Cp1002_1768	CpCIP5297_1785	Cp258_1783	-	PSE	Hypothetical protein
Cp1002_1820	CpCIP5297_1840	Cp258_1836	-	secreted	Hypothetical protein
Cp1002_1848	CpCIP5297_1869	Cp258_1864	*nrfC*	membrane	Cytochrome c nitrite reductase, small subunit
Cp1002_1893	CpCIP5297_1920	Cp258_1910	-	secreted	Membrane protein
Cp1002_1933	CpCIP5297_1961	Cp258_1951	-	PSE	VanW family protein
Cp1002_1954	CpCIP5297_1983	Cp258_1972	-	PSE	Hypothetical protein
Cp1002_1962	CpCIP5297_1991	Cp258_1981	-	PSE	Hypothetical protein
Cp1002_1976	CpCIP5297_2004	Cp258_1995	-	secreted	Hypothetical protein

[a] putative surface-exposed

Figure legends

Fig. 1. Genome map of *C. pseudotuberculosis* biovar *equi* strain 258.

Fig. 2. Genome alignment of *C. pseudotuberculosis*, *C. ulcerans*, *C. diphtheriae*, and *C. glutamicum* strains. The figure shows the alignment of *C. pseudotuberculosis* 258 (Cp 258), *C. pseudotuberculosis* CIP 52.97 (Cp CIP 52.97), *C. diphtheriae* NCTC 13129 (Cd NCTC 13129), *C. ulcerans* BR-AD 22 (Cu BR-AD22), and *C. glutamicum* ATCC 13032 (Cg ATCC 13032) using the genome of *C. pseudotuberculosis* 1002 as a reference sequence. The outermost circle highlights the eleven pathogenicity islands (PICP 1−11) in red.

Fig. 3. Electron microscopy of *C. pseudotuberculosis* strains 1002 (A), CIP 52.97 (B), and 258 (C).

Figure1

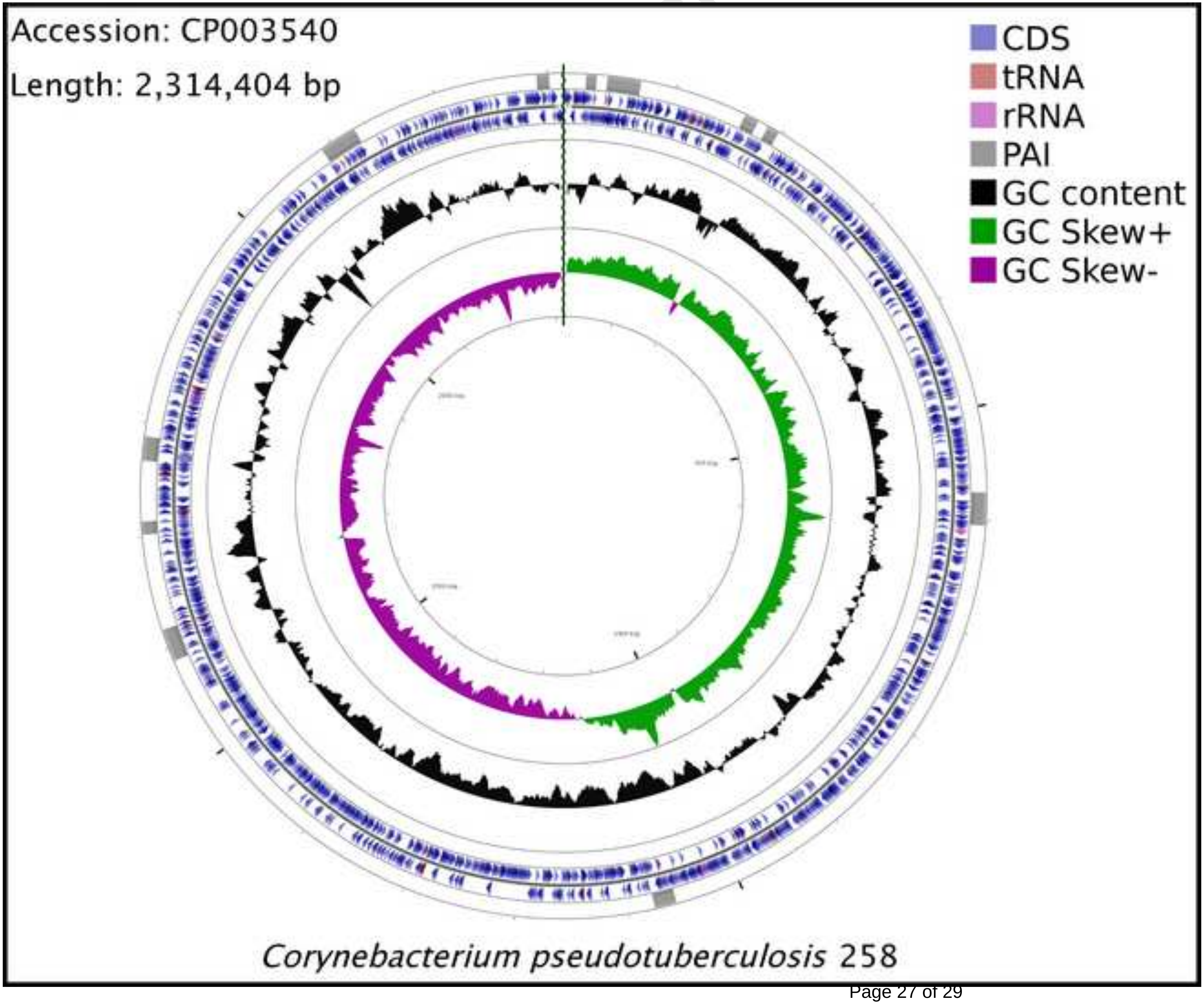

Accession: CP003540
Length: 2,314,404 bp
CDS
tRNA
rRNA
PAI
GC content
GC Skew+
GC Skew-
Corynebacterium pseudotuberculosis 258

Figure2

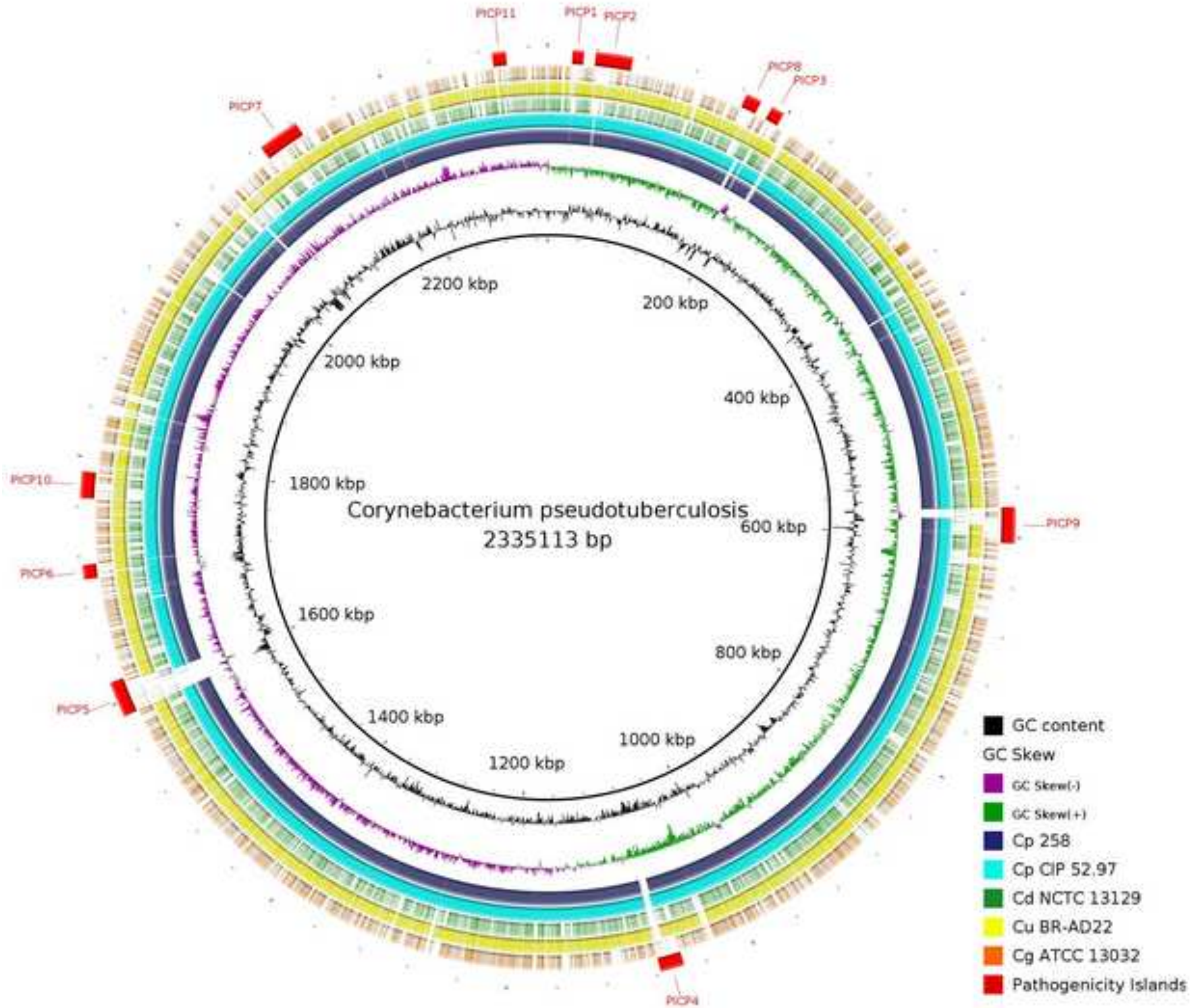
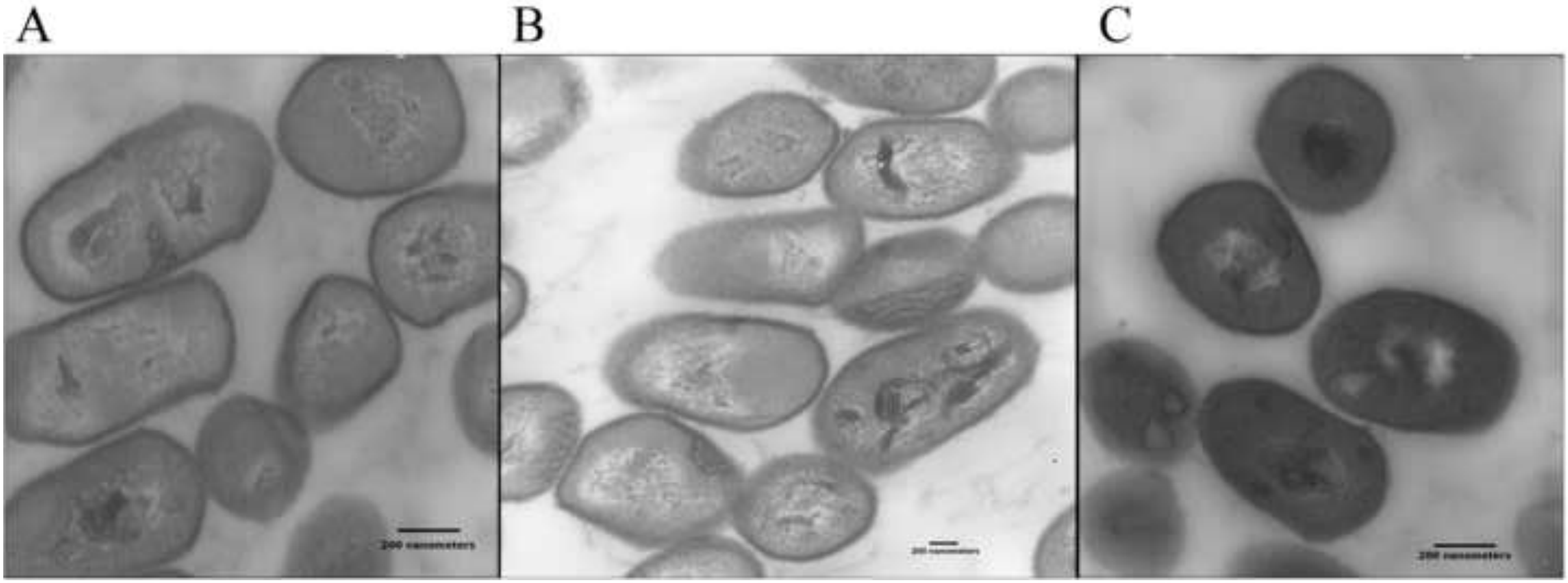

PICP11
PICP1
PICP2
PICP8
PICP3
PICP7
PICP10
PICP9
PICP6
PICP5
PICP4
2200 kbp
200 kbp
2000 kbp
400 kbp
1800 kbp
600 kbp
1600 kbp
800 kbp
1400 kbp
1000 kbp
1200 kbp
Corynebacterium pseudotuberculosis
2335113 bp
GC content
GC Skew
GC Skew(-)
GC Skew(+)
Cp 258
Cp CIP 52.97
Cd NCTC 13129
Cu BR-AD22
Cg ATCC 13032
Pathogenicity Islands

Figure3



II.I.4 The pan-genome of the animal pathogen *Corynebacterium pseudotuberculosis* reveals differences in genome plasticity between the biovar ovis and equi strains.
Soares SC, Silva A, Trost E, Blom J, Ramos R, Carneiro A, Ali A, Santos AR, Pinto AC, Diniz C, Barbosa EG, Dorella FA, Aburjaile F, Rocha FS, Nascimento KK, Guimarães LC, Almeida S, Hassan SS, Bakhtiar SM, Pereira UP, Abreu VA, Schneider MP, Miyoshi A, Tauch A, **Azevedo V**.
PLoS One. 2013;8(1):e53818. doi: 10.1371/journal.pone.0053818. Epub 2013 Jan 14.

A era de sequenciamento genômico de *C. pseudotuberculosis* finalmente culminou nas análises pan-genômicas de toda a espécie e de ambos os biovares separadamente. No artigo que segue, são descritas as análises filogenômicas do gênero *Corynebacterium* usando os software Gegenees e comparações adicionais da árvore filogenética gerada com dados da literatura. Além disso, são mostradas as análises do pan-genôma, core genoma e singletons de toda a espécie e de ambos os biovares. A evolução do pan-genoma foi extrapolada a partir de cada subconjunto de dados. Finalmente, foram preditas PAIs visando analisar a correlação da plasticidade no cluster de genes de pili com o comportamento intracelular facultativo de *C. pseudotuberculosis*.

OPEN ACCESS Freely available online

PLOS | ONE

The Pan-Genome of the Animal Pathogen *Corynebacterium pseudotuberculosis* Reveals Differences in Genome Plasticity between the Biovar *ovis* and *equi* Strains

Siomar C. Soares[1,2,3], Artur Silva[4], Eva Trost[2,3], Jochen Blom[2], Rommel Ramos[4], Adriana Carneiro[4], Amjad Ali[1], Anderson R. Santos[1], Anne C. Pinto[1], Carlos Diniz[1], Eudes G. V. Barbosa[1], Fernanda A. Dorella[1], Flávia Aburjaile[1], Flávia S. Rocha[1], Karina K. F. Nascimento[1], Luís C. Guimarães[1,2,3], Sintia Almeida[1], Syed S. Hassan[1], Syeda M. Bakhtiar[1], Ulisses P. Pereira[5], Vinicius A. C. Abreu[1], Maria P. C. Schneider[4], Anderson Miyoshi[1], Andreas Tauch[2]◑, Vasco Azevedo[1]*◑

[1] Department of General Biology, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, [2] Center for Biotechnology, Bielefeld University, Bielefeld, Nordrhein-Westfalen, Germany, [3] CLIB Graduate Cluster Industrial Biotechnology, Center for Biotechnology, Bielefeld University, Bielefeld, Nordrhein-Westfalen, Germany, [4] Department of Genetics, Federal University of Pará, Belém, Pará, Brazil, [5] Department of Veterinary Medicine, Federal University of Lavras, Lavras, Brazil

Abstract

Corynebacterium pseudotuberculosis is a facultative intracellular pathogen and the causative agent of several infectious and contagious chronic diseases, including caseous lymphadenitis, ulcerative lymphangitis, mastitis, and edematous skin disease, in a broad spectrum of hosts. In addition, *Corynebacterium pseudotuberculosis* infections pose a rising worldwide economic problem in ruminants. The complete genome sequences of 15 *C. pseudotuberculosis* strains isolated from different hosts and countries were comparatively analyzed using a pan-genomic strategy. Phylogenomic, pan-genomic, core genomic, and singleton analyses revealed close relationships among pathogenic corynebacteria, the clonal-like behavior of *C. pseudotuberculosis* and slow increases in the sizes of pan-genomes. According to extrapolations based on the pan-genomes, core genomes and singletons, the *C. pseudotuberculosis* biovar *ovis* shows a more clonal-like behavior than the *C. pseudotuberculosis* biovar *equi*. Most of the variable genes of the biovar *ovis* strains were acquired in a block through horizontal gene transfer and are highly conserved, whereas the biovar *equi* strains contain great variability, both intra- and inter-biovar, in the 16 detected pathogenicity islands (PAIs). With respect to the gene content of the PAIs, the most interesting finding is the high similarity of the pilus genes in the biovar *ovis* strains compared with the great variability of these genes in the biovar *equi* strains. Concluding, the polymerization of complete pilus structures in biovar *ovis* could be responsible for a remarkable ability of these strains to spread throughout host tissues and penetrate cells to live intracellularly, in contrast with the biovar *equi*, which rarely attacks visceral organs. Intracellularly, the biovar *ovis* strains are expected to have less contact with other organisms than the biovar *equi* strains, thereby explaining the significant clonal-like behavior of the biovar *ovis* strains.

Citation: Soares SC, Silva A, Trost E, Blom J, Ramos R, et al. (2013) The Pan-Genome of the Animal Pathogen *Corynebacterium pseudotuberculosis* Reveals Differences in Genome Plasticity between the Biovar *ovis* and *equi* Strains. PLoS ONE 8(1): e53818. doi:10.1371/journal.pone.0053818

Editor: Igor Mokrousov, St. Petersburg Pasteur Institute, Russian Federation

Received October 8, 2012; **Accepted** December 3, 2012; **Published** January 14, 2013

Funding: This work was supported by grants from the Brazilian funding agencies Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq; http://cnpq.br/) (grant number 578219/2008-5), Fundação de Apoio à Pesquisa de Minas Gerais (FAPEMIG; http://www.fapemig.br/) (grant number APQ-01004-08), Fundação Amazônia Paraense de Amparo à Pesquisa (FAPESPA; http://www.fapespa.pa.gov.br/), and Rede Paraense de Genômica e Proteômica (RPGP; http://www.genoma.ufpa.br/). The funding agencies had no role in the study design, the data collection and analysis, the decision to publish,or the preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: vasco@icb.ufmg.br

◑ These authors contributed equally to this work.

Introduction

The genus *Corynebacterium* belongs to the CMNR group from the supra-generic group of *Actinomycetes*, which includes genera of great medical, veterinary, and biotechnological importance, such as *Corynebacterium*, *Mycobacterium*, *Nocardia*, and *Rhodococcus*. These genera have specific features in common, such as a high DNA G+C content and a specific organization of the cell wall, which is mainly composed of peptidoglycans, arabinogalactans, and mycolic acids [1]. The genus *Corynebacterium* was originally created to include *Corynebacterium diphtheriae* and other pathogenic species [2]. Several other bacteria that differed in shape, pathogenicity and sporulation were later added to this group [3]. Currently, the genus is composed of pathogenic species such as *Corynebacterium diphtheriae*, the causative agent of diphtheria [4]; opportunistic pathogens such as *Corynebacterium jeikeium*, which is responsible for some nosocomial infections in humans [5]; and non-pathogenic species such as *Corynebacterium glutamicum*, which is highly utilized in industrial amino acid production [6].

Corynebacterium pseudotuberculosis is a facultative intracellular and pleomorphic member of the genus *Corynebacterium*. This bacterium is non-motile, although it does possess fimbriae, and it is the causative agent of caseous lymphadenitis (CLA) in sheep and goats [7]. A close taxonomic relationship between *C. pseudotuberculosis* and *Corynebacterium ulcerans* has been suggested because these organisms are the only corynebacteria that produce the exotoxin phospholipase D [8,9]. Moreover, some strains of *C. pseudotuberculosis* and *C. ulcerans* express the diphtheria toxin, which indicates a relationship between both species and *C. diphtheriae* [10]. This relationship has also been demonstrated by a phylogenetic analysis of the *rpoB* gene [1]. The initial classification of *C. pseudotuberculosis* was based on morphological and biochemical characteristics [7,11]: the results of the nitrate reduction test play an important role in distinguishing the biovar *ovis* (isolated from sheep and goats; negative nitrate reduction) from the biovar *equi* (isolated from horses and bovines; positive nitrate reduction) [12].

In sheep and goats, *C. pseudotuberculosis* biovar *ovis* strains are responsible for causing the aforementioned infectious, contagious, chronic disease CLA, which is mainly characterized by the presence of caseous necrosis on the lymphatic glands or abscess formation in superficial lymph nodes and subcutaneous tissues [13]. CLA is a widespread disease that has been reported in several countries, including Australia, Brazil, Canada, New Zealand, South Africa, and the United States, where sheep and goat farming are prevalent [1,14–18]. CLA produces economic losses for sheep and goat farmers by causing skin deterioration and reducing yields of milk and wool. In addition to these effects, the visceral form of the disease can affect internal organs, resulting in weight loss, carcass condemnation and death [19]. The disease is transmitted through direct contact with superficial wounds, which can be the result of common procedures such as castration and shearing [20]. The transmission and dissemination of *C. pseudotuberculosis* are also associated with the following: a high resistance to environmental conditions [21–23]; a low detection rate, with the visceral form of the disease usually being detected in the later stages or in the slaughterhouse [24]; the inefficacy of antibiotic therapies due to abscess formation and an intra-macrophagic lifestyle [25]; high variability in the severity of the disease in vaccinated animals and in the protection levels of the vaccines [26]; and the variable efficacy of licensed vaccines, which are intended for use in sheep, in goat immunizations [27].

Although *C. pseudotuberculosis* was initially identified as causing CLA in sheep and goats, this bacterium has also been isolated from other species that exhibit different symptoms, including horses, cows, camels, buffalo, and even humans [1,28–30]. Despite the broad host spectrum, natural cross-species transmission of *C. pseudotuberculosis* between small ruminants and cattle does not appear to occur [12], although infections of cattle with both biovars have been previously reported [31].

C. pseudotuberculosis infections in horses can display three different disease patterns: external abscesses (pigeon fever), ulcerative lymphangitis of the limbs, and a visceral form that affects the internal organs [32,33]. Additionally, several clinical symptons of the diseases caused by *C. pseudotuberculosis* have been described in cattle: pyogranulomatous reactions, abscess formation, mastitis, visceral commitment, and necrotic and ulcerative dermatitis on the heel of the foot, which is accompanied by edematous swelling and lameness [24]. In bulls and buffalo, there is evidence of the mechanical transmission of *C. pseudotuberculosis* by houseies or other diptera, in addition to transmission via skin contact between animals [23,24,34–37]. Moreover, all reported outbreaks of CLA in horses in the United States have been preceded by large populations of houseies and other diptera during the summer, a phenomenon promoted by high environmental temperatures and drought conditions [38] that may also be related to a rise in the number of affected herds in Israel [24].

Although the pathogenic mechanism of CLA is well understood, there remains a lack of information about the virulence factors of *C. pseudotuberculosis* and the pathogenic mechanisms of the other diseases caused by this bacterium [1,39,40]. Virulence factors play an important role in the adhesion, invasion, colonization, spread inside the host, and immune system evasion of pathogenic bacteria; they also allow contact, penetration and survival inside the host [41]. Billington *et al.* [42] reported four *C. pseudotuberculosis* genetic factors, the *fagABC* operon and the *fagD* gene, that play an important role in virulence; they are involved in iron acquirement and, therefore, enable the bacterium to survive in environments where iron is scarce. The *fagABC* operon and the *fagD* gene are found in a pathogenicity island along with the *pld* gene, which encodes phospholipase D (PLD) [43]. PLD is the primary virulence factor of *C. pseudotuberculosis*; it promotes the hydrolysis and degradation of sphingomyelin in endothelial cell membranes, which increases vascular permeability and contributes to the spread and persistence of the bacterium in the host [27,44,45]. More recently, Trost *et al.* [46] reported the presence of two pilus gene clusters in the *C. pseudotuberculosis* FRC41 strain, which is in agreement with the previously reported visualization of pilin structures in other strains of *C. pseudotuberculosis* [47]. Pili are helical, cylinder-shaped structures, which are observed attached to and protruding from the bacterial cell surface. Pili play an important role in virulence as they enable pathogens to bind to molecules on various host tissues. After attaching to the host cell surface, the pathogen is able to initiate specific biochemical processes, such as extracellular and intracellular invasion, that will result in its proliferation in and dissemination among the host tissues [48].

To better understand the different symptoms of *C. pseudotuberculosis* infections in the broad spectrum of hosts and how genome plasticity is related to the symptom patterns, we performed pan-genomic comparative analyses of 15 *C. pseudotuberculosis* strains. In the following sections, we present the phylogenomic correlations between *C. pseudotuberculosis* and other corynebacteria. Furthermore, we describe the content and extrapolations of the following gene subsets from *C. pseudotuberculosis*: the "pan-genome", which is the complete inventory of genes found in any member of the species; the "core genome", which is composed of the genes that are present in all the species strains and that are thus important for basic life processes; and the "singletons", which represent genes found only in a given strain. Finally, we provide insights into the specific subsets (singletons and the pan- and core genomes) of both biovars of *C. pseudotuberculosis*, *ovis* and *equi*, and we correlate these subsets with the plasticity of pathogenicity islands, virulence genes, and biovar-specific diseases.

Materials and Methods

Genome Sequences

The genome sequences of 15 *C. pseudotuberculosis* strains were retrieved from the NCBI database (http://www.ncbi.nlm.nih.gov/genbank/): 9 biovar *ovis* strains, which were isolated from sheep, goats, humans, llamas, antelopes, and cows, and 6 biovar *equi* strains, which were isolated from horses, camels, and buffalo (Table 1). The strains were isolated in Oceania (Australia), South America (Brazil and Argentina), North America (United States), Africa (South Africa, Egypt and Kenya), southwestern Asia (Israel),

Table 1. General information about the 15 *C. pseudotuberculosis* strains used in this work.

Strains	Biovar	Host	Country of isolation	Clinical description	Genome size	Number of genes	Singletons	GenBank accession N°	Reference
1002	*ovis*	Goat	Brazil	CLA abscess	2,335,113	2,203	0	CP001809	[43]
C231	*ovis*	Sheep	Australia	CLA abscess	2,328,208	2,204	3	CP001829	[43]
42/02-A	*ovis*	Sheep	Australia	CLA abscess	2,337,606	2,164	5	CP003062	[49]
PAT10	*ovis*	Sheep	Argentina	Lung abscess	2,335,323	2,200	1	CP002924	[50]
3/99-5	*ovis*	Sheep	Scotland	CLA	2,337,938	2,239	39	CP003152	[49]
267	*ovis*	Llama	USA	CLA abscess	2,337,628	2,249	8	CP003407	[51]
P54B96	*ovis*	Antelope	South Africa	CLA abscess	2,337,657	2,205	2	CP003385	–
I19	*ovis*	Cow	Israel	Bovine mastitis abscess	2,337,730	2,213	0	CP002251	[52]
FRC41	*ovis*	Human	France	Necrotizing lymphadenitis	2,337,913	2,171	12	CP002097	[46]
CIP52.97	*equi*	Horse	Kenya	Ulcerative lymphangitis	2,320,595	2,194	30	CP003061	[53]
316	*equi*	Horse	USA	Abscess	2,310,415	2,234	25	CP003077	[54,55]
258	*equi*	Horse	Belgium	Ulcerative lymphangitis	2,314,404	2,195	29	CP003540	[56]
1/06-A	*equi*	Horse	USA	Abscess	2,279,118	2,127	20	CP003082	[57]
Cp162	*equi*	Camel	UK	Neck abscess	2,293,464	2,150	13	CP003652	[58]
31	*equi*	Buffalo	Egypt	Abscess	2,297,010	2,170	50	CP003421	[59]

doi:10.1371/journal.pone.0053818.t001

and Europe (the United Kingdom, Belgium, France and Scotland). The clinical symptoms of infections with these strains vary broadly and include abscesses, mastitis, lymphangitis, necrogranuloma, and edematous skin disease (Table 1).

Corynebacterium Genus Phylogenomic Analyses

The Gegenees (version 1.1.4) software was used to perform the phylogenomic analyses at the genus level and to retrieve the GenBank sequences of all the complete *Corynebacterium* genomes from the NCBI ftp site. Briefly, Gegenees was used to divide the genomes into small sequences and to perform an all-versus-all similarity search to determine the minimum content shared by all the genomes. Next, the minimum shared content was subtracted from all the genomes, resulting in the variable content, which was compared with all the other strains to generate the percentages of similarity. Finally, these percentages were plotted in a heatmap chart with a spectrum ranging from red (low similarity) to green (high similarity) [60]. The Gegenees data can also be exported as a distance matrix file in nexus format. Here, we used the distance matrix as an input file for the SplitsTree (version 4.12.6) software to generate a phylogenomic tree using the UPGMA method [61,62].

Pan-genome, Core Genome and Singleton Analyses

This section describes the analyses that were performed for all of the following three datasets: A) all strains, using *C. pseudotuberculosis* strain 1002 as a reference; B) the biovar *ovis* strains, using *C. pseudotuberculosis* strain 1002 as a reference; and C) the biovar *equi* strains, using *C. pseudotuberculosis* strain CIP52.97 as a reference. To calculate the pan-genome, core genome and singletons of the *C. pseudotuberculosis* species, we used EDGAR (version 1.2), multiple-strain genome comparison software that performs homology analyses based on a specific cutoff that is automatically adjusted to the query dataset [63]. Initially, the genome sequences of *C. pseudotuberculosis* were retrieved from GenBank, and a new project was created on the annotation platform GenDB (version 2.4) to homogenize the genome annotations [64]. Subsequently, an EDGAR project was created based on the GenDB annotations, and homology calculations based on BLAST Score Ratio Values (SRVs) were performed. According to the SRV method, instead of using raw BLAST scores or E-values, a normalization of each BLAST bit score is calculated by considering the maximum possible bit score (i.e., the bit score of the subject gene against itself). This results in a value ranging from 0 to 1 [65], which is multiplied by 100 and rounded in a percentage value of homology. Finally, a sliding window on the SRV distribution pattern was used to automatically calculate the SRV cutoff with EDGAR [63]. For this work, a SRV cuttof of 59 was estimated. Pairs of genes exhibiting a Bidirectional Best Hit where both single hits have a SRV higher than the specific cutoff were considered to be orthologous genes.

The core genome was calculated as the subset of genes presenting orthologs in all the selected strains. The gene set of subject strain A was compared with the gene set of query strain B, and only genes with orthologs in both strains were members of core AB. The resulting subset was then compared with the gene set of query strain C, and the comparisons continued in a reductive manner. The pan-genome was calculated in the same way, but in an additive manner: the initial pan-genome was composed of strain A, and the non-orthologous genes of strain B were added to pan-genome A to create the pan-genome AB. The resulting set of genes was then compared with strain C, and the comparisons continued in the same manner. Finally, the singletons were calculated as genes that were present in only one strain and thus did not present orthologs in any other *C. pseudotuberculosis* sequenced strain.

The developments of the core genome, pan-genome and singletons of *C. pseudotuberculosis* were calculated based on permutations of all the sequenced genomes. The developments of the core genome and singletons were calculated using the least-squares fit of the exponential regression decay to the mean values. In contrast, the statistical computing language R was used to calculate the pan-genome extrapolation using Heaps' Law by estimating the parameters κ and γ using the nonlinear least-squares curve fit to the mean values [66,67].

The core genes of all the strains, including the biovar *ovis* strains and the biovar *equi* strains, were classified by their Cluster of Orthologous Genes (COG) functional category as the following: 1. information storage and processing; 2. cellular processes and signaling; 3. metabolism; and 4. poorly characterized. To perform this analysis, the query sets of core genes were submitted to BLAST protein (blastp) similarity searches against the COG database, the proteins with E-values higher than 10^{-6} were discarded, and the best BLAST results for each protein were considered for the COG functional category information retrieval. Finally, the whole-genome comparison maps were visualized using the software CGView Comparison Tool (CCT) [68]. All the strains were plotted against *C. pseudotuberculosis* strains 1002 and CIP52.97 to generate two genome comparison maps.

Pathogenicity Island Prediction

The plasticity of the 15 genomes was assessed using PIPS: Pathogenicity Island Prediction Software (version 1.1.2). PIPS is a multi-pronged approach that predicts pathogenicity islands (PAIs) based on common features, such as G+C content, codon usage deviation, high concentrations of virulence factors and hypothetical proteins, the presence of transposases and tRNA flanking sequences, and the absence of the query region in non-pathogenic organisms of the same genus or related species [69]. *C. glutamicum* strain ATCC 13032 was selected as the non-pathogenic organism of the same genus [6], and separate predictions were performed for each strain. The sizes of the islands were compared with those of all the other strains via ACT: Artemis Comparison Tool (version 10.2.0) and CCT [68,70]. Following the curation of the PAIs, the genes of all the islands in each strain were assessed for their presence/absence in all the other strains using the pan-genome data generated by EDGAR. The overall number of genes in the PAIs of the subject strain that were shared by the query strains was expressed as a percentage and plotted in a heatmap. The percentages were also converted into a nexus file, which was used in SplitsTree (version 4.12.6) to create a phylogenomic tree using the UPGMA method [61,62]. Finally, zoomed PAI figures were created using a script from CCT (create_zoomed_maps.sh) with the zoom option selected as 30×.

Results

Phylogenomics of the Genus *Corynebacterium* and *C. pseudotuberculosis* Biovars

To evaluate the phylogenomic relationships between *C. pseudotuberculosis* strains and other species of the genus *Corynebacterium*, the *Corynebacterium* shared gene content was automatically determined using Gegenees. Then, the shared gene content was subtracted from all genomes and the resulting variable content of each genome sequence was cross-compared to generate a phylogenomic tree and to plot a heatmap (Figure 1). According to the generated phylogenomic tree, the pathogenic species *C. diphtheriae*, *C. pseudotuberculosis*, and *C. ulcerans* formed three closely related clusters. Moreover, *C. glutamicum* and *Corynebacterium efficiens*, two non-pathogenic bacteria of great industrial importance as amino acid producers [6,71], appeared closely related in a different cluster. Additionally, *Corynebacterium kroppenstedtii*, another pathogenic bacterium of the *Corynebacterium* genus, was positioned between the clusters of pathogens (*C. pseudotuberculosis*, *C. diphtheriae* and *C. ulcerans*) and non-pathogens (*C. glutamicum* and *C. efficiens*). Finally, the opportunistic bacteria *C. jeikeium*, *Corynebacterium urealyticum* and *Corynebacterium resistens* [5,72,73] clustered together with the non-pathogenic *Corynebacterium variabile* [74], whereas *Corynebacterium aurimucosum* formed a new branch [75].

At the species level, the *C. pseudotuberculosis* genomes clustered in two separate groups representing the two biovars of the species: biovar *ovis*, with more than 99% similarity according to the heatmap; and biovar *equi*, with a similarity ranging from 95% to 100%. Moreover, the heatmap indicated an almost clonal-like behavior of *C. pseudotuberculosis* compared with the *C. diphtheriae* species, which presented similarities raging from 82% to 100%.

An alternative to assess the clonal-like behavior of species is the use of a circular genome comparison, which was performed with the software CCT. The results reveal regions of plasticity based on a chosen reference and, interestingly, plot the genomes from outer to inner circles by order of decreasing similarity. As shown in Figure 2, we plotted all the genomes using *C. pseudotuberculosis* strain 1002 (bv. *ovis*) and *C. pseudotuberculosis* strain CIP52.97 (bv. *equi*) as references. Figure 2A shows specific patterns of deletions in all the biovar *equi* strains compared with *C. pseudotuberculosis* 1002. In Figure 2B, however, the deletions in the comparison with *C. pseudotuberculosis* CIP52.97 are not specific to particular biovars, but rather are generalized. In both cases, the genomes that were classified as having the same biovar as the reference strain were clustered together in the outer circles, whereas the other strains were clustered in the inner circles.

The Pan-genome of the Species *C. pseudotuberculosis*

To achieve a global view of the genome repertoire of *C. pseudotuberculosis*, the pan-genome (i.e., the total number of non-redundant genes) was calculated using the abovementioned SRV method with the software EDGAR (Figure 3). The resulting pan-genome of *C. pseudotuberculosis* contained a total of 2,782 genes, which is 1.3-fold the average total number of genes in each of the 15 strains (2,078). However, when the pan-genomes of the biovars were calculated separately, a slightly different scenario emerged, in which the biovar *ovis* had a pan-genome of 2,403 genes, 1.14-fold the average total number of genes in each biovar *ovis* strain (2,098), and the biovar *equi* had a pan-genome with 2,521 genes, 1.23-fold the average total number of genes in each biovar *equi* strain (2,047).

Additionally, the extrapolation of the *C. pseudotuberculosis* pan-genome was calculated by curve fitting based on Heaps' Law, as represented by the formula $n = \kappa * N^{-\alpha}$, where n is the expected number of genes for a given number of genomes, N is the number of genomes, and the other terms are constants defined to fit the specific curve [67]. The variables κ and γ were determined to be 2,043.06 and 0.11, respectively, by using the statistical computing language R. According to Heaps' Law,1) an $\alpha \leq 1$ is representative of an open pan-genome, meaning that each added genome will contribute some new genes and the pan-genome will increase, and 2) an $\alpha > 1$ represents a closed pan-genome, in which the addition of new genomes will not significantly affect the pan-genome. Using the formula $\alpha = 1 - \gamma$, we inferred that the pan-genome of *C. pseudotuberculosis* is increasing with an α of 0.89, indicating that it has an open pan-genome. The extrapolation of the pan-genome was also separately calculated for both biovars, *ovis* and *equi*. Although the biovar *equi* had the same α as the entire pan-genome (0.89), the biovar *ovis* had a much-higher α of 0.94.

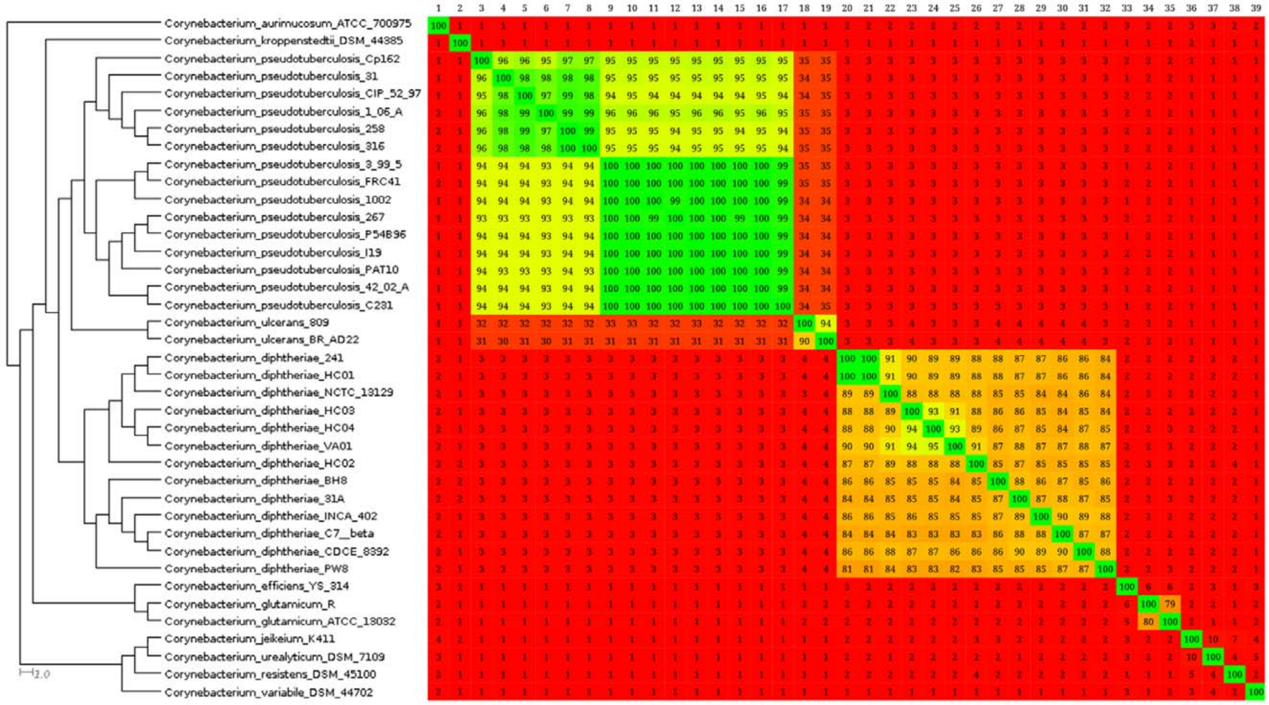


Figure 1. Phylogenomic tree and heatmap analyses of the genus *Corynebacterium*. All the complete genomes from the genus *Corynebacterium* were retrieved from the NCBI ftp site. Comparisons between the variable content of all the strains were plotted as percentages of similarity on the heatmap using Gegenees (version 1.1.4). The percentage of similarity was used to generate a phylogenomic tree with SplitsTree (version 4.12.6). Numbers from 1 to 39 (upper-left to upper-right corner) represent species from *Corynebacterium aurimucosum* ATCC 70097 to *Corynebacterium variable* DSM 44702 (upper-left to lower-left corner). Percentages were plotted with a spectrum ranging from red (low similarity) to green (high similarity). On the heatmap, the upper portion is not symmetrical to the lower portion because the variable contents of all genomes present different sizes. Therefore, considering a scenario where the variable content from genomes A and B are composed of 100 and 80 genes, respectively, with a common repertoire of 40 genes, genome A will present 40% of similarity to genome B and genome B will present 50% of similarity to genome A.
doi:10.1371/journal.pone.0053818.g001

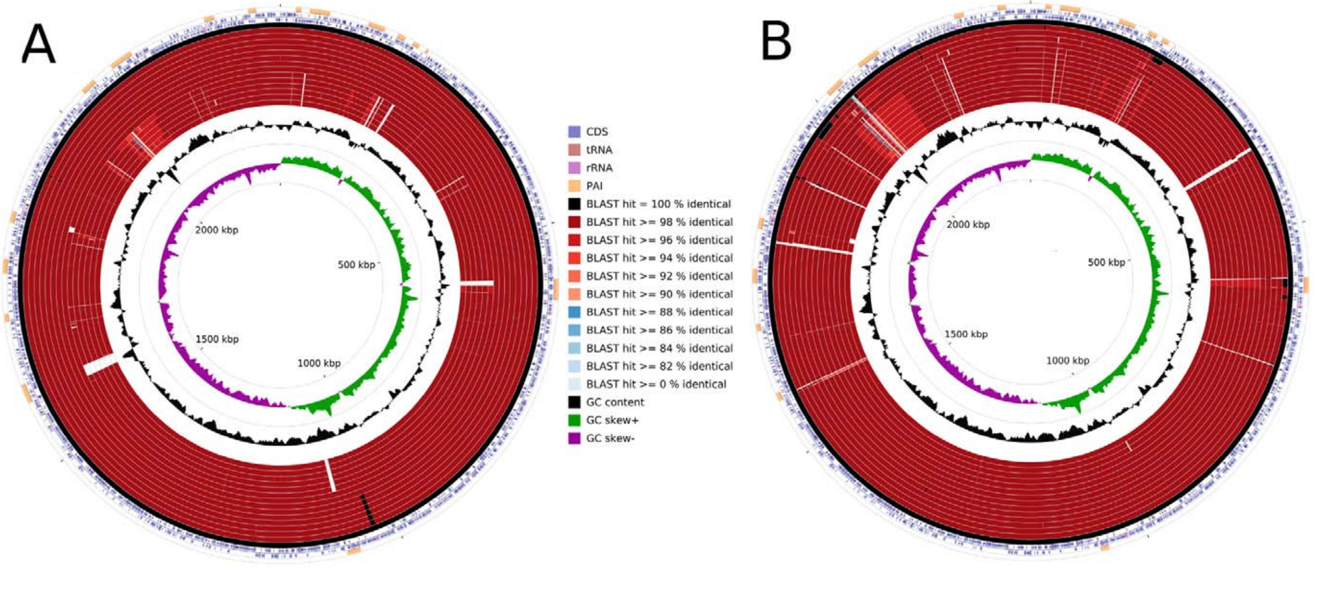


Figure 2. Comparative genomic maps of the *C. pseudotuberculosis* biovar *equi* and *ovis* strains. A, all the *C. pseudotuberculosis* strains were aligned using *C. pseudotuberculosis* strain 1002 as a reference. From the inner to outer circle on A: the biovar *equi* strains Cp31, Cp1/06-A, CpCp162, Cp258, Cp316 and CpCIP52.97; and, the biovar *ovis* strains CpC231, CpP54B96, Cp267, CpPAT10, CpI19, Cp42/02-A, Cp3/99-5, CpFRC41 and Cp1002. B, all the *C. pseudotuberculosis* strains were aligned using *C. pseudotuberculosis* strain CIP52.97 as a reference. From the inner to outer circle on B: the biovar *ovis* strains CpC231, Cp1002, CpPAT10, Cp267, CpP54B96, CpI19, Cp42/02-A, CpFRC41, Cp3/99-5; and, the biovar *equi* strains Cp1/06-A Cp31, CpCp162, Cp316, Cp258 and CpCIP52.97. CDS, coding sequences; tRNA, transfer RNA; rRNA, ribosomal RNA; and PAI, pathogenicity island.
doi:10.1371/journal.pone.0053818.g002

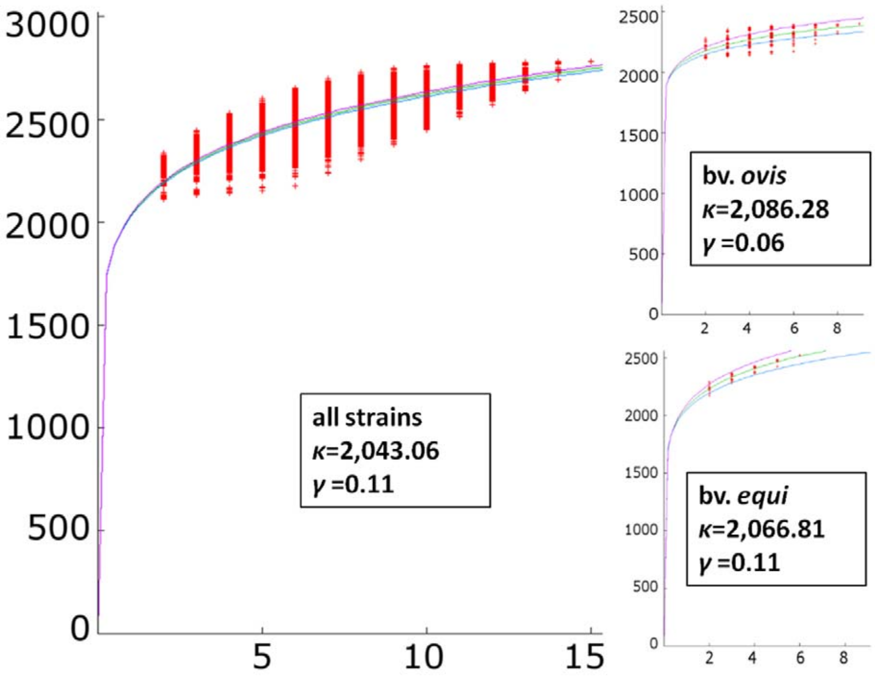


Figure 3. Pan-genome development of *C. pseudotuberculosis*. Center chart, the pan-genome development using permutations of all 15 strains of *C. pseudotuberculosis*; upper-right chart, the pan-genome development of the *C. pseudotuberculosis* biovar *ovis* strains; lower-right chart, the pan-genome development of the *C. pseudotuberculosis* biovar *equi* strains.
doi:10.1371/journal.pone.0053818.g003

Core Genome of the Species *C. pseudotuberculosis*

The core genome of a species is defined as the subset of genes from the pan-genome that are shared by all strains. Here, the core genome of *C. pseudotuberculosis* was calculated with the software EDGAR by defining the subset of genes that presented orthologs in all the strains using the SRV method. The subset of core genes of *C. pseudotuberculosis* contained 1,504 genes, which represented 54% of the entire pan-genome of the species (2,782 genes). This subset may decrease with the addition of new genomes, as shown by the tendency of the core genes in the blue curve (Figure 4). However, although this subset may slightly decrease, the extrapolation of the curve can be calculated by the least-squares fit of the exponential regression decay to the mean values, as represented by the formula $n = \kappa * exp[-x/\tau]+tg(\theta)$, where *n* is the expected subset of genes for a given number of genomes, *x* is the number of genomes, *exp* is Euler's number, and the other terms are constants defined to fit the specific curve. Interestingly, that formula can be used to predict that with a high number of genomes (*x*), the $\kappa * exp[-x/\tau]$ term will tend toward 0, where $tg(\theta)$ represents the convergence of the genome subset. Based on this observation, the core genome of *C. pseudotuberculosis* tended to converge to 1,347 genes, which represented 48% of the pan-genome of the species (2,782 genes).

The separate analyses of the core genomes of biovars *ovis* and *equi* (Figure 4) presented different scenarios. The core genome of the *C. pseudotuberculosis* biovar *ovis* strains contained 1,818 genes, and it tended to stabilize at approximately 1,719 genes, according to the exponential regression decay. The *C. pseudotuberculosis* biovar *equi* strains, however, presented a more compact core genome of 1,599 genes and tended to stabilize at 1,404 genes. Altogether, with a total *C. pseudotuberculosis* core genome of 1,504 genes and a biovar *ovis* core genome of 1,818 genes, the core genome of biovar *ovis* is predicted to contain 314 orthologous genes that are shared by all strains from this biovar and are absent from one or more strains of biovar *equi* (Figure 5). Additionally, using the same strategy, the biovar *equi*, with 1,599 genes, contained 95 core genes that were absent from one or more strains of biovar *ovis* (Figure 5).

The core genome of all the strains and the differential core genome of the biovar *ovis* and *equi* strains were classified by COG functional category. According to the chart in Figure 6, the core genome of all the strains had a large number of genes related to the categories "Metabolism" and "Information storage and processing". Moreover, a high proportion of the core genome of all the strains was classified as "Poorly characterized". However, when analyzing the differential core genes of the biovar *ovis* and *equi* strains separately, a higher proportion of "Poorly characterized" genes was clearly detected in the differential core genes when compared with the core genome of all the strains (Figure 6). Finally, the biovar *equi* had a larger number of genes classified under the functional category "Cellular processes and signaling" than biovar *ovis* strains.

Singletons: Strain-specific Genes Detected in the Species *C. pseudotuberculosis*

The singletons of a strain are defined as the subset of genes that are absent from all the other strains and are thus responsible for increases in the number of genes in the pan-genome. We used the SRV method and EDGAR to calculate the subset of *C. pseudotuberculosis* singletons as the genes that did not present orthologs in any other strain. Moreover, by the least-squares fit of the exponential regression decay to the mean values, as previously described by the formula $n = \kappa * exp[-x/\tau]+tg(\theta)$, we calculated the $tg(\theta)$ (Figure 4) for the three datasets: A) all the genomes, B) the biovar *ovis* genomes, and C) the biovar *equi* genomes. The $tg(\theta)$ for all the genomes was 18.805, meaning that each sequenced genome added approximately 19 genes to the total gene pool of the species *C. pseudotuberculosis*, i.e., the pan-genome. However, the individual analysis of each biovar revealed a scenario in which each

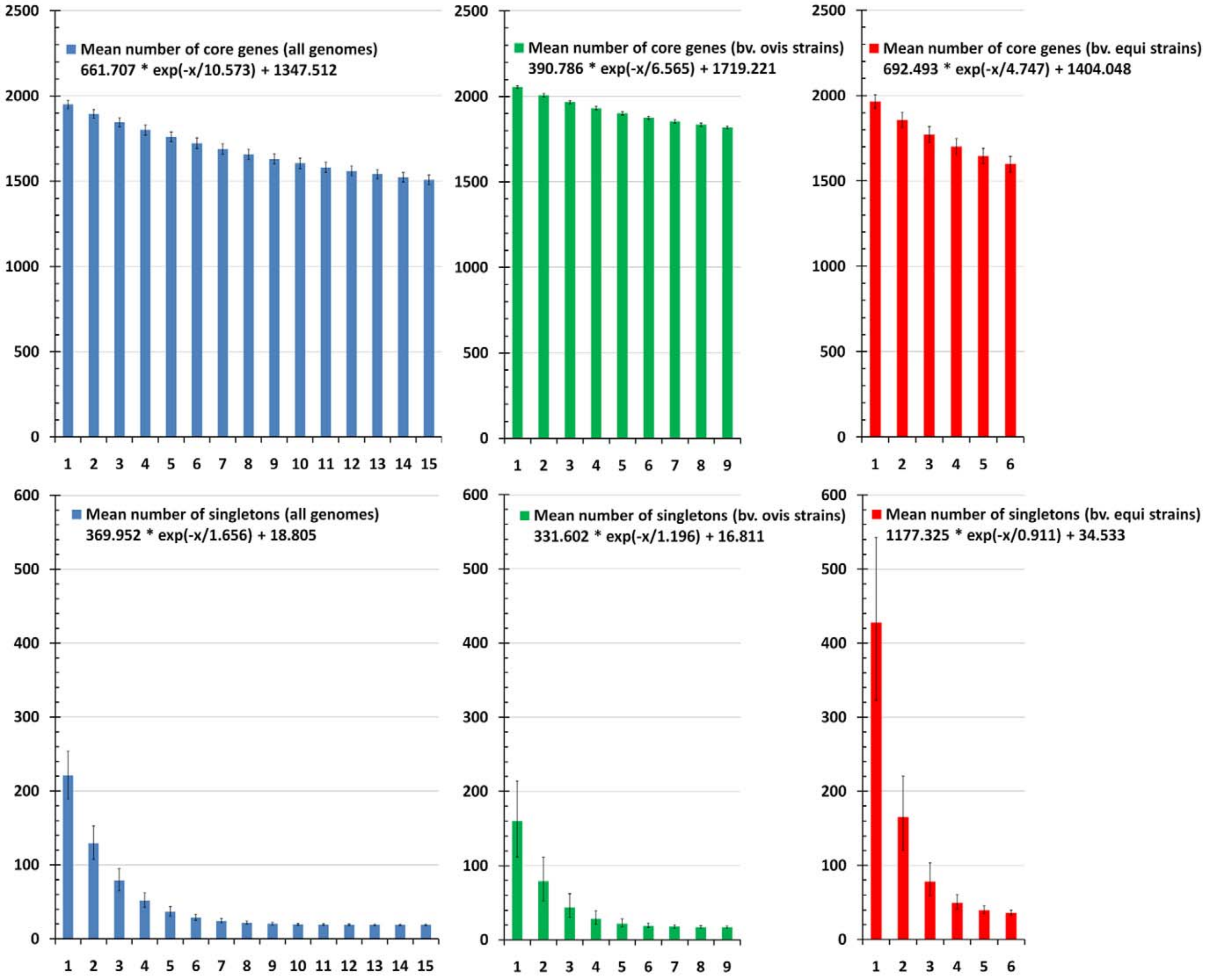


Figure 4. Core genome and singleton development of *C. pseudotuberculosis*. Upper-left, the core genome development using permutations of all 15 strains of *C. pseudotuberculosis*; upper-center, the core genome development of the *C. pseudotuberculosis* biovar *ovis* strains; upper-right, the core genome development of the *C. pseudotuberculosis* biovar *equi* strains; lower-left, the singleton development using permutations of all 15 strains of *C. pseudotuberculosis*; lower-center, the singleton development of the *C. pseudotuberculosis* biovar *ovis* strains; lower-right, the singleton development of the *C. pseudotuberculosis* biovar *equi* strains.
doi:10.1371/journal.pone.0053818.g004

sequenced biovar *ovis* strain contributed ~16 genes, but each sequenced biovar *equi* strain contributed ~34 genes.

Detection of PAIs in the *C. pseudotuberculosis* Genomes

Intraspecies genome plasticity may result from several events, of which horizontal gene transfer is particularly important because it can cause the acquisition of blocks of genes (genomic islands, or GEIs), producing evolution by quantum leaps [76]. PAIs are important in this context because they represent a class of GEIs that carry virulence genes, i.e., factors that enable or enhance the parasitic growth of an organism inside a host [77]. Therefore, high concentrations of the two following subsets of genes would be expected inside PAIs: 1) shared genes, which are shared by two or more, but not all, strains; and 2) singletons.

In previous studies, seven PAIs were identified in *C. pseudotuberculosis* biovar *ovis* strains 1002 and C231 (PiCps 1–7) [43], and four additional PAIs have been identified in *C. pseudotuberculosis* strain 1002 by further comparisons with *C. pseudotuberculosis* strains 316 and 258 (PiCps 8–11) [54–56]. The latter subset of PAIs was identified due to a better view of the two biovars and their specific patterns of plasticity. Here, we applied the same methodology used in the previous studies, using the software PIPS to achieve a global view of the PAIs in 15 *C. pseudotuberculosis* strains. Briefly, in addition to the previously identified 11 PAIs, we found 5 new PAIs, identified as PiCps 12–16. Although the 16 PAIs are present in all strains, they have different patterns of deletions, especially in the biovar *equi* strains (Figure 2). PiCp1, as previously described [43], harbors the *pld* gene and the *fag* operon and is present in all the strains. PiCp3 harbors the diphtheria toxin gene (*tox*) in *C. pseudotuberculosis* strain 31, and PiCps 7 and 15 harbor the *spaD* and *spaA* pilus gene clusters, respectively.

To assess the level of plasticity in the PAIs, we used the orthologous data predicted by EDGAR to calculate the percentage of PAIs (from each strain) present in each of the other strains. Using these data, we generated a phylogenomic tree of the strains with SplitsTree (Figure 7). The phylogenomic tree produced a clear

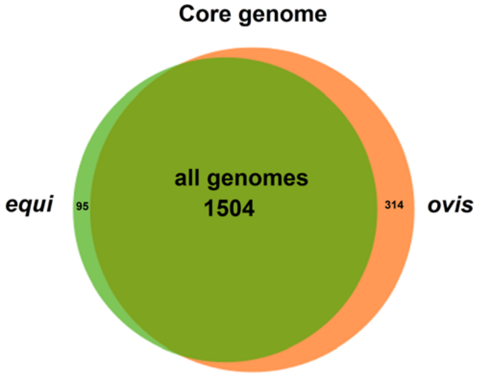


Figure 5. Venn diagram representing the core genomes of the *C. pseudotuberculosis* strains. All genomes, the number of genes composing the core genome of all the strains; *equi*, the number of genes of the core genome of the *C. pseudotuberculosis* biovar *equi* strains, which were absent in one or more of the *C. pseudotuberculosis* biovar *ovis* strains; *ovis*, the number of genes of the core genome of the *C. pseudotuberculosis* biovar *ovis* strains, which were absent in one or more of the *C. pseudotuberculosis* biovar *equi* strains.
doi:10.1371/journal.pone.0053818.g005

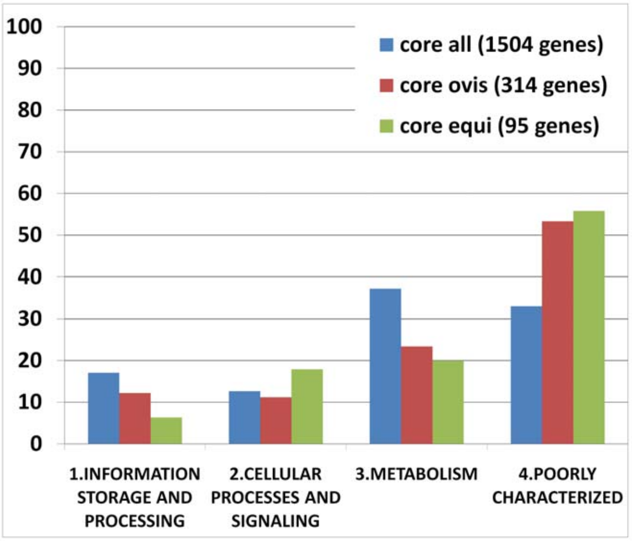


Figure 6. Core genes of the *C. pseudotuberculosis* strains classified by COG functional category. Core all, the genes composing the core genome of all the strains; core *ovis*, the genes of the core genome of the *C. pseudotuberculosis* biovar *ovis* strains, which were absent in one or more of the *C. pseudotuberculosis* biovar *equi* strains; core *equi*, the genes of the core genome of the *C. pseudotuberculosis* biovar *equi* strains, which were absent in one or more of the *C. pseudotuberculosis* biovar *ovis* strains.
doi:10.1371/journal.pone.0053818.g006

separation of the *ovis* and *equi* biovar strains, similar to the phylogenomic tree created using Gegenees (Figure 1). A further comparison of the Gegenees and PAI phylogenomic trees revealed that the latter strategy did not cluster *C. pseudotuberculosis* strains 42/02-A and C231 in the same branch as did the former. However, two other branches were in agreement with the phylogenomic tree created by Gegenees: *C. pseudotuberculosis* strains 258 and 316 clustered together in a biovar *equi* group, and *C. pseudotuberculosis* strains 3/99-5 and FRC41 clustered in a biovar *ovis* group.

Additionally, we used the comparison data generated by the PAI analyses to create a new heatmap (Figure 7), from which we deduced a high level of intra-biovar similarity in the *ovis* strains with respect to the PAI content (82–100%). Although biovar *ovis* showed a lower level of similarity to biovar *equi* with respect to the PAI content (78–91%), the former tended to present a similar deletion pattern in the same PAIs, independent of the strain. The biovar *equi* strains, however, contained large deletions and a lower level of similarity intra-biovar (77–88%) and also compared with the biovar *ovis* PAIs (62–74%) (Figure 2A).

Variations in Pathogenicity Islands Encoding Exotoxin Virulence Factors

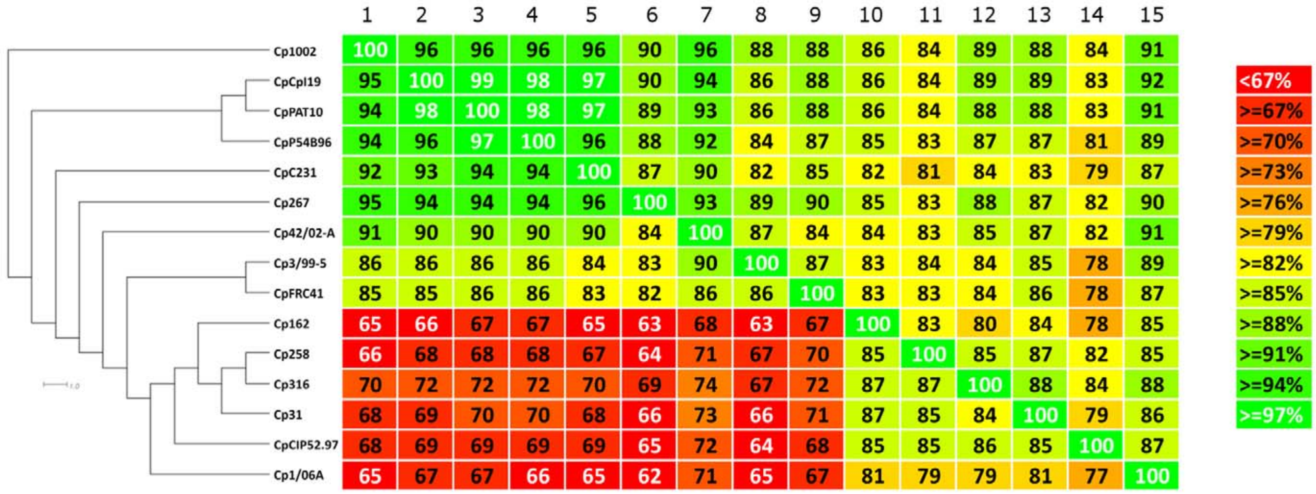
As described previously, the major toxin of *C. pseudotuberculosis* is phospholipase D (PLD), which is encoded by the *pld* gene and is strongly associated with the spread of bacteria throughout the host cells [1]. In a previous study, this toxin was shown to be harbored by a PAI (PiCp1) close to the *fag* operon, which also encodes important virulence factors that are responsible for iron acquisition in environments where this element is scarce [43]. Here, we found that the *pld* gene was present in 14 of 15 strains, with similarities ranging from 98–100%. This finding was expected due to the important role of PLD during the disease course; *pld* mutants present a diminished ability to spread throughout the host [1].

Although the *pld* gene plays a pivotal role in pathogenesis, *C. pseudotuberculosis* strain 31 contains a frameshift mutation near the 3′-end of this gene that could decrease the ability of this strain to spread throughout the host. However, *C. pseudotuberculosis* strain 31 was the only strain in our dataset to present another important virulence factor, the diphtheria toxin gene (*tox*) (Cp31_0135). The diphtheria toxin (DT) is an important virulence factor in *C. diphtheriae*, in which the gene was acquired through lysogenization by corynephages, meaning that the *tox* gene is also present in a PAI in this species and can be horizontally transferred to other organisms. Briefly, the *tox* gene is regulated by the chromosomal iron-dependent repressor DtxR [78], which blocks the transcription process by binding to the *tox* operator [79]. When gene transcription is activated, the toxin precursor is exported and cleaved into two fragments (A and B), which are joined by a disulfide bond [80]; fragment B binds the membrane of the host cell, mediating the internalization of fragment A, which exhibits ADP-ribosyltransferase activity [79,81].

The exotoxin catalizes the transfer of adenosine diphosphate ribose (ADP-ribosylation) from nicotinamide adenine dinucleotide (NAD) to a histidine residue of elongation factor 2 (EF-2), called diphthamide. This process leads to inactivation of EF-2 and inhibits chain elongation during protein synthesis [82]. This toxin has also been identified in *C. ulcerans* strains, where it causes diphtheria-like illness [83,84], and, interestingly, in two *C. pseudotuberculosis* strains isolated from buffalo in Egypt [10,85]. The *tox* gene from *C. pseudotuberculosis* 31 has 560 amino acids in length, does not present any frameshift and has ~96–97% similarity to the *tox* genes from several *C. diphtheriae* strains and from corynephage β, as well as ~94–95% similarity to the *tox* gene from *C. ulcerans* 0102 (data not shown). Given the absence of the *pld* gene, the similarity of the *tox* gene from *C. pseudotuberculosis* to those from the *C. diphtheriae* strains, the conservation of all the domains and the presence of the gene in other strains isolated from buffalo in Egypt, the following question can be raised: is DT required for *C. pseudotuberculosis* to infect buffalo or is this feature more closely related to the geographical location (Egypt) than to the host?

Variations and Deletions Detected in PiCps 4, 5 and 9

Specific patterns of deletions in PiCps 4, 5 and 9 of *C. pseudotuberculosis* CIP52.97, 316 and 258 (biovar *equi* strains) have



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Cp1002	100	96	96	96	96	90	96	88	88	86	84	89	88	84	91
CpCpI19	95	100	99	98	97	90	94	86	88	86	84	89	89	83	92
CpPAT10	94	98	100	98	97	89	93	86	88	86	84	88	88	83	91
CpP54B96	94	96	97	100	96	88	92	84	87	85	83	87	87	81	89
CpC231	92	93	94	94	100	87	90	82	85	82	81	84	83	79	87
Cp267	95	94	94	94	96	100	93	89	90	85	83	88	87	82	90
Cp42/02-A	91	90	90	90	90	84	100	87	84	84	83	85	87	82	91
Cp3/99-5	86	86	86	86	84	83	90	100	87	83	84	84	85	78	89
CpFRC41	85	85	86	86	83	82	86	86	100	83	83	84	86	78	87
Cp162	65	66	67	67	65	63	68	63	67	100	83	80	84	78	85
Cp258	66	68	68	68	67	64	71	67	70	85	100	85	87	82	85
Cp316	70	72	72	72	70	69	74	67	72	87	87	100	88	84	88
Cp31	68	69	70	70	68	66	73	66	71	87	85	84	100	79	86
CpCIP52.97	68	69	69	69	69	65	72	64	68	85	85	86	85	100	87
Cp1/06A	65	67	67	66	65	62	71	65	67	81	79	79	81	77	100

Figure 7. Phylogenomic tree and heatmap analyses of the *Corynebacterium pseudotuberculosis* strains based on pathogenicity island plasticity. Comparisons between the PAI contents of all the strains were plotted as percentages of similarity on the heatmap using Gegenees (version 1.1.4). The percentages of similarity were used to generate a phylogenomic tree with SplitsTree (version 4.12.6). Numbers from 1 to 15 (upper-left to upper-right corner) represent the strains from Cp1002 to Cp1/06-A (upper-left to lower-left corner). On the heatmap, the upper portion is not symmetrical to the lower portion because the pathogenicity islands contents of all genomes present different sizes. Therefore, considering a scenario where the pathogenicity islands content from genomes A and B are composed of 100 and 80 genes, respectively, with a common repertoire of 40 genes, genome A will present 40% of similarity to genome B and genome B will present 50% of similarity to genome A.
doi:10.1371/journal.pone.0053818.g007

been demonstrated [54–56]. Here, we detected the same deletions in all the biovar *equi* strains, which indicates that these deletion events were specific to the mentioned biovar (Figure S1). Although most of the deleted CDSs encoded hypothetical or phage proteins (integrases and phage-associated proteins), one gene of PiCp5 encoded a putative sigma 70 factor (Cp1002_1452) and deserves attention because it is most likely involved in the correct assembly of the transcription machinery at specific promoters and is therefore associated with the general transcription process [43].

Differences between Pilus Gene Clusters Located on PiCp15 and PiCp7

According to work performed by Yanagawa and Honda in 1976 [47], *C. pseudotuberculosis* cells possess pilus structures, although the number of pili per bacterial cell is small, and at times, a long bundle measuring more than several micrometers in length was the only pilus observed. In a more recent genomic study, two clusters of pilus genes were described in *C. pseudotuberculosis* FRC41 and were named according to their major pilin gene: the *spaA* (*srtB-spaA-srtA-spaB-spaX-spaC*) and *spaD* (*srtC-spaD-spaY-spaE-spaF*) clusters, where *srtA* and *srtB* are the specific sortases of the *spaA* cluster; *spaA*, *spaB* and *spaC* encode the major, base and tip pilin proteins, respectively, of the *spaA* cluster; *srtC* is the specific sortase of the *spaD* cluster; *spaD*, *spaE* and *spaF* encode the major, base and tip pilin proteins, respectively, of the *spaD* cluster; and *spaX* and *spaY* have currently unknown functions. Additionally, a housekeeping sortase (*srtD*) is likely responsible for anchoring the pili to the cellwall [46].

Interestingly, the *spaA* and *spaD* gene clusters were located in PAIs (PiCps 15 and 7, respectively) (Figure 8), which is in agreement with the presence of pilin genes in horizontally acquired regions of Gram-negative and Gram-positive bacteria, such as *Vibrio cholerae* and *C. diphtheriae*, respectively [86,87]. Moreover, although the biovar *ovis* strains had a complete *spaA* cluster, the biovar *equi* strains contained a large deletion at the position where the *spaA* and *srtB* genes should be located (PiCp15). Furthermore, the entire *srtA-spaB-spaX-spaC* region presented a low similarity to the same region in the biovar *ovis* strains, which was caused by small deletions, frameshift mutations and nucleotide substitutions (Figure 8).

With respect to the *spaD* cluster of the biovar *ovis* strains, the major pilin gene *spaD* contains a frameshift in *C. pseudotuberculosis* P54B96 and PAT10; and in *C. pseudotuberculosis* 267, the tip pilin gene *spaF* also contains a frameshift. In biovar *equi* strains, the *spaD* gene of all the strains had 99% similarity to the *spaD* gene of the biovar *ovis* strains. However, *C. pseudotuberculosis* CIP52.97 contains a frameshift mutation in the specific sortase gene *srtC*. Furthermore, the base and tip pilin genes, *spaE* and *spaF*, respectively, of *C. pseudotuberculosis* strains 258, 316, 1/06-A and Cp162 are merged into the same reading frame.

Discussion

Corynebacterium pseudotuberculosis – all Strains

According to the *rpoB* gene tree generated by Khamis *et al.* [88], *C. jeikeium*, *C. urealyticum*, *C. kroppenstedtii* and *C. variabile* cluster together in group 3, and *C. aurimucosum* appears in group 1. Moreover, *C. glutamicum* and *C. efficiens* cluster together in one branch, whereas *C. pseudotuberculosis*, *C. diphtheriae* and *C. ulcerans* appear closely related in another branch. Furthermore, *C. ulcerans* appears closer to *C. pseudotuberculosis* than to *C. diphtheriae*. Based on our results, we can deduce that although many variable regions exist between the pathogenic members of the genus *Corynebacterium*, these species tend to cluster together because they most likely share some core virulence determinants. Finally, although *C. kroppenstedtii* did not cluster with group 3, the other species were in perfect agreement with the *rpoB* analysis of Khamis *et al.* [88].

Two striking characteristics of *C. kroppenstedtii* are the absence of mycolic acids in the cell wall (due to the losses of a condensase gene cluster and a mycolate reductase gene) and a lipophilic phenotype (due to the absence of a microbial type I fatty acid synthase gene) [89]. Therefore, the transitional phylogenomic position of *C. kroppenstedtii* between the pathogenic and non-pathogenic species was in agreement with the lack of important

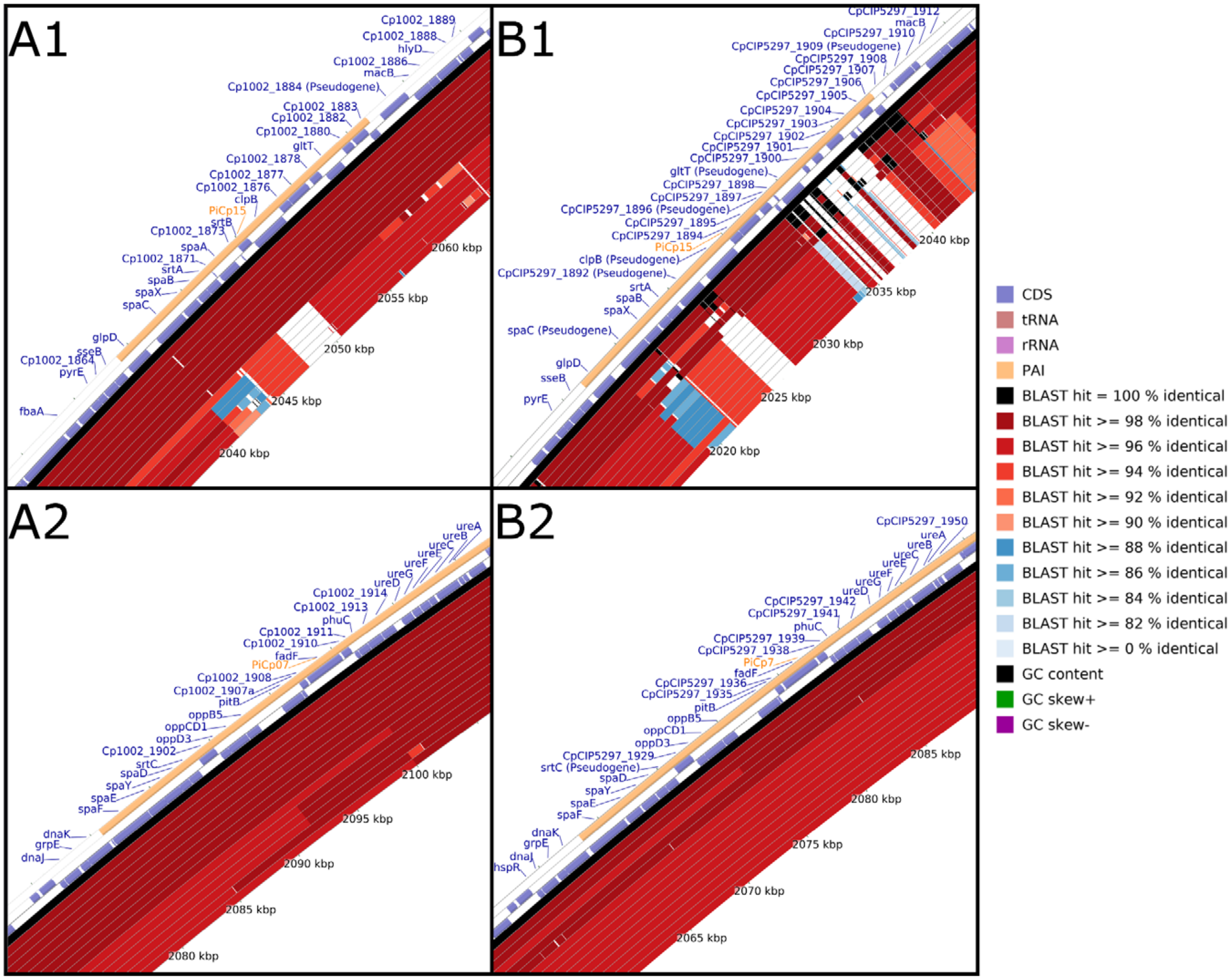


Figure 8. Plasticity of the pilus gene clusters *spaA* and *spaD* in *C. pseudotuberculosis*. A1 and B1, PiCp15 harboring the *spaA* cluster of genes; A2 and B2, PiCp7 harboring the *spaD* cluster of genes. A, all the *C. pseudotuberculosis* strains were aligned using *C. pseudotuberculosis* strain 1002 as a reference. From the inner to outer circle on A1 and A2: the biovar *equi* strains Cp31, Cp1/06-A, CpCp162, Cp258, Cp316, CpCIP52.97; and, the biovar *ovis* strains CpC231, CpP54B96, Cp267, CpPAT10, CpI19, Cp42/02-A, Cp3/99-5, CpFRC41 and Cp1002. B, all the *C. pseudotuberculosis* strains were aligned using *C. pseudotuberculosis* strain CIP52.97 as a reference. From the inner to outer circle on B1 and B2: the biovar *ovis* strains CpC231, Cp1002, CpPAT10, Cp267, CpP54B96, CpI19, Cp42/02-A, CpFRC41, Cp3/99-5, Cp1/06-A; and, the biovar *equi* strains Cp31, CpCp162, Cp316, Cp258 and CpCIP52.97. CDS, coding sequences; tRNA, transfer RNA; rRNA, ribosomal RNA; and PAI, pathogenicity island.
doi:10.1371/journal.pone.0053818.g008

virulence genes and the low pathogenic potential characteristic of *C. kroppenstedtii* [89–91].

At the species level, the heatmap indicated a clonal-like behavior of *C. pseudotuberculosis* compared with the *C. diphtheriae* species. Trost *et al.* [87] have highlighted the high plasticity of the *C. diphtheriae* genome, which is mainly related to the 57 genomic islands identified in this species. With respect to the clonal-like behavior of *C. pseudotuberculosis*, Bolt [92] have identified 10 STs among 73 strains of *C. pseudotuberculosis* typed by MLST, where 7 and 4 STs were associated with 64 and 9 strains of biovar *ovis* and *equi*, respectively. The few number of STs identified by MLST was in agreement with previous typing studies [17,93,94] in that the strains of *C. pseudotuberculosis* are clonally related. Moreover, although there were 7 STs identified for biovar *ovis* strains, 6 and 7 of them were clustered in one sole eBURST group when considering single locus variants (SLVs) and double locus variants (DLVs), respectively; and, all the STs identified for biovar *equi* shared two alleles with the biovar *ovis* strains [92]. Finally, the MLST findings indicate that: 1) biovar *ovis* and *equi* strains share a common evolutionary origin, although they are now relatively distinct genotypic clusters; and, 2) biovar *ovis* is a clonal-like organism. Our results with respect to this clonal-like behavior of *C. pseudotuberculosis* are also in agreement with PFGE data from Connor *et al.* [17] and can also be inferred from the extrapolation of the pan-genome data, in which *C. pseudotuberculosis* had a slightly higher α value of 0.89 compared with the *C. diphtheriae* α value of 0.69; and, from the total number of genes in the pan-genome of *C. pseudotuberculosis* (2,782 genes), which is compact compared with that of the closely related species *C. diphtheriae*, which contains 4,786 genes [87].

Although *C. pseudotuberculosis* displays some clonal-like behavior, the resulting α of 0.89 from the extrapolation of the pan-genome indicates that it has an open pan-genome. Moreover, considering that α is inversely proportional to the pan-genome increasing rate, in contrast to the *C. diphtheriae* α of 0.69, the α of 0.89 of the *C. pseudotuberculosis* pan-genome indicates that the latter is increasing at a slower rate. This slow increase is related to the low number of singletons (~19) added to the pan-genome of *C. pseudotuberculosis* by each newly sequenced strain, whereas each strain of *C. diphtheriae* added ~65 genes to the entire pan-genome [87]. Moreover, the slow increase and higher α value are in agreement with the intracellular facultative behavior of this species. Because strictly intracellular organisms tend to have closed pan-genomes due to their limited contact with potential gene donors, an intracellular facultative organism such as *C. pseudotuberculosis*, even when it has different hosts, can be expected to have an α that is closer to 1 than that of *C. diphtheriae* [95,96].

With respect to the core genome of all the strains, a large number of genes are related to the categories "Metabolism" and "Information storage and processing". The "Information storage and processing" category contains genes involved in translation, ribosomal structure and biogenesis, RNA processing and modification, transcription, replication, recombination and repair, and other important functions; the "Metabolism" category contains genes involved in the production and conversion of energy, as well as the transport and metabolism of carbohydrates, amino acids, nucleotides, coenzymes, lipids, inorganic ions and secondary metabolites. Given the importance of the core genome, these two functional categories are expected to be highly represented in the analyzed subset. Finally, although a large number of "Poorly characterized" genes were identified in the core gene subset, this result is in agreement with previous core genome analyses of *Aggregatibacter actinomycetemcomitans*, in which one-third of the genes were categorized as "Poorly characterized" and approximately one-third were classified under "Metabolism" [97].

Corynebacterium pseudotuberculosis – Biovars *Ovis* and *Equi*

Connor *et al.* [17] and Bolt [92] have investigated the clonal aspect of *C. pseudotuberculosis* using PFGE and MLST, respectively, which enabled them to differentiate the *equi* and *ovis* biovars. On the phylogenomic tree, the *C. pseudotuberculosis* genomes also clustered in two separate groups representing the two biovars of the species: biovar *ovis*, with more than 99% similarity according to the heatmap; and biovar *equi*, with a similarity ranging from 95% to almost 100%. This result highlights the higher plasticity of *C. pseudotuberculosis* biovar *equi* compared with the biovar *ovis* strains, although this plasticity is not as high as that described for *C. diphtheriae* strains. Moreover, the same conclusion (regarding the relative plasticity of the two biovars) may be drawn from the number of singletons, in which the biovar *equi* strains presented higher levels of variability in the number of singletons, compared with the biovar *ovis* strains (Table 1). The circular genome comparison generated by CCT also revealed the clonal-like behavior of biovar *ovis*, with all the *ovis* strains containing minor deletions compared with *C. pseudotuberculosis* strain 1002 (Figure 2A); and the presence of a higher number of singletons in biovar *equi*, with all the strains from both biovars presenting similar deletion patterns when compared with *C. pseudotuberculosis* strain CIP52.97 (Figure 2B). Finally, the majority of the genomic variations on the circular genome comparison were found in PAI regions, which are very important for virulence potential and host adaptation and are known as mosaic and unstable [69].

Interestingly, the analysis of the pan-genome subsets revealed that the *ovis* and *equi* biovar strains contain major variations of the data found in the entire pan-genome. Although the pan-genome of biovar *equi* had an invariable α value of 0.89, the pan-genome of the biovar *ovis* had a higher α value of 0.94, which was strictly correlated to the higher clonal-like behavior of this biovar compared with biovar *equi* [92]. Moreover, its high α value and the pan-genome curve suggest that the pan-genome of biovar *ovis* is increasing at a slower rate than that of biovar *equi*.

The same conclusion may be drawn from the development of singletons: each biovar *ovis* strain added ~16 singletons to the pan-genome, but each biovar *equi* strain added ~34 singletons to the gene pool. Moreover, although the core genome subset of the biovar *ovis* strains (1,818 CDS) was slightly higher than that of the biovar *equi* strains (1,599 CDS), most of the variable genes of the biovar *ovis* strains were acquired in blocks through horizontal gene transfer and are highly conserved throughout the entire biovar, as shown in Figure 2A. In contrast, the biovar *equi* strains presented great variability, both intra- and inter-biovar, in the content of the detected pathogenicity islands (Figure 2B). Finally, a comparison of the similarity levels on the two heatmaps, generated by Gegenees (93–100%, Figure 1) and from PAI contents (62–100%, Figure 7), also revealed that most of the variability defining the biovars *ovis* and *equi* arose from the gene content of the PAIs.

In view of this, one possible explanation for the large number of "Poorly characterized" genes in the differential core subsets of both biovars *ovis* and *equi* is the abovementioned acquisition of these subsets by horizontal gene transfer, which tends to involve a large number of hypothetical proteins [98], and the maintenance of these acquired regions in different biovars because they enabled the biovars to colonize specific hosts. Finally, the higher proportion of the functional category "Cellular processes and signaling" in biovar *equi* is most likely related to host adaptation because many genes in this cluster had functions such as defense mechanisms, signal transduction mechanisms, cell wall/membrane/envelope biogenesis, cell motility, and extracellular structures.

Variations in Pilus Gene Clusters

With respect to the gene content of the PAIs, the most interesting finding is the high similarity of the pilus genes in the biovar *ovis* strains, which is in contrast to the large variability of these genes in the biovar *equi* strains. Pilus gene clusters are normally acquired in a block through horizontal gene transfer and are composed of a specific sortase gene and the major, base and tip pilin genes. Briefly, the specific sortase protein of each cluster is responsible for cleaving the LPxTG motif of the major, base and tip pilin proteins of that cluster between the threonine (T) and glycine (G) amino acids, capturing the cleaved polypeptides, polymerizing the monomers, and transferring the final product to the housekeeping sortase of the bacterium for its final incorporation into the cell wall [99,100]. In the absence of a housekeeping sortase, the pilus-specific sortase can mediate the incorporation of the polymer into the cell wall. However, the presence of both housekeeping and specific sortases is necessary to efficiently anchor the pilus to the cell wall [101]. Moreover, although the expression of the major pilin is absolutely required for the specific pilus polymerization, the base and tip pilin monomers may still attach to the cell wall in its absence [100–103].

Although the biovar *ovis* strains present a complete *spaA* cluster, the biovar *equi* were shown to present large deletions in this cluster. Because of the deletion of the major pilin SpaA in the biovar *equi*, the base and tip pilin monomers would be expected to be the only pilin structures that could attach to the cell wall in a non-polymerized manner. Moreover, the deletion of one of the specific sortase genes in biovar *equi*, *srtB*, could also interfere in the efficient cell wall-anchoring of these monomers, causing them to be secreted [101]. Finally, even the production and sizes of these proteins may vary among the biovar *equi* strains because these proteins contain small deletions and frameshift mutations. Altogether, the differences in the *spaA* cluster of the biovar *equi* strains could account for the different levels of host cell attachment compared with the biovar *ovis* strains and even among the biovar *equi* strains, as found in the *C. diphtheriae* species [87,104,105].

In contrast to the high similarity found between the *spaA* clusters of the biovar *ovis* strains, the *spaD* clusters presented differences in three strains of this biovar. In *C. pseudotuberculosis* P54B96 and PAT10, a frameshit in the major pilin gene *spaD* impairs the coding of the entire protein and, thus, the polymerization of the pilin structure; and, in *C. pseudotuberculosis* 267, the tip pilin gene *spaF* also contains a frameshift. Although the tip pilin is not required for the polymerization of the pilin structure and adhesion to the host cell wall, its absence can slightly decrease the degree of adherence, which could reduce the spread of *C. pseudotuberculosis* strain 267 [106]. With respect to the *spaD* cluster of the biovar *equi* strains, a frameshift mutation in the specific sortase gene *srtC* of *C. pseudotuberculosis* CIP52.97 prevents the polymerization of the pilin structure. Moreover, the base and tip pilin genes, *spaE* and *spaF*, respectively, of *C. pseudotuberculosis* strains 258, 316, 1/06-A and Cp162 are merged into the same reading frame. Overall, these results suggest that although *C. pseudotuberculosis* 258, 316, 1/06-A and Cp162 can polymerize the major pilin, *C. pseudotuberculosis* strain 31 is most likely the only biovar *equi* strain able to polymerize an entire pilin structure from the *spaD* cluster, whereas all the biovar *ovis* strains are likely capable of producing one or two types of pilin structures (*spaA* and *spaD*).

Summarizing, all the *C. pseudotuberculosis* biovar *ovis* strains likely contain a functional *spaA* cluster of pilus genes; only three strains (267, P54B96 and PAT10) are unable to polymerize an entire *spaD* pilin structure (most likely, they instead attach monomers or incompletely polymerized pilin structures). In contrast, all the biovar *equi* strains contain deletions, which render them unable to polymerize any *spaA* pilin structures; within this biovar, only *C. pseudotuberculosis* 31 appears to be able to polymerize an entire *spaD* pilin structure. Given the pivotal role played by pili in the processes of adhesion and internalization, the polymerization of complete pilin structures in the biovar *ovis* strains could be responsible for the great ability of these strains to spread throughout host tissues and penetrate cells to grow intracellularly [48,101,106,107]. Based on this observation, the biovar *ovis* strains are expected to have less contact with other organisms than the biovar *equi* strains and to therefore show more clonal-like behavior. Finally, these results could also explain the distinct pattern of the diseases caused by *C. pseudotuberculosis* in horses, which involves ulcerative lymphangitis that rarely evolves to a visceral form [108]. However, more studies are needed to assess whether the *C. pseudotuberculosis* biovars *equi* and *ovis* truly present different patterns of pilin formation and, thus, variable degrees of host tissue adhesion, spreading and cell internalization.

Supporting Information

Figure S1 Plasticity of PiCps 4, 5 and 9. A1 and B1, PiCp9; A2 and B2, PiCp4; A3 and B3, PiCp5. A, all the *C. pseudotuberculosis* strains were aligned using *C. pseudotuberculosis* strain 1002 as a reference. From the inner to outer circle on A1, A2 and A3: the biovar *equi* strains Cp31, Cp1/06-A, CpCp162, Cp258, Cp316, CpCIP52.97; and, the biovar *ovis* strains CpC231, CpP54B96, Cp267, CpPAT10, CpI19, Cp42/02-A, Cp3/99-5, CpFRC41 and Cp1002. B, all the *C. pseudotuberculosis* strains were aligned using *C. pseudotuberculosis* strain CIP52.97 as a reference. From the inner to outer circle on B1, B2 and B3: the biovar *ovis* strains CpC231, Cp1002, CpPAT10, Cp267, CpP54B96, CpI19, Cp42/02-A, CpFRC41, Cp3/99-5, Cp1/06-A; and, the biovar *equi* strains Cp31, CpCp162, Cp316, Cp258 and CpCIP52.97. CDS, coding sequences; tRNA, transfer RNA; rRNA, ribosomal RNA; and PAI, pathogenicity island.
(TIFF)

Acknowledgments

The authors thank the CAPES/DAAD international cooperation for financing an exchange scholarship (http://www.capes.gov.br/cooperacao-internacional) (grant number 5117119) and CLIB - Graduate Cluster Industrial Biotechnology (http://www.graduatecluster.net/), where SCS and LCG are currently participating as guest students.

Author Contributions

Read and gave insights about the manuscript: SCS AS ET JB RR AC AA ARS ACP CD EGVB FAD FA FSR KKFN LCG SA SSH SMB UPP VACA MPCS AM AT VA. Conceived and designed the experiments: AT VA. Performed the experiments: SCS ET JB RR AC AA ARS ACP CD EGVB FAD FA FSR KKFN LCG SA SSH SMB UPP VACA. Analyzed the data: SCS ET JB. Contributed reagents/materials/analysis tools: SCS AS ET JB AT VA. Wrote the paper: SCS AS MPCS AM AT VA.

References

1. Dorella FA, Pacheco LGC, Oliveira SC, Miyoshi A, Azevedo V (2006) *Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. Vet Res 37: 201–218.
2. Lehman KB, Neumann R (1896) Atlas und grundriss der bakeriologie und lehrbuch der speziellen bakteriologischen diagnositk. 1st ed. J.F. Lehmann, Munchen.
3. Pascual C, Lawson PA, Farrow JA, Gimenez MN, Collins MD (1995) Phylogenetic analysis of the genus *Corynebacterium* based on 16S rRNA gene sequences. Int J Syst Bacteriol 45: 724–728.
4. Cerdeño-Tárraga AM, Efstratiou A, Dover LG, Holden MTG, Pallen M, et al. (2003) The complete genome sequence and analysis of *Corynebacterium diphtheriae* NCTC13129. Nucleic Acids Res 31: 6516–6523.
5. Tauch A, Kaiser O, Hain T, Goesmann A, Weisshaar B, et al. (2005) Complete genome sequence and analysis of the multiresistant nosocomial pathogen *Corynebacterium jeikeium* K411, a lipid-requiring bacterium of the human skin flora. J Bacteriol 187: 4671–4682.
6. Kalinowski J, Bathe B, Bartels D, Bischoff N, Bott M, et al. (2003) The complete *Corynebacterium glutamicum* ATCC 13032 genome sequence and its impact on the production of L-aspartate-derived amino acids and vitamins. J Biotechnol 104: 5–25.
7. Jones D, Collins MD (1986) Irregular, nonsporing gram-positive rods, section 15. pages 1261–1579 in bergey's manual of systematic bacteriology. Williams & Wilkins, Co., Baltimore, MD.
8. Buck GA, Cross RE, Wong TP, Loera J, Groman N (1985) DNA relationships among some tox-bearing corynebacteriophages. Infect Immun 49: 679–684.
9. Groman N, Schiller J, Russell J (1984) *Corynebacterium ulcerans* and *Corynebacterium pseudotuberculosis* responses to DNA probes derived from corynephage beta and *Corynebacterium diphtheriae*. Infect Immun 45: 511–517.
10. Wong TP, Groman N (1984) Production of diphtheria toxin by selected isolates of *Corynebacterium ulcerans* and *Corynebacterium pseudotuberculosis*. Infect Immun 43: 1114–1116.
11. Muckle CA, Gyles CL (1982) Characterization of strains of *Corynebacterium pseudotuberculosis*. Can J Comp Med 46: 206–208.
12. Biberstein EL, Knight HD, Jang S (1971) Two biotypes of *Corynebacterium pseudotuberculosis*. Vet Rec 89: 691–692.
13. Ayers JL (1977) Caseous lymphadenitis in goat and sheep: Review of diagnosis, pathogenesis, and immunity. JAVMA n. 171: 1251–1254.
14. Ben Saïd MS, Ben Maitigue H, Benzarti M, Messadi L, Rejeb A, et al. (2002) Epidemiological and clinical studies of ovine caseous lymphadenitis. Arch Inst Pasteur Tunis 79: 51–57.
15. Arsenault J, Girard C, Dubreuil P, Daignault D, Galarneau JR, et al. (2003) Prevalence of and carcass condemnation from maedi-visna, paratuberculosis and caseous lymphadenitis in culled sheep from Quebec, Canada. Prev Vet Med 59: 67–81.
16. Binns SH, Bailey M, Green LE (2002) Postal survey of ovine caseous lymphadenitis in the United Kingdom between 1990 and 1999. Vet Rec 150: 263–268.
17. Connor KM, Quirie MM, Baird G, Donachie W (2000) Characterization of United Kingdom isolates of *Corynebacterium pseudotuberculosis* using pulsed-field gel electrophoresis. J Clin Microbiol 38: 2633–2637.

18. Paton MW, Walker SB, Rose IR, Watt GF (2003) Prevalence of caseous lymphadenitis and usage of caseous lymphadenitis vaccines in sheep flocks. Aust Vet J 81: 91–95.
19. Hodgson AL, Carter K, Tachedjian M, Krywult J, Corner LA, et al. (1999) Efficacy of an ovine caseous lymphadenitis vaccine formulated using a genetically inactive form of the *Corynebacterium pseudotuberculosis* phospholipase D. Vaccine 17: 802–808.
20. Pugh DG (2002) Caseous Lymphadenitis. In: Sheep & Goat Medicine Saunders 207–208.
21. Radostits OM, Gay CC, Blood DC, Hinchcliff KW (2002) Clínica veterinária. um tratado de doenças dos bovinos, ovinos, suínos, caprinos e eqüinos. Ed. Guanabara, Koogan, 9ª edição.
22. Augustine JL, Renshaw HW (1986) Survival of *Corynebacterium pseudotuberculosis* in axenic purulent exudate on common barnyard fomites. Am J Vet Res 47: 713–715.
23. Yeruham I, Friedman S, Perl S, Elad D, Berkovich Y, et al. (2004) A herd level analysis of a *Corynebacterium pseudotuberculosis* outbreak in a dairy cattle herd. Vet Dermatol 15: 315–320.
24. Yeruham I, Elad D, Friedman S, Perl S (2003) *Corynebacterium pseudotuberculosis* infection in Israeli dairy cattle. Epidemiol Infect 131: 947–955.
25. Collett MG, Bath GF, Cameron CM (1994) *Corynebacterium pseudotuberculosis* infections. In: Infections diseases of livestock with special reference to Southern Africa. Oxford University Press 2: 1387–1395.
26. Dorella FA, Pacheco LG, Seyffert N, Portela RW, Meyer R, et al. (2009) Antigens of *Corynebacterium pseudotuberculosis* and prospects for vaccine development. Expert Rev Vaccines 8: 205–213.
27. Williamson LH (2001) Caseous lymphadenitis in small ruminants. Vet. Clin. North Am. Food Anim. Pract 17: 359–371.
28. Liu DTL, Chan W, Fan DSP, Lam DSC (2005) An infected hydrogel buckle with *Corynebacterium pseudotuberculosis*. Br J Ophthalmol 89: 245–246.
29. Mills AE, Mitchell RD, Lim EK (1997) *Corynebacterium pseudotuberculosis* is a cause of human necrotising granulomatous lymphadenitis. Pathology 29: 231–233.
30. Peel MM, Palmer GG, Stacpoole AM, Kerr TG (1997) Human lymphadenitis due to *Corynebacterium pseudotuberculosis*: report of ten cases from Australia and review. Clin Infect Dis 24: 185–191.
31. Barakat AA, Selim SA, Atef A, Saber MS, Nafie EK, et al. (1984) Two serotypes of *Corynebacterium pseudotuberculosis* isolated from different animal species. Revue Scientifique et Technique de l'OIE 3(1): 151–163.
32. Aleman M, Spier SJ, Wilson WD, Doherr M (1996) *Corynebacterium pseudotuberculosis* infection in horses: 538 cases (1982–1993). J Am Vet Med Assoc 209: 804–809.
33. Pratt SM, Spier SJ, Carroll SP, Vaughan B, Whitcomb MB, et al. (2005) Evaluation of clinical characteristics, diagnostic test results, and outcome in horses with internal infection caused by *Corynebacterium pseudotuberculosis*: 30 cases (1995–2003). J Am Vet Med Assoc 227: 441–448.
34. Braverman Y, Chizov-Ginzburg A, Saran A, Winkler M (1999) The role of houseflies (Musca domestica) in harbouring *Corynebacterium pseudotuberculosis* in dairy herds in Israel. Revue Scientifique et Technique de l'OIE 18 n° 3: 681–690.
35. Addo P (1983) Role of the common house fly (Musca domestica) in the spread of ulcerative lymphangitis. Vet Rec 113(21): 496–497.
36. Selim SA (2001) Oedematous skin disease of buffalo in Egypt. J Vet Med B Infect Dis Vet Public Health 48: 241–258.
37. Yeruham I, Braverman Y, Shpigel NY, Chizov-Ginzburg A, Saran A, et al. (1996) Mastitis in dairy cattle caused by *Corynebacterium pseudotuberculosis* and the feasibility of transmission by houseflies. I. Vet Q 18: 87–89.
38. Spier S (2008) *Corynebacterium pseudotuberculosis* infection in horses: An emerging disease associated with climate change? Equine Veterinary Education 20: 37–39.
39. McKean S, Davies J, Moore R (2005) Identification of macrophage induced genes of *Corynebacterium pseudotuberculosis* by differential fluorescence induction. Microbes Infect 7: 1352–1363.
40. McKean SC, Davies JK, Moore RJ (2007) Expression of phospholipase D, the major virulence factor of *Corynebacterium pseudotuberculosis*, is regulated by multiple environmental factors and plays a role in macrophage death. Microbiology 153: 2203–2211.
41. Schumann W (2007) Thermosensors in eubacteria: role and evolution. J Biosci 32: 549–557.
42. Billington SJ, Esmay PA, Songer JG, Jost BH (2002) Identification and role in virulence of putative iron acquisition genes from *Corynebacterium pseudotuberculosis*. FEMS Microbiol : Lett. 208, 41–45.
43. Ruiz JC, D'Afonseca V, Silva A, Ali A, Pinto AC, et al. (2011) Evidence for reductive genome evolution and lateral acquisition of virulence functions in two *Corynebacterium pseudotuberculosis* strains. PLoS One 6: e18551.
44. Alves FSF, Olander H (1999) Uso de vacina toxóide no controle da linfadenite caseosa em caprinos. Veterinária Notícias, Uberlândia n° 5: 69–75.
45. Songer JG, Libby SJ, Iandolo JJ, Cuevas WA (1990) Cloning and expression of the phospholipase D gene from *Corynebacterium pseudotuberculosis* in *Escherichia coli*. Infect Immun 58: 131–136.
46. Trost E, Ott L, Schneider J, Schröder J, Jaenicke S, et al. (2010) The complete genome sequence of *Corynebacterium pseudotuberculosis* FRC41 isolated from a 12-year-old girl with necrotizing lymphadenitis reveals insights into gene-regulatory networks contributing to virulence. BMC Genomics 11: 728.
47. Yanagawa R, Honda E (1976) Presence of pili in species of human and animal parasites and pathogens of the genus *Corynebacterium*. Infect Immun 13: 1293–1295.
48. Wilson JW, Schurr MJ, LeBlanc CL, Ramamurthy R, Buchanan KL, et al. (2002) Mechanisms of bacterial pathogenicity. Postgrad Med J 78: 216–224.
49. Pethick FE, Lainson AF, Yaga R, Flockhart A, Smith DGE, et al. (2012) Complete Genome Sequences of *Corynebacterium pseudotuberculosis* Strains 3/99–5 and 42/02-A, Isolated from Sheep in Scotland and Australia, Respectively. J Bacteriol 194: 4736–4737.
50. Cerdeira LT, Pinto AC, Schneider MPC, de Almeida SS, dos Santos AR, et al. (2011) Whole-genome sequence of *Corynebacterium pseudotuberculosis* PAT10 strain isolated from sheep in Patagonia, Argentina. J Bacteriol 193: 6420–6421.
51. Lopes T, Silva A, Thiago R, Carneiro A, Dorella FA, et al. (2012) Complete Genome Sequence of *Corynebacterium pseudotuberculosis* Strain Cp267, Isolated from a Llama. J Bacteriol 194: 3567–3568.
52. Silva A, Schneider MPC, Cerdeira L, Barbosa MS, Ramos RTJ, et al. (2011) Complete genome sequence of *Corynebacterium pseudotuberculosis* I19, a strain isolated from a cow in Israel with bovine mastitis. J Bacteriol 193: 323–324.
53. Cerdeira LT, Schneider MPC, Pinto AC, de Almeida SS, dos Santos AR, et al. (2011) Complete genome sequence of *Corynebacterium pseudotuberculosis* strain CIP 52.97, isolated from a horse in Kenya. J Bacteriol 193: 7025–7026.
54. Ramos RTJ, Silva A, Carneiro AR, Pinto AC, Soares SDC, et al. (2012) Genome Sequence of the *Corynebacterium pseudotuberculosis* Cp316 Strain, Isolated from the Abscess of a Californian Horse. J Bacteriol 194: 6620–6621.
55. Ramos RTJ, Carneiro AR, Soares SC, Santos AR, Almeida SS, et al. (2013) Tips and tricks for the assembly a *Corynebacterium pseudotuberculosis* genome using a semiconductor sequencer. Microbial Biotechnology in press.
56. Soares SC, Trost E, Ramos RTJ, Carneiro AR, Santos AR, et al. (2012) Genome sequence of *Corynebacterium pseudotuberculosis* biovar equi strain 258 and prediction of antigenic targets to improve biotechnological vaccine production. J Biotechnol in press.
57. Pethick FE, Lainson AF, Yaga R, Flockhart A, Smith DGE, et al. (2012) Complete Genome Sequence of *Corynebacterium pseudotuberculosis* Strain 1/06-A, Isolated from a Horse in North America. J Bacteriol 194: 4476.
58. Hassan SS, Schneider MPC, Ramos RTJ, Carneiro AR, Ranieri A, et al. (2012) Whole-Genome Sequence of *Corynebacterium pseudotuberculosis* Strain Cp162, Isolated from Camel. J Bacteriol 194: 5718–5719.
59. Silva A, Ramos RTJ, Ribeiro Carneiro A, Cybelle Pinto A, de Castro Soares S, et al. (2012) Complete Genome Sequence of *Corynebacterium pseudotuberculosis* Cp31, Isolated from an Egyptian Buffalo. J Bacteriol 194: 6663–6664.
60. Agren J, Sundström A, Håfström T, Segerman B (2012) Gegenees: fragmented alignment of multiple genomes for determining phylogenomic distances and genetic signatures unique for specified target groups. PLoS One 7: e39107.
61. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23: 254–267.
62. Kloepper TH, Huson DH (2008) Drawing explicit phylogenetic networks and their integration into SplitsTree. BMC Evol Biol 8: 22.
63. Blom J, Albaum SP, Doppmeier D, Pühler A, Vorhölter F, et al. (2009) EDGAR: a software framework for the comparative analysis of prokaryotic genomes. BMC Bioinformatics 10: 154.
64. Meyer F, Goesmann A, McHardy AC, Bartels D, Bekel T, et al. (2003) GenDB–an open source genome annotation system for prokaryote genomes. Nucleic Acids Res 31: 2187–2195.
65. Lerat E, Daubin V, Moran NA (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. PLoS Biol 1: E19.
66. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, et al. (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial pan-genome. Proc Natl Acad Sci U S A 102: 13950–13955.
67. Tettelin H, Riley D, Cattuto C, Medini D (2008) Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 11: 472–477.
68. Grant JR, Arantes AS, Stothard P (2012) Comparing thousands of circular genomes using the CGView Comparison Tool. BMC Genomics 13: 202.
69. Soares SC, Abreu VAC, Ramos RTJ, Cerdeira L, Silva A, et al. (2012) PIPS: pathogenicity island prediction software. PLoS One 7: e30848.
70. Carver TJ, Rutherford KM, Berriman M, Rajandream M, Barrell BG, et al. (2005) ACT: the Artemis Comparison Tool. Bioinformatics 21: 3422–3423.
71. Nishio Y, Nakamura Y, Kawarabayasi Y, Usuda Y, Kimura E, et al. (2003) Comparative complete genome sequence analysis of the amino acid replacements responsible for the thermostability of *Corynebacterium efficiens*. Genome Res 13: 1572–1579.
72. Schröder J, Maus I, Meyer K, Wördemann S, Blom J, et al. (2012) Complete genome sequence, lifestyle, and multi-drug resistance of the human pathogen *Corynebacterium resistens* DSM 45100 isolated from blood samples of a leukemia patient. BMC Genomics 13: 141.
73. Tauch A, Trost E, Tilker A, Ludewig U, Schneiker S, et al. (2008) The lifestyle of *Corynebacterium urealyticum* derived from its complete genome sequence established by pyrosequencing. J Biotechnol 136: 11–21.
74. Schröder J, Maus I, Trost E, Tauch A (2011) Complete genome sequence of *Corynebacterium variabile* DSM 44702 isolated from the surface of smear-ripened cheeses and insights into cheese ripening and flavor generation. BMC Genomics 12: 545.

75. Trost E, Götker S, Schneider J, Schneiker-Bekel S, Szczepanowski R, et al. (2010) Complete genome sequence and lifestyle of black-pigmented *Corynebacterium aurimucosum* ATCC 700975 (formerly C. nigricans CN-1) isolated from a vaginal swab of a woman with spontaneous abortion. BMC Genomics 11: 91.
76. Schmidt H, Hensel M (2004) Pathogenicity islands in bacterial pathogenesis. Clin Microbiol Rev 17: 14–56.
77. Karaolis DK, Johnson JA, Bailey CC, Boedeker EC, Kaper JB, et al. (1998) A *Vibrio cholerae* pathogenicity island associated with epidemic and pandemic strains. Proc Natl Acad Sci U S A 95: 3134–3139.
78. Oram DM, Avdalovic A, Holmes RK (2002) Construction and characterization of transposon insertion mutations in *Corynebacterium diphtheriae* that affect expression of the diphtheria toxin repressor (DtxR). J Bacteriol 184: 5723–5732.
79. Nakao H, Pruckler JM, Mazurova IK, Narvskaia OV, Glushkevich T, et al. (1996) Heterogeneity of diphtheria toxin gene, tox, and its regulatory element, dtxR, in *Corynebacterium diphtheriae* strains causing epidemic diphtheria in Russia and Ukraine. J Clin Microbiol 34: 1711–1716.
80. Hadfield TL, McEvoy P, Polotsky Y, Tzinserling VA, Yakovlev AA (2000) The pathology of diphtheria. J Infect Dis (Suppl 1): S116–20.
81. Murphy JR (2011) Mechanism of Diphtheria Toxin Catalytic Domain Delivery to the Eukaryotic Cell Cytosol and the Cellular Factors that Directly Participate in the Process. Toxins (Basel) 3: 294–308.
82. Holmes RK (2000) Biology and molecular epidemiology of diphtheria toxin and the tox gene. J Infect Dis 181 Suppl 1: S156–67.
83. Sekizuka T, Yamamoto A, Komiya T, Kenri T, Takeuchi F, et al. (2012) *Corynebacterium ulcerans* 0102 carries the gene encoding diphtheria toxin on a prophage different from the *C. diphtheriae* NCTC 13129 prophage. BMC Microbiol 12: 72.
84. Sing A, Bierschenk S, Heesemann J (2005) Classical diphtheria caused by *Corynebacterium ulcerans* in Germany: amino acid sequence differences between diphtheria toxins from *Corynebacterium diphtheriae* and *C. ulcerans*. Clin Infect Dis 40: 325–326.
85. Maximescu P, Oprişan A, Pop A, Potorac E (1974) Further studies on *Corynebacterium* species capable of producing diphtheria toxin (*C. diphtheriae*, *C. ulcerans*, *C. ovis*). J Gen Microbiol 82: 49–56.
86. LeMieux J, Hava DL, Basset A, Camilli A (2006) RrgA and RrgB are components of a multisubunit pilus encoded by the *Streptococcus pneumoniae* rlrA pathogenicity islet. Infect Immun 74: 2453–2456.
87. Trost E, Blom J, Soares SDC, Huang I, Al-Dilaimi A, et al. (2012) Pangenomic study of *Corynebacterium diphtheriae* that provides insights into the genomic diversity of pathogenic isolates from cases of classical diphtheria, endocarditis, and pneumonia. J Bacteriol 194: 3199–3215.
88. Khamis A, Raoult D, La Scola B (2004) rpoB gene sequencing for identification of *Corynebacterium* species. J Clin Microbiol 42: 3925–3931.
89. Tauch A, Schneider J, Szczepanowski R, Tilker A, Viehoever P, et al. (2008) Ultrafast pyrosequencing of *Corynebacterium kroppenstedtii* DSM44385 revealed insights into the physiology of a lipophilic corynebacterium that lacks mycolic acids. J Biotechnol 136: 22–30.
90. Collins MD, Falsen E, Akervall E, Sjöden B, Alvarez A (1998) *Corynebacterium kroppenstedtii* sp. nov., a novel *Corynebacterium* that does not contain mycolic acids. Int J Syst Bacteriol 48 Pt 4: 1449–1454.
91. Paviour S, Musaad S, Roberts S, Taylor G, Taylor S, et al. (2002) *Corynebacterium species* isolated from patients with mastitis. Clin Infect Dis 35: 1434–1440.
92. Bolt F (2009) The population structure of the *Corynebacterium diphtheriae* group. University of Warwick. PhD thesis. Available: http://wrap.warwick.ac.uk/1759/. Accessed 26 November 2012.
93. Songer JG, Beckenbach K, Marshall MM, Olson GB, Kelley L (1988) Biochemical and genetic characterization of *Corynebacterium pseudotuberculosis*. Am J Vet Res 49: 223–226.
94. Sutherland SS, Hart RA, Buller NB (1993) Ribotype analysis of *Corynebacterium pseudotuberculosis* isolates from sheep and goats. Aust Vet J 70: 454–456.
95. Halachev MR, Loman NJ, Pallen MJ (2011) Calculating orthologs in bacteria and Archaea: a divide and conquer approach. PLoS One 6: e28388.
96. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005) The microbial pan-genome. Curr Opin Genet Dev 15: 589–594.
97. Kittichotirat W, Bumgarner RE, Asikainen S, Chen C (2011) Identification of the pangenome and its components in 14 distinct *Aggregatibacter actinomycetemcomitans* strains by comparative genomic analysis. PLoS One 6: e22420.
98. Hsiao WWL, Ung K, Aeschliman D, Bryan J, Finlay BB, et al. (2005) Evidence of a large novel gene pool associated with prokaryotic genomic islands. PLoS Genet 1: e62.
99. Ton-That H, Schneewind O (2004) Assembly of pili in Gram-positive bacteria. Trends Microbiol 12: 228–234.
100. Ton-That H, Marraffini LA, Schneewind O (2004) Sortases and pilin elements involved in pilus assembly of *Corynebacterium diphtheriae*. Mol Microbiol 53: 251–261.
101. Mandlik A, Swierczynski A, Das A, Ton-That H (2008) Pili in Gram-positive bacteria: assembly, involvement in colonization and biofilm development. Trends Microbiol 16: 33–40.
102. Ton-That H, Marraffini LA, Schneewind O (2004) Protein sorting to the cell wall envelope of Gram-positive bacteria. Biochim Biophys Acta 1694: 269–278.
103. Ton-That H, Schneewind O (2003) Assembly of pili on the surface of *Corynebacterium diphtheriae*. Mol Microbiol 50: 1429–1438.
104. Hirata Jr R, Pereira GA, Filardy AA, Gomes DLR, Damasco PV, et al. (2008) Potential pathogenic role of aggregative-adhering *Corynebacterium diphtheriae* of different clonal groups in endocarditis. Braz J Med Biol Res 41: 986–991.
105. Hirata RJ, Souza SMS, Rocha-de-Souza CM, Andrade AFB, Monteiro-Leal LH, et al. (2004) Patterns of adherence to HEp-2 cells and actin polymerisation by toxigenic *Corynebacterium diphtheriae* strains. Microb Pathog 36: 125–130.
106. Mandlik A, Swierczynski A, Das A, Ton-That H (2007) *Corynebacterium diphtheriae* employs specific minor pilins to target human pharyngeal epithelial cells. Mol Microbiol 64: 111–124.
107. Zasada AA, Formińska K, Rzeczkowska M (2012) Occurence of pili genes in *Corynebacterium diphtheriae* strains. Med Dosw Mikrobiol 64(1): 19–27.
108. Hall K, McCluskey BJ, Cunningham W (2001) *Corynebacterium pseudotuberculosis* infections (Pigeon Fever) in horses in Western Colorado: An epidemiological investigation. Journal of Equine Veterinary Science 21(6): 284–286.

II.I.5 PIPS: pathogenicity island prediction software.
Soares SC, Abreu VA, Ramos RT, Cerdeira L, Silva A, Baumbach J, Trost E, Tauch A, Hirata R Jr, Mattos-Guaraldi AL, Miyoshi A, **Azevedo V**.
PLoS One. 2012;7(2): e30848. doi: 10.1371/journal.pone.0030848. Epub 2012 Feb 15.

Em uma tentativa inicial de predizer vários fatores de virulência de *C. pseudotuberculosis* 1002, foram utilizados diversos softwares preditores de Ilhas de Patogenicidade (PAIs). Contudo, a maioria dos softwares são limitados à análise de características específicas de PAIs, negligenciando a análise em um contexto amplo que considere todas as características inerentes a estas regiões. Somente os softwares PredictBias e IslandViewer predizem as PAIs de uma maneira multidimensionada. Contudo, ambos os softwares apresentam outras limitações, como: processos com alto custo computacional e dependências não resolvidas que impedem a instalação do IslandViewer; e, uma arquitetura online que requer a submissão das sequencias genômicas no PredictBias, o que impede a análise de sequencias genômicas que ainda não foram publicadas e, portanto, não podem ser submetidas online antes da publicação. Para contornar estes problemas, nosso grupo criou um software para a predição de PAIs, chamado PIPS, que está disponível publicamente para instalação em computadores pessoais e que supera outros softtwares disponíveis publicamente em termos de performance. O seguinte paper descreve a implementação do PIPS e também apresenta comparações deste software com os programas PredictBias e IslandViewer.

PIPS: Pathogenicity Island Prediction Software

Siomar C. Soares[1], Vinícius A. C. Abreu[2], Rommel T. J. Ramos[3], Louise Cerdeira[3], Artur Silva[3], Jan Baumbach[4], Eva Trost[5], Andreas Tauch[5], Raphael Hirata Jr.[6], Ana L. Mattos-Guaraldi[6], Anderson Miyoshi[1], Vasco Azevedo[1,2]*

1 Department of General Biology, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, **2** Department of Biochemistry and Immunology, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, **3** Department of Genetics, Federal University of Pará, Belém, Pará, Brazil, **4** Department of Computer Science, Max-Planck-Institut für Informatik, Saarbrücken, Saarland, Germany, **5** Center for Biotechnology, Bielefeld University, Bielefeld, Nordrhein-Westfalen, Germany, **6** Microbiology and Immunology Discipline, Medical Sciences Faculty, State University of Rio de Janeiro, Rio de Janeiro, Brazil

Abstract

The adaptability of pathogenic bacteria to hosts is influenced by the genomic plasticity of the bacteria, which can be increased by such mechanisms as horizontal gene transfer. Pathogenicity islands play a major role in this type of gene transfer because they are large, horizontally acquired regions that harbor clusters of virulence genes that mediate the adhesion, colonization, invasion, immune system evasion, and toxigenic properties of the acceptor organism. Currently, pathogenicity islands are mainly identified *in silico* based on various characteristic features: (1) deviations in codon usage, G+C content or dinucleotide frequency and (2) insertion sequences and/or tRNA genetic flanking regions together with transposase coding genes. Several computational techniques for identifying pathogenicity islands exist. However, most of these techniques are only directed at the detection of horizontally transferred genes and/or the absence of certain genomic regions of the pathogenic bacterium in closely related non-pathogenic species. Here, we present a novel software suite designed for the prediction of pathogenicity islands (pathogenicity island prediction software, or PIPS). In contrast to other existing tools, our approach is capable of utilizing multiple features for pathogenicity island detection in an integrative manner. We show that PIPS provides better accuracy than other available software packages. As an example, we used PIPS to study the veterinary pathogen *Corynebacterium pseudotuberculosis*, in which we identified seven putative pathogenicity islands.

Citation: Soares SC, Abreu VAC, Ramos RTJ, Cerdeira L, Silva A, et al. (2012) PIPS: Pathogenicity Island Prediction Software. PLoS ONE 7(2): e30848. doi:10.1371/journal.pone.0030848

Editor: Igor Mokrousov, St. Petersburg Pasteur Institute, Russian Federation

Received July 12, 2011; **Accepted** December 22, 2011; **Published** February 15, 2012

Funding: This work was supported by grants from the Brazilian Funding Agencies Conselho Nacional de Desenvolvimento Científico e Tecnológico (grant CNPq/MAPA; http://cnpq.br/) [grant number 578219/2008-5] and Fundação de Apoio à Pesquisa de Minas Gerais (FAPEMIG; http://www.fapemig.br/) [grant number APQ-01004-08]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: vasco@icb.ufmg.br

Introduction

Bacteria are the most abundant and diverse organisms on Earth [1]. This diversity is mainly the result of the remarkable genomic plasticity of bacteria, which allows bacteria to adapt to a wide range of environments, enhancing their pathogenic potential [2,3]. Various mechanisms can promote genome plasticity, including point mutations, gene conversion, chromosome rearrangements (inversions and translocations), deletions, and the acquisition of DNA from other cells through horizontal gene transfer (HGT). Those mobile elements can be acquired via plasmids, bacteriophages, transposons, insertion sequences and genomic islands (GEIs) [4].

GEIs play a major role in the fast and dramatic adaptation of species phenotypes to different environments by carrying clusters of genes that can cooperate to confer a cell with novel and useful phenotypes, such as the ability to survive inside a host. GEIs are large genomic regions that present deviations in codon usage, G+C content or dinucleotide frequency compared to other parts of the organism's genome; these characteristics are hallmarks of chromosome regions that were acquired horizontally from other species in a single block. GEIs are often flanked by insertion sequences or tRNA genes and transposase coding genes; these segments are responsible for the genomic incorporation of alien DNA obtained through transformation, conjugation or bacteriophage infection [5].

Horizontally acquired genes

GEIs acquired by transposase-mediated insertion have inverted repeats (IR) or insertion sequences (IS) in their flanking regions and often harbor tRNA coding sequences [6]. Genes coding for tRNA and tmRNA (hereafter tRNA genes) are "hot spots" for the insertion of genetic elements; they possess a 3′-terminal sequence that is recognized by integrases and are frequently found in *selC* and *leuX* tRNA genes (selenocysteine and leucine, respectively) [6,7].

The identification of horizontally acquired regions is usually based on the detection of a chromosome region's G+C content and codon usage that differs from that found in the rest of the genome. Clusters of horizontally acquired genes may have a skewed G+C content and codon usage, reflecting a distinct genomic signature from a donor organism [8]. Although these G+C content-skewed regions within an acceptor organism genome remain functional to some extent, there is selective pressure for the acquired region to adapt its codon usage to that of the acceptor organism to enhance expression. This adaptation in codon usage is driven by selective forces, such as codon/anticodon linkage and a greater frequency of a certain codon for the tRNA gene [9]. Codon usage bias in bacteria is closely related to base composition, and the adoption of preferential G+C- or A+T-rich codons may lead to a similar G+C content of genes throughout the genome [10]. Given the high density of coding regions in prokaryotic genomes, codon usage adaptation, in addition to point mutations and other evolutionary forces, can lead to homogeneity in the base composition of bacteria. Consequently, the identification of mobile genomic regions based solely on their discrepant genomic signature is usually only possible for regions that were recently acquired from distant organisms [11,12].

In addition to the aforementioned features, Hsiao *et al.* [13] demonstrated that GEIs have a high frequency of hypothetical proteins (putative proteins with unknown function) when compared to the rest of the genome. These investigators indicated that this higher frequency could result from gene acquisition from organisms that have not yet been sequenced, including non-culturable bacteria.

Virulence factors and pathogenicity islands

GEIs may carry a number of coding regions that are useful for a cell. The GEIs that carry gene coding for virulence factors are collectively known as pathogenicity islands (PAIs). PAIs are characterized by the high frequency of genes that code for factors that enable or enhance the parasitic growth of the microorganism within a host [14]. Virulence factors mediate adhesion, colonization, invasion, immune system evasion and toxigenesis, which are necessary for infection [15].

Hacker *et al.* [5] first described PAIs after observing the loss of virulence of pathogenic varieties of *Escherichia coli* through deletions of hemolysin and fimbrial adhesin genes. They demonstrated that these genes are located in the same chromosomal region and can be removed by deletion events, both *in vitro* and *in vivo*. PAI identification using traditional molecular biology techniques without genomic information services is laborious and time-consuming because of the need for phenotypic analyses of the strains and the delimitation of the target genes. Additionally, PAIs often present variable stability, mosaic structure and uncharacterized genes.

In silico analysis of pathogenicity islands

PAI analysis is becoming more feasible with the increasing number of sequenced prokaryotic genomes and the development of new bioinformatics methods that can assemble data retrieved from next-generation sequencers (NGS). NGS plataforms have the potential to increase the number of completed genome projects orders of magnitude more rapidly than the earlier Sanger method and at a small fraction of the cost. Consequently, the need for the development of genomic data retrieval softwares is increasing. Several computational programs have been specifically designed for spotting PAIs and other HGTs. However, most of the programs use criteria that are not sufficiently stringent to provide useable sensitivity and specificity. Overall, existing software only screens for horizontal gene transfer, through G+C content or dinucleotide deviations (e.g., wavelet analysis of the G+C content, cumulative GC profile, δ_P-web, IVOM, IslandPath and PAI-IDA) [16–23] and codon usage deviation (SIGI-HMM and PAI-IDA) [16,24] or for the absence of elements of the putative PAI in non-pathogenic species (IslandPath, Islander, IslandPick and tRNAcc) [7,8,20,25], which may result in the detection of false-positive PAIs [8,26]. Pundhir *et al.* [27] affirm that "Although efficient in the detection of GIs, these tools give much false positive results for PAIs. This is because a region showing distinct nucleotide content may be alien to the host genome but may not necessarily be involved in Pathogenicity". Therefore, these tools may detect a metabolic island, a GEI associated with secondary metabolite biosynthesis, as a false-positive PAI if it exhibits all of the PAI features except for the virulence factors. Finally, some PAIs may exhibit deviations only in the G+C content or codon usage, demonstrating the importance of using more than one software system in a multi-pronged approach.

Two currently available PAI detection programs use a multi-pronged strategy for the detection of PAIs, accounting for several characteristics of the genome. One of these programs, PredictBias, identifies PAIs by its genomic signature, its absence in taxonomically related organisms and the presence of genes coding for virulence factors, classifying them as either biased-composition PAIs if they present horizontal transfer characteristics or unbiased-composition PAIs otherwise [27]. Another program, IslandViewer, performs a combined analysis using three other programs: ColomboSIGI-HMM, based on codon usage analysis of each coding sequence (CDS) of the genome; IslandPick, which characterizes PAIs by their absence in phylogenetically closely related organisms; and IslandPath-DIMOB, which finds regions that have dinucleotide content deviation and harbor genes related to mobility [8,28,29].

Although PredictBias and IslandViewer are robust programs that use multi-pronged strategies, they have some restrictions. For example, PredictBias can only be used in a web-based interface; the genome sequence must be sent to the server to be analyzed. A web-based interface can be a limitation, such as when the genome sequence is not yet published and, thus, the data cannot be sent to third parties. Island Viewer, on the other hand, includes a source code for installation on a personal server. However, IslandPick, one of the programs that Island Viewer requires, is strongly dependent on an in-house MySQL database of all published bacterial genomes, which make its use very time-consuming. Moreover, this program requires a very fast server with an unconventional configuration.

Our main goal in this work was to develop new software to predict PAIs with more efficiently than currently available software and to make the software easier to install on a personal computer. Our software, PIPS (pathogenicity island prediction software), predicts PAIs using a novel and more complete approach based on the detection of multiple PAI features: atypical G+C content, codon usage deviation, virulence factors, hypothetical proteins, transposases, flanking tRNA and its absence in non-pathogenic organisms.

In the next sections, we describe the implementation of this software, which is used with several other tools. Model organisms of the genera *Corynebacterium* and *Escherichia* were used in the validation process. The results and discussion section includes data derived from the analyses of *Corynebacterium diphtheriae* and *Escherichia coli* that validate and prove the superior efficiency of this program over other multi-pronged tools. We also performed a case study on *Corynebacterium pseudotuberculosis* that demonstrates the importance of examining various PAI features along with comparisons of PAIs between closely related species.

Materials and Methods

The steps that are required to use PIPS and the necessary input information are represented in the flowchart in Figure 1.

Genomic signature

Putatively acquired regions are identified based on the analysis of G+C content and codon usage patterns, as described below.

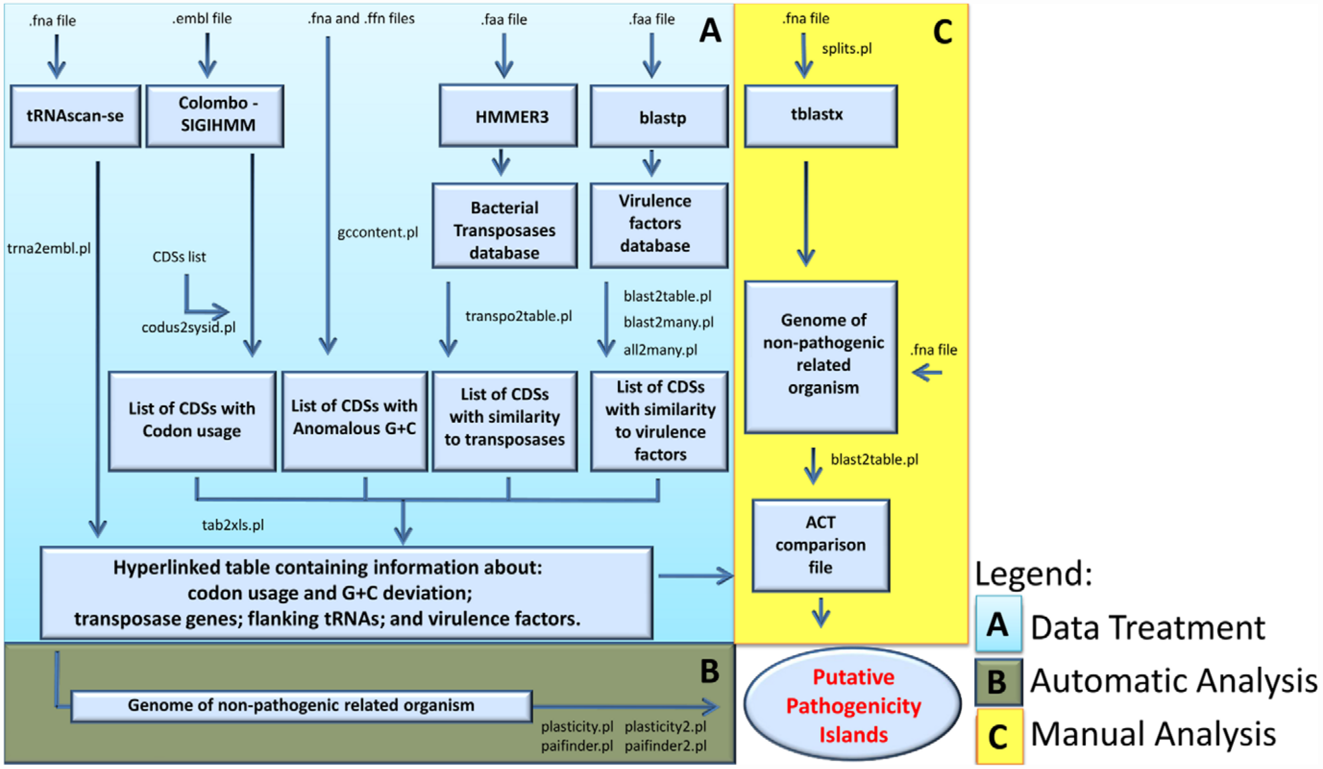

.fna file
.embl file
.fna and .ffn files
.faa file
.faa file
A
.fna file
splits.pl
C
tRNAscan-se
Colombo - SIGIHMM
HMMER3
blastp
tblastx
Bacterial Transposases database
Virulence factors database
trna2embl.pl
CDSs list
codus2sysid.pl
gccontent.pl
transpo2table.pl
blast2table.pl
blast2many.pl
all2many.pl
Genome of non-pathogenic related organism
.fna file
List of CDSs with Codon usage
List of CDSs with Anomalous G+C
List of CDSs with similarity to transposases
List of CDSs with similarity to virulence factors
blast2table.pl
tab2xls.pl
ACT comparison file
Hyperlinked table containing information about: codon usage and G+C deviation; transposase genes; flanking tRNAs; and virulence factors.
Legend:
A Data Treatment
B Automatic Analysis
C Manual Analysis
B
Genome of non-pathogenic related organism
plasticity.pl
paifinder.pl
plasticity2.pl
paifinder2.pl
Putative Pathogenicity Islands

Figure 1. Flowchart presenting each PAI analysis step performed by PIPS. The procedure is divided into the following steps: (A) data treatment; (B) automatic analyses; and (C) manual analyses.
doi:10.1371/journal.pone.0030848.g001

Codon usage deviation. The Colombo SIGI-HMM software was used to predict acquired genes and their putative origins based on taxon-specific differences in codon usage [29]. This software analyzes sequences of predicted proteins of an .embl input file using a hidden Markov model (HMM). This method considers a pattern of observations issued from a hidden Markov chain structure. Additionally, Colombo SIGI-HMM allows the parameter sensitivity to be configured. We pre-configured the parameter sensitivity to 95% to detect any minor anomalies in codon usage because the data are subjected to other major analyses at later stages.

G+C deviation. The Artemis software includes a tool that detects regions with atypical G+C content. This tool calculates the mean G+C content of the genome along with its standard deviation and uses 2.5 standard deviations (SD) as a boundary limit (cutoff) to predict regions with atypical G+C content [30]. The high accuracy of this tool is due to its 1,000-base window size, which identifies even intergenic regions. However, the standard deviation boundary cannot be configured in this program. The base composition of the genome and its coding sequences (CDSs) were analyzed with a Perl script, using input files in .fna and .ffn formats. The script also analyzes the G+C content of the genome and each CDS using 1.5 SD as a boundary to identify putatively acquired regions, as described by Jain *et al.* [31].

To validate the script, the complete *C. diphtheriae* genome was analyzed using Artemis to generate a positive dataset of all genome CDSs with atypical G+C; the sensitivity and specificity of the method were calculated with configurations varying from 0.1 to 3.0 SD. These data were plotted and analyzed in a receiver operating characteristic (ROC) curve (Figure 2) [32].

Based on the ROC curve, the boundary is located between 1.0 and 1.5 SD. The area under the curve (AUC) was then analyzed to determine the most precise value, i.e., the value that gives the largest AUC (Figure 2) [32], which corresponds to the output data generated by the script with a 1.5 SD boundary configuration.

Transposases

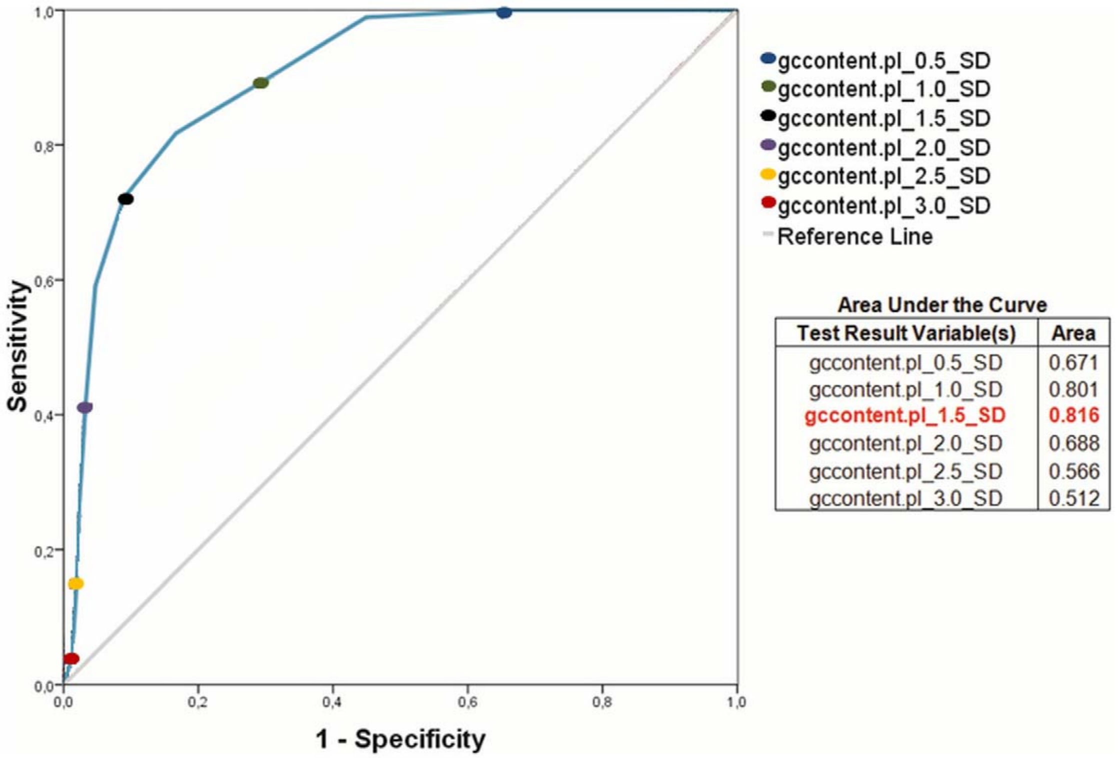
Putative transposase genes are identified by PIPS, which uses HMMER3 [33] to search a bacterial transposase protein database that was retrieved from the Pfam protein families database [34]. The HMMsearch only considers alignments with an e-value of 1e-5 to avoid erroneous alignments that could result in false-positive prediction of transposase genes. A Perl script was created to process the HMMER3 output file and generate a list of putative transposases.

Virulence factors

Virulence genes are identified using BLASTP (BLAST-NCBI [35]) searches with an e-value of 1e-5 against a virulence factor database, mVIRdb. This database contains proteins from eight sources, including toxin, virulence factor and antibiotic resistance gene sequences [36].

Hypothetical proteins

The term "hypothetical protein" is used to identify putative coding sequences without significant matches against non-redundant protein and protein domain databases during genome annotation. Data from annotation in the genome .embl file are used to identify hypothetical proteins. Alternatively, automatic annotation of a whole genome nucleotide file can be processed on our website using an annotation tool (Annotatiohmm). Annotatiohmm is an additional software system that is specifically designed to predict ORFs using the software genemark [37],



Area Under the Curve

Test Result Variable(s)	Area
gccontent.pl_0.5_SD	0.671
gccontent.pl_1.0_SD	0.801
gccontent.pl_1.5_SD	0.816
gccontent.pl_2.0_SD	0.688
gccontent.pl_2.5_SD	0.566
gccontent.pl_3.0_SD	0.512

Figure 2. ROC curve showing the sensitivity and specificity of the Perl script for the identification of regions with GC content deviation. Y-axis: sensitivity; X-axis: 100-specificity. The higher the accuracy is, the closer the curve is to the upper-left corner.
doi:10.1371/journal.pone.0030848.g002

based on a closely related species HMM profile. After the prediction, it performs HMM searches in the Pfam protein families database to create an .embl file, which can be used by PIPS [33,34].

Transfer RNAs

Transfer RNA genes are identified by the software tRNAscan-SE [38], and the output file is parsed by a Perl script to generate a file that can be used in Artemis and ACT (Artemis comparison tool) software to identify flanking tRNAs.

Genomic plasticity

Genomic plasticity analyses are performed using the premise that most pathogenicity islands are absent in non-pathogenic organisms of the same genus or other related species [4]. PIPS analyses may also be performed with a closely related pathogenic organism. However, the pathogenicity islands shared by the two organisms will not be detected during the identification process. In addition, it may erroneously identify other classes of GEIs (e.g., resistance islands and metabolic islands) as PAIs. Therefore, the use and careful choice of the non-pathogenic species is crucial.

PIPS performs two different analyses to identify regions with genomic plasticity. First, an automatic analysis generates a list of putative pathogenicity islands. Second, it creates files that can be manually analyzed to complement and curate the automatic analysis.

Automatic analysis. After the identification of genes that are related to virulence and CDSs presenting characteristics that suggest horizontal transfer, PIPS performs a protein similarity search using BLASTP with the pathogenic bacterium (query) against a non-pathogenic species (subject). The input file in this step contains the predicted protein sequences from the two genomes, and the BLASTP is performed with an e-value of 1e-5. The blastp output file is parsed by Perl scripts that find regions of the non-pathogenic bacterium (subject) that are absent in the pathogenic bacterium (query). Finally, the CDSs are clustered in major regions using their genome coordinates and are identified as "putative pathogenicity islands" based on the finding of virulence factors and characteristics that indicate horizontal transfer, i.e., G+C content deviation or codon usage deviation at higher frequencies than found in the whole genome sequence.

Manual analysis. A second protein search is performed using tblastx against the non-pathogenic species with an e-value of 1e-5. The output file is parsed by a Perl script, generating a comparison file that can be used in the ACT software. This tool permits the visualization of protein similarity areas and insertion, deletion, translocation and inversion regions [39].

The Corynebacterium genus

Corynebacterium diphtheriae strain NCTC 13129 [GenBank: BX248353] – This microorganism is the etiological agent of diphtheria, an infectious disease of the upper respiratory tract, which has been largely controlled by widespread vaccination. Diphtheria has re-emerged in some regions, however, especially in Europe, causing considerable mortality because of the appearance of new biotypes and inadequate vaccination [40].

C. diphtheriae was chosen to validate PIPS because it is a pathogenic species with 13 putative PAIs that is closely related to *C. pseudotuberculosis*. These 13 PAIs were identified by performing analyses based on the following: anomalies in nucleotide composition (e.g., G+C content, GC skew and/or dinucleotide frequency); their absence in *Corynebacterium glutamicum* and *Corynebacterium efficiens*; flanking tRNAs; and the presence of genes encoding virulence factors, such as fimbrial and fimbria-related genes, iron-uptake systems, a potential siderophore biosynthesis system, a lantibiotic biosynthesis system, exported proteins, two-component-system proteins, insertion sequence transposases and the *tox* gene, which is located in a corynephage-acquired region and is responsible for the pathognomonic symptoms of diphtheria [41].

C. glutamicum strain ATCC 13032 [GenBank: BX927147] was chosen for the comparison analyses, which is non-pathogenic and of biotechnological interest, being widely used for the industrial production of amino acids such as L-glutamic acid and L-lysine [42].

C. pseudotuberculosis strains 1002 [GenBank: CP001809] and C231 [GenBank: CP001829] were chosen to test PIPS after validation, both of which are facultative intracellular pathogens. This species is the etiological agent of the globally distributed disease known as caseous lymphadenitis (CLA), which mainly affects small ruminants. However, this bacterial species can affect a wide range of host species, causing different diseases. *C. pseudotuberculosis* is less well studied than *C. diphtheriae*. The virulence factors of *C. pseudotuberculosis* that lead to CLA have not yet been exhaustively characterized, making studies concerning PAIs in this species invaluable [43].

The Escherichia coli species

Among the *E. coli* species, we chose the uropathogenic *E. coli* (*UPEC*) strain *CFT073* [GenBank: AE014075], a pyelonephritogenic *UPEC* isolate that has a wide range of putative and known virulence genes that are responsible for survival in the host. The *UPEC* strains deserve great attention because they are responsible for up to 90% of uncomplicated urinary tract infections. In addition, using comparative genomic hybridization analysis and combining genomics, bioinformatics, and microarray technologies, 13 pathogenicity islands larger than 30 kb have already been described in *E. coli* strain CFT073 [44].

Escherichia coli strain *K-12*, substrain *MG1655* [GenBank: U00096], was chosen for the genomic plasticity comparison with the *UPEC* strain *CFT073* because it is the best-studied non-pathogenic strain of this species. In addition, the genomic sequence of this strain undergoes constant curation and updating, reducing erroneous annotations [45,46].

Results and Discussion

Software validation using *C. diphtheriae* PAIs

A genomic region was identified as a putative PAI of *C. diphtheriae* (PICD) when it had the following properties. First, it presented most of the PAI features in *C. diphtheriae* (e.g., higher concentration inside the genomic region than in the whole genome of virulence factors and/or hypothetical proteins and CDSs with codon usage deviation and/or atypical G+C content). Second, it was absent in *C. glutamicum*. PIPS found 12 of the 13 *C. diphtheriae* PAIs; except for *C. diphtheriae* PICDs 10 and 13, all of the islands were 1–7 CDSs larger than the published sequences (Figure S1).

Comparison between PIPS and other programs

To compare the efficiency of PIPS in identifying PAIs with the results of other available programs, we analyzed the sensitivity and specificity using published data, with *C. diphtheriae* PAIs as a positive dataset (Table 1). For this task, each CDS in a genome was labeled as "positive" when it was harbored by a PAI and "negative" otherwise. For more detailed information concerning the composition of PAIs predicted by the programs, see Table S1.

Table 1. Comparison between the software used to identify pathogenicity islands in the *C. diphtheriae* strain *NCTC 13129*.

Software	Sensitivity (%)	Specificity(%)	Accuracy(%)
IslandPath_DIMOB	13.6	98.3	89.2
IslandPick	65.2	81.9	80.1
SIGI_HMM	14.0	94.9	86.2
IslandViewer	74.4	76.4	76.2
PredictBias_GEI	30.8	84.4	78.6
PredictBias_PAI	2.4	88.7	79.4
PIPS_Auto	86.4	85.0	85.1
PIPS_Manual	96.8	87.1	88.1

doi:10.1371/journal.pone.0030848.t001

PredictBias showed good specificity (88.7%), at the cost of sensitivity (2.4%), when using only predicted PAIs (PredictBias_PAI) as a positive dataset for the test (Table 1). The sensitivity was higher (30.8%) when GEIs identified by the program (Table 1) were used as a positive dataset (PredictBias). The classification errors may be a consequence of the virulence factor database used by the program. The database was created using an NCBI search with the following keywords: 'Virulence', 'Adhesin', 'Siderophore', 'Invasin', 'Endotoxin' and 'Exotoxin' [36]. The size of the database is a determining factor in discerning PAIs from GEIs. The larger the database is, the higher the probability of correct classification of a gene as a virulence factor and, consequently, the higher the probability of correct PAI identification.

IslandViewer identified 10 *C. diphtheriae* PAIs; however, their sizes varied from those of the published PAIs. Two of the three programs used in IslandViewer, IslandPath-DIMOB and Colombo/SIGI-HMM, had low sensitivity for PAI prediction (13.6% and 14%, respectively). However, the poor performance of Colombo/SIGI-HMM mainly results from the high stringency of its parameters. In our case, setting the program's "sensitivity" parameter to 95% resulted in higher sensitivity and proved to be an efficient approach for the identification of regions with codon usage deviation.

IslandPick had a higher sensitivity (65.2%) than the other programs used in IslandViewer (Table 1). This software performs analyses that are based on the premise that PAIs are absent in related non-pathogenic organisms. The superior performance of this strategy corroborates the importance of genomic comparisons between the bacterium to be analyzed and a non-pathogenic strain or species of the same genus. Finally, the programs IslandPick, IslandPath-DIMOB and Colombo/SIGI-HMM, when combined in IslandViewer, gave a higher sensitivity for predicting PAIs (74.4%) than when used alone (65.2%, 13.6% and 14.0%, respectively), which demonstrates the importance of a combined analysis instead solely analyzing a single PAI feature.

PIPS correctly identified 12 of the 13 PAIs. Based on *C. diphtheriae* genomic annotation, the only PAI that was not identified by PIPS, PICD 5 of *C. diphtheriae*, has an atypical G+C content of 52.2%. However, when a boundary value of 1.5 standard deviations was used to identify atypical G+C content, we found reference values that varied from 45.95 to 60.04%. In addition, when using Artemis, the annotation tool did not indicate any atypical G+C in this PAI, which is in agreement with PIPS. Moreover, except for its absence in *C. glutamicum*, PICD 5 of *C. diphtheriae* did not show any other PAI feature. Additionally, the

Table 2. Comparison between the software used to identify pathogenicity islands in the uropathogenic *E. coli* strain *CFT 073*.

Software	Sensitivity (%)	Specificity(%)	Accuracy(%)
IslandPath_DIMOB	44.5	99.3	90.2
IslandPick	7.5	99.7	84.5
SIGI_HMM	21.9	96.9	84.5
IslandViewer	55.8	96.2	89.5
PredictBias_GEI	60.0	93.7	88.1
PredictBias_PAI	39.2	96.2	86.8
PIPS_Auto	94.8	93.7	93.9

doi:10.1371/journal.pone.0030848.t002

IslandViewer and PredictBias results also indicate that the classification of PICD 5 of *C. diphtheriae* as a PAI is erroneous.

Finally, automatic analysis using PIPS gave better performance than the previously available techniques (86.4% sensitivity, 85.0% specificity). However, manual analysis of PIPS results in improved identification of the PAIs (96.8% sensitivity, 87.1 specificity), showing the importance of manual curation of the data based on biological knowledge.

Identification of the well-studied pathogenicity islands of the uropathogenic *E. coli* strain *CFT 073*

After the validation of PIPS with a Gram-positive bacterium, we analyzed the *UPEC* strain CFT073 to determine how well PIPS performs with a Gram-negative bacterium. Gram-negative bacteria are important in this context because their PAIs tend to present all of the PAI features concurrently; additionally, *E. coli* PAIs have been extensively described in the literature [5,7,44,47–51]. The *UPEC* strain CFT073 was chosen because it possesses several known PAIs. We used 13 PAIs described by Lloyd *et al.* [44] as our gold standard and compared the accuracy of PIPS with IslandViewer and PredictBias, as we had performed with *C. diphtheriae*. The *E. coli* strain *K-12* was used as the non-pathogenic closely related organism for validation in this step. The sensitivity and specificity of the methods are shown in Table 2.

The specificity achieved by the other methods (93.7–99.3%) was greater than that of PIPS (93.7%), although PIPS had a much higher sensitivity (94.8%) than the other methods (7.5–60%). This reduced specificity may result from novel pathogenicity islands that were not previously identified rather than false-positive results. In addition, the higher accuracy of PIPS (93.9%) when compared to the other methods (84.5–90.2%) supports our previous conclusion that PIPS gives the best performance when identifying true positive and true negative CDSs, based on the analysis of PAIs of the *UPEC* strain *CFT073*.

Case study: C. pseudotuberculosis

After validating PIPS, we identified putative PAIs of *C. pseudotuberculosis*. The underlying properties (i.e., codon usage, G+C content, virulence factors and hypothetical proteins) of the *C. pseudotuberculosis* (PICPs) and *C. diphtheriae* (PICDs) PAIs are given in Table 3. For further details, please refer to Figure S2.

G+C content. *C. pseudotuberculosis* PICPs had similar frequencies of CDSs with G+C content deviations to those identified in *C. diphtheriae* PICDs. Compared to the frequency in their respective genomes, the frequency of CDSs with G+C content deviation in *C. pseudotuberculosis* PICPs and *C. diphtheriae* PICDs was approximately doubled.

Codon usage. The frequency of CDSs with codon usage deviation was found to be higher in the *C. diphtheriae* PICDs than in the *C. pseudotuberculosis* PICPs, reflecting the patterns found in the genomes of *C. diphtheriae* and *C. pseudotuberculosis* (Table 3). However, the frequency of CDSs with codon usage deviation in *C. pseudotuberculosis* PICPs, although lower than the frequency in *C. diphtheriae* PICDs, was three times that in the *C. pseudotuberculosis* genome (Table 3). In PICDs, the frequency of this feature was twice that in the whole genome.

Virulence factors. The frequency of virulence factors in *C. pseudotuberculosis* PICPs is approximately twice that in other parts of the *C. pseudotuberculosis* genome, in contrast to findings in *C. diphtheriae* PICDs (Table 3). When looking at PAIs separately, the frequencies of virulence factors in *C. pseudotuberculosis* PICPs were also higher than in *C. diphtheriae* PICDs; however, *C. diphtheriae* PICDs had higher frequencies of hypothetical proteins, i.e., putative proteins without significant similarity to any previously described protein (Table 3). These proteins may have an unknown role in pathogenicity, possibly explaining the low frequencies of the possible virulence factors found in these regions.

Frequencies of the features in each *C. pseudotuberculosis* PICP

The properties that were analyzed in a global genomic view in the previous section (i.e., codon usage, G+C content, virulence factors and hypothetical proteins) were assessed for each *C. pseudotuberculosis* PICP to compare their contributions to the classification. To plot this graph, we used the frequency, in percent, of the CDSs, presenting the chosen feature in the *C. pseudotuberculosis* PICP relative to the total number of CDSs in the same PICP.

Table 3. Percentage of PAI features along the genome and the pathogenicity islands of *C. pseudotuberculosis* and *C. diphtheriae*.

	Codon usage deviation (%)	GC content deviation (%)	Virulence factors (%)	Hypothetical proteins (%)
C. diphtheriae NCTC 13129 PICDs	45.20	20.80	18.40	39.20
C. diphtheriae NCTC 13129 genome	26.89	9.52	17.45	27.19
C. pseudotuberculosis 1002 PICPs	14.79	23.08	30.77	31.95
C. pseudotuberculosis 1002 genome	3.52	11.65	17.27	31.95
C. pseudotuberculosis C231 PICPs	19.62	20.25	32.91	31.65
C. pseudotuberculosis C231 genome	3.80	10.76	17.77	31.64

doi:10.1371/journal.pone.0030848.t003

In a comparison of the frequency of CDSs with codon usage deviation, *C. pseudotuberculosis* PICPs 3, 5, 6 and 7 had higher frequencies than those found in the whole genome of *C. pseudotuberculosis* 1002. In *C. pseudotuberculosis* C231, together with the previously described PAIs (PICPs 3, 5, 6 and 7), *C. pseudotuberculosis* PICP1 also had a greater frequency of CDSs with codon usage deviation than that of the whole genome (Figure 3). This observation may mean that *C. pseudotuberculosis* PICP1 has become more adapted to the acceptor's codon usage in strain 1002 when compared to the same PAI in strain C231. The frequency of CDSs with G+C content deviation in strains 1002 and C231 was higher in *C. pseudotuberculosis* PICPs 1, 3, 5 and 6 (Figure 3).

In general, the frequency of genes with similarity to virulence factors in PAIs was greater than that in the rest of the genome, except for *C. pseudotuberculosis* PICP5. However, this island, along with *C. pseudotuberculosis* PICPs 3 and 6, had higher frequencies of hypothetical proteins.

No single characteristic was consistent throughout all *C. pseudotuberculosis* PICPs. However, the absence of *C. pseudotuberculosis* PICPs in non-pathogenic bacteria, in addition to a high frequency of at least one of the classic PAI features, and the finding of virulence genes were used as determining factors for the characterization of a PAI.

Co-occurrence of pathogenicity islands in *C. pseudotuberculosis* and *C. diphtheriae*

C. pseudotuberculosis PICPs were compared to the genome of *C. diphtheriae* NCTC 13129 to determine whether these islands are present in this organism.

Interestingly, most *C. pseudotuberculosis* PICP3 genes are found in the genome of *C. diphtheriae* NCTC 13129, with the same gene order, identified as *C. diphtheriae* PICD 3 (Figure 4). The presence of this PAI in two pathogenic species and its absence in non-pathogenic *C. glutamicum* provide evidence for the importance of this region for determining the virulence of *C. pseudotuberculosis* and *C. diphtheriae*.

Moreover, the flanking regions of the PICP5 of *C. pseudotuberculosis* are the same as those of PICD8 of *C. diphtheriae* (Figure 5). This pattern highlights this region as a putative "hotspot" for the insertion of transposons and, most likely, GEIs.

Conclusions

Pathogenicity islands play a major role in the virulence of pathogenic bacteria, and therefore, their correct identification and characterization may provide valuable data.

We developed software (PIPS) that accurately identifies pathogenicity islands; it is easy to install, which makes it accessible even to researchers with little computational knowledge. In addition, this software has a web-based interface that is platform and installation independent, facilitating fast analysis. Moreover, PIPS uses a complete approach that is based on the detection of multiple PAIs, i.e., atypical G+C content, codon usage deviation, virulence factors, hypothetical proteins, transposases, flanking tRNA and its absence in non-pathogenic organisms.

During the validation, this software identified 12 of the 13 previously described *C. diphtheriae* PAIs, demonstrating its superior efficiency compared to the other currently available software systems, which identified 6 and 10 PAIs (PredictBias and IslandViewer, respectively). Furthermore, PIPS achieved a high

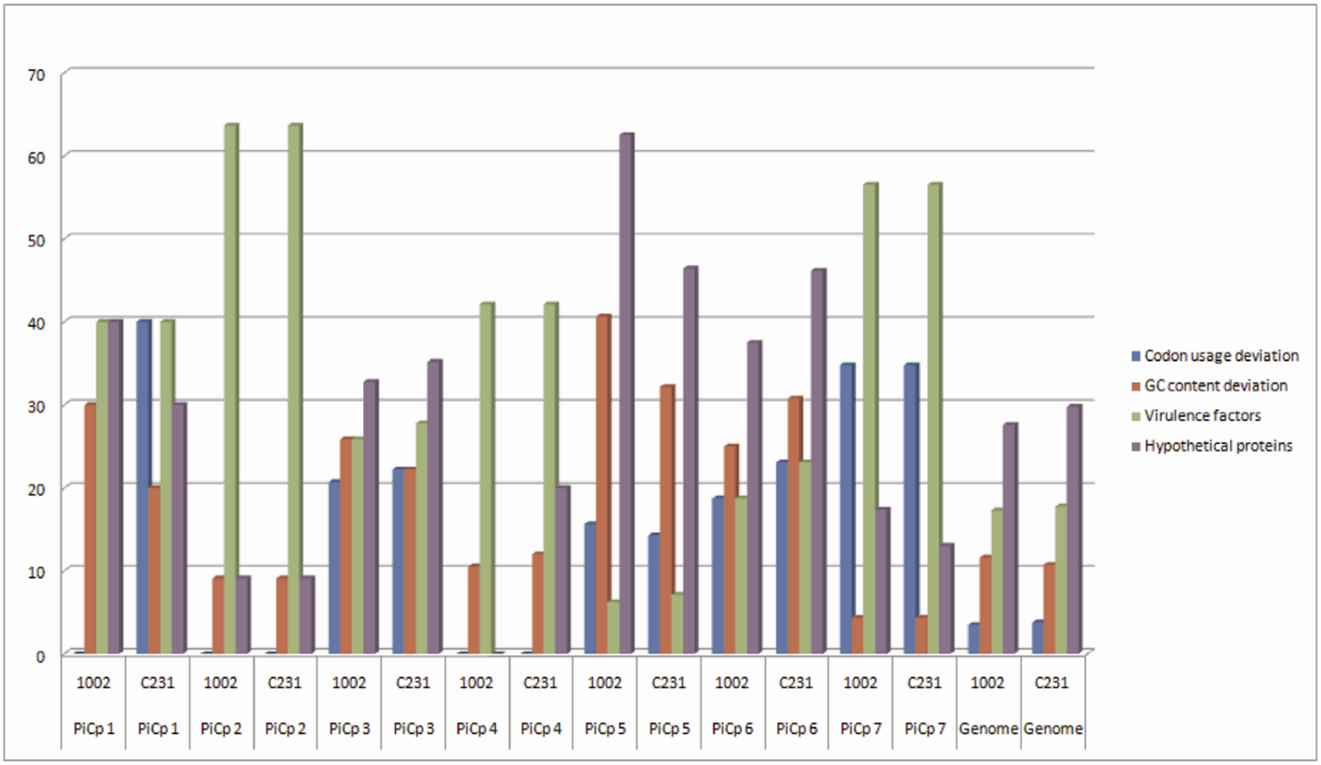


Figure 3. Frequencies of PAI features within the PICPs and in the full genomes of *C. pseudotuberculosis* strains 1002 and C231. Y-axis: frequency in percentage; X-axis: PICPs and genomes of *C. pseudotuberculosis* strains 1002 and C231. The frequencies of the features in each PICP and in the whole genomes of the two strains are represented in the following colors: blue for codon usage deviation; red for GC content deviation; green for virulence factors; and purple for hypothetical proteins.
doi:10.1371/journal.pone.0030848.g003

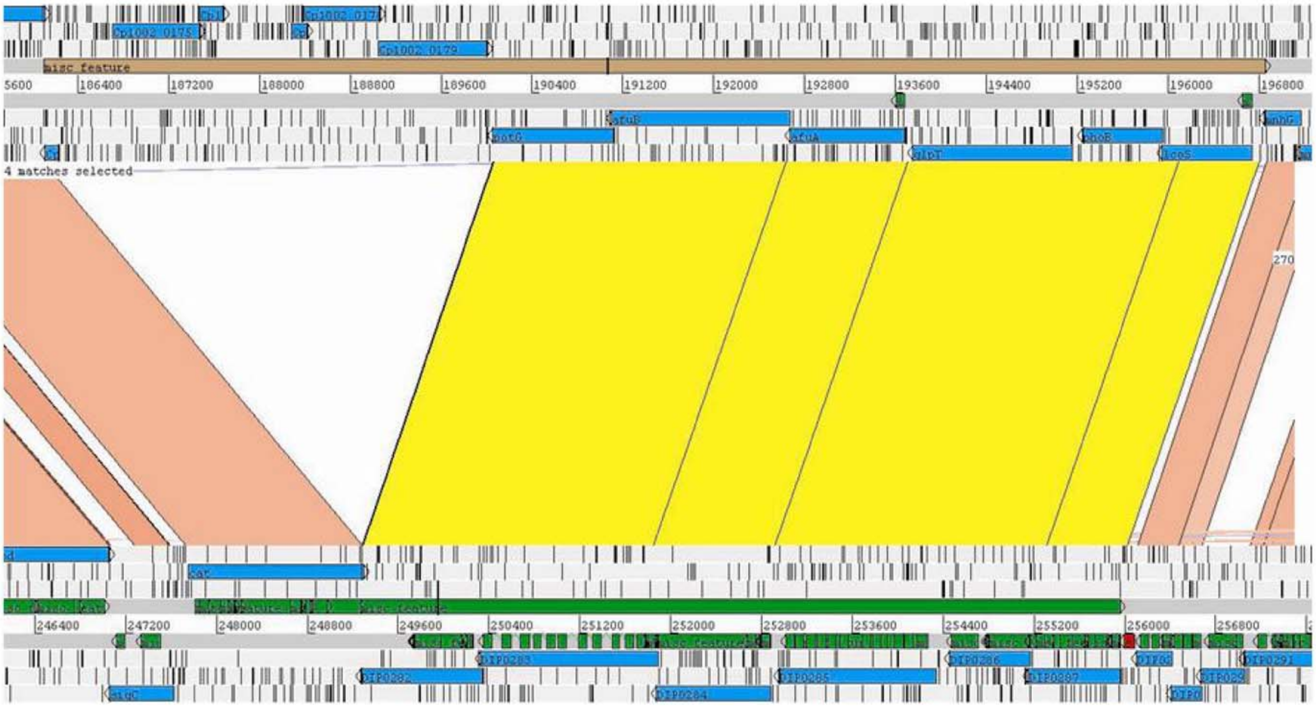


Figure 4. PICP3 and PICD3 (top and bottom, respectively) in the *C. pseudotuberculosis* and *C. diphtheriae* genomes. Cp1002 and *C. diphtheriae* NCTC 13129 are shown at the top and bottom, respectively. Regions of similarity between the two genomes are marked in pink. Regions of similarity between two PAIs are marked in yellow, showing the presence of PICD3 in *C. pseudotuberculosis* with an insertion. Image generated by ACT (the Artemis Comparison Tool).
doi:10.1371/journal.pone.0030848.g004

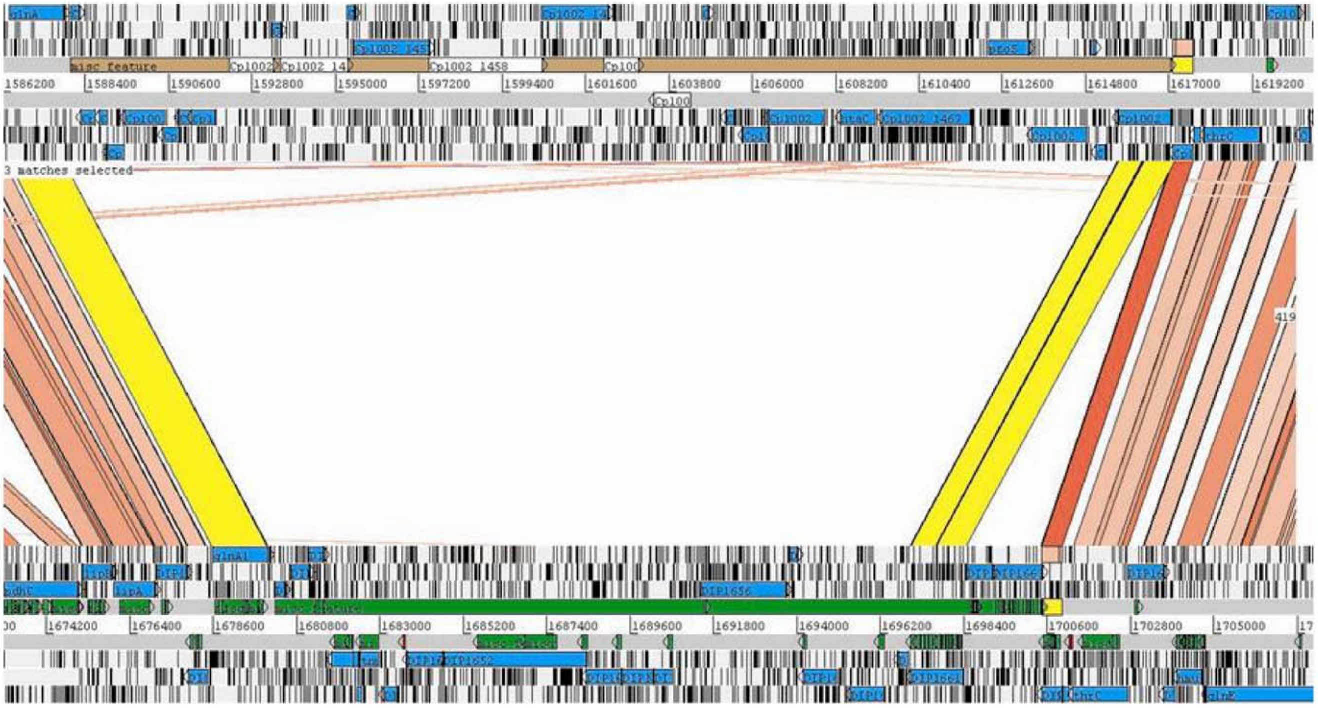


Figure 5. Replacement of the *C. diphtheriae* PICD8 (bottom) with *C. pseudotuberculosis* PICP5 (top). Regions of similarity are represented by lines between the two genomes. The flanking regions of PICD8 and PICP5 are highlighted in yellow, showing the region of replacement. Image generated by ACT (the Artemis Comparison Tool).
doi:10.1371/journal.pone.0030848.g005

overall sensitivity, specificity and accuracy in identifying PAIs in *C. diphtheriae* NCTC13129 and *E. coli* CFT073. Moreover, we predicted 7 PAIs in *C. pseudotuberculosis* and showed that no single characteristic was consistent throughout all of the *C. pseudotuberculosis* PICPs. This latter finding, in addition to our success with this program, highlights the need for a multi-pronged strategy toward PAI identification that heavily weights the absence in a closely related non-pathogenic organism in addition to signs of HGT and the presence of virulence factors.

Finally, the identification of *C. pseudotuberculosis* PICP3, an island that is shared by *C. pseudotuberculosis* and *C. diphtheriae*, along with the identification of *C. pseudotuberculosis* PICP5, an island that is located in a putative "hotspot", corroborates the accuracy of the program for correct identification of PAIs.

Future PIPS development will focus on increasing the software speed in searches for insertion sequences. The next versions will also aim to facilitate analysis through the implementation of a graphic interface and minimization of the required programs (Availability and requirements are described in Appendix S1).

Supporting Information

Figure S1 Prediction of PICD12 of *C. diphtheriae* with a different size than the literature prediction. At the top, the *C. diphtheriae* genome; at the bottom, the *C. glutamicum* genome. In green, highlighted by an orange box, *C. diphtheriae* PICD12 as described in the literature; in red, an additional region identified by PIPS. This image was generated by ACT.
(DOC)

Figure S2 Graphic representation of PAI features in the genome (A) and in the pathogenicity islands (B) of C. pseudotuberculosis and C. diphtheriae. Y-axis: frequency as a percentage; X-axis: codon usage deviation, GC content deviation, virulence factors and hypothetical proteins. C. diphtheriae strain NCTC 13129 is in blue, and C. pseudotuberculosis strains 1002 and C231 are in red and green, respectively. (A) Frequency of the PAI features in the genomes and (B) frequency of the PAI features in the pathogenicity islands of the bacteria.
(DOC)

Table S1 PAI composition. The PAIs composition of the *C. diphtheriae* strain NCTC 13129, as described in the literature and as identified by PIPS, IslandViewer and PredicBias.
(DOC)

Appendix S1 Availability and Requirements.
(DOC)

Acknowledgments

The authors thank all of the collaborators of the Genome Network of the State of Minas Gerais and the Genomics and Proteomics Network of the state of Pará for their contributions.

Author Contributions

Conceived and designed the experiments: AM VA. Performed the experiments: SCS VACA RTJR LC AS. Analyzed the data: SCS VACA RTJR LC AS JB ET AT RH ALMG AM VA. Contributed reagents/materials/analysis tools: SCS VACA RTJR LC AS JB ET AT RH ALMG AM VA. Wrote the paper: SCS VACA RTJR LC AS JB ET AT RH ALMG AM VA. Read and gave insights about the software: SCS VACA RTJR LC AS JB ET AT RH ALMG AM VA.

References

1. Oren A (2004) Prokaryote diversity and taxonomy: current status and future challenges. Philos Trans R Soc Lond B Biol Sci 359: 623–638.
2. Dobrindt U, Hacker J (2001) Whole genome plasticity in pathogenic bacteria. Curr Opin Microbiol 4: 550–557.
3. Maurelli AT, Fernández RE, Bloch CA, Rode CK, Fasano A (1998) "Black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of Shigella spp. and enteroinvasive Escherichia coli. Proc Natl Acad Sci U S A 95: 3943–3948.
4. Schmidt H, Hensel M (2004) Pathogenicity islands in bacterial pathogenesis. Clin Microbiol Rev 17: 14–56.
5. Hacker J, Bender L, Ott M, Wingender J, Lund B, et al. (1990) Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extraintestinal Escherichia coli isolates. Microb Pathog 8: 213–225.
6. Hou YM (1999) Transfer RNAs and pathogenicity islands. Trends Biochem Sci 24: 295–298.
7. Ou H, Chen L, Lonnen J, Chaudhuri RR, Thani AB, et al. (2006) A novel strategy for the identification of genomic islands by comparative analysis of the contents and contexts of tRNA sites in closely related bacteria. Nucleic Acids Res 34: e3.
8. Langille MGI, Hsiao WWL, Brinkman FSL (2008) Evaluation of genomic island predictors using a comparative genomics approach. BMC Bioinformatics 9: 329.
9. Karlin S, Mrázek J, Campbell AM (1998) Codon usages in different gene classes of the Escherichia coli genome. Mol Microbiol 29: 1341–1355.
10. Hershberg R, Petrov DA (2009) General rules for optimal codon choice. PLoS Genet 5: e1000556.
11. Dufraigne C, Fertil B, Lespinats S, Giron A, Deschavanne P (2005) Detection and characterization of horizontal transfers in prokaryotes using genomic signature. Nucleic Acids Res 33: e6.
12. Lawrence JG, Ochman H (1997) Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol 44: 383–397.
13. Hsiao WWL, Ung K, Aeschliman D, Bryan J, Finlay BB, et al. (2005) Evidence of a large novel gene pool associated with prokaryotic genomic islands. PLoS Genet 1: e62.
14. Karaolis DK, Johnson JA, Bailey CC, Boedeker EC, Kaper JB, et al. (1998) A Vibrio cholerae pathogenicity island associated with epidemic and pandemic strains. Proc Natl Acad Sci U S A 95: 3134–3139.
15. Schumann W (2007) Thermosensors in eubacteria: role and evolution. J Biosci 32: 549–557.
16. Tu Q, Ding D (2003) Detecting pathogenicity islands and anomalous gene clusters by iterative discriminant analysis. FEMS Microbiol Lett 221: 269–275.
17. Vernikos GS, Parkhill J (2006) Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. Bioinformatics 22: 2196–2203.
18. van Passel MWJ, Bart A, Waaijer RJA, Luyf ACM, van Kampen AHC, et al. (2004) An in vitro strategy for the selective isolation of anomalous DNA from prokaryotic genomes. Nucleic Acids Res 32: e114.
19. Liò P, Vannucci M (2000) Finding pathogenicity islands and gene transfer events in genome data. Bioinformatics 16: 932–940.
20. Hsiao W, Wan I, Jones SJ, Brinkman FSL (2003) IslandPath: aiding detection of genomic islands in prokaryotes. Bioinformatics 19: 418–420.
21. Zhang CT, Wang J, Zhang R (2001) A novel method to calculate the G+C content of genomic DNA sequences. J Biomol Struct Dyn 19: 333–341.
22. Zhang C, Zhang R (2004) Genomic islands in Rhodopseudomonas palustris. Nat Biotechnol 22: 1078–1079.
23. Zhang R, Zhang C (2004) A systematic method to identify genomic islands and its applications in analyzing the genomes of Corynebacterium glutamicum and Vibrio vulnificus CMCP6 chromosome I. Bioinformatics 20: 612–622.
24. Merkl R (2004) SIGI: score-based identification of genomic islands. BMC Bioinformatics 5: 22.
25. Mantri Y, Williams KP (2004) Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities. Nucleic Acids Res 32: D55–8.
26. Gao J, Chen L (2010) Theoretical methods for identifying important functional genes in bacterial genomes. Res Microbiol 161: 1–8.
27. Pundhir S, Vijayvargiya H, Kumar A (2008) PredictBias: a server for the identification of genomic and pathogenicity islands in prokaryotes. In Silico Biol 8: 223–234.
28. Langille MGI, Brinkman FSL (2009) IslandViewer: an integrated interface for computational identification and visualization of genomic islands. Bioinformatics 25: 664–665.
29. Waack S, Keller O, Asper R, Brodag T, Damm C, et al. (2006) Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. BMC Bioinformatics 7: 142.
30. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, et al. (2000) Artemis: sequence visualization and annotation. Bioinformatics 16: 944–945.
31. Jain R, Ramineni S, Parekh N (2008) Integrated Genomic Island Prediction Tool (IGIPT). Proceedings of the 2008 International Conference on Information Technology. pp 131–132.

32. Zweig MH, Campbell G (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 39: 561–577.
33. Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39: W29–37.
34. Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. (2010) The Pfam protein families database. Nucleic Acids Res 38: D211–22.
35. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.
36. Zhou CE, Smith J, Lam M, Zemla A, Dyer MD, et al. (2007) MvirDB–a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. Nucleic Acids Res 35: D391–4.
37. Lukashin AV, Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res 26: 1107–1115.
38. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25: 955–964.
39. Carver TJ, Rutherford KM, Berriman M, Rajandream M, Barrell BG, et al. (2005) ACT: the Artemis Comparison Tool. Bioinformatics 21: 3422–3423.
40. Hadfield TL, McEvoy P, Polotsky Y, Tzinserling VA, Yakovlev AA (2000) The pathology of diphtheria. J Infect Dis 181(Suppl 1): S116–20.
41. Cerdeño-Tárraga AM, Efstratiou A, Dover LG, Holden MTG, Pallen M, et al. (2003) The complete genome sequence and analysis of Corynebacterium diphtheriae NCTC13129. Nucleic Acids Res 31: 6516–6523.
42. Kalinowski J, Bathe B, Bartels D, Bischoff N, Bott M, et al. (2003) The complete Corynebacterium glutamicum ATCC 13032 genome sequence and its impact on the production of L-aspartate-derived amino acids and vitamins. J Biotechnol 104: 5–25.
43. Dorella FA, Pacheco LGC, Oliveira SC, Miyoshi A, Azevedo V (2006) Corynebacterium pseudotuberculosis: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. Vet Res 37: 201–218.
44. Lloyd AL, Rasko DA, Mobley HLT (2007) Defining genomic islands and uropathogen-specific genes in uropathogenic Escherichia coli. J Bacteriol 189: 3532–3546.
45. Blattner FR, Plunkett G, III, Bloch CA, Perna NT, Burland V, et al. (1997) The complete genome sequence of Escherichia coli K-12. Science 277: 1453–1462.
46. Riley M, Abe T, Arnaud MB, Berlyn MKB, Blattner FR, et al. (2006) Escherichia coli K-12: a cooperatively developed annotation snapshot–2005. Nucleic Acids Res 34: 1–9.
47. Hochhut B, Dobrindt U, Hacker J (2005) Pathogenicity islands and their role in bacterial virulence and survival. Contrib Microbiol 12: 234–254.
48. Hacker J, Blum-Oehler G, Mühldorfer I, Tschäpe H (1997) Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol Microbiol 23: 1089–1097.
49. Blum G, Ott M, Lischewski A, Ritter A, Imrich H, et al. (1994) Excision of large DNA regions termed pathogenicity islands from tRNA-specific loci in the chromosome of an Escherichia coli wild-type pathogen. Infect Immun 62: 606–614.
50. Hochhut B, Wilde C, Balling G, Middendorf B, Dobrindt U, et al. (2006) Role of pathogenicity island-associated integrases in the genome plasticity of uropathogenic Escherichia coli strain 536. Mol Microbiol 61: 584–595.
51. Tsai N, Wu Y, Chen J, Wu C, Tzeng C, et al. (2006) Multiple functions of l0036 in the regulation of the pathogenicity island of enterohaemorrhagic Escherichia coli O157:H7. Biochem J 393: 591–599.

II.I.6 GIPSy: Genomic island prediction software.
Soares SC, Geyik H, Ramos RT, de Sá PH, Barbosa EG, Baumbach J, Figueiredo HC, Miyoshi A, Tauch A, Silva A, **Azevedo V**.
J Biotechnol. 2015 Sep 12. pii: S0168-1656(15)30115-2. doi:10.1016/j.jbiotec.2015.09.008. [Epub ahead of print]

Após a implementação do programa PIPS, vimos a necessidade de criar uma nova ferramenta que fosse capaz de predizer outras classes de ilhas genômicas além das ilhas de patogenicidade (PAI), sejam elas ilhas de resistência (RI), ilhas metabólicas (MI) e ilhas simbióticas (SI). Neste contexto, nosso grupo desenvolveu o software GIPSy: Genomic Island Prediction Software, que realiza as mesmas análises que o software PIPS, acrescentando a predição de fatores relacionados a metabolismo, resistência a antibióticos e simbiose com plantas. GIPSy predisse corretamente as seguintes ilhas genômicas previamente descritas: 13 PAIs maiores que 30kb em *Escherichia coli* CFT073; 1 MI de *Burkholderia pseudomallei* K96243, que aparenta ser uma ilha diversa (miscellaneous island); 1 RI de *Acinetobacter baumannii* AYE, chamada AbaR1; e, 1 SI de *Mesorhizobium loti* MAFF303099 que apresenta uma estrutura mosaica. O software apresentou acurácia na predição de cada uma das classes de ilhas genômicas utilizando referências da literatura além de ser um software totalmente gráfico, o que facilita a instalação e utilização do mesmo por biólogos computacionais. GIPSy foi o primeiro software desenvolvido para a predição de todas as classes de ilhas genômicas de forma específica, foi implementado em Java e está disponível publicamente no site http://www.bioinformatics.org/groups/?group_id=1180

Journal of Biotechnology 232 (2016) 2–11

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

GIPSy: Genomic island prediction software

Siomar C. Soares[a,b,*], Hakan Geyik[b,c], Rommel T.J. Ramos[d], Pablo H.C.G. de Sá[d], Eudes G.V. Barbosa[b,e], Jan Baumbach[e], Henrique C.P. Figueiredo[f], Anderson Miyoshi[b], Andreas Tauch[c], Artur Silva[d], Vasco Azevedo[b]

[a] *Department of Immunology, Microbiology and Parasitology, Federal University of Triângulo Mineiro, Uberaba, Minas Gerais, Brazil*
[b] *Department of General Biology, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil*
[c] *Institute for Genome Research and Systems Biology, Center for Biotechnology (CeBiTec), Bielefeld University, Bielefeld, Germany*
[d] *Department of Genetics, Federal University of Pará, Belém, Pará, Brazil*
[e] *Computational Biology laboratory, Institute for Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark*
[f] *AQUACEN, National Reference Laboratory for Aquatic Animal Diseases, Ministry of Fisheries and Aquaculture, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil*

ARTICLE INFO

Article history:
Received 6 May 2015
Received in revised form 28 August 2015
Accepted 11 September 2015
Available online 12 September 2015

Keywords:
Genomic islands
Pathogenicity islands
Resistance islands
Symbiotic islands
Metabolic islands
Horizontal gene transfer
Genome plasticity

ABSTRACT

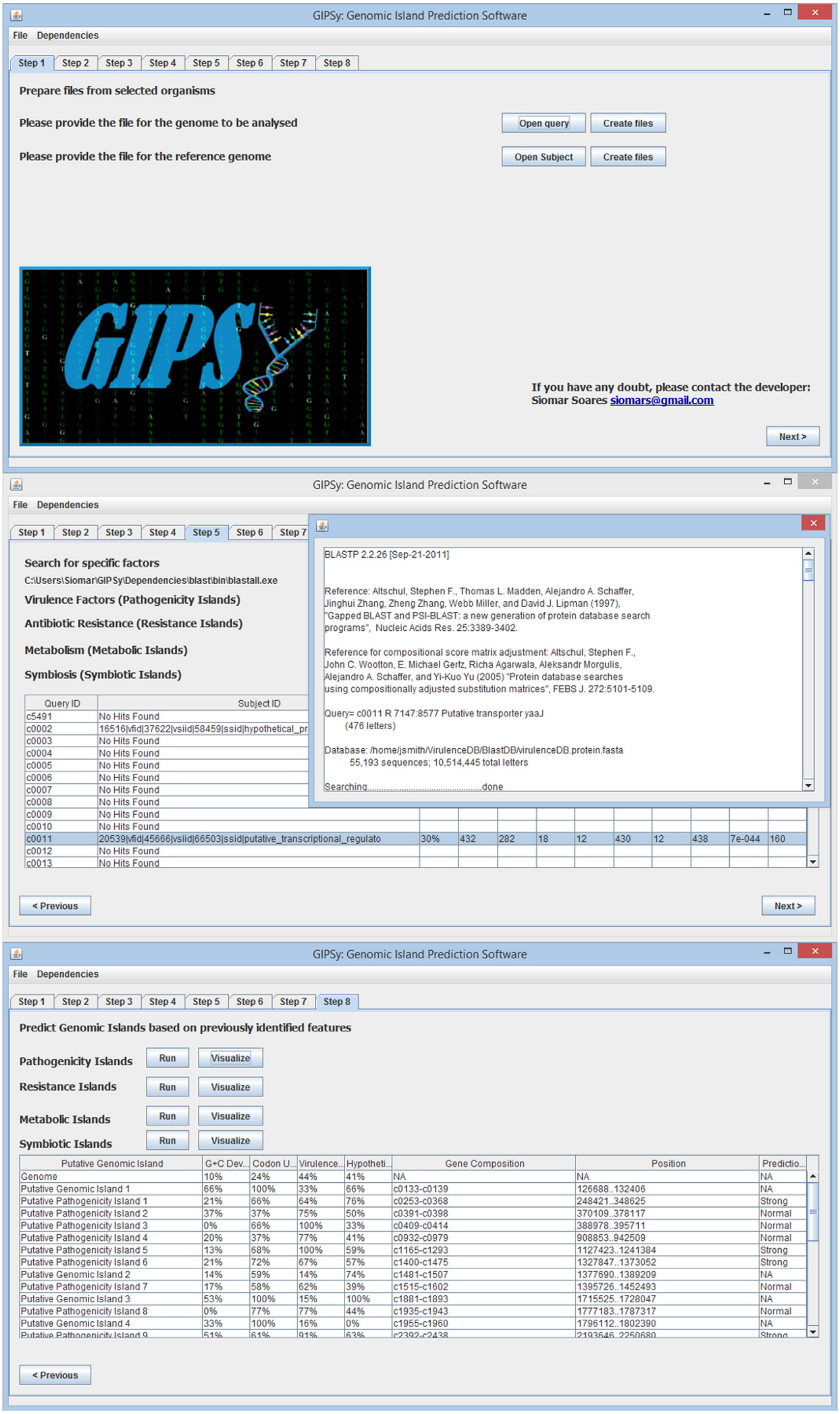
Bacteria are highly diverse organisms that are able to adapt to a broad range of environments and hosts due to their high genomic plasticity. Horizontal gene transfer plays a pivotal role in this genome plasticity and in evolution by leaps through the incorporation of large blocks of genome sequences, ordinarily known as genomic islands (GEIs). GEIs may harbor genes encoding virulence, metabolism, antibiotic resistance and symbiosis-related functions, namely pathogenicity islands (PAIs), metabolic islands (MIs), resistance islands (RIs) and symbiotic islands (SIs). Although many software for the prediction of GEIs exist, they only focus on PAI prediction and present other limitations, such as complicated installation and inconvenient user interfaces. Here, we present GIPSy, the genomic island prediction software, a standalone and user-friendly software for the prediction of GEIs, built on our previously developed pathogenicity island prediction software (PIPS). We also present four application cases in which we crosslink data from literature to PAIs, MIs, RIs and SIs predicted by GIPSy. Briefly, GIPSy correctly predicted the following previously described GEIs: 13 PAIs larger than 30 kb in *Escherichia coli* CFT073; 1 MI for *Burkholderia pseudomallei* K96243, which seems to be a miscellaneous island; 1 RI of *Acinetobacter baumannii* AYE, named AbaR1; and, 1 SI of *Mesorhizobium loti* MAFF303099 presenting a mosaic structure. GIPSy is the first life-style-specific genomic island prediction software to perform analyses of PAIs, MIs, RIs and SIs, opening a door for a better understanding of bacterial genome plasticity and the adaptation to new traits.

1. Introduction

Bacteria are highly diverse organisms broadly distributed over the world (Oren, 2004). The adaptability of bacteria to different environments mainly results from their high genomic plasticity upon which selection will finally act allowing evolution to occur (Barbosa et al., 2014; Dobrindt and Hacker, 2001). Genome plasticity is achieved through diverse mechanisms like point mutations, rearrangements (inversions and translocations), deletions and insertion of DNA from other organisms (including plasmids, prophages, transposons and noncanonical classes of mobile genetic elements) (Bellanger et al., 2014; Brüssow et al., 2004; Dobrindt and Hacker, 2001; Schmidt and Hensel, 2004). In this context, genomic islands (GEIs) play an important role. They are large regions in the DNA sequence (~6–200 kb) acquired from other organisms through one of the above mentioned mechanisms. They promote the bacterial evolution by providing blocks of genes involved in

* Corresponding author.
E-mail addresses: siomars@gmail.com, siomar@icbn.uftm.edu.br (S.C. Soares), hakan.geyik@uni-bielefeld.de (H. Geyik), rommelramos@ufpa.br (R.T.J. Ramos), pablogomesdesa@gmail.com (P.H.C.G. de Sá), eudesgvb@gmail.com (E.G.V. Barbosa), jan.baumbach@imada.sdu.dk (J. Baumbach), figueiredoh@yahoo.com (H.C.P. Figueiredo), miyoshi@icb.ufmg.br (A. Miyoshi), tauch@cebitec.uni-bielefeld.de (A. Tauch), asilva@ufpa.br (A. Silva), vascoariston@gmail.com (V. Azevedo).

http://dx.doi.org/10.1016/j.jbiotec.2015.09.008
0168-1656/© 2015 Elsevier B.V. All rights reserved.



Prepare files from selected organisms

Please provide the file for the genome to be analysed

Please provide the file for the reference genome

If you have any doubt, please contact the developer:
Siomar Soares siomars@gmail.com

Search for specific factors

C:\Users\Siomar\GIPSy\Dependencies\blast\bin\blastall.exe

Virulence Factors (Pathogenicity Islands)

Antibiotic Resistance (Resistance Islands)

Metabolism (Metabolic Islands)

Symbiosis (Symbiotic Islands)

```
BLASTP 2.2.26 [Sep-21-2011]


Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Reference for compositional score matrix adjustment: Altschul, Stephen F.,
John C. Wootton, E. Michael Gertz, Richa Agarwala, Aleksandr Morgulis,
Alejandro A. Schaffer, and Yi-Kuo Yu (2005) "Protein database searches
using compositionally adjusted substitution matrices", FEBS J. 272:5101-5109.

Query= c0011 R 7147:8577 Putative transporter yaaJ
       (476 letters)

Database: /home/jsmith/VirulenceDB/BlastDB/virulenceDB.protein.fasta
          55,193 sequences; 10,514,445 total letters

Searching..........................................done
```

Query ID	Subject ID										
c5491	No Hits Found										
c0002	16516|vfid|37622|vsiid|58459|ssid|hypothetical_pr										
c0003	No Hits Found										
c0004	No Hits Found										
c0005	No Hits Found										
c0006	No Hits Found										
c0007	No Hits Found										
c0008	No Hits Found										
c0009	No Hits Found										
c0010	No Hits Found										
c0011	20539|vfid|45666|vsiid|66503|ssid|putative_transcriptional_regulato	30%	432	282	18	12	430	12	438	7e-044	160
c0012	No Hits Found										
c0013	No Hits Found										

Predict Genomic Islands based on previously identified features

Pathogenicity Islands

Resistance Islands

Metabolic Islands

Symbiotic Islands

Putative Genomic Island	G+C Dev...	Codon U...	Virulence...	Hypotheti...	Gene Composition	Position	Predictio...
Genome	10%	24%	44%	41%	NA	NA	NA
Putative Genomic Island 1	66%	100%	33%	66%	c0133-c0139	126688..132406	NA
Putative Pathogenicity Island 1	21%	66%	64%	76%	c0253-c0368	248421..348625	Strong
Putative Pathogenicity Island 2	37%	37%	75%	50%	c0391-c0398	370109..378117	Normal
Putative Pathogenicity Island 3	0%	66%	100%	33%	c0409-c0414	388978..395711	Normal
Putative Pathogenicity Island 4	20%	37%	77%	41%	c0932-c0979	908853..942509	Normal
Putative Pathogenicity Island 5	13%	68%	100%	59%	c1165-c1293	1127423..1241384	Strong
Putative Pathogenicity Island 6	21%	72%	67%	57%	c1400-c1475	1327847..1373052	Strong
Putative Genomic Island 2	14%	59%	14%	74%	c1481-c1507	1377690..1389209	NA
Putative Pathogenicity Island 7	17%	58%	62%	39%	c1515-c1602	1395726..1452493	Normal
Putative Genomic Island 3	53%	100%	15%	100%	c1881-c1893	1715525..1728047	NA
Putative Pathogenicity Island 8	0%	77%	77%	44%	c1935-c1943	1777183..1787317	Normal
Putative Genomic Island 4	33%	100%	16%	0%	c1955-c1960	1796112..1802390	NA
Putative Pathogenicity Island 9	51%	61%	91%	63%	c2392-c2438	2193646..2250680	Strong

Fig. 1. Screenshots of GIPSy's Java-based graphical user interface.
In the top of the figure is step 1, where the GenBank or EMBL files for the query and subject genomes are provided for the software to create the additional files. In the middle of the figure is step 5, where the software searches the query genome for GEI specific factors (*i.e.*, virulence, antibiotic resistance, metabolism-related or symbiosis-related factors). In the bottom of the figure is step 8, where the software predicts the specific class of GEI (pathogenicity, resistance, metabolic or symbiotic island) based on the resulting files from previous steps.

correlated traits, *e.g.,* by the incorporation of integrative conjugative elements carrying antibiotic resistance genes, like the Tn*916* element in *Streptococcus* species (Holden et al., 2009).

Because GEIs are acquired through horizontal gene transfer (HGT), they normally exhibit the genomic signature of the donor organism. The adoption of G + C- or A + T-rich codons will ultimately lead to a similar G + C content of the genes throughout the genome of the donor organism (Hershberg and Petrov (2009). After incorporation of this region by an acceptor organism, this genomic signature of the donor organism will result in a GEI with anomalous G + C content and codon usage deviation in the acceptor genome. Furthermore, GEIs often harbor transposase, tyrosine recombinase and serine recombinase genes and flanking direct repeats that are remnants from the incorporation event. They may also have flanking tRNA genes that harbor a specific recombination site in their 3'end (Hou, 1999). Finally, they may as well exhibit a so-called mosaic structure resulting from several independent insertion events at the same hotspot (Bellanger et al., 2014; Hacker and Carniel, 2001; Schmidt and Hensel, 2004; Soares et al., 2012).

We classify GEIs into four different categories: (i) pathogenicity islands (PAIs), which carry virulence factor genes (Dobrindt et al., 2000); (ii) metabolic islands (MIs), which harbor genes associated to the biosynthesis of (secondary) metabolites (Tumapa et al., 2008); (iii) resistance islands (RIs) with resistance factor genes, *i.e.,* genes encoding antibiotic resistance-related proteins (Krizova and Nemec, 2010); and (iv) symbiotic islands (SIs) providing the bacterium with a genomic repertoire for sustaining a host-bacterium symbiotic relationship (Barcellos et al., 2007).

The term PAI was first introduced by Hacker and colleagues. They identified and experimentally validated large unstable genomic regions in the genome of *Escherichia coli* by using a combination of gene cloning, gel electrophoresis, pulse-field gel electrophoresis and southern blots (Hacker et al., 1990). Without a previous prediction of putative GEIs, such *in vitro* approaches would be time-consuming and expensive. Nowadays, many software tools for identifying putative GEIs exist. They typically search for commonly shared features in a set of sequenced bacterial genomes and concentrate on pathogenicity islands (Soares et al., 2012). There is still a lack, however, of tools for finding other classes of GEIs (*i.e.,* MIs, RIs and SIs). See (Soares et al., 2012) for a comparison of available software packages and their advantages and limitations.

Here, we present GIPSy, the genomic island prediction software, which aims for providing the community with a standalone tool and user-friendly interface to accurately predict all four classes of GEIs. We built on our previously developed pathogenicity island prediction software (PIPS) (Soares et al., 2012) and extended it to also detect the remaining three genomic island types: MIs, RIs and SIs. The emerging GIPSy software is the first computational tool for comprehensive detection of life-style-specific genomic islands of four different kinds, rather than concentrating on pathogenicity and virulence only. It is programmed in Java and, thus, platform-independent. It is publicly available online at the project web site: http://www.bioinformatics.org/groups/?group_id=1180http://www.bioinformatics.org/groups/?group_id=1180.

2. Materials and methods

GIPSy itself is developed in Java. Database and software dependencies are automatically downloaded and properly set up by the GIPSy installer, which is also developed in Java. Fig. 1 shows the graphical user interface. It features a set of steps to be executed guided by the GIPSy front-end in a tutorial/wizard style. In the following, we explain and justify each of the island detection steps that are reflected by the step-by-step data analysis pipeline of GIPSy.

2.1. Pipeline overview

GEI predictions are based on commonly shared features: genomic signature deviation (G + C content and codon usage); presence of transposase genes; factors for virulence, metabolism, antibiotic resistance, or symbiosis; flanking tRNA genes; and absence in other organisms of the same genus or closely related species. Eight steps, schematically depicted in the flowchart in Fig. 2, are necessary to evaluate the presence of these genomic features.

2.2. Step 1: data preprocessing

GIPSy accepts genome files in embl (".embl") and genbank formats (".genbank", ".gbk" or ".gb") as input. The whole genome sequence and the nucleotide and amino acid sequences of all coding sequences (CDS) are retrieved from the input files and saved in ".fna", ".ffn" and ".faa" files, respectively. Whenever a genbank file is provided by the user, it is converted into an embl file, the input of Colombo/SIGIHMM (step 3). In addition, GIPSy creates a list of all genes ("locus_tag") with information on CDS location and CDS product.

2.3. Step 2: G + C content deviation

The G + C content deviation is analyzed in accordance with the PAI detection approach of PIPS (Soares et al., 2012). In brief, the G + C content of the whole genome retrieved from the ".fna" file is calculated as the rate: *genome G + C/genome length*. Next, the G + C content of each CDS retrieved from the ".ffn" file is calculated as the rate: *CDS G + C/CDS length*. Finally, the mean value is considered as being the G + C content of the whole genome, and the standard deviation (SD) value is calculated using the total number of CDSs in the genome as dataset. Following previous evaluations (Jain et al., 2008; Soares et al., 2012), CDSs with a G + C content outside of the range mean ± 1.5 SD are assigned as anomalous regions, as this shows the best correlation between sensitivity and specificity. Hence, we offer "1.5" as standard parameter to the user here.

2.4. Step 3: codon usage deviation

Colombo/SIGIHMM is used to calculate the codon usage deviation by using Hidden Markov Models (Waack et al., 2006) and the embl file as input. In GIPSy, the set of CDSs predicted by Colombo/SIGIHMM is not taken as the sole analyzed feature for GEI prediction, but rather as one of a cascade of features that, altogether, reveal the horizontally acquired regions. In view of this, GIPSy offers the highest possible sensitivity parameter of Colombo/SIGIHMM (0.95) as standard value.

2.5. Step 4: transposase genes

In GIPSy, the prediction of transposase genes is performed using the software HMMER3 (Eddy, 2011), which searches in the ".faa" file of the query genome for hidden patterns using a transposase database. This database contains transposase, tyrosine recombinase and serine recombinase genes from bacteria, retrieved from PFAM (Finn et al., 2010). The *E*-value used in HMMER3 analyses may be configured by the user in GIPSy and the standard value was set to 1E-04.

2.6. Step 5: class specific factors

In order to correctly predict the GEIs and to classify them into their specific classes, we used five protein databases with specific factors for each class. For these five databases, protein similarity

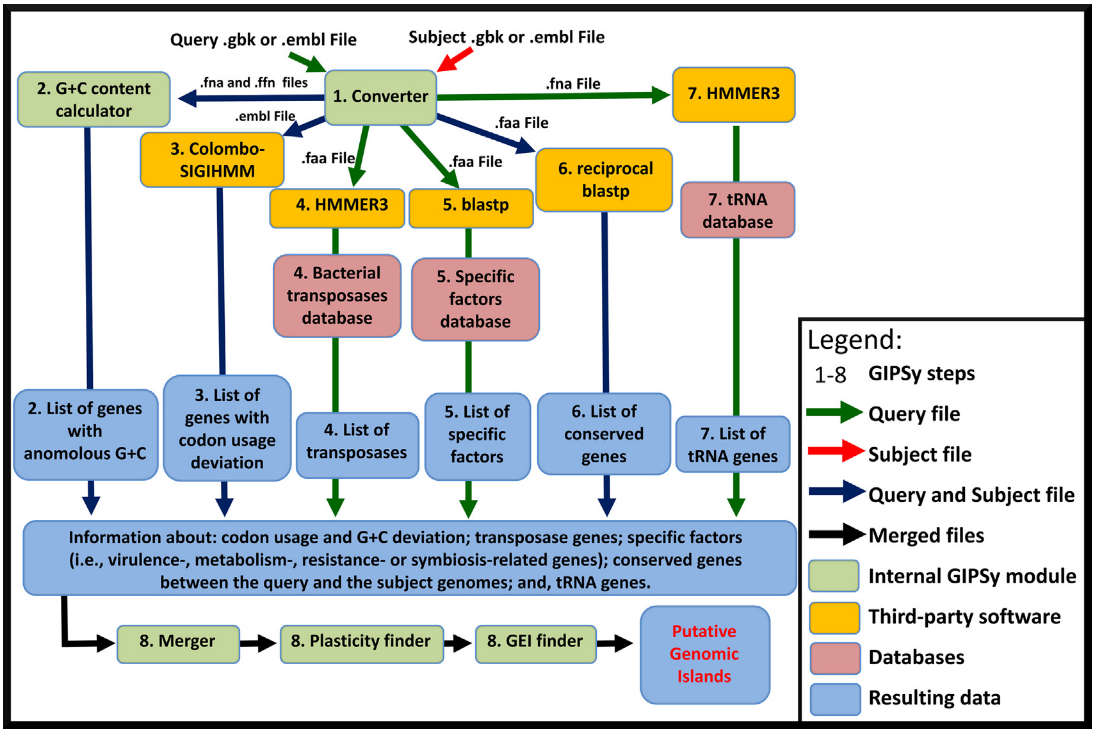

Query .gbk or .embl File
Subject .gbk or .embl File
1. Converter
.fna and .ffn files
.embl File
.fna File
.faa File
.faa File
.faa File
2. G+C content calculator
3. Colombo-SIGIHMM
4. HMMER3
5. blastp
6. reciprocal blastp
7. HMMER3
7. tRNA database
4. Bacterial transposases database
5. Specific factors database
2. List of genes with anomolous G+C
3. List of genes with codon usage deviation
4. List of transposases
5. List of specific factors
6. List of conserved genes
7. List of tRNA genes
Information about: codon usage and G+C deviation; transposase genes; specific factors (i.e., virulence-, metabolism-, resistance- or symbiosis-related genes); conserved genes between the query and the subject genomes; and, tRNA genes.
8. Merger
8. Plasticity finder
8. GEI finder
Putative Genomic Islands
Legend:
1-8 GIPSy steps
Query file
Subject file
Query and Subject file
Merged files
Internal GIPSy module
Third-party software
Databases
Resulting data

Fig. 2. Flowchart showing all the steps performed by GIPSy. (For interpretation of the reference to colour in this figure legend, the reader is referred to the web version of this article.)
Green arrows represent the query genome. Red arrow represents the subject genome. Blue arrows represent both the query and subject genomes. Numbers from 1 to 8 represent the eight steps performed by GIPSy during GEI analysis.

searches are performed using BLAST on the ".faa" file from the query genome, offering a standard *E*-value cutoff parameter of 1E-06 to the user.

Virulence genes are identified by performing protein similarity searches with blastp algorithm in the virulence factors database mVIRdb. This database harbors protein sequences from ten different databases: Tox-Prot, SCORPION, the PRINTS virulence factors, VFDB, TVFac, Islander, ARGO, CONUS and a subset of VIDA (Zhou et al., 2007).

To predict antibiotic resistance factors, we merged two databases: ARDB, antibiotic resistance genes database (Liu and Pop, 2009); and CARD, the Comprehensive Antibiotic Resistance Database (McArthur et al., 2013).

The database of metabolism-related factors was created by retrieving all genes under the metabolism category of the cluster of orthologous groups of proteins database. The metabolism category of the COG database harbors genes related to energy production and conversion and also genes for metabolism and transport of carbohydrates, amino acids, nucleotides, co-enzymes, lipids, inorganic ions and secondary metabolites.

To perform symbiosis island analyses, we used the NodMutDB database, which harbors 2,834 symbiosis-related genes collected through literature review and public database searches (Mao et al., 2005). The symbiosis-related genes in NodMutDB are mainly composed of nitrogen-fixation and nodulation-related genes.

2.7. *Step 6: absence in organism of the same genus or related species*

Due to environmental pressures and codon adaptation, very closely related bacterial species may drastically differ in nucleotide content, although still sharing high similarity levels at amino acid level. In view of this, the prediction of commonly shared orthologous genes through reciprocal protein similarity searches is much more reliable for inferring gene synteny and HGT events than simply performing nucleotide alignments of the whole genome sequence. In step 6, GIPSy performs reciprocal BLASTs between CDSs from the query and subject genomes to predict regions of similarity and putative genes that are present in the query genome and absent from the subject genome. The additional clustering of genes and the identification of putative horizontally acquired genes is performed in step 8 and is further explained in the appropriate section.

2.8. *Step 7: tRNA genes*

In our previously released software PIPS, we used tRNASCAN-SE to predict tRNA genes in the query genome (Lowe and Eddy, 1997). However, tRNASCAN-SE was specifically developed for Linux-based systems and does not have a Windows-based counterpart. In order to circumvent this problem, during GIPSy implementation, we used HMMER3 searches against a database of bacterial tRNA genes retrieved from tRNAdb (Jühling et al., 2009) to identify tRNA sequences in the genomes. Although the HMMER3 searches do not provide the putative codon/anticodon sequences of tRNAs, it can be performed with the same efficiency as tRNASCAN-SE in tRNA genes identification and is also able to predict disrupted tRNA genes, which are very important for GEI analysis as they may be indicative of HGT events. Moreover, to avoid time-consuming processing of whole genome sequences during reverse/complementing the genome to find tRNA genes on the reverse strand, we created the reverse/complement alignments of the tRNA gene sequences in order to have both possibilities for HMM profiles, forward and reverse. Finally, the searches are performed in the ".fna" file from the query genome with a standard *E*-value of 1E-04, which may be adjusted by the user.

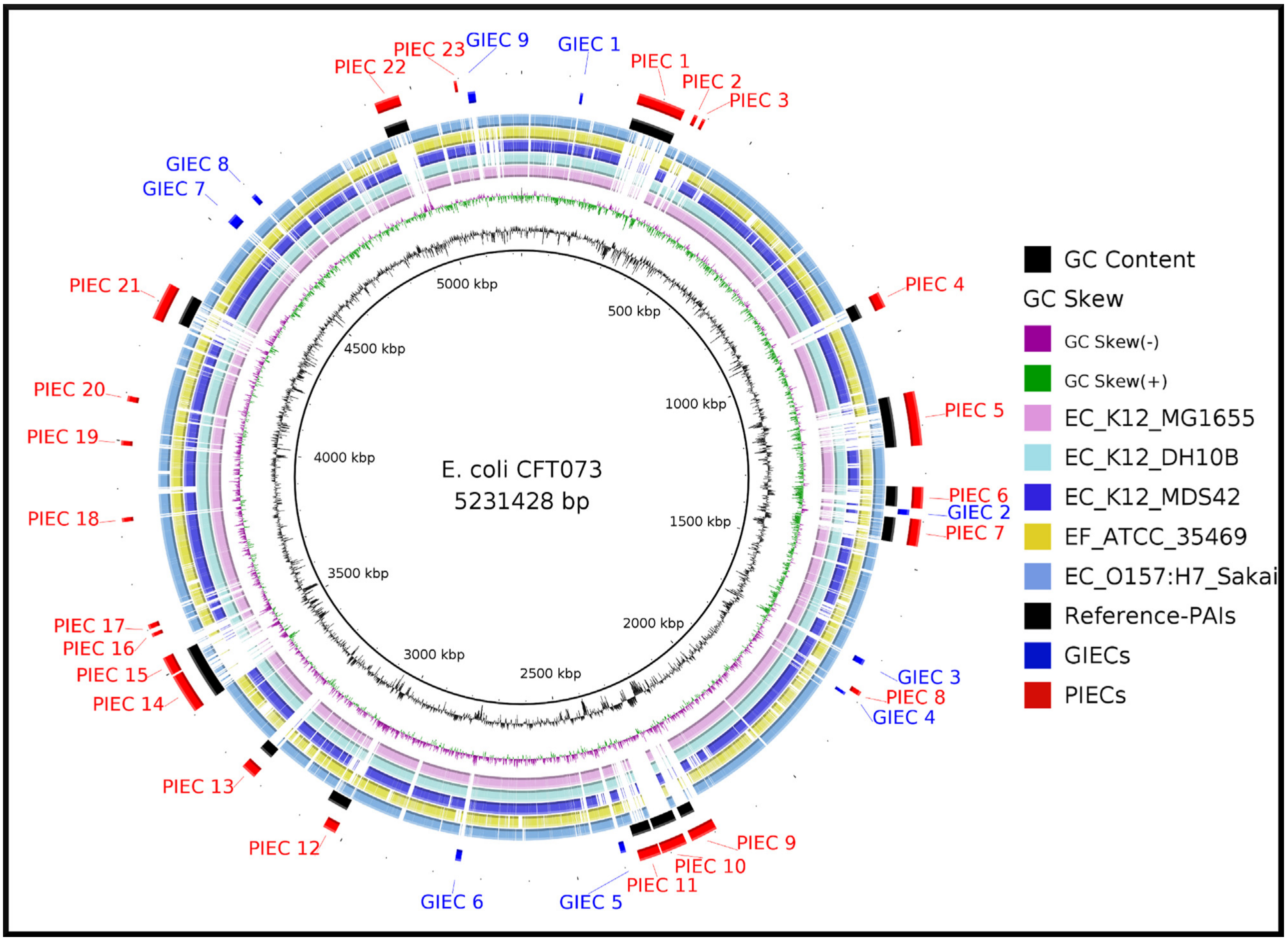


Fig. 3. Circular genome comparison showing PAIs predicted for *E. coli* CFT073 and regions of genome plasticity.
The figure was plotted using *E. coli* CFT073 as reference. EC, *E. coli*; EF, *E. fergusonii*; reference-PAIs, PAIs previously described; PIEC, putative pathogenicity island predicted by GIPSy in *E. coli* CFT073; GIEC, putative genomic island predicted by GIPSy in *E. coli* CFT073. The analyses were performed using *E. coli* K12 substr. MG1655 as non-pathogenic closely-related strain. The figure was generated using the software BRIG: blast ring image generator.

2.9. Step 8: merging and clustering of information and prediction of genomic islands

Once all previous steps have been performed, all the information is merged in a tab-delimited file using the locus_tag as an unique key (Merger—Fig. 2). The information on CDS similarity from the reciprocal BLAST is then used to cluster neighboring genes that are commonly shared by both organisms in blocks according to their relative position in both genomes. Then, regions longer than 6 kb that are only present in the query organism are separated as candidate regions for GEI analyses (Plasticity finder—Fig. 2). Putative GEIs are predicted by the presence of a combination of features higher in number than those in the whole genome sequence (GEI finder—Fig. 2).

For an easier interpretation of the data, GEIs are assigned a prediction score as “Weak”, “Normal” or “Strong” according to the following rules: (I) strong, for regions presenting very high concentrations of the chosen specific factor and all other features when compared to the whole genome sequence; (II) strong, also for regions presenting low to normal concentration of one genomic signature feature, high concentration of the other and very high concentration of the chosen specific factor; (III) normal, for regions presenting a low concentration of one genomic signature feature and high concentration of the other, with high concentration of the chosen specific factor; (IV) normal, also for regions with normal to high concentration of the chosen specific factor and, also, high concentrations of hypothetical proteins and genomic signature deviation; and, (V) weak otherwise. Additionally, the presence of transposase genes may change the prediction score from “Weak” to “Normal” and from “Normal” to “Strong”. Finally, regions presenting low concentration of specific factors and high concentrations of any genomic signature deviation are denoted as putative genomic islands.

2.10. Dataset and circular genome comparisons

The dataset for GEI predictions was composed of strains from *E. coli*, *Burkholderia pseudomallei*, *Acinetobacter baumannii* and *Mesorhizobium loti* for PAIs, MIs, RIs, and SIs, respectively. Briefly, *E. coli* CFT073 was chosen for PAI analysis because of its large number of PAIs already described in literature and also for the previous *in vitro* identification of genomic plasticity in 13 PAIs longer than 30 kb. The genome sequence of *E. coli* K-12 substr. MG1655 was the reference organism used in PAI analyses (Lloyd et al., 2007).

For MI predictions, we selected *B. pseudomallei* strains K96243 and 688 as query and subject genomes, respectively. Briefly, 16 GIs have already been identified in the genome of *B. pseudomallei* K96243, which vary from phage sequence and miscellaneous islands to one large MI. Moreover, these GEIs are highly variable in content when compared to several other strains of *B. pseudomallei* as shown by multiplex PCR assay (Holden et al., 2004). *B. pseudomallei* K96243 has seven putative GEIs (PGEIs) yet to be ana-

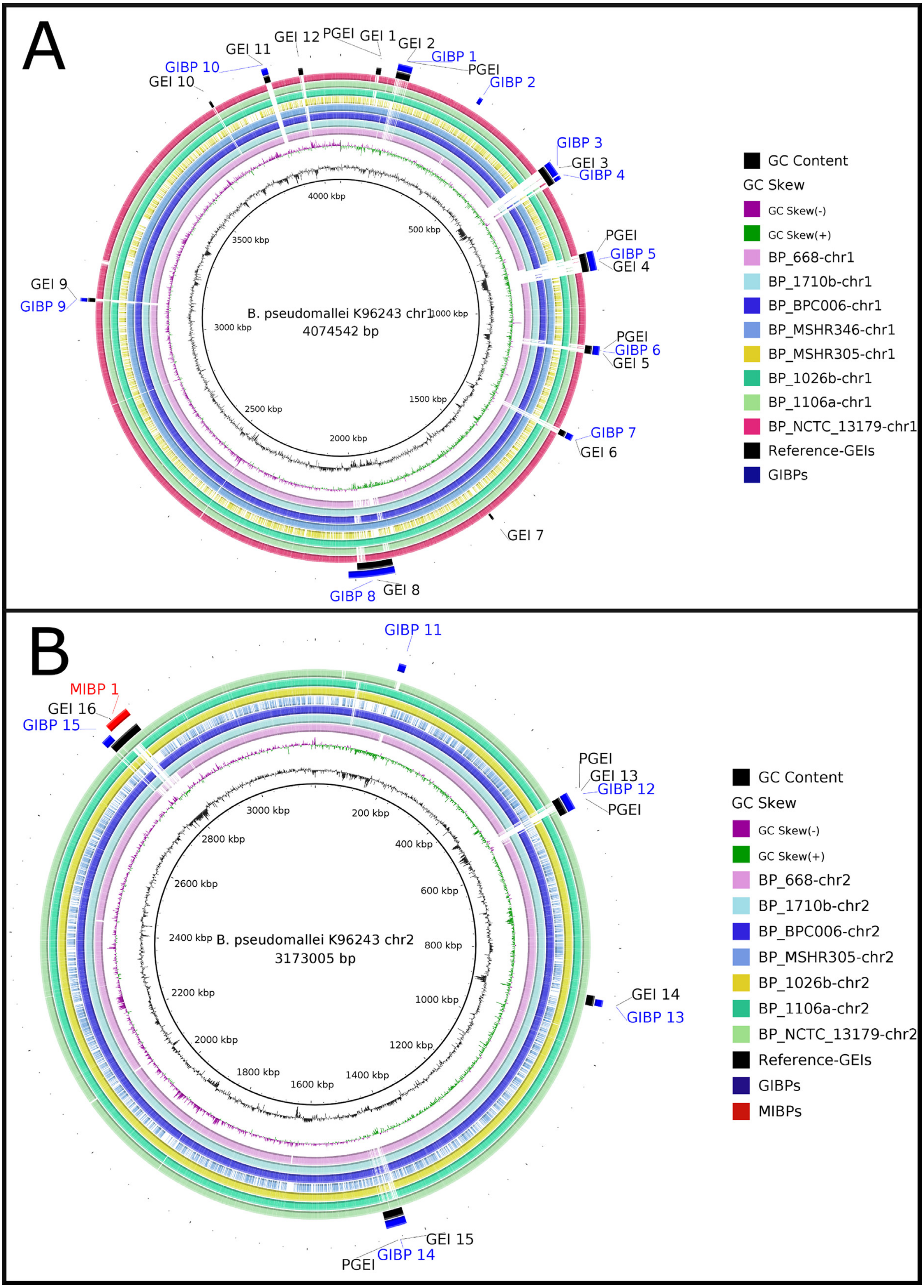


Fig. 4. Circular genome comparison showing MIs predicted for *B. pseudomallei* K96243 chromosomes 1 and 2 and regions of genome plasticity. (A) Figure plotted using *B. pseudomallei* K96243 chromosome 1 as a reference. (B) Figure plotted using *B. pseudomallei* K96243 chromosome 2 as a reference. BP, *B. pseudomallei*; reference-GEIs, GEIs previously described in literature; PGEI, regions previously described in literature as putative GEIs; MIBP, putative MIs predicted by GIPSy in *B. pseudomallei* K96243 chromosomes 1 and 2; GIBP, putative GEIs predicted by GIPSy in *B. pseudomallei* K96243 chromosomes 1 and 2. The analyses were performed using *B. pseudomallei* 668 chromosomes 1 and 2, respectively, as closely-related strain. The figure was generated using the software BRIG: blast ring image generator.

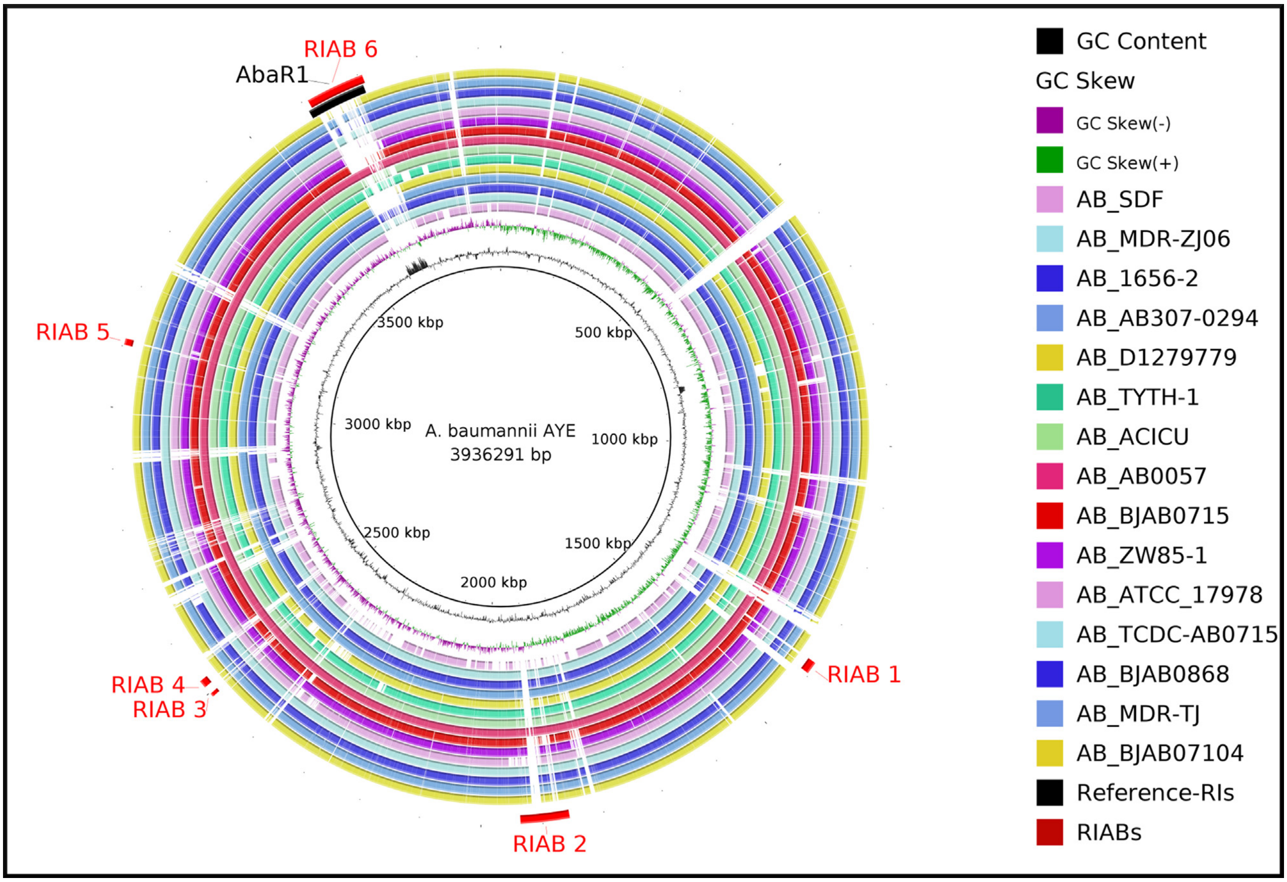


Fig. 5. Circular genome comparison showing RIs predicted for *A. baumannii* AYE and regions of genome plasticity.
The figure was plotted using *A. baumannii* AYE as a reference. AB, *A. baumannii*; reference-RIs, RIs previously published; RIAB, putative RIs predicted by GIPSy in *A. baumannii* AYE. The analyses were performed using *A. baumannii* SDF as closely-related strain. The figure was generated using the software BRIG: blast ring image generator.

lysed. Finally, *B. pseudomallei* 668 was used as a subject genome due to the genomic plasticity that exists between this strain and *B. pseudomallei* K96243 (Tumapa et al., 2008).

For RIs predictions, we used *A. baumannii* strains AYE and SDF as query and subject genome, respectively. *A. baumannii* AYE is an invasive multidrug resistant strain, which carries a large antibiotic resistance island of ~86 kb named AbaR1, whereas *A. baumannii* SDF is a non-invasive and susceptible strain (Sahl et al., 2011). Finally, for SI prediction, we used the genome sequences of *M. loti* MAFF303099 and *Mesorhizobium sp.* BNC1 as query and subject genomes, respectively. *M. loti* MAFF303099 carries a large SI of ~611 kb named SI MAFF309999 (Kasai-Maita et al., 2013) spanning from 4,644,702 bp to 5,255,766 bp (Uchiumi et al., 2004), whereas *Mesorhizobium* sp. BNC1 is a non-symbiotic species isolated from sewage (Black et al., 2012). For the locations of the reference GEIs for each of the above mentioned organisms, please refer to Supplementary file 1.

Finally, all circular genome comparisons were visualized using the software BRIG: blast ring image generator (Alikhan et al., 2011).

3. Results and discussion

After implementing GIPSy, we applied the software to specifically predict PAIs, MIs, RIs, or SIs in *Escherichia coli* CFT073, *Burkholderia pseudomallei* K96243, *Acinetobacter baumannii* AYE and *Mesorhizobium loti* MAFF303099. In order to verify the results generated by GIPSy, we have only chosen species for which the coordinates of the underlying class of GEI were already published. The comparisons of the results generated by GIPSy for each class of GEI with the data retrieved from literature are summarized in the next sections.

3.1. Application case 1: GIPSy prediction of pathogenicity islands from E. coli CFT073

We have predicted PAIs for *E. coli* CFT073 using *E. coli* K12 substr. MG1655 as non-pathogenic closely-related strain. In *E. coli* CFT073, GIPSy has predicted 23 putative PAIs and 9 putative GEIs of *E. coli* (PIECs and GIECs, respectively), which are distributed through the genome sequence (Fig. 3 and Supplementary file 1). Out of these 23 PIECs, 14 are located in regions previously defined as PAIs by (Lloyd et al., 2007) with only small differences in sizes (PIECs 9, 12 and 21) and a disruption in the middle of PIECs 14 and 15. The 9 additional PAIs predicted by GIPSy in *E. coli* present lengths ranging from 6 kb (PIEC 23) to 11 kb (PIEC 20). Although comparative genomic hybridization analysis and combined genomics, bioinformatics, and microarray-based assays performed previously (Lloyd et al., 2007) have successfully predicted PAIs in *E. coli*, they have only focused on the prediction of regions longer than 30 kb, which prevented the identification of the additional PIECs predicted by GIPSy. Moreover, most of the 9 additional PIECs present recombinations (*i.e.*, insertions in the query or deletions in the subject genomes) in at least one strain of *E. coli*, which are highlighted by the white regions in the BRIG circular genomic comparison. Once GEIs may harbor a remnant functional transposase, serine recombinase or tyrosine recombinase and target insertion sequences, they are very unstable and may delete or rearrange at different rates. In previous predictions of PAIs of *Corynebacterium diphtheriae* NCTC13129 using the

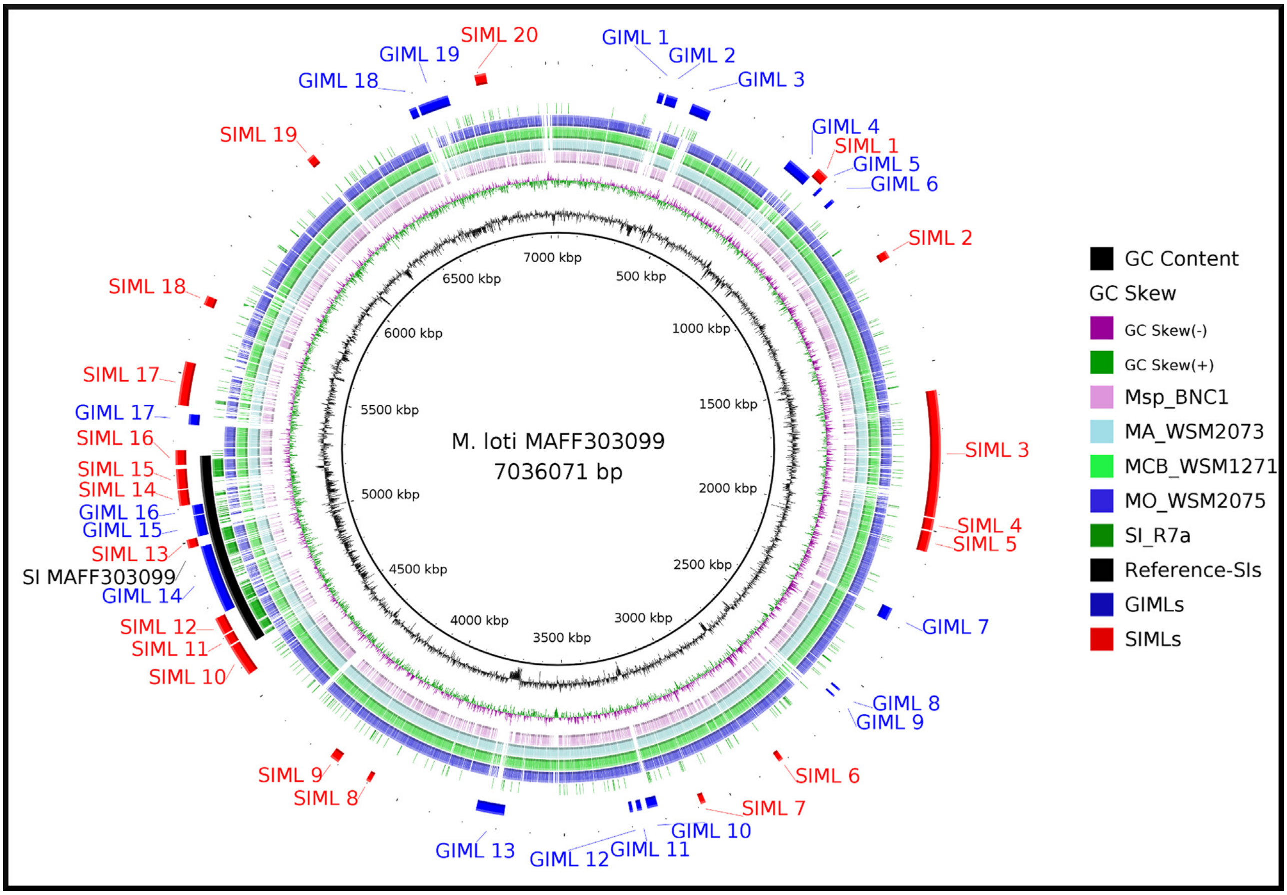


Fig. 6. Circular genome comparison showing SIs predicted for *M. loti* MAFF303099 and regions of genome plasticity.
The figure was plotted using *M. loti* MAFF303099 as reference. Msp, *Mesorhizobium* sp.; Ma, *M. australicum*; Mcb, *M. ciceri* biovar *bisserrulae*; Mo, *M. opportunistum*; SI_R7a, SI from *M. loti* R7a; SI MAFF303099, SI from *M. loti* MAFF303099; reference-SIs, SIs previously published; SIML, putative SI of *M. loti* MAFF303099 predicted by GIPSy; GIML, putative GEIs of *M. loti* MAFF303099 predicted by GIPSy. The analyses were performed using *Mesorhizobium* BNC1 as closely-related species. The figure was generated using the software BRIG: blast ring image generator.

software PIPS (Soares et al., 2012), we have also found 11 additional PAIs not previously cited in literature, where 10 of them have been shown in posterior pan-genomics analyses to have deletions, transpositions and duplications across the 13 different strains of *C. diphtheriae* analyzed (Trost et al., 2012). Taking together, the lack of reference PAIs smaller than 30 kb in *E. coli* CFT073 and the high number of recombinations in the additional PAIs both corroborate for the correct prediction performed by GIPSy. Likewise the 9 additional PIECs, the 9 GIECs predicted by GIPSy were also shorter than 30 kb, ranging in size from 6 kb (GIEC 1) to 23 kb (GIEC 7), and they all presented recombinations in other closely-related genomes.

3.2. *Application case 2: GIPSy prediction of metabolic islands from* B. pseudomallei *K96243 chromosomes 1 and 2*

In order to establish how well GIPSy performs in predicting MIs, we used the genomes of *B. pseudomallei* K96243 (chromosomes 1 and 2) with *B. pseudomallei* 668 (chromosomes 1 and 2) as a reference. In a previous work (Holden et al., 2004), it was presented that *B. pseudomallei* harbors 16 GEIs (GEI 1-16) and 7 other putative GEIs distributed across the two chromosomes (Fig. 4 and Supplementary file 1). Additionally, GEI 16 was reported as a putative metabolic island. In the present study, GIPSy has predicted 1 MI (MIBP 1) and a total of 15 GEIs (GIBPs 1-15) in chromosomes 1 and 2 of *B. pseudomallei* K96243.

Out of these 16 islands, 12 GIBPs are located in the genomic locations where GEIs and additional PGEIs were previously described (Holden et al., 2004). The location of these GIBPs in highly variable regions, *i.e.,* presenting different patterns of recombinations in one or more strains, is indicative of their correct prediction as horizontally acquired regions. Interestingly, the MI known from literature was both predicted as MIBP 1 and GIBP 15, which points this region as also being a putative miscellaneous island, like previously described for GEIs 1, 4, 8 and 14 (Holden et al., 2004). Finally, 6 of the 7 PGEIs and 11 of the 16 GEIs previously described in literature were predicted as GIBPs.

3.3. *Application case 3: GIPSy prediction of resistance islands from* A. baumannii *AYE*

To evaluate the performance of GIPSy in predicting RIs, we have used the genome sequence of *A. baumannii* AYE and compared it to the reference species *A. baumannii* SDF. Additionally, for comparison of results, we plotted the approximate genomic coordinates of the resistance island AbaR1 in *A. baumannii* AYE, as previously showed (Sahl et al., 2011). In this analysis, GIPSy has predicted 6 putative RIs of *A. baumannii* AYE (RIABs), which are depicted in Fig. 5 and described in Supplementary file 1. Interestingly, RIAB 6 was predicted to be located in the same region of AbaR1. Although several additional RIABs, others than AbaR1, have been predicted for *A. baumannii* AYE, all of them showed recombinations in other strains of *A. baumannii*, ranging from small ones in RIABs 1, 3, 4, and 5 to larger ones in RIAB 2. This could be the result of recent acquirement of the region by *A. baumannii* AYE or even the excision of a part of the region in other strains and also corroborate the correct prediction of these regions as putative RIs.

3.4. Application case 4: GIPSy prediction of symbiosis islands from M. loti *MAFF303099*

We used *M. loti* MAFF303099 as the query genome for the prediction of SIs and *Mesorhizobium sp.* BNC1 as the reference non-symbiotic closely-related species. Additionally, we plotted the genomic coordinates of the previously described 611-kb SI of *M. loti* MAFF303099 (Uchiumi et al., 2004) and also compared the genome sequence of the SI R7A with the reference strain *M. loti* MAFF303099. From these analyses, we predicted 20 putative SIs and 19 putative GEIs of *M. loti* MAFF303099 (SIMLs and GIMLs, respectively), where the SI MAFF303099 is represented by the SIMLs 11 to 16 and GIMLs 14 to 16 (Fig. 6 and Supplementary file 1).

In literature, it was reported that *Mesorhizobium* species may shift their lifestyle from non-symbiotic to symbiotic by a single HGT event, the incorporation of a 500-kb symbiosis island from *M. loti* R7A (Sullivan et al., 1995). However, from an evolutionary point of view, it is more likely that *Mesorhizobium* has shaped its genome in a cascade of HGT events, where the acquirement of the previously described SIs of ~500-kb and ~611-kb in *M. loti* strains R7A and MAFF303099, respectively, were only the triggering events for the symbiotic relationships of those strains with leguminous plants.

3.5. Limitations of GIPSy

In our previous work, we demonstrated that the comparison of a query genome with a closely related organism chosen by the user improves the identification of GEIs (Soares et al., 2012), and we therefore implemented this feature in GIPSy. For a better prediction, bacteria that best fit as subject organisms are strains or closely related species that do not share the searched trait with the query species. For instance, for PAI analysis it is highly recommended to use a non-pathogenic closely-related subject organism, whereas for SI predictions, the subject organism shall be a non-symbiotic closely related organism. However, as a drawback, the lack of a genome from a closely related organism may impair the GEIs analyses. For instance, PAI identification may still be performed using a pathogenic closely-related organism in case no genome sequence of a non-pathogenic close relative is available. However, PAIs that are commonly shared by both organisms will be missed by the software, generating false negative results. In order to solve this problem, one may predict PAIs for the same organism using diverse reference genomes, consolidate the results and compare the putative plasticity of the given region with that of other genomes from organisms of the same species or genus, where the absence of the region is indicative of its putative horizontal acquirement.

Additionally, many genes may have dual purposes in different contexts. For instance, adhesins are widely known as virulence factors in pathogenic organisms once they help the bacteria in having contact with the host cells and also mediate internalization (Wilson et al., 2002). However, they also play important roles in probiotic bacteria where they help the organism, for example, in attaching to intestine cells to create a physical barrier against pathogenic bacteria and also in exerting the probiotic effect (Johnson and Klaenhammer, 2014; Von Ossowski et al., 2011). The complete set of proteins, expressed in a coordinated fashion and interacting with other species-specific proteins, will ultimately determine whether the bacteria will be pathogenic or not. In view of this, the same GEI predicted as PAI in a pathogenic organism may also be present as a GEI in a non-pathogenic organism and will also be missed by the software. In summary, as in genome annotation, the manual curation of the predicted GEIs in the light of the biological aspects of the studied organism plays a pivotal role in the correct identification of the searched class of GEI.

4. Conclusions

We created a software that implements the previously described methodology of PIPS in Java for easy-of-use, and we updated its architecture to harbor additional databases for the identification of different classes of GEIs, like MIs, SIs, and RIs. Here, we used data from literature to demonstrate the performance of GIPSy in predicting PAIs, MIs, RIs and SIs, and in verifying the correct prediction of: 13 PAIs larger than 30 kb in *E. coli* CFT073; 1 MI previously described for *B. pseudomallei* K96243, which seems to be a miscellaneous island; 1 RI of *A. baumannii* AYE; and, 1 SI of *M. loti* MAFF303099, which presents a mosaic structure with symbiotic-related and -unrelated regions.

Although some regions of the MI of *B. pseudomallei* K96243 and SI of *M. loti* MAFF303099 were predicted as putative GEIs, they may be a product of the mosaic structure of GEIs created due to the additional incorporation of other genomic regions in remnant mobile elements or, also, due to the lack of information related to genes located at this specific region. However, the presence of these putative GEIs in highly variable regions supports their horizontal acquirement and their classification as GEIs. Moreover, most of the putative GEIs identified during the analyses of all four classes (PAIs, MIs, RIs and SIs) also presented regions of recombinations in other strains, corroborating the efficacy of GIPSy.

Finally, although GEIs are highly studied, most of the previous works were related to the identification of PAIs, resulting in a lack of information for MIs, RIs and SIs from *in vitro* and *in silico* analyses, which may be used for deeper validation of GIPSy and also for the elucidation of additional gene content harbored by those other classes that could be used in future research. In this scenario, GIPSy arises as the first GEI prediction software to specifically target all classes of GEIs, opening a door for a better understanding of bacterial genome plasticity and the adaptation to new traits.

Competing interests

'The author(s) declare that they have no competing interests'.

Acknowledgements

The work of HG at UFMG (Belo Horizonte) was supported by a mobility grant from Bielefeld University (Germany). The authors thank the CNPq for the post-doc scholarship of SCS (grant number 151262/2013-6).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbiotec.2015.09.008.

References

Alikhan, N., Petty, N.K., Ben Zakour, N.L., Beatson, S.A., 2011. Blast ring image generator (brig): simple prokaryote genome comparisons. BMC Genom. 12, 402.

Barbosa, E., Röttger, R., Hauschild, A., Azevedo, V., Baumbach, J., 2014. On the limits of computational functional genomics for bacterial lifestyle prediction. Brief Funct Genom. 13, 398–408.

Barcellos, F.G., Menna, P., da Silva Batista, J.S., Hungria, M., 2007. Evidence of horizontal transfer of symbiotic genes from a *Bradyrhizobium japonicum* inoculant strain to indigenous *Diazotrophs sinorhizobium* (*ensifer*) *fredii* and *Bradyrhizobium elkanii* in a brazilian savannah soil. Appl. Environ. Microbiol. 73, 2635–2643.

Bellanger, X., Payot, S., Leblond-Bourget, N., Guédon, G., 2014. Conjugative and mobilizable genomic islands in bacteria: evolution and diversity. FEMS Microbiol. Rev. 38, 720–760.

Black, M., Moolhuijzen, P., Chapman, B., Barrero, R., Howieson, J., Hungria, M., Bellgard, M., 2012. The genetics of symbiotic nitrogen fixation: comparative genomics of 14 *Rhizobia* strains by resolution of protein clusters. Genes (Basel) 3, 138–166.

Brüssow, H., Canchaya, C., Hardt, W., 2004. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiol. Mol. Biol. Rev. 68, 560–602.

Dobrindt, U., Janke, B., Piechaczek, K., Nagy, G., Ziebuhr, W., Fischer, G., Schierhorn, A., Hecker, M., Blum-Oehler, G., Hacker, J., 2000. Toxin genes on pathogenicity islands: impact for microbial evolution. Int. J. Med. Microbiol. 290, 307–311.

Dobrindt, U., Hacker, J., 2001. Whole genome plasticity in pathogenic bacteria. Curr. Opin. Microbiol. 4, 550–557.

Eddy, S.R., 2011. Accelerated profile HMM searches. PLoS Comput. Biol. 7, e1002195.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L.L., Eddy, S.R., Bateman, A., 2010. The pfam protein families database. Nucleic Acids Res. 38, D211–D222.

Hacker, J., Bender, L., Ott, M., Wingender, J., Lund, B., Marre, R., Goebel, W., 1990. Deletions of chromosomal regions coding for fimbriae and hemolysins occur *in vitro* and *in vivo* in various extraintestinal *E. coli* isolates. Microb. Pathog. 8, 213–225.

Hacker, J., Carniel, E., 2001. Ecological fitness, genomic islands and bacterial pathogenicity. A darwinian view of the evolution of microbes. EMBO Rep. 2, 376–381.

Hershberg, R., Petrov, D.A., 2009. General rules for optimal codon choice. PLoS Genet 5, e1000556.

Holden, M.T.G., Titball, R.W., Peacock, S.J., Cerdeño-Tárraga, A.M., Atkins, T., Crossman, L.C., Pitt, T., Churcher, C., Mungall, K., Bentley, S.D., Sebaihia, M., Thomson, N.R., Bason, N., Beacham, I.R., Brooks, K., Brown, K.A., Brown, N.F., Challis, G.L., Cherevach, I., Chillingworth, T., Cronin, A., Crossett, B., Davis, P., DeShazer, D., Feltwell, T., Fraser, A., Hance, Z., Hauser, H., Holroyd, S., Jagels, K., Keith, K.E., Maddison, M., Moule, S., Price, C., Quail, M.A., Rabbinowitsch, E., Rutherford, K., Sanders, M., Simmonds, M., Songsivilai, S., Stevens, K., Tumapa, S., Vesaratchavest, M., Whitehead, S., Yeats, C., Barrell, B.G., Oyston, P.C.F., Parkhill, J., 2004. Genomic plasticity of the causative agent of melioidosis, *Burkholderia pseudomallei*. Proc. Natl. Acad. Sci. U. S. A. 101, 14240–14245.

Holden, M.T.G., Hauser, H., Sanders, M., Ngo, T.H., Cherevach, I., Cronin, A., Goodhead, I., Mungall, K., Quail, M.A., Price, C., Rabbinowitsch, E., Sharp, S., Croucher, N.J., Chieu, T.B., Mai, N.T.H., Diep, T.S., Chinh, N.T., Kehoe, M., Leigh, J.A., Ward, P.N., Dowson, C.G., Whatmore, A.M., Chanter, N., Iversen, P., Gottschalk, M., Slater, J.D., Smith, H.E., Spratt, B.G., Xu, J., Ye, C., Bentley, S., Barrell, B.G., Schultsz, C., Maskell, D.J., Parkhill, J., 2009. Rapid evolution of virulence and drug resistance in the emerging zoonotic pathogen *Streptococcus suis*. PLoS One 4, e6072.

Hou, Y.M., 1999. Transfer rnas and pathogenicity islands. Trends Biochem. Sci. 24, 295–298.

Jain, R., Ramineni, S., Parekh, N., 2008. Integrated genomic island prediction tool (IGIPT). Proceedings of the 2008 International Conference on Information Technology, 131–132.

Johnson, B.R., Klaenhammer, T.R., 2014. Impact of genomics on the field of probiotic research: historical perspectives to modern paradigms. Antonie Van Leeuwenhoek 106, 141–156.

Jühling, F., Mörl, M., Hartmann, R.K., Sprinzl, M., Stadler, P.F., Pütz, J., 2009. TRNAdb 2009: compilation of tRNA sequences and tRNA genes. Nucleic Acids Res. 37, D159–D162.

Kasai-Maita, H., Hirakawa, H., Nakamura, Y., Kaneko, T., Miki, K., Maruya, J., Okazaki, S., Tabata, S., Saeki, K., Sato, S., 2013. Commonalities and differences among symbiosis islands of three *Mesorhizobium loti* strains. Microbes. Environ. 28, 275–278.

Krizova, L., Nemec, A., 2010. A 63 kb genomic resistance island found in a multidrug-resistant *Acinetobacter baumannii* isolate of european clone I from 1977. J. Antimicrob. Chemother. 65, 1915–1918.

Liu, B., Pop, M., 2009. Ardb—antibiotic resistance genes database. Nucleic Acids Res. 37, D443–7.

Lloyd, A.L., Rasko, D.A., Mobley, H.L.T., 2007. Defining genomic islands and uropathogen-specific genes in uropathogenic *E. coli*. J. Bacteriol. 189, 3532–3546.

Lowe, T.M., Eddy, S.R., 1997. TRNAscan-se: a program for improved detection of transfer rna genes in genomic sequence. Nucleic Acids Res. 25, 955–964.

Mao, C., Qiu, J., Wang, C., Charles, T.C., Sobral, B.W.S., 2005. Nodmutdb: a database for genes and mutants involved in symbiosis. Bioinformatics 21, 2927–2929.

McArthur, A.G., Waglechner, N., Nizam, F., Yan, A., Azad, M.A., Baylay, A.J., Bhullar, K., Canova, M.J., De Pascale, G., Ejim, L., Kalan, L., King, A.M., Koteva, K., Morar, M., Mulvey, M.R., O'Brien, J.S., Pawlowski, A.C., Piddock, L.J.V., Spanogiannopoulos, P., Sutherland, A.D., Tang, I., Taylor, P.L., Thaker, M., Wang, W., Yan, M., Yu, T., Wright, G.D., 2013. The comprehensive antibiotic resistance database. Antimicrob. Agents Chemother. 57, 3348–3357.

Oren, A., 2004. Prokaryote diversity and taxonomy: current status and future challenges. Philos. Trans. R. Soc. Lond. B Biol. Sci. 359, 623–638.

Sahl, J.W., Johnson, J.K., Harris, A.D., Phillippy, A.M., Hsiao, W.W., Thom, K.A., Rasko, D.A., 2011. Genomic comparison of multi-drug resistant invasive and colonizing *Acinetobacter baumannii* isolated from diverse human body sites reveals genomic plasticity. BMC Genom. 12, 291.

Schmidt, H., Hensel, M., 2004. Pathogenicity islands in bacterial pathogenesis. Clin. Microbiol. Rev. 17, 14–56.

Soares, S.C., Abreu, V.A.C., Ramos, R.T.J., Cerdeira, L., Silva, A., Baumbach, J., Trost, E., Tauch, A., Hirata, R.J., Mattos-Guaraldi, A.L., Miyoshi, A., Azevedo, V., 2012. PIPS: pathogenicity island prediction software. PLoS One 7, e30848.

Sullivan, J.T., Patrick, H.N., Lowther, W.L., Scott, D.B., Ronson, C.W., 1995. Nodulating strains of *Rhizobium loti* arise through chromosomal symbiotic gene transfer in the environment. Proc. Natl. Acad. Sci. U. S. A. 92, 8985–8989.

Trost, E., Blom, J., Soares, S.D.C., Huang, I., Al-Dilaimi, A., Schröder, J., Jaenicke, S., Dorella, F.A., Rocha, F.S., Miyoshi, A., Azevedo, V., Schneider, M.P., Silva, A., Camello, T.C., Sabbadini, P.S., Santos, C.S., Santos, L.S., Hirata, R.J., Mattos-Guaraldi, A.L., Efstratiou, A., Schmitt, M.P., Ton-That, H., Tauch, A., 2012. Pangenomic study of *Corynebacterium diphtheriae* that provides insights into the genomic diversity of pathogenic isolates from cases of classical diphtheria, endocarditis, and pneumonia. J. Bacteriol. 194, 3199–3215.

Tumapa, S., Holden, M.T.G., Vesaratchavest, M., Wuthiekanun, V., Limmathurotsakul, D., Chierakul, W., Feil, E.J., Currie, B.J., Day, N.P.J., Nierman, W.C., Peacock, S.J., 2008. *Burkholderia pseudomallei* genome plasticity associated with genomic island variation. BMC Genom. 9, 190.

Uchiumi, T., Ohwada, T., Itakura, M., Mitsui, H., Nukui, N., Dawadi, P., Kaneko, T., Tabata, S., Yokoyama, T., Tejima, K., Saeki, K., Omori, H., Hayashi, M., Maekawa, T., Sriprang, R., Murooka, Y., Tajima, S., Simomura, K., Nomura, M., Suzuki, A., Shimoda, Y., Sioya, K., Abe, M., Minamisawa, K., 2004. Expression islands clustered on the symbiosis island of the *Mesorhizobium loti* genome. J. Bacteriol. 186, 2439–2448.

Waack, S., Keller, O., Asper, R., Brodag, T., Damm, C., Fricke, W.F., Surovcik, K., Meinicke, P., Merkl, R., 2006. Score-based prediction of genomic islands in prokaryotic genomes using hidden markov models. BMC Bioinf. 7, 142.

Wilson, J.W., Schurr, M.J., LeBlanc, C.L., Ramamurthy, R., Buchanan, K.L., Nickerson, C.A., 2002. Mechanisms of bacterial pathogenicity. Postgrad Med. J. 78, 216–224.

Zhou, C.E., Smith, J., Lam, M., Zemla, A., Dyer, M.D., Slezak, T., 2007. Mvirdb—a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. Nucleic Acids Res. 35, D391–D394.

Von Ossowski, I., Satokari, R., Reunanen, J., Lebeer, S., De Keersmaecker, S.C.J., Vanderleyden, J., de Vos, W.M., Palva, A., 2011. Functional characterization of a mucus-specific LPxTGsurface adhesin from probiotic *Lactobacillus rhamnosus* gg. Appl. Environ. Microbiol. 77, 4465–4472.

II.I.7 *Corynebacterium pseudotuberculosis* may be under anagenesis and biovar Equi forms biovar Ovis: a phylogenic inference from sequence and structural analysis.
Oliveira A, Teixeira P, Azevedo M, Jamal SB, Tiwari S, Almeida S, Silva A, Barh D, Dorneles EM, Haas DJ, Heinemann MB, Ghosh P, Lage AP, Figueiredo H, Ferreira RS, **Azevedo V**.
BMC Microbiol. 2016 Jun 2;16(1):100. doi: 10.1186/s12866-016-0717-4.PMID:

Diante da crescente quantidade de dados biológicos obtidos por técnicas de sequenciamento genômico, nosso grupo se deparou com uma variedade de genomas da espécie *C. pseudotuberculosis*, aos quais foram significantes na compreensão dos subtipos existentes dentro dessa espécie: os biovares Ovis e Equi. Esse trabalho foi pioneiro no sentido de explorar questões de convergências e divergências evolutivas, a fim de obter respostas da possível formação de nova(s) espécie(s), considerando esses subtipos. Apesar dos nossos resultados não apontarem o aparecimento de novas espécies, foi possível observar a influência de especiação biológica no sentido de entender que os biovares estão em processo de anagênese (i.e. processo de evolução progressiva que ocorre no interior de uma espécie, levando à especiação). Os resultados obtidos nesse trabalho permitiram a interpretação, por técnicas de evolução molecular, que o biovar Ovis está se originando na natureza pela existência do biovar Equi, no qual poderá ser possível em uma análise futura observar a separação em duas espécies (ou não) por cladogênese.

Oliveira *et al. BMC Microbiology* (2016) 16:100
DOI 10.1186/s12866-016-0717-4

BMC Microbiology

RESEARCH ARTICLE Open Access

Corynebacterium pseudotuberculosis may be under anagenesis and biovar Equi forms biovar Ovis: a phylogenic inference from sequence and structural analysis

Alberto Oliveira[1], Pammella Teixeira[1], Marcela Azevedo[1], Syed Babar Jamal[1], Sandeep Tiwari[1], Sintia Almeida[1], Artur Silva[3], Debmalya Barh[4], Elaine Maria Seles Dorneles[5], Dionei Joaquim Haas[5], Marcos Bryan Heinemann[6], Preetam Ghosh[7], Andrey Pereira Lage[5], Henrique Figueiredo[8], Rafaela Salgado Ferreira[2] and Vasco Azevedo[1*]

Abstract

Background: *Corynebacterium pseudotuberculosis* can be classified into two biovars or *biovars* based on their nitrate-reducing ability. Strains isolated from sheep and goats show negative nitrate reduction and are termed biovar Ovis, while strains from horse and cattle exhibit positive nitrate reduction and are called biovar Equi. However, molecular evidence has not been established so far to understand this difference, specifically if these *C. pseudotuberculosis* strains are under an evolutionary process.

Results: The ERIC 1 + 2 Minimum-spanning tree from 367 strains of *C. pseudotuberculosis* showed that the great majority of biovar Ovis strains clustered together, but separately from biovar Equi strains that also clustered amongst themselves. Using evolutionarily conserved genes (*rpoB, gapA, fusA, and rsmE*) and their corresponding amino acid sequences, we analyzed the phylogenetic relationship among eighteen strains of *C. pseudotuberculosis* belonging to both biovars Ovis and Equi. Additionally, conserved point mutation based on structural variation analysis was also carried out to elucidate the genotype-phenotype correlations and speciation. We observed that the biovars are different at the molecular phylogenetic level and a probable anagenesis is occurring slowly within the species *C. pseudotuberculosis*.

Conclusions: Taken together the results suggest that biovar Equi is forming the biovar Ovis. However, additional analyses using other genes and other bacterial strains are required to further support our anagenesis hypothesis in *C. pseudotuberculosis*.

Keywords: *Corynebacterium pseudotuberculosis*, Evolution, Molecular phylogeny, Structural biology

Background

The genus *Corynebacterium* belongs to the bacterial phylum *Actinobacteria*, also known as *Actinomycetes*. This phylum comprises *Mycobacterium*, *Nocardia* and *Rhodococcus* genera, which together form a supra-generic group known by their initials as CMNR [1–3]. These organisms share some common features, such as:

(i) A specific well-organized cell wall mainly characterized by the presence of vast components of peptidoglycan, mycolic acid, and arabinogalactan [4–7];
(ii) high G + C content (47 %–74 %) [5];
(iii) Gram-positive [8].

Within the genus *Corynebacterium*, the species *Corynebacterium pseudotuberculosis* is reported to be a facultative intracellular pathogen in mammals [9, 10]. The pathologies associated with *C. pseudotuberculosis* are of great importance to veterinary medicine because this bacterium is considered the main etiologic agent of

* Correspondence: vasco@icb.ufmg.br
[1] Departamento de Biologia Geral, Laboratório de Genética Celular e Molecular, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
Full list of author information is available at the end of the article

caseous lymphadenitis (CLA). CLA is characterized by abscess formation in the thorax and abdomen, or in major lymph nodes, causing dermonecrosis and finally resulting in hypertrophy in the affected region [9, 11]. CLA is mostly found in small ruminants (mainly sheep and goats), but other mammals, such as cattle, pigs, deer, sheep, horses, camels and even human beings, although with very rare incidences, can be affected [12, 13]. Curiously, infection by *C. pseudotuberculosis* may cause other diseases,such as ulcerative lymphangitis (UL), a pathology of lymphatic vessels of the lower extremities, particularly hind legs, which is most frequent in horses [14, 15].

In order to understand differences in clinical presentation by infection of *C. pseudotuberculosis*, some studies have proposed to classify this microorganism from genetic, morphological and biochemical points of view [16–18]. Especially based on its ability to breakdown nitrate [19, 20], *C. pseudotuberculosis* was classified into two biovars, Ovis and Equi. Strains isolated from sheep and goats, which are usually negative in nitrate reductase activity, were classified as biovar Ovis; whereas the strains isolated from horse and cattle, which are usually positive in the nitrate reduction test, were classified as biovar Equi.

Additionally, studies have attempted to define these two biovars (Ovis and Equi) using restriction endonucleases (*Eco*RV and *Pst*I) on chromosomal DNA or focusing on nitrate reduction determination methods [21]. Discrimination of both isolates was also possible using other methodologies, such as restriction fragment length polymorphism analysis of 16S ribosomal DNA [19, 20, 22] and pulsed-field gel electrophoresis (PFGE) in combination with biochemical analysis [23]. Other studies have investigated the possible evolutionary divergences of the genus *Corynebacterium* using 16S rRNA sequences [24–26], the preferred genetic tool used to characterize organisms taxonomically [27]. Although 16S rRNA (*rmsE*) gene sequencing is highly useful with regards to bacterial classification, published data has proven analysis of the partial nucleotide sequences of the RNA polymerase β-subunit gene (*rpoB*) is more accurate for *Corynebacterium* species [18, 28].

There is phenotypic evidence (nitrate test) and genotypic evidence (Enterobacterial repetitive intergenic consensus sequence-based - ERIC-PCR and Single-nucleotide polymorphism (SNP) analysis [29, 30]) showing differences between the two biovars. However, our goal in this paper is to investigate the evolutionary differences between biovars Equi and Ovis of *C. pseudotuberculosis*. Thus, in this study, we performed ERIC 1 + 2 PCR and evolutionary analysis, using maximum likelihood method in combination with gene and protein structural analysis, of genes *rsmE,* a vital component of the ribosome, and *rpoB*, a region of strong influence of RNA polymerase activity, to find a kinship and phylogenetic distances. To study the molecular divergences between biovar Ovis and Equi of *C. pseudotuberculosis,* we also considered other genes like *gapA*, which has been used as a target in taxonomic comparisons of bacteria, taking into account the functions that this gene infers in possible differences in the metabolism of carbohydrates, glycolysis and cell survival [31, 32], and *fusA*, that has been used for phylogenetic analysis and taxonomic classification of bacterial species of the genus *Pantoea*, a pathogen to humans and plants [33].

Methods

ERIC 1 + 2 PCR Minimum-spanning tree

The minimum-spanning tree (MST) was generated using ERIC1 + 2-PCR data from 367 *C. pseudotuberculosis* strains, including 226 biovar Ovis field strains [17, 29, 34], 139 biovar Equi field strains (34 strains with published data [29] and 105 strains isolated from equines in USA – unpublished data), type strain ATCC 19410^{T} and vaccine strain1002. The *C. pseudotuberculosis* ATCC 19410^{T} type strain and 1002 vaccine strain were genotyped by ERIC1 + 2-PCR one time in each of the four different assays gathered in the present study ([17, 29, 34] and unpublished data). The MST was built using UPGMA, to calculate the distance matrix, and Prim's algorithm associated with the priority rule and permutation resampling [35, 36]. The MST presented is the top scoring tree, i.e., the tree with the highest overall reliability score.

Dataset

Eighteen *C. pseudotuberculosis* strains with available genome sequences (Additional file 1) had their sequences of genes/proteins *rpoB, gapA, fusA, and rsmE* (Additional file 2) submitted to phylogenetic and structural analyses. ERIC1 + 2-PCR results were available for 13 of those 18 strains. All genome sequences were available from the NCBI database [37, 38]. Protein functional information was from UniProt database annotation [39].

Alignment and phylogenetic analysis

We have used Clustal-X [40] for multiple sequence alignment (MSA) and Jalview [41] to visualize and edit the MSA**,** create phylogenetic trees, explore molecular structures and annotation. The analyses on the transition and transversion mutations were done using the software MEGA 6 [42]. Average Nucleotide Identity (ANI) [43] was employed to evaluate relatedness among strains in substitution of the labour-intensive DNA-DNA hybridization (DDH) technique.

In order to create the phylogenetic trees, we first obtained an evolutionary model adapted to the MSA. Therefore, we used Adaptive Server Evolution (http://www.datamonkey.org/) portal to define one evolutionary model. The outcome of these tests indicate the TN93 model [44]. Seaview [45] was then used to construct the tree based on the model previously presented by the Adaptive Server Evolution portal. The tree was created using the maximum likelihood method performed by PHYML [46], which is available in Seaview. Branch support consistencies were evaluated using the nonparametric bootstrap test [47] with 250 replicates and the approximate likelihood ratio test (ALRT) [48]. The viewing and editing of the tree was carried out using the Figtree tool [49, 50] that enabled us to either characterize different gene groups based on the bootstrap values calculation or to represent the evolutionary time scale. The multiple sequence alignments contain all four genes used in this work.

Statistical analysis

The timetree was generated using the RelTime method [51]. Divergence times for all branching points in the user-supplied topology were calculated using the Maximum Likelihood method based on the General Time Reversible model [52]. Bars around each node represent 95 % confidence intervals which were computed using the method described by Tamura et al. [42]. The estimated log likelihood value of the topology shown is −34964.1082. Similar evolutionary rates were merged between ancestors and descendants so that many clocks were identified in the topology. The rate variation model allowed for some sites to be evolutionarily invariable ([+*I*], 29.1188 % sites). The tree is drawn to scale, with branch lengths proportional to the relative number of substitutions per site. Also, the statistic Tajima's D [53] was used in order to compare the average number of pairwise differences with the number of segregating sites. The Tajima's D statistical can be understood below:

$$E[\pi] = \theta = E\left[\frac{S}{\Sigma_{i=1}^{n-1}\frac{1}{i}}\right] = 4N\mu$$

Where S means the number of segregations sites, n the number of samples and i is the index of summations. Follows the Tajima's D statistical test, there are factors that can change the expected values of S and π . The crux of Tajima's D test statistic is the difference in the expectations for these two variables (which can be positive or negative). Finally, D is calculated by considering the differences between the two estimates of the population genetics parameter θ. The D value is obtained by dividing these differences, that is called d by the square root of its variance $\sqrt{\hat{V}(d)}$.

$$D = \frac{d}{\sqrt{\hat{V}(d)}}$$

The alignment with all 4 genes from two *C. pesudotuberculosis* strains, vaccine strain 1002 from biovar Ovis and strain E19 from biovar Equi, and one *Arcanobacterium haemolyticum* strain, AH 20595 employed as an external group, were used for this statistical analysis. *P*-value less than 0.05 were used to reject the null hypothesis of equal rates between lineages.

Estimation of the pattern of nucleotide substitution

In order to observe the probability of transition (G↔A) and transversion (A↔C) substitutions, the MSA, of all 4 genes from biovars Equi and Ovis, was taken into account, with the aim to determine the types of molecular changes that were occurring. The Maximum Composite Likelihood Estimate of the Pattern of Nucleotide Substitution method was used with the gamma model that corrects for multiple hits, taking into account the rate substitution differences between nucleotides and the inequality of nucleotide frequencies [54]. The analysis involved 19 (external group inside) nucleotide sequences. The transition/tranversion ratio (R), that is the number of these replacement, was calculated by $R = [A^*G^*k_1 + T^*C^*k_2]/[(A + G)^*(T + C)]$ with rate ratios of k_1 evaluating purines and rate ratios of k_2 evaluating pyrimidines. Codon positions included were 1st + 2nd + 3rd + Noncoding. All positions containing gaps and missing data were eliminated. There were a total of 6542 positions in the final dataset. Evolutionary analyses were conducted in MEGA6 [42].

Structural analysis

For structural analysis, the gene sequences were translated to amino acid sequences using the Transeq program: http://www.ebi.ac.uk/Tools/emboss/. After the amino acid alignment, we constructed 3-D models of the proteins and interpreted the molecular differences and consequences. Comparative molecular modeling of proteins was performed with the software Modeller [55]. Additional file 3 presents the details of the template structures using information from NCBI [52] and PDB (Protein Data Bank) [56]. Twenty models were built for each of the templates using MODELLER and one model for each template. PyMOL V.1.5.0.4 was used to visualize three-dimensional protein structures (Schrödinger, LLC.). All homology models were evaluated using several different model evaluation tools such as PROCHECK [55], evaluating the stereochemistry quality, Discrete Optimized Protein Energy (DOPE) score [56], a statistical potential able to provide a energetic validation and RMSD obtained from a structural alignment with a protein with similar function, to provide an functional validation. The latter was carried out in the software PyMOL. For all analysis we selected amino acids that are closer than 4 angstroms (Å) from the variant amino acid.

*We want to inform that this work does not use any participants, children, parent or guardian.

Results

C. pseudotuberculosis biovar Ovis and biovar Equi strains clustered separately by ERIC 1 + 2 PCR

The ERIC 1 + 2 Minimum-spanning tree from 367 strains (365 field strains, type strain ATCC 19410^T and vaccine strain 1002) of *C. pseudotuberculosis* showed that the great majority of biovar Ovis strains clustered together, but separately from biovar Equi strains that also clustered amongst themselves (Fig. 1-a) (Additional file 4). The same was observed when the minimum-spanning tree was constructed only from strains with complete genome sequences (Fig. 1-b) (Additional file 4). Moreover, analysis of all *C. pseudotuberculosis* strains depicted the existence of five major clonal complexes; three that clustered around biovar Ovis and two around biovar Equi strains.

The clonal complex labelled 1 was almost totally composed of *C. pseudotuberculosis* strains from Minas Gerais State, Brazil, whereas clonal complex 2 contained strains from the states of Minas Gerais, Pernambuco and São Paulo, Brazil. The last clonal complex related to *C. pseudotuberculosis* biovar Ovis strains (3) showed the most diverse composition, including strains from Argentina, Australia, Brazil (Minas Gerais, Bahia, and São Paulo), Egypt, Israel, Scotland and USA. *C. pseudotuberculosis* biovar Equiclonal complexes were mainly composed of strains from Egypt (4) and USA (5).

The repeatability observed for the *C. pseudotuberculosis* ATCC 19410^T and 1002 vaccine strain in ERIC1 + 2-PCR was 84 % considering the four different experiments conducted [17, 29, 34] and also the unpublished data from 105 strains.

Statistical analysis and molecular phylogeny from *C. pseudotuberculosis* Ovis and Equi biovars

The phylogenetic tree was carried out to understand the taxonomic distribution of biovars observed from ERIC-PCR. This result showed a small fragmentation of branches among the 18 *C. pseudotuberculosis* strains. For the analysis we added an external group using the gene sequences from *Arcanobacterium haemolyticum* (AH 20595) that are homologous to the four *C. pseudotuberculosis* genes analyzed. This can be visualized in the tree by a red edge (Fig. 2). The fragmented edges that belong to biovar Equi are highlighted in green and black (only strain 162), while the ones that belongs to biovar Ovis are highlighted in orange. The bootstrap value appears along the tree as well as in the legend. The result for ANI was 98 % (Additional file 5). The results of the timetree (Additional files 6 and 7) showed a difference among *C. pseudotuberculosis* strains of biovars Equi and Ovis, taking into account the number of genetic distances to a fraction of the time. Biovar Equi strains trees have different branch length and taxonomic separation. On the other hand, biovar Ovis strain trees present more similar genetic distances between them, without differences in branch lengths. In addition, the results obtained in Tajima's D, shown in Table 1, point to the existence of differences among average numbers of pairwise differences with the number of segregating sites. These data can be supported in regards to a Value of Tajima's $D < 0$, which was in our results −2.07.

Statistical differences and transition and transversion point mutations detected in the multiple sequence alignment from biovares Ovis and Equi

The statistical results of transitional and transversion replacement has pointed to the existence of specific points mutations for both biovars. For example, the Additional file 8 shows one of the arrows that for biovar Equi there is the presence of guanine instead of adenine, a nitrogenous base found only in biovar Ovis. Table 2 shows the distribution of the nucleotide variation in the complete alignment. We can observe that the percentage of transition mutations is higher than that of the transversion mutations. The nucleotide frequencies are 21.51 % (A), 22.95 % (T/U), 27.59 % (C), and 27.95 % (G). The transition/transversion rate ratios are $k_1 = 13.99$ (purines) and $k_2 = 3.89$ (pyrimidines). The overall transition/transversion bias is $R = 4.35$.

Specific amino acid variation among biovar Ovis and Equi

Next, we evaluated the location of these point mutations on protein structure and whether the physicochemical characteristics of the mutated residues differ between the two biovars. We considered the protein sequences from the four genes used in this study, and amino acid differences were evaluated based on a sequence alignment with proteins (Additional file 9) possessing the same function and known. After the alignment, the position of each mutation in the structure was analyzed in comparison to the active site of each protein. In general, mutations were located in the surface and distant from the active site region (Additional file 10).

For *fusA* protein observed two residue modifications (Val177Ile and Asp371Glu) when comparing the two biovars. In both cases the amino acids within each pair have the same physicochemical characteristics: both valine and isoleucine are nonpolar and hydrophobic, and aspartate and glutamate are both polar and negatively charged. Therefore, both mutations are conservative and

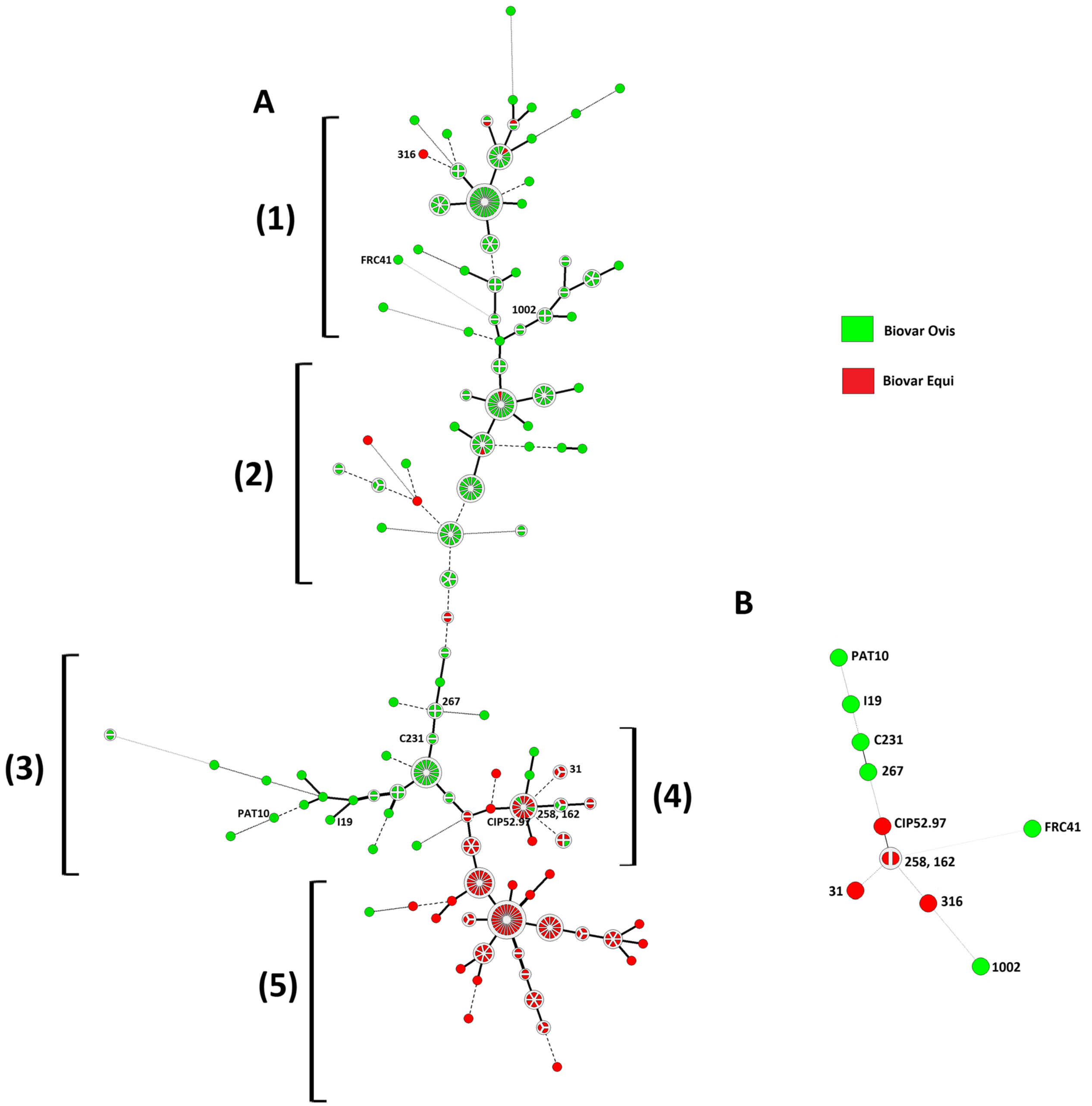


Fig. 1 Minimal spanning tree by ERIC 1 + 2-PCR of 367 (**a**) and eleven (**b**) *C. pseudotuberculosis* strains. The branch length indicates the distance between the nodes as follows: (━━) up to 1 %; (——) up to 5 %; (— —) up to 10 %; (······) up to 15 %; and (▬▬) above 15 %. The sizes of the nodes depend on the number of strains (their population size). Wedges in circles represent the proportion of *C. pseudotuberculosis* isolates from respective sources. The MST presented is the tree with the highest overall reliability score and was calculated using the UPGMA associated with the priority rule and permutation resampling using Bionumerics 7.1 (Applied Maths, Sint-Martens-Latem, Belgium)

the mutations most likely don't affect the function of this protein (Additional file 11).

For *gapA* protein, we observed that Asparagine (Asn) was changed to Aspartic acid (Asp) at position 97 when in both the biovars. Both are polar amino acids, however they differ in physicochemical terms, since Asn is neutral and contains a hydrogen bond donor, while Asp is usually negatively charged at neutral pH. Additionaly, at position 207 another amino acid change was observed, from threonine (Thr) to isoleucin (Ile). In this case, the amino acids differ in size (Ile has a bigger side chain) and polarity (while Ile is nonpolar, Thr is a neutral polar amino acid). Interestingly, *C. pseudotuberculosis* strain 162 that belongs to biovar Equi does not show the same

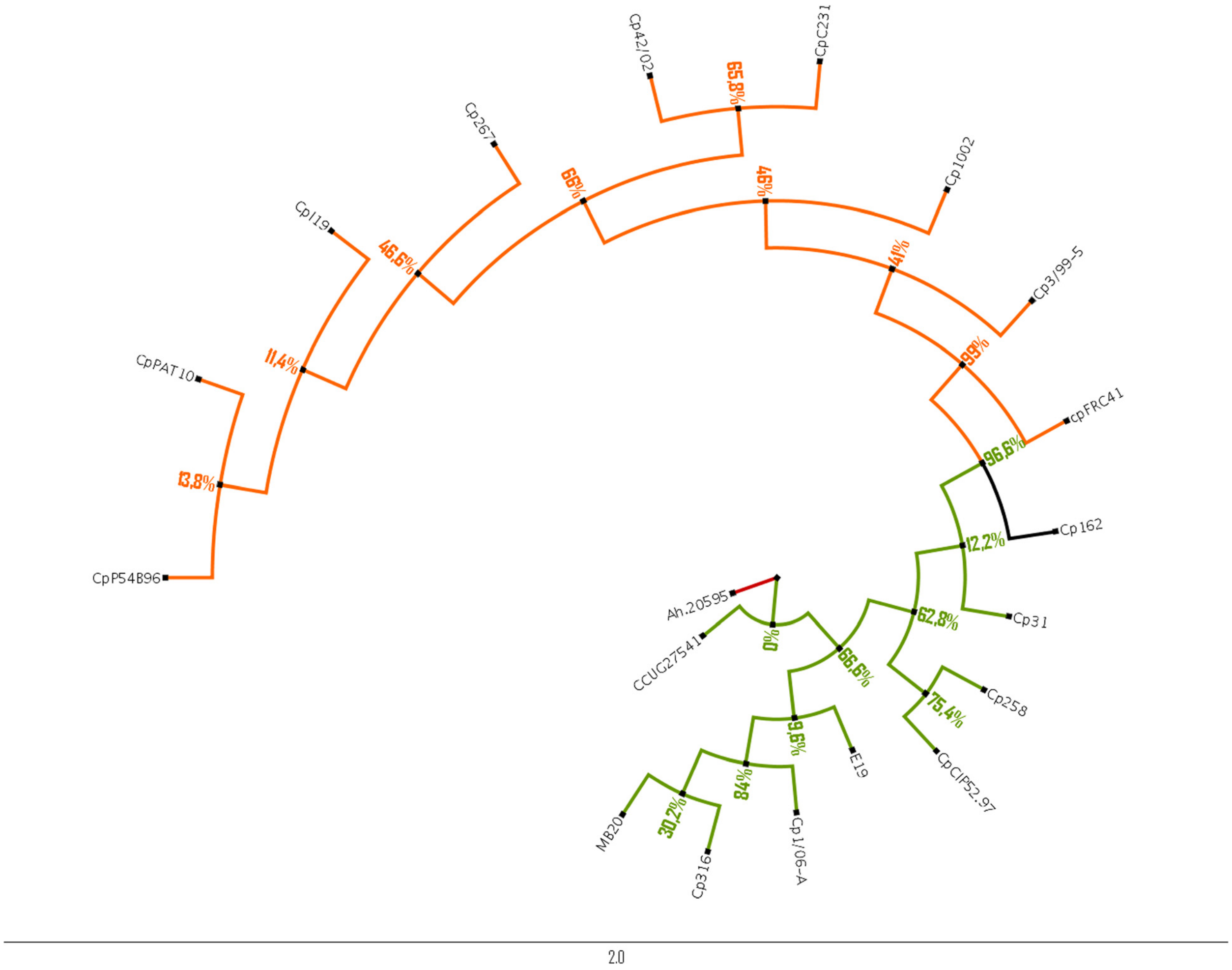


Fig. 2 Phylogenetic tree of Equi and Ovis biovars determined by maximum likelihood method. Phylogenetic tree demonstrating the relationships of the *C. pseudotuberculosis* strains represented by biovars Ovis (orange) and Equi (green) showing their evolutionary differences. The tree is based on the results of distance matrix analyses of all 4 genes explored in this work. The topology of the tree was determined by performing maximum likelihood analyses. The outside group is highlighted in the brown edge. Boostrap values can be identified by the label on the left side and the nodes in the tree

changes when compared to the other strains from the same biovar.

Comparative analysis of the 16S ribosomal RNA methyltransferase (*rsmE*) from *C. pseudotuberculosis* biovars Equi and Ovis showed a His18Arg variation at the N-terminal region and a Thr132Ala variation, at the C-terminal region. Both Histidine (His) and Arginine (Arg) located at position 18 have polar side chains and may be positively charged at neutral pH, however Arg has a much higher pKa. At position 132, the amino acids Threonine (Thr) and Alanine (Ala) differ functionally as Thr is a neutral polar amino acid and Ala is nonpolar (Additional file 10). The Table 3 shows all variations for these proteins.

Finally, for the beta subunit of RNA polymerase (*rpoB*) only one residue difference, Ala979Thr, was observed when comparing biovars Equi and Ovis. In this case there is a change in polarity for these two amino acids (Additional file 10). Curiously *C. pseudotuberculosis* strains 1002, 3/99, and FRC41, belonging to biovar Ovis, shared the same amino acids when compared to biovar Equi strains (Table 4).

Discussion

We earlier suggested from data of 102 *C. pseudotuberculosis* strains that *C. pseudotuberculosis* biovar Ovis and Equi exhibited different ERIC1 + 2-PCR clustering pattern (Dorneles et al., [29]). The present data from 373 strains also showed a clear difference in clustering pattern between *C. pseudotuberculosis* biovar Ovis and Equi, with a few exceptions (Fig. 1-a). However, as the present MST results were based on a representative number of genotyped strains (139 *C. pseudotuberculosis* biovars Equi field strains, 226 *C. pseudotuberculosis* biovar Ovis field strains, from twelve countries and isolated from eight different hosts), it allows us to make more

Table 1 Results from Tajima's neutrality test

m	S	p_s	Θ	π	D
19	841	0.52	0.15	0.07	−2.07

The analysis involved 19 amino acid sequences. The coding data was translated assuming a genetic code table. All positions containing gaps and missing data were eliminated. There were a total of 1599 positions in the final dataset. Evolutionary analyses were conducted in MEGA6
Abbreviations: m = number of sequences, n = total number of sites, S = Number of segregating sites, $p_s = S/n$, $\Theta = p_s/a_1$, π = nucleotide diversity, and D is the Tajima test statistic [52]

reliable inferences. The five largest clonal complexes were observed in the MST, three were mainly related to *C. pseudotuberculosis* biovar Ovis strains (1, 2, and 3) and two with *C. pseudotuberculosis* biovar Equi strains (4 and 5), from which other clonally related isolated groups emerge. The distinct clustering pattern of *C. pseudotuberculosis* biovar Ovis and Equi strains might reflect the number of genes specific to each biovar [57].

In order to understand the clustering formation, we have used algorithms and techniques to construct qualitative phylogenetic trees to explain the best possible evolutionary relationship between biovars Ovis and Equi from *C. pseudotuberculosis* strains. The difference in the biovars can be clearly observed from the ERIC 1 + 2-PCR MST and phylogenetic tree (Figs. 1 and 2). However, it was not possible to determine if the two biovars belongs to different species based on the results obtained after the phylogenetic tree and average nucleotide identity (ANI) analyses. Although there are some genetic and biochemical differences between them, it was not clear whether the branches, observed at phylogenetic tree, have distinct groups of organisms. Based on our data, we hypothesized that biological speciation may be occurring, for instance, anagenesis, which is a process of progressive evolution of species involving changes in the gene frequency of a population [58].

Based on the anagenesis mode of speciation, we have hypothesized that speciation is occurring within the *C. pseudotuberculosis* at a molecular level. The genes *rmsE*, *rpoB*, *fusA*, and *gapA* that we used are strong candidates for phylogenetic analysis because they are involved in many important cellular processes, such as maintenance of cellular integrity, cell survival and several metabolic reactions [31–33]. According to Dorela et al. [8], *rpoB* is a relevant gene used to explore evolutionary routes serving as a tool to search for new species, differently from 16S rRNA gene sequencing data, that presents resolution problems at genus and/or species level. In our study, it was observed that the 16S sequence may be better in characterizing the molecular differences (mainly in structural analysis) between the two biovars, when compared to the *rpoB* gene. Our multiple sequence alignments analysis showed that all the genes have transition and transversion point mutations, which is defining and separating *C. pseudotuberculosis* biovar Equi from Ovis. In addition, it was observed from the timetree analysis that differences between the two *C. pseudotuberculosis*

Table 2 Maximum composite likelihood estimate of the pattern of nucleotide substitution

	A	T	C	G
A	-	*2.11*	*2.53*	**35.92**
T	*1.98*	-	**9.86**	*2.57*
C	*1.98*	**8.2**	-	*2.57*
G	**27.65**	*2.11*	*2.53*	-

Each entry shows the probability of substitution (r) from one base (row) to another base (column) [54]. For simplicity, the sum of *r* values is made equal to 100. Rates of different transitional substitutions are shown in bold and those of transversionsal substitutions are shown in *italics*

Table 3 Distribution of amino acids from proteins in biovars. The amino acid variants positions present physicochemical differences, and are observed to be specific to a biovar type. Some variations exhibit an increase or decrease in the number of interactions between amino acids that influence the stability of the protein

Protein	Position	Amino acid	Biovar
fusA	177	Valine (V)	Equi
		Isoleucine (I)	Ovis
	371	Glutamine (E)	Equi
		Aspartic acid (D)	Ovis
gapA	97	Asparagine (N)	Equi
		Aspartic acid (D)	Ovis
	207	Threonine (T)	Equi
		Isoleucine (I)	Ovis
rmsE	18	Histidine (H)	Ovis
		Arginine (R)	Equi
	132	Threonine (T)	Ovis
		Alanine (A)	Equi

Table 4 Distribution of amino acids from protein *rpoB* in biovars. The position 979 has variation in amino acids: Alanine for some strains as 1002, 106, 3/99, 31, 258, 52.97, 162 and FRC41; while the same position is Threonine for other biovars as 42/02, 267, 231, I19, P54B96 and PAT10

Alanine (A)^{979}		Threonine (T)^{979}	
Strain	Biovar	Strain	Biovar
1002	Ovis	42/02	Ovis
106	Equi	267	
3/99	Ovis	231	
31	Equi	I19	
258	Equi	P54B96	
52.97	Equi	PAT10	
162	Equi		
FRC41	Ovis		

biovars were due to time of acquired mutations. It is possible to observe that the strains of biovar Equi are more prone to have mutations while those of *C. pseudotuberculosis* biovar Ovis were more stable. It is not possible, even with these findings, to state that the two groups constitute different species. However, taking into account the time slice versus genetic differences, *C. pseudotuberculosis* strains belonging to biovar Equi are under a constant mutational process, regarding the lengths of the branches. Considering the analyses of *C. pseudotuberculosis* strain 162, that have phylogenetic features approaching biovar Ovis, we interpret that some changes are being stabilized giving rise to strains of biovar Ovis. This event may indicate that a biological speciation may be occurring slowly in *C. pseudotuberculosis*. Our hypothesis is that an anagenesis process is happening from *C. pseudotuberculosis* biovar Equi to *C. pseudotuberculosis* biovar Ovis. Such anagenesis is observed when a sufficient number of mutations are fixed in a population, which makes the emergence of a new phenotype possible in the future. Based on our data, we dismissed the possibility of a cladogenesis event because we did not observe the transformation of an organism into two others, but rather the possible genesis of one organism into another. Therefore, it is highly likely that the mutations described in this study strengthen the idea of anagenesis.

Also, the results from Tajima's D statistical test give us the biological interpretation that the proportion of mutations that alter codons for amino acids is higher than expected in both biovars and the population is evolving as per mutation-drift equilibrium. Furthermore, also is possible be interpreted rare alleles present at low frequencies in both biovars and population expansion after a recent bottleneck. Our structural analysis showed that none of the described mutations occurred within these sites or in positions previously described as critical for activity. Therefore, probably they do not affect protein function, but they are useful to indicate phylogenetic distance. For example, it was observed that *C. pseudotuberculosis* strain 162, which belongs to biovar Equi, is phylogenetically closer to biovar Ovis strains than to biovar Equi strains, as observed in the phylogenetic tree (Fig. 2). Probably this phylogenetic rapprochement is due to *C. pseudotuberculosis* strain 162 present Ovis specific mutations, although it is included in Equi biovar based on the nitrate test. Regarding this observation we hypothesize that biovar Ovis is being originated from biovar Equi, as mentioned before, as *C. pseudotuberculosis* 162 strain shares point mutations from both biovars.

Taking into account the reports discussed in the literature, many differences within the genomes of some *C. pseudotuberculosis* strains were reported [8, 29, 57]. Recently, Almeida et al. [30] (unpublished data) found a 99 % concordance in detection by PCR of gene *narG*, responsible for nitrate reduction, and nitrate reduction test, suggesting that *C. pseudotuberculosis* biovars Equi could be identified by the presence of gene *narG*, whereas biovar Ovis does not present it. Guimarães et al. [34] showed a typing method based on PCR (ERIC-PCR), which proved to be a good method to discriminate genetic differences among *C. pseudotuberculosis* strains. Using this method it was possible to observe that *C. pseudotuberculosis* biovar Ovis and biovar Equi strains clustered separately [29]. The differences in the clustering pattern of *C. pseudotuberculosis* biovar Ovis and biovar Equi strains could reflect the number of specific genes in each biovar [57]. This fact is evident from the complete genome analyses of 18 *C. pseudotuberculosis* strains; among the total 1504 genes, it was shown that *C. pseudotuberculosis* biovar Ovis contains 314 orthologous genes that are shared by all strains from this biovar but are absent from one or more strains of *C. pseudotuberculosis* biovar Equi [57]. Furthermore, *C. pseudotuberculosis* biovar Equi strains have 95 core genes that are absent from one or more strains of *C. pseudotuberculosis* biovar Ovis [57]. Soares et al. [57], working with pathogenicity islands (PAI), found differences in some genes related to *pilus* formation in *C. pseudotuberculosis* biovar Equi and biovar Ovis strains. *Pilus* gene clusters are acquired normally in block through horizontal gene transfer and are composed of a specific sortase gene and the major, base and tip pilin genes. Genetic variation was found in genome analyses between the two *C. pseudotuberculosis* biovars (Almeida et al., [30]). *C. pseudotuberculosis* biovar Ovis strain VD57 had its genome analysed for the presence of SNP's, and the average variation found when it was compared to *C. pseudotuberculosis* biovar Ovis strains was 823 nucleotides, whereas when compared to the *C. pseudotuberculosis* biovar Equi strains that number increased to 25285.3 SNP's.

Regarding the anagenesis theory, we suggest that *C. pseudotuberculosis* biovar Equi is forming, after some genomic changes, *C. pseudotuberculosis* biovar Ovis. Hence, we postulated that *C. pseudotuberculosis* biovar Equi strains could be evolutionarily older compared to *C. pseudotuberculosis* biovar Ovis strains.

Conclusions

C. pseudotuberculosis biovar Equi and biovar Ovis contain different molecular characteristics, although they belong to the same species. The formation of a new species is an event that happens very slowly in nature. It is possible to interpret that a speciation process is occurring from the anagenesis event. With regard to the statistical data, analysis of sequence and structure of proteins addressed in this study, we conclude that *C. pseudotuberculosis* biovar Ovis is being formed from *C. pseudotuberculosis* biovar Equi through anagenesis.

Additional files

Additional file 1: Information about the strains of *C. pseudotuberculosis* in this work. In total, 18 strains were used of which nine strains were Equi and nine strains were Ovis as tabulated below. (PDF 19 kb)

Additional file 2: Information about the genes of *C. pseudotuberculosis*. All four genes were used in this work. These were united in the same alignment for phylogenetic analysis. (PDF 7 kb)

Additional file 3: Ratio templates used for molecular modeling. € Resolution of the structure was determined by experimental methods of electron microscopy. (PDF 35 kb)

Additional file 4: Bionumerics 7.1 (Applied Maths, Sint-Martens-Latem, Belgium) (BMP 1582 kb)

Additional file 5: Average Nucleotide Identity (ANI) from biovar ovis and biovar equi. Typically, the ANI values between genomes of the same species are above 95 %. Our results show the value 98.76 %, which we expected due to the results of phylogeny. One-way ANI 1: 98.64 % (SD: 1.15 %), from 67439 fragments. One-way ANI 2: 98.69 % (SD: 1.20 %), from 101821 fragments. Two-way ANI: 98.76 % (SD: 0.92 %), from 34975 fragments. (PNG 435 kb)

Additional file 6: Molecular clock analysis by Maximum Likelihood method. Relationship between molecular divergence and time. The analysis involved 19 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 6514 positions in the final dataset. Evolutionary analyses were conducted in MEGA6. (PDF 55 kb)

Additional file 7: Results from comparative test with molecular clocks using the Maximum Likelihood method and without molecular clock of strains. (PDF 8 kb)

Additional file 8: Fragment of multiple sequence alignment. The figure shows transition (arrows in blue) and transversion (red arrow) point mutations observed in nucleotide sequences from both biovars Equi and Ovis. (TIF 164 kb)

Additional file 9: Fragment of multiple sequence alignment of protein. The figure shows the substitutions of amino acids, in which most of the cases the changes modify the physicochemical characteristics. (TIF 216 kb)

Additional file 10: Representation of the structure of *gapA*, *rsmE* and *rpoB* by molecular modeling. The molecular differences between the amino acid from biovars Equi and Ovis induce an increase in the number of chemical bonds between amino acids that are close to the variant residue. (TIF 4028 kb)

Additional file 11: Structure of the *fusA* protein in the biovars Equi and Ovis analyzed by molecular modeling [55]. Expansion of 4 angstroms (Å) from the variant amino acid where it is possible to identify clusters of neighboring residues interacting among themselves. (A) Variation between V ↔ I, (B) Variation between D ↔ E. (TIF 2499 kb)

Abbreviations

CLA, caseous lymphadenitis; DOPE, discrete optimized protein energy; ERIC-PCR, enterobacterial repetitive intergenic consensus sequence-based; MSA, multiple sequence alignment; MST, minimum-spanning tree; NCBI, National Center Biotechnology Institute; PDB, Protein Data Bank; PFGE, pulsed-field gel electrophoresis; UL, ulcerative lymphangitis; Uniprot, Universal Protein Resource

Acknowledgements

The authors would like to acknowledge the research institution UFMG and Professor Raphael Hirata from University from State of Rio de Janeiro.

Funding

This work was supported by grants from Conselho nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes), Brazil.

Availability of data and materials

The datasets supporting the conclusions of this article are included within the article and its additional files. The phylogenetic tree was deposited in TreeBASE under the URL http://purl.org/phylo/treebase/phylows/study/TB2:S19338.

Authors' contributions

AO: Analysis and interpretation from this work, PT: support in modeling and phylogenetic processes, MA, SJ, ST, DB, PG: support with writing the paper, SA: support in obtaining sequences, DB, HF and RF: supports analysis, EMSD, DJH, MBH, APL: discussion on the ERIC PCR method, AS, HF and VA: Research support. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

[1]Departamento de Biologia Geral, Laboratório de Genética Celular e Molecular, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil. [2]Departamento de Bioquímica e Imunologia, Laboratório de Genética Celular e Molecular, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil. [3]Departmento de Genética, Universidade Federal do Pará, Pará, Brazil. [4]Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba, Medinipur WB-721172, India. [5]Departamento de Medicina Veterinária Preventiva, Escola de Veterinária – Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil. [6]Departamento de Medicina Veterinária Preventiva e Saúde Animal, Faculdade de Medicina Veterinária e Zootecnia – Universidade de São Paulo, São Paulo, Brazil. [7]Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA. [8]Aquacen, National Reference Laboratory for Aquatic Animal Diseases, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil.

Received: 2 April 2015 Accepted: 25 May 2016
Published online: 02 June 2016

References

1. Hard GC. Corynebacterium ovis Electron Microscopic Examination of Corynebacterium ovis. J Bacteriol. 1969;97:1480–5.
2. Paule BJ A, Meyer R, Moura Costa LF, Bahia RC, Carminati R, Regis LF, Vale VLC, Freire SM, Nascimento I, Schaer R, Azevedo V. Three-phase partitioning as an efficient method for extraction/concentration of immunoreactive excreted-secreted proteins of Corynebacterium pseudotuberculosis. Protein Expr Purif. 2004;34:311–6.
3. Songer JG. Bacterial phospholipases and their role in virulence. Trends Microbiol. 1997;5:156–61.
4. Bayan N, Houssin C, Chami M, Leblon G. Mycomembrane and S-layer: two important structures of Corynebacterium glutamicum cell envelope with promising biotechnology applications. J Biotechnol. 2003;104:55–67.
5. Funke G, Lawson PA, Collins MD. Heterogeneity within human-derived centers for disease control and prevention (CDC) coryneform group ANF-1-like bacteria and description of Corynebacterium auris sp. nov. Int J Syst Bacteriol. 1995;45:735–9.
6. Hall V. Corynebacterium atypicum sp. nov., from a human clinical source, does not contain corynomycolic acids. Int J Syst Evol Microbiol. 2003;53:1065–8.
7. Hard GC. Comparative toxic effect of the surface lipid of Corynebacterium ovis on peritoneal macrophages. Infect Immun. 1975;12:1439–49.
8. Dorella FAD, Gustavo L, Achecoa CP, Liveirab SCO, Iyoshia AM, Zevedoa VA. Corynebacterium pseudotuberculosis: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. Vet Res. 2006;37:201–218
9. Williamson LH. Caseous lymphadenitis in small ruminants. Vet Clin North Am Food Anim Pract. 2001;17:359–71. vii.
10. Trost E, Ott L, Schneider J, Schröder J, Jaenicke S, Goesmann A, Husemann P, Stoye J, Dorella FA, Rocha FS, Soares SDC, D'Afonseca V, Miyoshi A, Ruiz J, Silva A, Azevedo V, Burkovski A, Guiso N, Join-Lambert OF, Kayal S, Tauch A. The complete genome sequence of Corynebacterium pseudotuberculosis

FRC41 isolated from a 12-year-old girl with necrotizing lymphadenitis reveals insights into gene-regulatory networks contributing to virulence. BMC Genomics. 2010;11:728.

11. Dorella FA, Pacheco LG, Seyffert N, Portela RW, Meyer R, Miyoshi A, Azevedo V. Antigens of Corynebacterium pseudotuberculosis and prospects for vaccine development. Expert Rev Vaccines. 2009;8:205–13.
12. Marchand CH, Salmeron C, Bou Raad R, Méniche X, Chami M, Masi M, Blanot D, Daffé M, Tropis M, Huc E, Maréchal P, Decottignies P, Bayan N. Biochemical disclosure of the mycolate outer membrane of Corynebacterium glutamicum. J Bacteriol. 2012;194:587–97.
13. Brown CC, Olander HJ, Alves SF. Synergistic hemolysis-inhibition titers associated with caseous lymphadenitis in a slaughterhouse survey of goats and sheep in Northeastern Brazil. Can J Vet Res. 1987;51:46–9.
14. Doherr MG, Carpenter TE, Wilson WD, Gardner IA. Application and evaluation of a mailed questionnaire for an epidemiologic study of Corynebacterium pseudotuberculosis infection in horses. Prev Vet Med. 1998;35:241–53.
15. Britz E, Spier SJ, Kass PH, Edman JM, Foley JE. The relationship between Corynebacterium pseudotuberculosis biovar equi phenotype with location and extent of lesions in horses. Vet J. 2014;200:282–6.
16. Judson R, Songer JG. Corynebacterium pseudotuberculosis: in vitro susceptibility to 39 antimicrobial agents. Vet Microbiol. 1991;27:145–50.
17. Dorneles EMS, Santana JA, Andrade GI, Santos ELS, Guimaraes AS, Mota RA, Santos AS, Miyoshi A, Azevedo V, Gouveia AMG, Lage AP, Heinemann MB. Molecular characterization of Corynebacterium pseudotuberculosis isolated from goats using ERIC-PCR. Genet Mol Res. 2012;11:2051–9.
18. Khamis A, Raoult D, Scola BLA. Comparison between rpoB and 16S rRNA Gene Sequencing for Molecular Identification of 168 Clinical Isolates of Corynebacterium Comparison between rpoB and 16S rRNA Gene Sequencing for Molecular Identification of 168 Clinical Isolates of Corynebacterium. J Clin Microbiol. 2005;43:1934–6.
19. Costa LR, Spier SJ, Hirsh DC. Comparative molecular characterization of Corynebacterium pseudotuberculosis of different origin. Vet Microbiol. 1998; 62:135–43.
20. Sutherland SS, Hart RA, Buller NB. Genetic differences between nitrate-negative and nitrate-positive C. pseudotuberculosis strains using restriction fragment length polymorphisms. Vet Microbiol. 1996;49:1–9.
21. Songer JG, Beckenbach K, Marshall MM, Olson GB, Kelley L. Biochemical and genetic characterization of Corynebacterium pseudotuberculosis. Am J Vet Res. 1988;49:223–6.
22. Vaneechoutte M, Riegel P, de Briel D, Monteil H, Verschraegen G, De Rouck A, Claeys G. Evaluation of the applicability of amplified rDNA-restriction analysis (ARDRA) to identification of species of the genus Corynebacterium. Res Microbiol. 1995;146:633–41.
23. Connor KM, Quirie MM, Baird G, Donachie W. Characterization of United Kingdom Isolates of Corynebacterium pseudotuberculosis Using Pulsed-Field Gel Electrophoresis Characterization of United Kingdom Isolates of Corynebacterium pseudotuberculosis Using Pulsed-Field Gel Electrophoresis. J Clin Microbiol. 2000;38:2633–7.
24. Khamis A, Raoult D, La Scola B. Comparison between rpoB and 16S rRNA gene sequencing for molecular identification of 168 clinical isolates of Corynebacterium. J Clin Microbiol. 2005;43:1934–6.
25. Bujnicki JM. Phylogenomic analysis of 16S rRNA:(guanine-N2) methyltransferases suggests new family members and reveals highly conserved motifs and a domain structure similar to other nucleic acid amino-methyltransferases. FASEB J. 2000;14:2365–8.
26. Pascual C, Lawson PA, Farrow JA, Gimenez MN, Collins MD. Phylogenetic analysis of the genus Corynebacterium based on 16S rRNA gene sequences. Int J Syst Bacteriol. 1995;45:724–8.
27. Balch WE, Fox GE, Magrum LJ, Woese CR, Wolfe RS. Methanogens: reevaluation of a unique biological group. Microbiol Rev. 1979;43:260–96.
28. Khamis A, Raoult D, La Scola B. rpoB gene sequencing for identification of Corynebacterium species. J Clin Microbiol. 2004;42:3925–31.
29. Dorneles EMS, Santana JA, Ribeiro D, Dorella FA, Guimaraes AS, Moawad MS, Selim SA, Garaldi ALM, Miyoshi A, Ribeiro MG, Gouveia AMG, Azevedo V, Heinemann MB, Lage AP. Evaluation of ERIC-PCR as Genotyping Method for Corynebacterium pseudotuberculosis Isolates. PLoS One. 2014;9:e98758.
30. Almeida, s., Sandeep Tiwari, Mariano, d., rocha, f. s., Jamal, Syed Babar, Coimbra, n. a. r., Raittz, r. t., Dorella, f. a., Carvalho, a. f., Pereira, f. l., Leal, c. a. g., Debmalya Barh, Ghosh, p., Figueiredo, h. c. p., Moura-Costa, l. f., Portela, r. w V: The Genome Anatomy of Corynebacterium pseudotuberculosis VD57 a Highly Virulent Strain Causing Caseous lymphadenitis. Stand Genomic Sci 2015;57:1-8.
31. Toyoda K, Teramoto H, Inui M, Yukawa H. Involvement of the LuxR-type transcriptional regulator RamA in regulation of expression of the gapA gene, encoding glyceraldehyde-3-phosphate dehydrogenase of Corynebacterium glutamicum. J Bacteriol. 2009;191:968–77.
32. Toyoda K, Teramoto H, Inui M, Yukawa H. Expression of the gapA gene encoding glyceraldehyde-3-phosphate dehydrogenase of Corynebacterium glutamicum is regulated by the global regulator SugR. Appl Microbiol Biotechnol. 2008;81:291–301.
33. Delétoile A, Decré D, Courant S, Passet V, Audo J, Grimont P, Arlet G, Brisse S. Phylogeny and identification of Pantoea species and typing of Pantoea agglomerans strains by multilocus gene sequencing. J Clin Microbiol. 2009; 47:300–10.
34. Guimarães ADS, Dorneles EMS, Andrade GI, Lage AP, Miyoshi A, Azevedo V, Gouveia AMG, Heinemann MB. Molecular characterization of Corynebacterium pseudotuberculosis isolates using ERIC-PCR. Vet Microbiol. 2011;153:299–306.
35. Feil EJ, Li BC, Aanensen DM, William P, Spratt BG, Hanage WP. eBURST : Inferring Patterns of Evolutionary Descent among Clusters of Related Bacterial Genotypes from Multilocus Sequence Typing Data eBURST : Inferring Patterns of Evolutionary Descent among Clusters of Related Bacterial Genotypes from Multilocus Sequen. J Bacteriol. 2004;186:1518–30.
36. Salipante SJ, Hall BG. Inadequacies of minimum spanning trees in molecular epidemiology. J Clin Microbiol. 2011;49:3568–75.
37. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Mizrachi I, Ostell J, Pruitt KD, Schuler JD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Souvorov A, Starchenko G, Tatusova TA. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2009; 37(Database issue):D5–D15.
38. Benson DA, Karsch Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2009;37(Database issue):D26–31.
39. Apweiler R, Bateman A, Martin MJ, O'Donovan C, Magrane M, Alam-Faruque Y, Alpi E, Antunes R, Arganiska J, Casanova EB, Bely B, Bingley M, Bonilla C, Britto R, Bursteinas B, Chan WM, Chavali G, Cibrian-Uhalte E, Silva A, Giorgi M, Fazzini F, Gane P, Castro LG, Garmiri P, Hatton-Ellis E, Hieta R, Huntley R, Legge D, Liu W, Luo J. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res. 2014;42:D191–8.
40. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. Clustal W and Clustal X version 2.0. Bioinformatics. 2007;23:2947–8.
41. Waterhouse AM, Procter JB, Martin DM A, Clamp M, Barton GJ. Jalview Version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009;25:1189–91.
42. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol. 2013;30:2725–9.
43. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme PTJ. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol. 2007;57:1.
44. Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol. 1993;10:512–26.
45. Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol. 2010;27:221–4.
46. Guindon S, Gascuel O. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. Syst Biol. 2003;52:696–704.
47. Felsenstein J. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution (N Y). 1985;39:783.
48. Anisimova M, Gascuel O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst Biol. 2006;55:539–52.
49. Suchard MA, Rambaut A. Many-Core Algorithms for Statistical Phylogenetics. Bioinformatics. 2009;25:1370–6.
50. Pybus OG, Rambaut A, Harvey PH. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics. 2000;155:1429–37.
51. Tamura K, Battistuzzi FU, Billing-Ross P, Murillo O, Filipski A, Kumar S. Estimating divergence times in large molecular phylogenies. Proc Natl Acad Sci U S A. 2012;109:19333–8.

52. Nei M, Kumar S, Nei M, Kumar S. Molecular Evolution and Phylogenetics. New York: Oxford University Press; 2000. p. 333. 2000(August).
53. Tajima F. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics. 1989;123:585–95.
54. Tamura K, Nei M, Kumar S. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc Natl Acad Sci U S A. 2004;101:11030–5.
55. Eswar N, Webb B, Marti-Renom MA, et al. Comparative protein structure modeling using MODELLER. Curr Protoc Protein Sci. 2007;Chapter 2:Unit 2.9. doi:10.1002/0471140864.ps0209s50.
56. Bernstein FC, Koetzle TF, Williams GJ, Meyer EE Jr., Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: A Computer-based Archival File For Macromolecular Structures. J of Mol Biol. 1977;112(535).
57. Soares SC, Silva A, Trost E, Blom J, Ramos R, Carneiro A, Ali A, Santos AR, Pinto AC, Diniz C, Barbosa EG V, Dorella FA, Aburjaile S, Rocha FS, Nascimento KKF, Guimaraes LC, Almeida S, Hassan SS, Bakhtiar SM, Pereira UP, Abreu VAC, Schneider MPC, Miyoshi A, Tauch A, Azevedo V. The pan-genome of the animal pathogen Corynebacterium pseudotuberculosis reveals differences in genome plasticity between the biovar ovis and equi strains. PLoS One. 2013;8:e53818.
58. Sons JW. The hypercycle: A principle of natural self-organization. Am J Vet Res. 1978;65:23.

II.I.8 Whole-genome optical mapping reveals a mis-assembly between two rRNA operons of *Corynebacterium pseudotuberculosis* strain 1002.
Mariano DC, Sousa Tde J, Pereira FL, Aburjaile F, Barh D, Rocha F, Pinto AC, Hassan SS, Saraiva TD, Dorella FA, de Carvalho AF, Leal CA, Figueiredo HC, Silva A, Ramos RT, Azevedo VA.
BMC Genomics. 2016 Apr 30;17(1):315. doi: 10.1186/s12864-016-2673-7.

Neste trabalho utilizou-se a técnica de mapeamento óptico (WGM) para realizar uma análise do genoma de *Corynebacterium pseudotuberculosis* 1002 (Cp1002). WGM é uma técnica de visualização de sítios de restrição e sua detecção por meio de imagens de alta resolução. Com os resultados de WGM foi possível realizar uma nova montagem de alta acurácia do genoma de Cp1002. A nova montagem demonstrou uma grande inversão de metade do genoma ocorrendo entre duas regiões codificadoras de rRNAs ribossomais. A causa da inversão não pode ser detectada, mas o uso de WGM pode ser destacado para melhoria de montagens de genomas de bactérias.

BMC Genomics

RESEARCH ARTICLE Open Access

Whole-genome optical mapping reveals a mis-assembly between two rRNA operons of *Corynebacterium pseudotuberculosis* strain 1002

Diego César Batista Mariano[1], Thiago de Jesus Sousa[1], Felipe Luiz Pereira[2], Flávia Aburjaile[1], Debmalya Barh[3], Flávia Rocha[1], Anne Cybelle Pinto[1], Syed Shah Hassan[1], Tessália Diniz Luerce Saraiva[1], Fernanda Alves Dorella[2], Alex Fiorini de Carvalho[2], Carlos Augusto Gomes Leal[2], Henrique César Pereira Figueiredo[2], Artur Silva[4], Rommel Thiago Jucá Ramos[4] and Vasco Ariston Carvalho Azevedo[1*]

Abstract

Background: Studies have detected mis-assemblies in genomes of the species *Corynebacterium pseudotuberculosis*. These new discover have been possible due to the evolution of the Next-Generation Sequencing platforms, which have provided sequencing with accuracy and reduced costs. In addition, the improving of techniques for construction of high accuracy genomic maps, for example, Whole-genome mapping (WGM) (OpGen Inc), have allow high-resolution assembly that can detect large rearrangements.

Results: In this work, we present the resequencing of *Corynebacterium pseudotuberculosis* strain 1002 (Cp1002). Cp1002 was the first strain of this species sequenced in Brazil, and its genome has been used as model for several studies *in silico* of caseous lymphadenitis disease. The sequencing was performed using the platform Ion PGM and fragment library (200 bp kit). A restriction map was constructed, using the technique of WGM with the enzyme *Kpn*I. After the new assembly process, using WGM as scaffolder, we detected a large inversion with size bigger than one-half of genome. A specific analysis using BLAST and NR database shows that the inversion occurs between two homology RNA ribosomal regions.

Conclusion: In conclusion, the results showed by WGM could be used to detect mismatches in assemblies, providing genomic maps with high resolution and allow assemblies with more accuracy and completeness. The new assembly of *C. pseudotuberculosis* was deposited in GenBank under the accession no. CP012837.

Keywords: Genomics, Sequencing, Optical mapping, Mis-assembly

Background

Corynebacterium pseudotuberculosis (*Cp*) is a Gram-positive, pleomorphic, facultative intracellular pathogenic bacteria that belongs to the group *Corynebacterium*, *Mycobacterium*, *Nocardia* and *Rhodococcus* (CMNR) [1]. *Cp* can be classified into two biovars: *equi* and *ovis*. Biovar *equi* is characterized by its capacity to nitrate-reductase production, while the biovar *ovis*, cannot [2]. Genomic plasticity analysis using 15 *Cp* strains demonstrates that the group of strains belonging to the *ovis* biovar are highly similar [3]. *Cp* is the etiological agent of the caseous lymphadenitis (CLA) disease, that affects mainly sheep and goat causing huge economic losses by affecting meet and wool production [4, 5]. It is also capable to cause diseases in cattle and humans. However, so far there is no proper diagnosis method or effective treatment available for *Cp* infection.

With the advent of next-generation sequencing (NGS) platforms [6–8], so far 37 *Cp* genomes have been

* Correspondence: vasco@icb.ufmg.br
[1]Laboratory of Cellular and Molecular Genetics, Department of General Biology, Institute of Biological Sciences, Federal University of Minas Gerais, CEP 31270-901 Belo Horizonte, Minas Gerais, Brazil
Full list of author information is available at the end of the article

completely sequenced of which Cp1002 is the first sequenced genome [3, 9–14]. Sequencing of several new strains are ongoing in our laboratory.

Recently the Cp31 strain that was originally sequenced using the SOLiD v3 platform and mate-pair library [9], was re-sequenced using Ion PGM platform [15]. This new sequencing discovered a new ~91 Kbp fragment in the Cp31 genome that is not present in NCBI. Therefore, there are possibilities that some of the available *Cp* genomes in NCBI may be incomplete and warns resequencing, reassembly, and minimization or closing gaps.

Due to the presence of highly repetitive regions that code for phage sequences, transposons, plasmid, and ribosomal RNA (rRNA) [16] in genomes and lack of good assemble software, finishing of assemblies is most critical step in genome assembly process [17]. Several strategies have been used to perform the scaffold based assemble process, for example: (i) scaffolding by reference, (ii) scaffolding by mate-pair libraries, or (iii) scaffolding by optical maps.

In the reference strategy, the contigs are oriented and positioned based on similar regions in a reference genome. This is a cost effective and a totally *in silico* method that can be executed through scaffolding software such as CONTIGuator [18] or Mauve [19], in addition to closing gaps software, like MapRepeat [20]. However, this strategy is not able to detect large sequence modifications, *e.g.*, large inversions detected between operons rRNA [21] or large chromosomal rearrangement [22] among others. The scaffolding by mate-pair libraries uses the distance of paired reads present in the contigs extremities to detect their orders. SSPACE [23] and GapFiller [24] like software can perform scaffolding and gap closing using paired data. The typical values for paired distances are 3 Kbp, 6 Kbp, 8 Kbp or 20 Kbp. However, if the length of the repetitive regions is bigger than the paired reads distance, the software cannot perform the scaffolding process [25].

On the other hand, whole-genome mapping (WGM), also known as optical mapping, uses images of unique DNA molecules immobilized in a polarized glass surface. The molecules are digested *in situ* by restriction enzymes, fragments sizes are calculated, and the high-resolution physical restriction map are used to determine the fragments order [26, 27]. Thus, optical mapping is considered one of the most accurate techniques to perform contigs scaffolding and it has been used to finishing several bacterial genomes [28]. The WGM technique uses Argus system (OpGen Inc, Gaithersburg, MD) that can be divided into four steps: (i) Extraction of chromosomal DNA, (ii) immobilization and *in situ* restriction digestion, (iii) image capture and measurement, and (iv) map assembly and analysis [26].

Recently, optical mapping has been largely used with success to detect genetic inversions in bacterial genomes. For example, WGM was used to detect a large genetic inversion between two Methicillin-resistant *Staphylococcus aureus* strains [29]. In a long-term evolution experiment, WGM was combined with genome sequencing (WGS) and PCR to analyze rearrangements in twelve *Escherichia coli* populations propagated in a glucose-limited environment for over 25 years [22]. In this experiment, they detected 19 inversions where three inversions found to have sizes larger than one-half of the chromosome. Thus, WGM can be considered to detect large rearrangements and mismatches in assemblies.

Corynebacterium pseudotuberculosis strain 1002

Corynebacterium pseudotuberculosis strain 1002 (Cp1002) was isolated from a *Caprine caseosus* in Curaça county, state of Bahia (Brazil) in 1971 [30]. Cp1002 was the first strain of this species sequenced in Brazil and its genome is used as a model for several studies of caseous lymphadenitis. Thus, this strain is considered to be representative for the *ovis* biovar and important for caseous lymphadenitis researches in Brazil.

The first sequencing of Cp1002 was performed using 454 Roche and Sanger that showed a circular genome with ~2.35 Mbp, G + C content of 52.2 %, 12 rRNA, 48 tRNA, 2,095 CDS, and 47 pseudogenes [13]. To finish the Cp1002 assembly, it was used the genetic order of *Corynebacterium* species with high similarity [13]. None experimental strategy was used to contigs scaffolding. Therefore, it is possible that mis-assemblies remained in the submitted genome of Cp1002 available in NCBI. Because of its importance in studies of caseous lymphadenitis, and after the results obtained previous studies [15], we consider Cp1002 as the candidate for a new sequencing in order to detect possible mis-assemblies.

In this work, we perform a resequencing of Cp1002 using the platform Ion PGM. We also construct a restriction map using the WGM technique (OpGen Inc, Gaithersburg, MD), and new assembly and annotation are performed. We also compared the newly obtained genome sequence with the first genome available at NCBI.

Methods

Strain and DNA isolation

Cp1002 was grown in brain-heart-infusion (BHI-HiMedia Laboratories Pvt. Ltd., India) at 37 °C under rotation. Extraction of chromosomal DNA was performed using 30 mL of 48–72 h culture of *C. pseudotuberculosis*, centrifuged at 4 °C and 4000 rpm for 15 minutes. Re-suspension of cell pellets was done in 600 μL Tris/EDTA/NaCl [10 mM Tris/HCl (pH 7.0), 10 mM EDTA (pH 8.0), and 300 mM NaCl], and transferred to tubes with beads for cell lysis using Precellys (2 cycles of 15 seconds at 6500 rpm with 30 seconds between them). Purification of DNA with phenol/chloroform/isoamyl alcohol (25:24:1) was followed by precipitation with ethanol/NaCl/glycogen (2.5 v, 10 % NaCl and 1 % glycogen). The DNA was resuspended in 30 μL MilliQ water, the concentration was determined by spectrophotometer, and the DNA was visualized using 1 % agarose gel electrophoresis.

Optical mapping

First, the DNA was extracted and isolated using Argus Sample Preparation Kit and Agencourt Genfind v2 DNA Isolation Kit. The DNA was immobilized and digested *in situ* in a MapCard Processor using the restriction enzyme (*Kpn*I). Thereafter, the molecules were imaged by fluorescence microscopy, and processed to detect restriction sites using the image acquisition software of Argus WGM system (OpGen Inc). Lastly, the Argus assembly software (OpGen Inc) was used to calculate a consensus of a restriction map and Argus MapSolver™ software (OpGen Inc, Gaithersburg, MD) was employed to import the DNA sequence and converted to *in silico* data.

Sequencing, assembly and annotation

The genome of Cp1002 was sequenced using Ion Torrent PGM System with 200 bp sequencing kit. The analysis of quality of the reads was performed using the FastQC software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and showed a Phred value, in most cases, greater than 20. Hence, it was not applied trimming or quality steps to raw reads before assembly. The *de novo* assembly was performed using Mira 3.9.18 [31] applying the parameters "-GE:not = 16 IONTOR_SETTINGS -AS:mrpc = 100". The scaffolding and gap closing were performed with SIMBA software (http://ufmg-simba.sourceforge.net) using the report generated by the software MapSolver™ (http://opgen.com/genomic-services/softwares/mapsolver) as reference to the scaffolder. The finishing of the genome was done using CLC Genomics Workbench 7.0 (Qiagen, USA) and the Website BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi). The annotation was performed using in-house scripts to fetch the annotations of a manually curated *C. pseudotuberculosis* genome annotation database obtained in the UniProt database (http://uniprot.org). Finally, the pseudogenes were curated manually using the Artemis software [32] and the UniProt database.

Comparing assemblies

To validate and to compare the new assembly (we called as Cp1002B) with the old genome of *C. pseudotuberculosis* 1002 available at NCBI (NC_017300) (we termed as Cp1002A), we performed the alignment between the experimental restriction map (obtained by WGM) of *C. pseudotuberculosis* 1002 with Cp1002B and with Cp1002A using MapSolver™ software (default parameters were used).

Thereafter, we used a modified version of the software CONTIGuator [18] to generate a syntenic comparison between Cp1002A and Cp1002B. For this comparison, we used the complete genome in a FASTA format for both the assemblies. Additionally, the annotation file (GenBank file) of Cp1002, the Website BLAST and NR database were used to detect repetitive regions that could be involved in possible genomic rearrangements.

Results

De novo assembly and annotation

The new assembly Cp1002B on Mira showed 9 contigs through 731,481 reads, with a N50 value of 402,955 bp and a deep coverage of ~58-fold (Table 1). The genome represents a circular chromosome of 2,335,107 bp, 52.2 % of G + C content, 12 rRNA, 48 tRNA, 2,071 CDS, and 43 pseudogenes.

Comparison between assemblies of Cp1002

The alignment between the experimental restriction map of Cp1002 (obtained by WGM) and the *in silico* restriction map of Cp1002B (obtained by MapSolver™) shows that the new assembly presents a high accuracy (Fig. 1). On the other side, the alignment between the experimental restriction map of Cp1002 and the *in silico* restriction map of Cp1002A shows a large inversion with a size larger than one-half of the genome (Fig. 2).

The syntenic comparison between Cp1002A and Cp1002B (Fig. 3) shows a genetic inversion that occurs between two regions encoding ribosomal RNA. The inversion occurs between the first rRNA operon (Fig. 3c) and the last rRNA operon (Fig. 3d), both highlighted in blue color in the figures.

Discussion

Our results showed that, in the new assembly, the number of CDS and pseudogenes are less in number as compared to the first assembly (Table 2). However, we believe that the new annotations are more accurate since bigger and improved databases are used. For instance, in

Table 1 Statistics of the *C. pseudotuberculosis* 1002 new assembly

Assembler	Mira 3.9.18
Reads assembled	731,481
Contigs	9
Shortest contig	4,133
Largest contig	542,891
N50	402,955
N90	218,254
N95	147,989
Total coverage	58.63

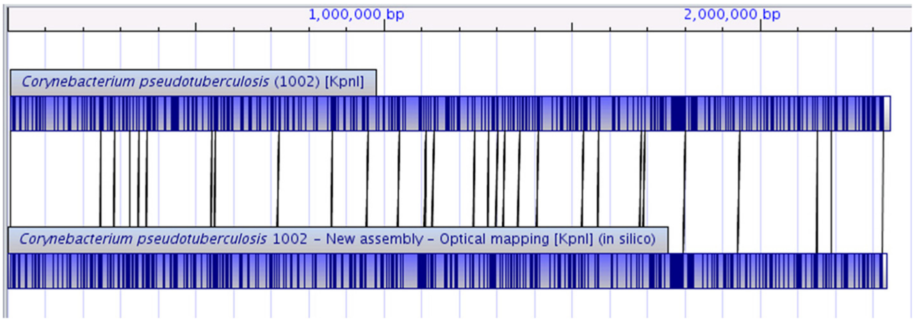

1,000,000 bp
2,000,000 bp
Corynebacterium pseudotuberculosis (1002) [KpnI]
Corynebacterium pseudotuberculosis 1002 – New assembly – Optical mapping [KpnI] (in silico)

Fig. 1 Alignment between the restriction map of *C. pseudotuberculosis* 1002 (*above*) and the *in silico* map of the new assembly of *C. pseudotuberculosis* 1002 (*below*). Both restriction maps were generated using the restriction enzyme *Kpn*I. The alignment shows a high similarity between the two restriction maps, indicating a high probability of a correct assembly

Cp1002A we detected 592 CDS as hypothetical proteins, with an average length of 617 bp. However, in Cp1002B we detected 551 hypothetical proteins, with an average length of 632 bp; thus improving the annotation. In some cases, we observed that two small hypothetical proteins join to form one large hypothetical protein. The results also showed that there is only 6 bp difference between these two assembled genomes Cp1002A and Cp1002B. Although, this value can be considered insignificant, this difference can be due to the homopolymer errors undetected in the manual frameshift curation.

Previously, it was predicted that the Cp1002 genome presented high similarity in genomic architecture, gene content and genetic order when compared to other *Corynebacterium* species [13]. Indeed, the assembly of Cp1002A was performed using reference-based assemblies techniques with short reads as well as other *Cp* strains [14]. The large inversion detected here is a mis-assembly caused by the limitations of the reference-based assembly strategies. Although genomes of the same specie tend to show high synteny, reference-based strategies cannot detect large inversions, as the mis-assembly detected in this work. Mis-assemblies in *Cp* genomes have been detected previously using mate-pair libraries [15], however it is the first time that WGM was used to correct *Cp* genome assemblies. The WGM technique is efficient to provide high accurate assemblies [22, 28, 29], and in this work, it was important to correct the assembly of Cp1002.

Furthermore, we detected a large inversion between two operons that encodes rRNA. The genome of Cp1002A presents a high synteny with other *Cp* strains [13]. However, Cp1002B shows a large inversion. Occurrences of large inversions are reported in several bacterial species [21, 22, 29]. Before the age of modern techniques for constructions of optical mapping, it was established the genome map of *Salmonella paratyphi* A using four endonucleases, *XbaI*, *I-CeuI*, *AvrII* (*Bln*I), and *SpeI* to generate fragments that could be compared [21]. They also compare the results with maps of other *Salmonella* species, and detect an inversion of half the genome between rRNA operons *rrnH* and *rrnG*. They postulated that the presence of this inversion is due to homologous recombination between the ribosomal genes. Another work proposed that the mechanism of producing chromosomal

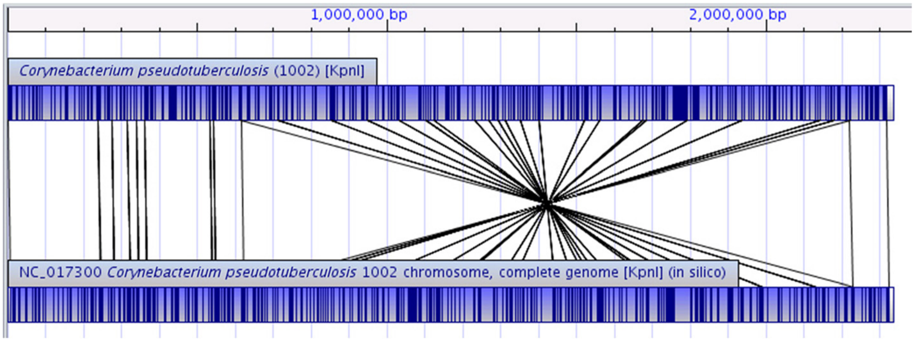

1,000,000 bp
2,000,000 bp
Corynebacterium pseudotuberculosis (1002) [KpnI]
NC_017300 *Corynebacterium pseudotuberculosis* 1002 chromosome, complete genome [KpnI] (in silico)

Fig. 2 Alignment between the restriction map of *C. pseudotuberculosis* 1002 (*above*) and the *in silico* map of the complete genome of *C. pseudotuberculosis* 1002 (NC_017300) obtained from NCBI database (*below*). Both the restriction maps were generated using the restriction enzyme *Kpn*I. The alignment shows a large inversion between the two restriction maps. A detailed analysis using CLC Genomics Workbench 7, BLAST and NR database shows that the inversion occurs between two rRNA regions

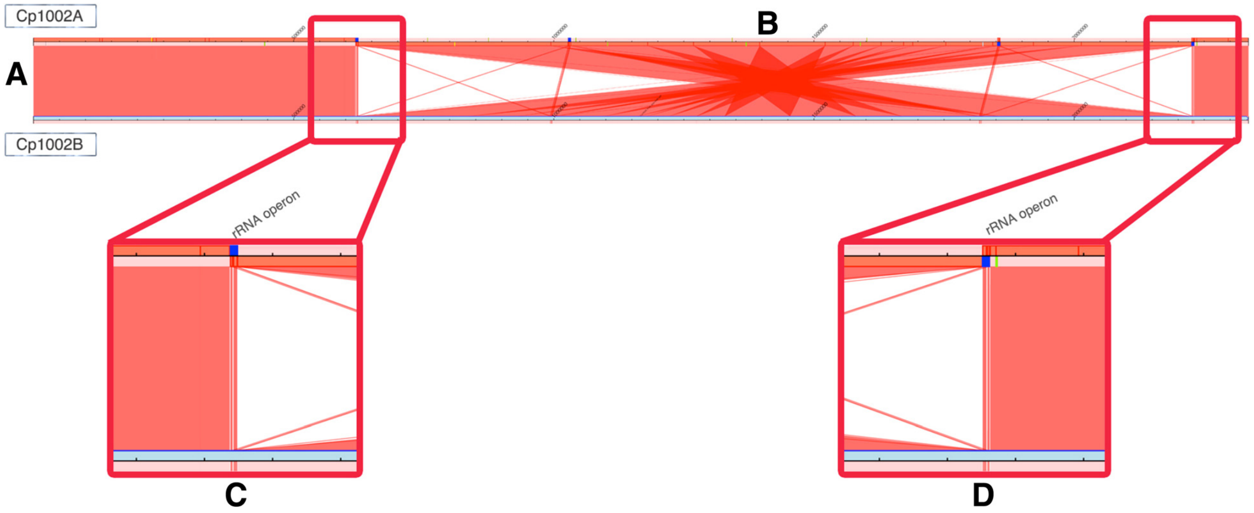


Fig. 3 Syntenic comparison between the first assembly (Cp1002A) and the new assembly (Cp1002B). **a** The genome of Cp1002A is showed above, while the genome of Cp1002B in shown below. Red lines linking the line above and the line below indicate syntenic regions. The annotation of Cp1002A was used to insert color targets in the graph that detect repetitive regions: blue for rRNA operons, light blue for transposons, yellow for plasmids and green for phages. **b** The genomes are highly similar, except by a genetic inversion larger than 1 Mbp between two rRNA operons. **c** rRNA operon in the left side of the genetic inversion. It is possible to detect a change in the sense strand after the rRNA operon that indicates an inversion. **d** rRNA operon in the right side of the inversion sequence

rearrangements is recombinational exchanges between homologous sequences, as found in ribosomal operon, similar to our observation here [33]. The large inversion detected between two rRNA operons in Cp1002 is not reported in *Cp* genome strains belong to *ovis* biovar.

Conclusions

Our new assembly (GenBank accession no. CP012837) was performed through a *de novo* strategy validated by experimental evidence (WGM), while the older assembly was performed by reference strategy. Thus, the new assembly corrected a large mis-assemble in Cp1002 genome that was not detected in the previous sequencing and assembly projects. Our optical mapping detected a large inversion between two rRNA operons in *Corynebacterium pseudotuberculosis* strain 1002. Inversion in *Cp* genome strains belong to *ovis* biovar are not reported so far but may be detected if we use WGM technique. However, the real effects of such major changes in the bacterial DNA need further evaluation.

Table 2 Comparison between the assemblies of *C. pseudotuberculosis* 1002: Cp1002A (first assembly) and Cp1002B (new assembly)

	Cp1002A	Cp1002B
Genome length	2,335,113 bp	2,335,107 bp
CDS	2,095	2,071
Hypothetical proteins	592	551
Pseudogenes	47	43
Depth coverage	31x	58x
GC %	52.2 %	52.2 %
rRNAs	12	12
tRNAs	48	48

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The genome sequence for *C. pseudotuberculosis* 1002 (Cp1002B) has been deposited in the GenBank database (accession no. CP012837).

The WGM dataset used to the Cp1002B sequence placements by MapSolver™ is included within the article (Additional file 1).

Additional file

Additional file 1: *C. pseudotuberculosis* 1002 sequence placement. This XML file contains the restriction site positions used to the sequence placements by MapSolver™. (XML 88 kb)

Abbreviations

CLA: caseous lymphadenitis; CDS: coding sequence; Cp: *Corynebacterium pseudotuberculosis*; Cp1002: *Corynebacterium pseudotuberculosis* strain 1002; Cp1002A: *Corynebacterium pseudotuberculosis* strain 1002 (first assembly); Cp1002B: *Corynebacterium pseudotuberculosis* strain 1002 (new assembly); Cp31: *Corynebacterium pseudotuberculosis* strain 31; NCBI: National Center for Biotechnology Information; PCR: polymerase chain reaction; WGM: whole-genome mapping; WGS: whole-genome sequencing.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DCBM: wrote the manuscript; TJS, FLP, FA, DB, ACP, SSH, TDLS, AFC, CAGL, HCPF, AS, RTJR and VACA gave insights about the manuscript; FR and FAD performed the experiments; DCBM, TJS and FLP: performed bioinformatics analysis; VACA, RTJR, AS and HCPF: designed and coordinate of experiments; all authors read and approved the final manuscript.

Acknowledgements

The authors thank the Ministério da Pesca e Aquicultura da Republica Federativa do Brasil and the funding agencies: Coordenação de Aperfeiçoamento de Pessoal de Pessoal de Nível Superior (CAPES), Fundação Amazônia de Amparo a Estudos e Pesquisas do Pará (FAPESPA), Fundação de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Author details

[1]Laboratory of Cellular and Molecular Genetics, Department of General Biology, Institute of Biological Sciences, Federal University of Minas Gerais, CEP 31270-901 Belo Horizonte, Minas Gerais, Brazil. [2]National Reference Laboratory for Aquatic Animal Diseases of Ministry of Fisheries and Aquaculture, Federal University of Minas Gerais, CEP 31270-901 Belo Horizonte, Minas Gerais, Brazil. [3]Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, WB 721172, India. [4]Institute of Biological Sciences, Federal University of Pará, Belém, Pará, Brazil.

Received: 17 December 2015 Accepted: 22 April 2016
Published online: 30 April 2016

References

1. Dorella FA, Carvalho Pacheco L, Oliveira SC, Miyoshi A, Azevedo V. *Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. Vet Res. 2006;37:201–18.
2. Aleman M, Spier SJ, Wilson WD, Doherr M. *Corynebacterium pseudotuberculosis* infection in horses: 538 cases (1982–1993). J Am Vet Med Assoc. 1996;209:804–9.
3. Soares SC, Silva A, Trost E, Blom J, Ramos R, Carneiro A, Ali A, Santos AR, Pinto AC, Diniz C, Barbosa EGV, Dorella FA, Aburjaile F, Rocha FS, Nascimento KKF, Guimarães LC, Almeida S, Hassan SS, Bakhtiar SM, Pereira UP, Abreu VAC, Schneider MPC, Miyoshi A,Tauch A, Azevedo V. The Pan-Genome of the Animal Pathogen *Corynebacterium pseudotuberculosis* Reveals Differences in Genome Plasticity between the Biovar ovis and equi Strains. PLoS One. 2013;8:e53818.
4. Paton M, Walker S, Rose I, Watt G. Prevalence of caseous lymphadenitis and usage of caseous lymphadenitis vaccines in sheep flocks. Aust Vet J. 2003;81:91–5.
5. Williamson L. Caseous lymphadenitis in small ruminants. Vet Clin North Am Food Anim Pract. 2001;17:359–71. vii.
6. El-Metwally S, Hamza T, Zakaria M, Helmy M. Next-Generation Sequence Assembly: Four Stages of Data Processing and Computational Challenges. PLoS Comput Biol. 2013;9:e1003345.
7. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol. 2009;10:R32.
8. Metzker ML. Emerging technologies in DNA sequencing. Genome Res. 2005;15:1767–76.
9. Silva A, Ramos RTJ, Ribeiro Carneiro A, Cybelle Pinto A, de Castro Soares S, Rodrigues Santos A, Silva Almeida S, Guimaraes LC, Figueira Aburjaile F, Vieira Barbosa EG, Alves Dorella F, Souza Rocha F, Souza Lopes T, Kawasaki R, Gomes Sa P, da Rocha Coimbra NA, Teixeira Cerdeira L, Silvanira Barbosa M, Cruz Schneider MP, Miyoshi A, Selim SAK, Moawad MS, Azevedo V. Complete Genome Sequence of *Corynebacterium pseudotuberculosis* Cp31, Isolated from an Egyptian Buffalo. J Bacteriol. 2012;194:6663–4.
10. Sousa TJ, Mariano D, Parise D, Parise M, Viana MVC, Guimarães LC, Benevides LJ, Rocha F, Bagano P, Ramos R, Silva A, Figueiredo H, Almeida S, Azevedo V. Complete Genome Sequence of *Corynebacterium pseudotuberculosis* Strain 12C. Genome Announc. 2015;3:e00759–15.
11. Baraúna RA, Guimarães LC, Veras AAO, de Sá PHCG, Graças DA, Pinheiro KC, Silva ASS, Folador EL, Benevides LJ, Viana MVC, Carneiro AR, Schneider MPC, Spier SJ, Edman JM, Ramos RTJ, Azevedo V, Silva A. Genome Sequence of *Corynebacterium pseudotuberculosis* MB20 bv. equi Isolated from a Pectoral Abscess of an Oldenburg Horse in California. Genome Announc. 2014;2:e00977–14.
12. Håvelsrud OE, Sørum H, Gaustad P. Genome Sequences of *Corynebacterium pseudotuberculosis* Strains 48252 (Human, Pneumonia), CS_10 (Lab Strain), Ft_2193/67 (Goat, Pus), and CCUG 27541. Genome Announc. 2014;2:e00869–14.
13. Ruiz JC, D'Afonseca V, Silva A, Ali A, Pinto AC, Santos AR, Rocha AAMC, Lopes DO, Dorella FA, Pacheco LGC, Costa MP, Turk MZ, Seyffert N, Moraes PMRO, Soares SC, Almeida SS, Castro TLP, Abreu VAC, Trost E, Baumbach J, Tauch A, Schneider MPC, McCulloch J, Cerdeira LT, Ramos RTJ, Zerlotini A, Dominitini A, Resende DM, Coser EM, Oliveira LM, et al. Evidence for Reductive Genome Evolution and Lateral Acquisition of Virulence Functions in Two *Corynebacterium pseudotuberculosis* Strains. PLoS One. 2011;6:e18551.
14. Cerdeira LT, Carneiro AR, Ramos RTJ, de Almeida SS, D'Afonseca V, Schneider MPC, Baumbach J, Tauch A, McCulloch JA, Azevedo VAC, Silva A. Rapid hybrid de novo assembly of a microbial genome using only short reads: *Corynebacterium pseudotuberculosis* I19 as a case study. J Microbiol Methods. 2011;86:218–23.
15. Ramos RTJ, Carneiro AR, de Castro SS, Barbosa S, Varuzza L, Orabona G, Tauch A, Azevedo V, Schneider MP, Silva A. High efficiency application of a mate-paired library from next-generation sequencing to postlight sequencing: *Corynebacterium pseudotuberculosis* as a case study for microbial de novo genome assembly. J Microbiol Methods. 2013;95:441–7.
16. Bashir A, Klammer AA, Robins WP, Chin C-S, Webster D, Paxinos E, Hsu D, Ashby M, Wang S, Peluso P, Sebra R, Sorenson J, Bullard J, Yen J, Valdovino M, Mollova E, Luong K, Lin S, LaMay B, Joshi A, Rowe L, Frace M, Tarr CL, Turnsek M, Davis BM, Kasarskis A, Mekalanos JJ, Waldor MK, Schadt EE. A hybrid approach for the automated finishing of bacterial genomes. Nat Biotechnol. 2012;30:701–7.
17. Ribeiro FJ, Przybylski D, Yin S, Sharpe T, Gnerre S, Abouelleil A, Berlin AM, Montmayeur A, Shea TP, Walker BJ, Young SK, Russ C, Nusbaum C, MacCallum I, Jaffe DB. Finished bacterial genomes from shotgun sequence data. Genome Res. 2012;22:2270–7.
18. Galardini M, Biondi EG, Bazzicalupo M, Mengoni A. CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. Source Code Biol Med. 2011;6.
19. Darling AC, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004;14:1394–403.
20. Mariano DC, Pereira FL, Ghosh P, Barh D, Figueiredo HC, Silva A, Ramos RT, Azevedo VA. MapRepeat: an approach for effective assembly of repetitive regions in prokaryotic genomes. Bioinformation. 2015;11:276.
21. Liu S-L, Sanderson KE. The chromosome of *Salmonella paratyphi* A is inverted by recombination between rrnH and rrnG. J Bacteriol. 1995;177:6585–92.
22. Raeside C, Gaffe J, Deatherage DE, Tenaillon O, Briska AM, Ptashkin RN, Cruveiller S, Medigue C, Lenski RE, Barrick JE, Schneider D. Large Chromosomal Rearrangements during a Long-Term Evolution Experiment with *Escherichia coli*. mBio. 2014;5:e01377–14–e01377–14.
23. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2011;27:578–9.
24. Boetzer M, Pirovano W. Toward almost closed genomes with GapFiller. Genome Biol. 2012;13:R56.
25. Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M, Penn CW, Robinson ER, Pallen MJ. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat Rev Microbiol. 2012;10:599–606.
26. Onmus-Leone F, Hang J, Clifford RJ, Yang Y, Riley MC, Kuschner RA, Waterman PE, Lesho EP. Enhanced De Novo assembly of high throughput pyrosequencing data using whole genome mapping. PLoS One. 2013;8:e61762.
27. Neely RK, Deen J, Hofkens J. Optical mapping of DNA: single-molecule-based methods for optical mapping of D. Wiley Online Libr. 2011;95:298–311.
28. Latreille P, Norton S, Goldman BS, Henkhaus J, Miller N, Barbazuk B, Bode HB, Darby C, Du Z, Forst S, Gaudriault S, Goodner B, Goodrich-Blair H, Slater S. Optical mapping as a routine tool for bacterial genome sequence finishing. BMC Genomics. 2007;8:321.
29. Shukla SK, Kislow J, Briska A, Henkhaus J, Dykes C. Optical Mapping Reveals a Large Genetic Inversion between Two Methicillin-Resistant *Staphylococcus aureus* Strains. J Bacteriol. 2009;191:5717–23.

30. Meyer R, Carminati R, Cerqueira RB, Vale V, Viegas S, Martinez T, Nascimento I, Schaer R, Silva JA, Ribeiro M. Evaluation of the goats humoral immune response induced by the *Corynebacterium pseudotuberculosis* lyophilized live vaccine. Rev Ciênc Médicas E Biológicas. 2002;1:42–8.
31. Chevreux B, Wetter T, Suhai S. Genome sequence assembly using trace signals and additional sequence information. In: German conference on bioinformatics. 1999. p. 45–56.
32. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics. 2012;28:464–9.
33. Anderson P, Roth J. Spontaneous tandem genetic duplications in *Salmonella typhimurium* arise by unequal recombination between rRNA (rrn) cistrons. Proc Natl Acad Sci. 1981;78:3113–7.

II.I.9 A novel comparative genomics analysis for common drug and vaccine targets in *Corynebacterium pseudotuberculosis* and other CMN group of human pathogens.
Barh D, Jain N, Tiwari S, Parida BP, D'Afonseca V, Li L, Ali A, Santos AR, Guimarães LC, de Castro Soares S, Miyoshi A, Bhattacharjee A, Misra AN, Silva A, Kumar A, **Azevedo V**.
Chem Biol Drug Des. 2011 Jul;78(1):73-84. doi: 10.1111/j.1747-0285.2011.01118.x. Epub 2011 May 25.

Chem Biol Drug Des 2011; **78:** 73–84

Research Article

© 2011 John Wiley & Sons A/S
doi: 10.1111/j.1747-0285.2011.01118.x

A Novel Comparative Genomics Analysis for Common Drug and Vaccine Targets in *Corynebacterium pseudotuberculosis* and other CMN Group of Human Pathogens

Debmalya Barh[1,2,*], Neha Jain[1,3], Sandeep Tiwari[1,3], Bibhu Prasad Parida[1,2,], Vivian D'Afonseca[4], Liwei Li[5], Amjad Ali[4], Anderson Rodrigues Santos[4], Luís Carlos Guimarães[4], Siomar de Castro Soares[4], Anderson Miyoshi[4], Atanu Bhattacharjee[6], Amarendra Narayan Misra[2], Artur Silva[7], Anil Kumar[3] and Vasco Azevedo[4]

[1]*Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology, Nonakuri, Purba Medinipur, West Bengal, India*
[2]*Department of Biosciences and Biotechnology, School of Biotechnology, Fakir Mohan University, Jnan Bigyan Vihar, Balasore, Orissa, India*
[3]*School of Biotechnology, Devi Ahilya University, Khandwa Rd., Indore, India*
[4]*Laboratório de Genética Celular e Molecular, Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, CP 486, CEP 31270-901, Belo Horizonte, Minas Gerais, Brazil*
[5]*Department of Biochemistry and Molecular Biology, Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN, USA*
[6]*Department of Biotechnology and Bioinformatics, Bioinformatics Laboratory, North Eastern Hill University, Shillong, India*
[7]*Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém-PA, Brazil*
**Corresponding author: Debmalya Barh, dr.barh@gmail.com*

Caseous lymphadenitis is a chronic goat and sheep disease caused by *Corynebacterium pseudotuberculosis* (*Cp*) that accounts for a huge economic loss worldwide. Proper vaccination or medication is not available because of the lack of understanding of molecular biology of the pathogen. In a recent approach, four *Cp* (CpFrc41, *Cp1002*, *CpC231*, and *CpI-19*) genomes were sequenced to elucidate the molecular pathology of the bacteria. In this study, using these four genome sequences along with other eight genomes (total 12 genomes) and a novel subtractive genomics approach (first time ever applied to a veterinary pathogen), we identified potential conserved common drug and vaccine targets of these four *Cp* strains along with other *Corybacterium, Mycobacterium* and *Nocardia* (CMN) group of human pathogens (*Corynebacterium diphtheriae* and *Mycobacterium tuberculosis*) considering goat, sheep, bovine, horse, and human as the most affected hosts. The minimal genome of *Cp1002* was found to consist of 724 genes, and 20 conserved common targets (to all *Cp* strains as well as CMN group of pathogens) from various metabolic pathways (13 from host-pathogen common and seven from pathogen's unique pathways) are potential targets irrespective of all hosts considered. ubiA from host-pathogen common pathway and an ABC-like transporter from unique pathways may serve dual (drug and vaccine) targets. Two *Corynebacterium*-specific (mscL and resB) and one broad-spectrum (rpmB) novel targets were also identified. Strain-specific targets are also discussed. Six important targets were subjected to virtual screening, and one compound was found to be potent enough to render two targets (cdc and nrdL). We are currently validating all identified targets and lead compounds.

Key words: *Corynebacterium pseudotuberculosis*, drug and vaccine targets, minimal genome, subtractive genomics

Received 13 December 2010, revised 3 March 2011 and accepted for publication 6 March 2011

Corynebacterium pseudotuberculosis (*Cp*) is a gram-positive bacteria and an important veterinary pathogen under the genus *Corynebacterium*. Other species under this genus are *Corynebacterium diphtheriae*, an important human pathogen, and *Corynebacterium glutamicum*, an important bacterium widely used in biotechnology (1). Owing to the pathogenic impacts and biological relevance, several *Corynebacterium* genomes including *C. diphtheriae*, *C. efficiens*, *C. urealyticum*, *C. aurimucosum* have been sequenced long back. The genus *Corynebacterium* belongs to the CMN group (2,3) that harbors species physiologically and ecologically heterogeneous although they share some common characteristics including a specific cell wall organization composed of peptidoglycan, arabinogalactan, and mycolic acid polymers (1,4), and having high G + C content in their genome (5,6).

Corynebacterium pseudotuberculosis infection causes the disease known as caseous lymphadenitis (CLA) in goat and sheep, a chronic contagious disease characterized by abscess formation in superficial lymph nodes and in subcutaneous tissues, and in severe cases, it infects the lungs, kidneys, liver, and spleen, threatening the life of the infected animal (7,8). CLA is prevalent around the world but extensively present in regions of intensive husbandry (9,10) including Australia (11), Brazil (12), New Zealand, South Africa, the USA, Israel (13), and the UK (14–16). CLA accounts for a significant economic loss by hindering the production of wool, leather, meat, and milk yields (8,11), decreasing reproductive efficiencies of affected animals, condemnation of carcasses and skins (17,18), culling of affected animals, and mortality from the internal environment (8).

Although *C. pseudotuberculosis* was originally identified as the causative microorganism of CLA in sheep and goats, this bacterium has also been isolated from other species, including horses, in which it causes ulcerative lymphangitis and pigeon fever in cattle, camels, swine, buffaloes, and humans (7,9,19). Ulcerative lymphangitis is one of the most common and economically deleterious infectious diseases of horses in California and is increasing in other dry, western states of the USA. Its onset is slow, leading to painful inflammation, nodules, and ulcers, especially in the regions of fetlock.

The pathogen also infects humans although very few cases are reported and most are because of occupational exposure with symptoms of lymphadenitis and abscesses. About 25 human cases have been reported in Australia (6,20).

In cattle, the pathogen is transmitted through the ingestion of contaminated food and water, through wounds on the skin surface, or by aerosol infection of the lungs (21). Early diagnosis is quite difficult, and no specific diagnostic test is available, increasing the severity of the disease owing to delay in treatment. Notably, an enzyme-linked immunoassay for the detection of phospholipase D is used for diagnosis, although sensitivity of the test remains unsatisfactory (22). There is also no available vaccine that can effectively prevent CLA. Some available bacterin-toxoid vaccines are capable of decreasing the prevalence and number of abscesses in the host; however, it is very difficult to maintain the efficiency of the vaccine for a long time, and the immunization efficacy also varies depending on the type of host (23,24).

Similarly, there is no drug available to effectively control the infection. Although the pathogen shows sensitivity to many drugs *in vitro*, *Cp* is reported to be highly resistant to penicillin (25) and several other drugs. The treatment approaches are difficult as the bacteria remains protected inside the host abscesses. Therefore, antibiotics are generally ineffective and are not recommended. Approach with antimicrobial chemotherapy is also not fully effective because of an inadequate drug delivery system that cannot cross the abscesses layer (26).

In a recent approach to elucidate the molecular biology and the pathogenicity of *Cp*, three strains (*Cp1002*:isolated from Brazilian goat, *CpC231*: isolated from Australian sheep, and *Cp I-19*: isolated from Israel dairy cow) were completely sequenced, and based on comparative genomics study, it is reported that both strains share common features including similar G + C content, gene density, and some distinguishing features including genome size, number of genes, and pseudogenes (27,28). The human isolate *CpFrc41* genome is already available in NCBI (NC_014329) that was also sequenced by the same group.

Dorella *et al.* in 2006 (6) have employed some genomic approaches to develop effective control measures to prevent the disease, but CLA still remains uncontrolled nevertheless. Therefore, we took advantage of the recently sequenced genome to identify common and effective drug and vaccine targets at the genomic level against strains of *Cp* that could be useful to design drugs and vaccines against the pathogen and to prevent or treat the disease especially in goat, sheep, bovine, horse, and human hosts. We have also included CMN group of two human pathogens (*C. diphtheriae* and *Mycobacterium tuberculosis*) and one industrially important microbe (*C. glutamicum*) in this analysis to identify broad-spectrum conserved potential targets that can be useful to develop a common drug and/or vaccine regardless of the pathogen or host.

Subtractive genomics approaches have successfully been used to identify targets in various human bacterial pathogens including *Pseudomonas aeruginosa* (29), *Helicobacter pylori* (30), *Burkholderia pseudomalleii* (31), *M. tuberculosis* (32), *Neisseria gonorrhoeae* (33), and *Salmonella typhi* (34) among others. This study is the first ever to apply the subtraction approach at a genomic level to identifying drug and vaccine targets specifically in *Cp*, a veterinary pathogen.

Materials and Methods

Genomes and identification of essential genes in Cp1002

We employed comparative and subtractive genomics approaches following a modified method as described by Barh and Kumar (33) on the strategy that a target will be an essential survival gene for the pathogen, which is non-homologous to its hosts. Twelve genomes were used in this study where the *Cp1002*, *CpC231,* and *CpI-19* were new, and other genomes (*CpFrc41*, *C. diphtheriae*, *C. glutamicum*, *M. tuberculosis*, goat, sheep, bovine, horse, and human) were accessed from NCBI genome server. We took the advantage of comparatively smallest genome of Brazilian strain *Cp1002* among the three *Cp* strains to identify essential genes and targets of the pathogen. In brief, each gene and protein sequence of the *Cp 1002* were subjected to BLASTx (35) and BLASTp (36), respectively, against the Database of Essential Genes (DEG: http://tubic.tju.edu.cn/deg) (37) to identify all essential genes of the strain and to map the minimal genome. Essential genes were shortlisted based on cutoff values for bit score, *E*-value, and percentage of identity at amino acid level, respectively, >100, $E = 0.0001$, and >35%. *E*-value is the 'Expect value' that describes the expected number of 'hits' to see by chance when searching a database of a particular size. The lower the *E*-value, the more significant the match is. In few cases, genes having <100 bits score and >25% to <35% identity were also selected where the query gene of *Cp1002* showed same gene name and functioned against a DEG listed essential gene hit. Proteins <100 amino acids were also included in the selection criteria. Selected essential genes were classified according to Clusters of Orthologous Groups of Proteins (COGs) nomenclature based on the comparative genomics with *CpFrc41*, *C. diphtheriae,* and *M. tuberculosis* using corresponding pathogen genomes available in NCBI.

Localization, pathogenic island (PAI), and core gene prediction

Membrane, potentially surface exposed (PSE), secreted, and cytoplasmic localization prediction of essential *Cp1002* proteins was carried out using SurfG+ (http://genome.jouy.inra.fr/surfgplus/) (a new tool under evaluation), and the results were cross-checked with tools used by Barh and Kumar (33). List of PAI-related *Cp* proteins and pangenomics-based identified core, accessory, and dispensable genes of *Cp* were prepared based on the study of D'Afonseca *et al.* (27) and PIPS software (http://www.genoma.ufpa.br/lgcm/pips) developed by Soares, S.C.; Abreu, V.A.C.; McCulloch, J.A.; D'Afonseca, V.; Ramos, R.T.J.; Silva, A.; Baumbach, J.; Trost, E.; Tauch, A.; Hirata-Jr., R.; Mattos-Guaraldi, A.L.; Miyoshi, A.; Azevedo, V. (unpublished data).

Genome subtraction for target identification in Cp1002

To subtract essential non-host homologs (potential targets) of *Cp1002,* we performed BLASTp against sheep, goat, and bovine genomes in NCBI BLAST server. Additionally, GoSh DB (http://www.itb.cnr.it/gosh) was used for goat and sheep. BLASTp was performed using each selected essential protein sequence of *Cp* at *E*-value cutoff $E = 1$ (for GoSh DB, 1e−1). Sequences that showed similarity with any of the selected hosts were eliminated, and sequences without homology (non-host homologs) were considered as putative targets at this initial stage of screening.

Identified targets were also screened against horse and human genomes using horse and human BLASTp at NCBI server with default parameters (*E*-value cutoff $E = 1$) to identify sequence similarity, respectively. The human genome was considered to avoid possible off targeting side-effects. In the results section, goat-, sheep-, and bovine-specific common targets have been grouped together and horse- and human-specific targets are represented separately as appropriate.

Common targets identification in Cp1002, CpC231, CpFrc41, CpI-19, *and other CMN group of pathogens*

To identify targets from the Australian sheep isolate *CpC231*, human isolate *CpFrc41*, and bovine isolate *CpI-19*, we employed a strategy to find whether the identified targets of goat isolate *Cp1002* were similar or identical to *CpC231*, *CpFrc41,* and *CpI-19* by aligning the amino acid sequences of identified essential proteins of *Cp1002* with the corresponding *CpC231* and *CpFrc41* sequences based on names and using BLAST. We also used the BLAST program available in http://corynecyc.cebio.org database for the same purpose. The selected *Cp1002* targets in the previous step that showed high similarity (∼80% identity at $E = 0.0001$) with corresponding *CpC231*, *CpFrc41,* and *CpI-19* protein sequence were selected as common targets for all these four strains (*Cp1002*, *CpC231*, *CpFrc41*, and *CpI-19*), while *Cp1002* proteins that did not show such homology were selected as putative targets for only *Cp1002*. Each identified *CpC231*, *CpFrc41*, and *CpI-19* target sequence was further subjected to DEG BLAST for cross-check. To assess whether the identified common *Cp* targets were essential genes or targets in other *Corynebacterium* and CMN group of species, the non-pathogenic *C. glutamicum* and human pathogens *C. diphtheriae* and *M. tuberculosis* genomes were analyzed following the method applied in the case of *CpC231*, *CpFrc41,* and *CpI-19*. Therefore, in this way, identified targets are common to all pathogens considered having a broad host range.

Metabolic pathway analysis

As goat, sheep, and horse metabolic pathways are not available, we presumed that the bovine pathways were sufficiently similar to these hosts. Host-pathogen common and pathogen-specific unique metabolic pathway–related targets were identified using a cross-species pathway comparison module available at http://corynecyc.cebio.org, selecting pathways for bovine, human, *Cp1002*, and *CpC231*. Owing to high similarities in genomic context with *Cp*, *C. diphtheriae* pathways from kyoto encyclopedia of genes and genomes (KEGG) (38) were utilized as reference for *Cp*. Bovine and human metabolic pathways from KEGG were also used as references for hosts. Pathways and related *Cp* targets were selected based on the following selection criteria: (1) The target must be an essential non-host homolog where hosts are goat, sheep, bovine, horse, or human. (2) Target should be a core gene of the pathogen. (3) A target is preferable if it is involved in pathogen's unique pathway. (4) A better target will be involved in more than one pathogen's unique pathways. (5) A pathway will be considered better if it consists of multiple targets. (6) An enzyme target should not be of same class of protein, and the EC. No. of the target should not match with any protein product of the host in host-pathogen's common pathways. (7) Pathogen-specific unique pathway targets that are common to all *Cp* strains as well as other pathogens considered are better for broad-spectrum targets. (8) Targets that are only present in *Cp1002*-specific pathway but not in *CpC231* or *CpFrc41* can be considered as *Cp1002*-specific targets and *vise versa*. (9) Non-host homolog PAI-related or virulence proteins are better targets. (10) Secreted, PSE, membrane-exposed enzymes or transporter targets can be considered for duel purpose, i.e., developing drug and vaccine where enzyme targets are more preferable. (11) Non-human homolog targets are considered to minimize possible off target side-effects and to avoid residual drug effect and absorption, distribution, metabolism, excretion, and toxicity (ADMET) as the products of all *Cp* hosts (except horse) are human consumable, and *CpFrc41* is a human isolate. (12) Targets should be common to most of the pathogen strains as well as its related species.

3D modeling and virtual screening

The three-dimensional (3D) protein structures for *C. pseudotuberculosis* genes were built using PRIME (Version 22), a protein modeling program from Schrodinger Inc. New York, NY, USA (http://www.schrodinger.com). *Cp1002* protein sequences were used. BLAST search was carried out against RCSB PDB (http://www.rcsb.org/pdb) to identify crystal structures that have high sequence similarity to CP proteins, which will be used as potential templates for model building. In addition to BLAST sequence alignment, secondary structures of CP proteins were predicted using the PSIPRED (39) program, which were further employed by the *Prime* program to adjust and optimize the alignment between CP protein and structural templates. The built CP models were energy minimized to remove any steric clashes.

To carry out structure-based virtual screening, a compound library containing lead-like small molecules was prepared. The compound structures were obtained from the ZINC (http://zinc.docking.org/) website (40), which are commercially available from the ChemDiv Inc. (San Diego, CA, USA) (http://www.chemdiv.com). The compounds were further processed in CANVAS (Version 13) from Schrodinger Inc. New York, NY, USA to eliminate structurally similar analogs and produce a structurally diverse set of 10 000 molecules. Compounds were then docked onto protein structures using GLIDE SP (Version 56) from Schrodinger Inc. with a rigid-receptor flexible-ligand protocol. Docking was focused on the pockets identified on protein models. The docked protein/ligand complex was scored using Schrodinger's proprietary *GlideScore* scoring function. It consists of eight empirical terms that are considered essential for the binding of a ligand to a protein, which includes van der Waals energy, Coulomb energy, lipophilic term for hydrophobic effects, hydrogen-bonding interaction, metal-binding term, penalty for buried polar groups, penalty for freezing rotatable bonds, and polar interaction excluding H-bonding. The scoring function was parameterized to best correlated with the experimentally determined thermodynamic binding data. The compounds were ranked by the Glide score. The ones showing on top of the list, which have the strongest predicted binding affinities, were considered hits from virtual screening.

Results

Minimal genome of Cp1002

Using DEG-based comparative genomics, we predicted the minimal genome of *Cp1002* to consist of 724 genes (*Cp1002* has 2098 genes); therefore, 34.0% of total protein coding sequences was found to be essential for the pathogen. The number can be further reduced using various criteria, but as it is not the goal of this analysis, we chose not to do so. Screened essential genes can be categorized into 19 functional groups based on COG classification (Figure 1). While translation machinery–related genes were found to be the largest group (113 genes), RNA processing and modification class were found to be the smallest (one gene). Using subtractive genomics, a total number of 118 non-host (goat, sheep, and bovine) essential genes belonging to various classes of COGs were predicted to be targeted in this pathogen. Essential genes to non-host homolog ratios within a functional group were highest (32 genes to 17) for the unknown function class and were lowest for the energy production and conversion group (67 genes to 1) (Figure 1).

Targets in the Cp1002 *genome*

At initial target screening, considering goat, sheep, and bovine as hosts, we identified 118 targets from *Cp1002* genome. However, after we screened targets based on our criteria 2 and 6, core gene, and EC numbers, only 100 targets were selected. Among them, 48 and 32 proteins, respectively, from host-pathogen common pathways and pathogen-specific unique pathways were found as potential targets. Two conserved membrane proteins (considered as other group, as they do not fall under any pathway) and a total of 18 hypothetical proteins were identified but not involved in any pathway (data not shown).

Common targets in Cp1002, CpC231, CpFrc41, *and* CpI-19 *with respect to goat, sheep, and bovine hosts*

Following the comparative genomics approach as described in the method, using goat, sheep, and bovine as hosts, we identified 76 putative targets common to *Cp1002* and *CpC231*. When we included the *CpFrc41*, the number of common targets further reduced to 56. Three targets from common as well as unique pathways, one from other group, and all hypothetical proteins that are present in *Cp1002* were absent in *CpC231*. Similarly, 13 and six targets, respectively, from common and pathogen's unique pathways of *Cp1002* are absent in *CpFrc41*. All 18 hypothetical and two other groups of targets of *Cp1002* were also not found in *CpFrc41*. Next, we added CpI-19 genome in pathogen list and found only 15 targets were common to all these four *Cp* isolates. However, two *Cp1002* proteins are named differentially in the case of *CpI-19* (Cp1002_1094 ABC-type transporter is CpI-19 putative membrane protein, and Cp1002_1959 phosphoribose diphosphate is *CpI-19* 4-hydroxybenzoate polyprenyltransferase-like prenyltransferase). Therefore, at this initial level of target screening, 51 targets were selected that are common to all three *Cp* strains with respect to goat, sheep, and bovine as hosts (data not shown).

Common Cp *targets with respect to human and horse*

As per our selection criteria 1, we next screened these 51 targets against horse and human genomes to identify targets that are common to all *Cp* strains with respect to all five hosts (goat, sheep, bovine, horse, and human) considered in this analysis. As found earlier, there was a decrease in the number of common targets with increase in the number of strains, and the similar trend was observed with increase in number of hosts. While we considered horse along with goat, sheep, and bovine, the total number of common targets decreased to 46 and when we further included human in the host list, the number was further reduced to 38 (26 in common and 13 in unique pathways). These 38 targets can be considered to develop drug for any *Cp* strains used in this analysis (Table S1).

Conserved common targets in other CMN species

Next, as per the selection criteria 12, to determine whether all these 38 targets were common in other species of *Corynebacterium*

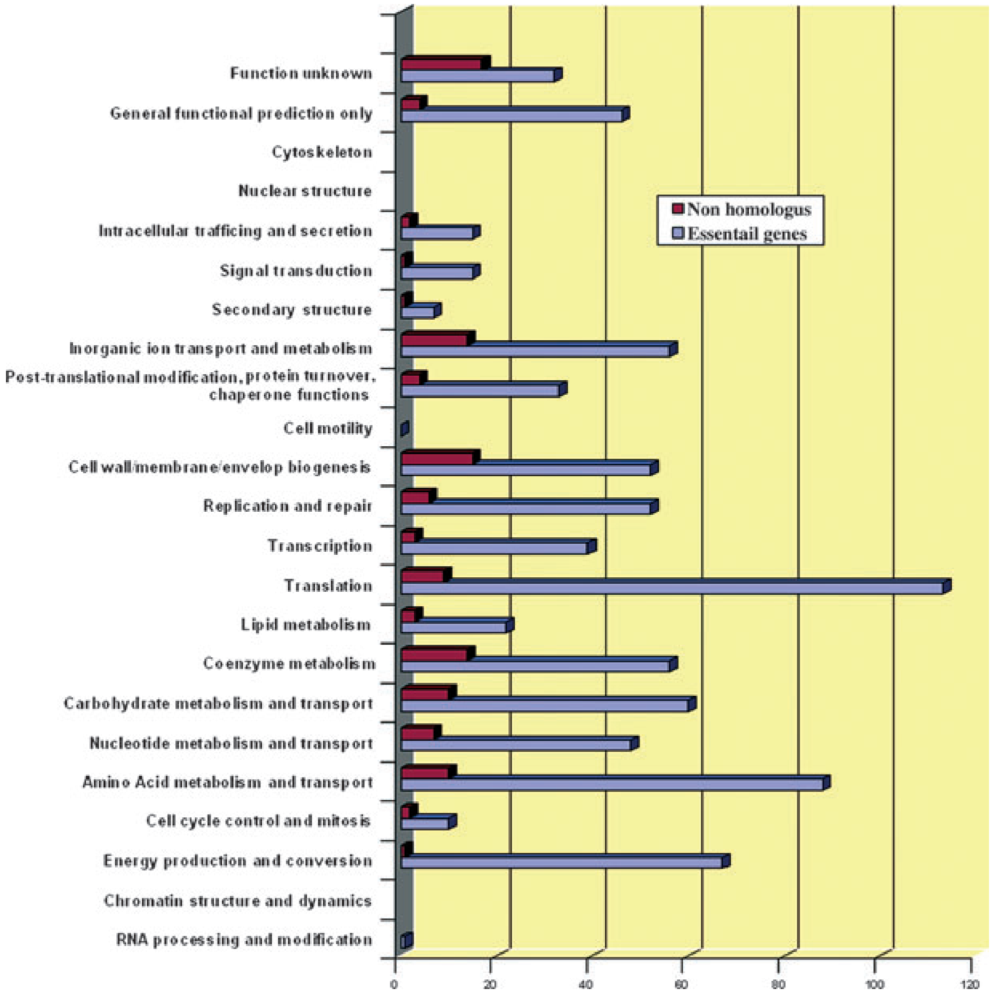


Figure 1: Clusters of Orthologous Groups of Proteins (COG) functional classification of *Cp1002* essential genes. The figure also shows the ratio of essential genes and non-host homolog genes of the species under each functional COG classes.

and CMN group of pathogens, we used similar comparative subtractive genomics approach that was used for target identification in *CpC231* and *Cp Frc41* from the list of targets identified in *Cp1002*. Selected species include the human pathogens *C. diphtheriae* (Cd) and *M. tuberculosis* (Mt). We have also considered non-pathogenic *Corynebacterium* species, *C. glutamicum* (Cg), for the same purpose. A drastic reduction in the number of targets was counted. Of 38 targets, only 20 targets were found to be common to all *Cp* strains and other CMN species. These 20 targets are non-homologous to any of the hosts (goat, sheep, bovine, horse, and human). Therefore, these targets may be used to develop broad-spectrum anti-*Cp* drugs irrespective of any host considered. Among these 20 targets, 13 (nine cytoplasmic enzymes, three ribosomal proteins, and one membrane enzyme, ubiA) belong to host-pathogen common pathways and rest (four cytoplasmic enzymes, one iron regulator ABC transporter (sufB), one membrane-located ABC-like transporter and one membrane protein) are involved in pathogens' unique metabolic pathways (Table S1).

Conserved common targets in pathogens' unique pathways

Of the 20 targets, seven targets are found to be involved in pathogen-specific unique metabolic pathways. When we applied our target selection criteria 5, as mentioned in the method, we found that peptidoglycan biosynthesis pathway was the most important pathogens' unique pathway that could be effectively targeted because of the presence of four cytoplasmic enzyme targets, namely murA, murD, murE, and murF. The next significant pathway was the transport system mainly ABC transporters. Important targets in this pathway are iron-regulated ABC-type transporter (sufB), a cytoplasmic ABC iron III transporter, and membrane-bound ABC-type transporter (*Cp1002*_1094). Membrane-localized enzyme putative lipoprotein signal peptidase (EC No: 3.4.23.36) that plays a crucial role in cell membrane/wall biogenesis and membrane transport was found to be an attractive target in all pathogens. This enzyme has been reported to be a putative target in *Aeromonas hydrophila* (41) and is also conserved in *Corynebacterium*; therefore, it can be better suited to develop anti-*Cp* drugs. As the enzyme is membrane localized, it can also be a good candidate to develop anti-*Cp* vaccine.

Conserved common targets in host-pathogen common pathways

Cytoplasmic translation machinery proteins constituted the highest number of targets (four of 13). These proteins are rpmB, rpmD, rpmL, and ribonuclease-P (rnpA). Among other targets, the most attractive one was homoserine dehydrogenase (thrA) from homoserine and lysine biosynthesis pathway. thrA is also a key enzyme in glycine, serine, and threonine metabolism pathways. Therefore, targeting thrA might block multiple essential metabolic pathways of the pathogen.

Imidazole glycerol-phosphate dehydrogenase (hisB) was identified from histidine metabolism pathway. Similarly, phosphoribose diphosphate (ubiA) in glycan metabolism pathway was found to be a broad-spectrum target. Being a membrane-located enzyme, ubiA may also serve dual purpose, i.e., drug and vaccine target.

Although, from biotin biosynthesis pathway, biotin synthase family transferase and biotin synthase (bioB) were identified as targets for all three *Cp* strains, only bioB was qualified to be a broad-spectrum target considering all pathogen genomes used in this analysis. Thiamine monophosphate kinase (thiL), dihydropteroate synthase (folP), and precorrin-4 c 11-methyl transferase (cobM), respectively, from thiamine, tetrahydrofolate, and adenosylcobalamine biosynthesis pathways were found to be attractive targets regardless any pathogen and host range considered. Two other important targets in common pathways were ribonucleotide reductase stimulatory protein (nrdL) and decxycitidine triphosphate deaminase (dcd) from, respectively, nucleotide metabolism and pyrimidine biosynthesis pathways.

Common novel targets in Cp strains

Extensive literature search was performed to identify novel targets. We considered novel targets that are not reported in any other pathogen but are common in all *Cp* strains with respect to all hosts considered. Therefore, we screened such novel targets from the list of 38 targets. In host-pathogen common metabolic pathways, such targets were cytoplasmic rplA, rpmB (from translation machinery), and membrane-located putative H+ antiporter subunit-c from ATP synthesis–coupled electron transport pathway. Although rplA and H+ antiporter subunit-c are absent in *M. tuberculosis*, rpmB was found to be a universal novel target for any pathogen considered in this analysis.

From pathogens' unique pathways, three novel targets, namely amino acid career protein (sodium and amino acid transport), mscL (cell wall biogenesis and transport), and resB (electron transport) were identified. These three targets are either membrane or PSE localized, conserved in *Corynebacteria*, and targets for all species. Therefore, these three can be used for dual purpose.

PAI-related targets

Pathogenicity island targets are attractive in developing drug/vaccine and as per our selection criteria (9), as mentioned in method, we scanned 38 targets for PAI, and only dcd was identified. dcd is found to be a common target for all pathogens and has been reported as a target in *M. tuberculosis* (42).

Targets selected for 3D modeling

To design drug, we selected some important and common targets. A total of six targets were selected. As found in the analysis, the peptidoglycan biosynthesis pathway is the most attractive pathogens' unique metabolic pathway; murA and murE were selected from this pathway. From host-pathogen common metabolic pathways, folP (tetrahydrofolate biosynthesis pathway), nrdL (nucleotide metabolism), and the sole PAI-related target, dcd (pyrimidine biosynthesis pathways) were selected. Although nrdH is not present in *C. diphtheriae*, we considered it because of its importance in redox pathway that is essential for the survival of any pathogen inside the host. nrdH has also been reported as an attractive target in *M. tuberculosis* (43).

As the experimentally determined 3D structures are not available, protein models were built using comparative modeling techniques. Protein structures are more conserved among evolutionary-related homologs. Generally, medium to high-resolution models can be obtained if the sequence identity is >30%. The sequence identity in this work ranges from 41% to 82% as shown in Table S2, which assures the quality of the models. The similarity between the model and the crystal structure on the binding site is generally even higher, especially for dcd.

Virtual screening and docking

Compounds identified from virtual screening with most favorable binding energy were considered as hits. Hits with strongest binding energy were depicted in sticks binding on the surface of the pocket (Figure 2), while the chemical structures of the top five hits for each protein were listed in Figure 3. The physicochemical properties of top five hits based on glide scores for each target protein are represented in Table S3. Important amino acid residues that interact with docked compound are list in Table S4. The hits were named from c1(gene) to c5(gene) in the order of predicted binding affinity. The inspection on the docked conformation shows that the binding cavity on the protein was explored very effectively by this top hits. Although the hits are not validated *in vitro*, it is interesting to see that the top one hit to folP, c1(folP), is actually a substructure of an antibiotic drug cefmetazole. Among the top five hits to each protein, there is one compound shared by two proteins, c5(dcd)/c1(nrdL). Although structurally speaking, the two cavities are not quite similar, because small molecules are flexible and could adopt different conformations during binding. It may render more potent antibiotic activity by targeting two essential bacterial proteins simultaneously.

Discussion

Although subtractive genomics is frequently used to identify drug targets in human pathogenic bacteria, in this study, for the first time, the approach was applied to identify drug and vaccine targets

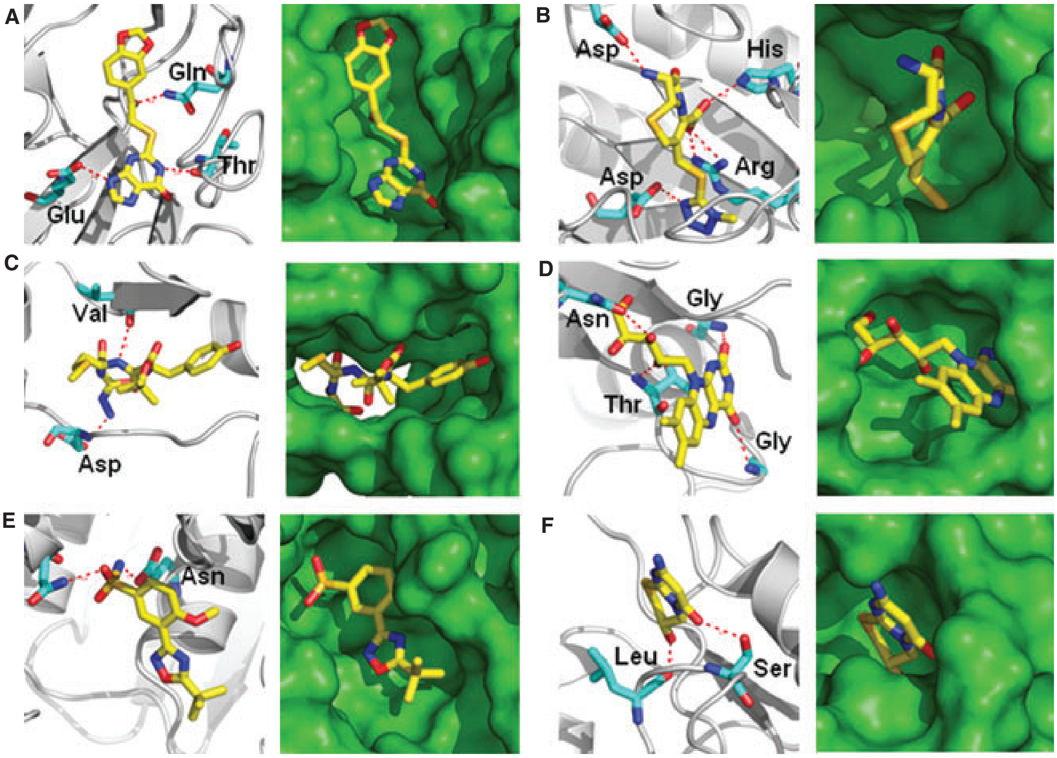


Figure 2: Ribbon and surface representation of the top compound bound to (A) dcd, (B) FolP, (C) nrdH, (D) nrdL, (E) murA, and (F) murE. The compounds are in stick representation with carbon, oxygen, and nitrogen atoms colored in yellow, red, and blue, respectively.

of a non-human pathogen. In DEG-based essential gene screening, most *Cp* hits were found with *M. tuberculosis*. During COG classification of essential *Cp* genes using two other *Corynebacterium* species, namely *C. diphtheriae* and *M. tuberculosis* proteomes available in NCBI, it was noted that *Cp* genes were shared by both species. A substantial number of *Cp* genes were conserved and present in *C. diphtheriae,* and a few genes that were not present were found in *M. tuberculosis* and *vice versa*. Essential genes for *C. diphtheriae* were not listed in DEG, but genes for *M. tuberculosis* were shown. It is interesting that while DEG listed 614 essential genes for *M. tuberculosis*, our analysis showed that the minimal genome of *Cp1002* consisted of approximately 724 genes. The higher number of essential genes in *Cp* relative to *M. tuberculosis* may be as a result of sharing and horizontal transfer of genes among CMN group of *Corynebacterium* species and other bacterial classes listed in DEG.

Polymorphic peptidoglycans are unique components that constitute the bacterial cell wall and play a vital role in bacterial defense, virulence, and survival. Therefore, the peptidoglycan biosynthesis pathways (I and II) that are unique to the bacteria are very crucial. Four cytoplasmic enzymes, murA, murD, murE, and murF, were identified as targets from this pathway. murA and murD were additionally involved in nucleotide sugar and glutamate metabolism pathways, respectively. murE and murF also were shown to play a vital role in lysine biosynthesis. While murD was conserved in *Corynebacterium*, murF was conserved in *Mycobacterium.* All four targets were previously reported in *Mycobacterium leprae* (44) and few other organisms (Table S1). In *Cp*, we found that this peptidoglycan biosynthesis pathway was the best targeting pathway as the above-mentioned four targets found here were essential non-host homologs with respect to all five hosts considered, and all targets were highly potential because of their additional involvement in multiple pathways. D-alanine is an essential component of the peptidoglycan layer in bacterial cell wall and D-alanine–D-alanine ligase (ddl) is a common target for various human pathogens in this pathway. But it is interesting that ddl was not found to be a target in *Cp.*

Bacterial transport system–related targets are attractive in developing antibiotics. Iron transport–related ABC transporters have been reported as essential genes and drug targets in *N. gonorrhoeae* (33) and *Clostridium perfringens* (45) among others. Such transporters were also predicted to be good vaccine targets because of their antigenic properties and exomembrane or PSE localization (46). We found membrane-localized ABC-type transporter (*Cp1002*_1094) and cytoplasmic iron-regulated ABC-type transporter (sufB) are broad-spectrum targets. Both of these targets have been identified in *C. perfringens* by Chhabra *et al.*, (45). *Cp1002*_1094, being a membrane protein, may be potential in developing drug as well as vaccine.

Putative lipoprotein signal peptidase (*Cp1002*_1377/lspA, EC: 3.4.23.36), which is conserved in *Corynebacterium*, was selected as an important target. It is a common target for all pathogens considered and also a non-homolog to all five hosts considered in this analysis. This target is involved in cell wall and membrane biogenesis, intracellular trafficking and secretion, membrane transport, protein export pathways, and an enzyme with the same EC number

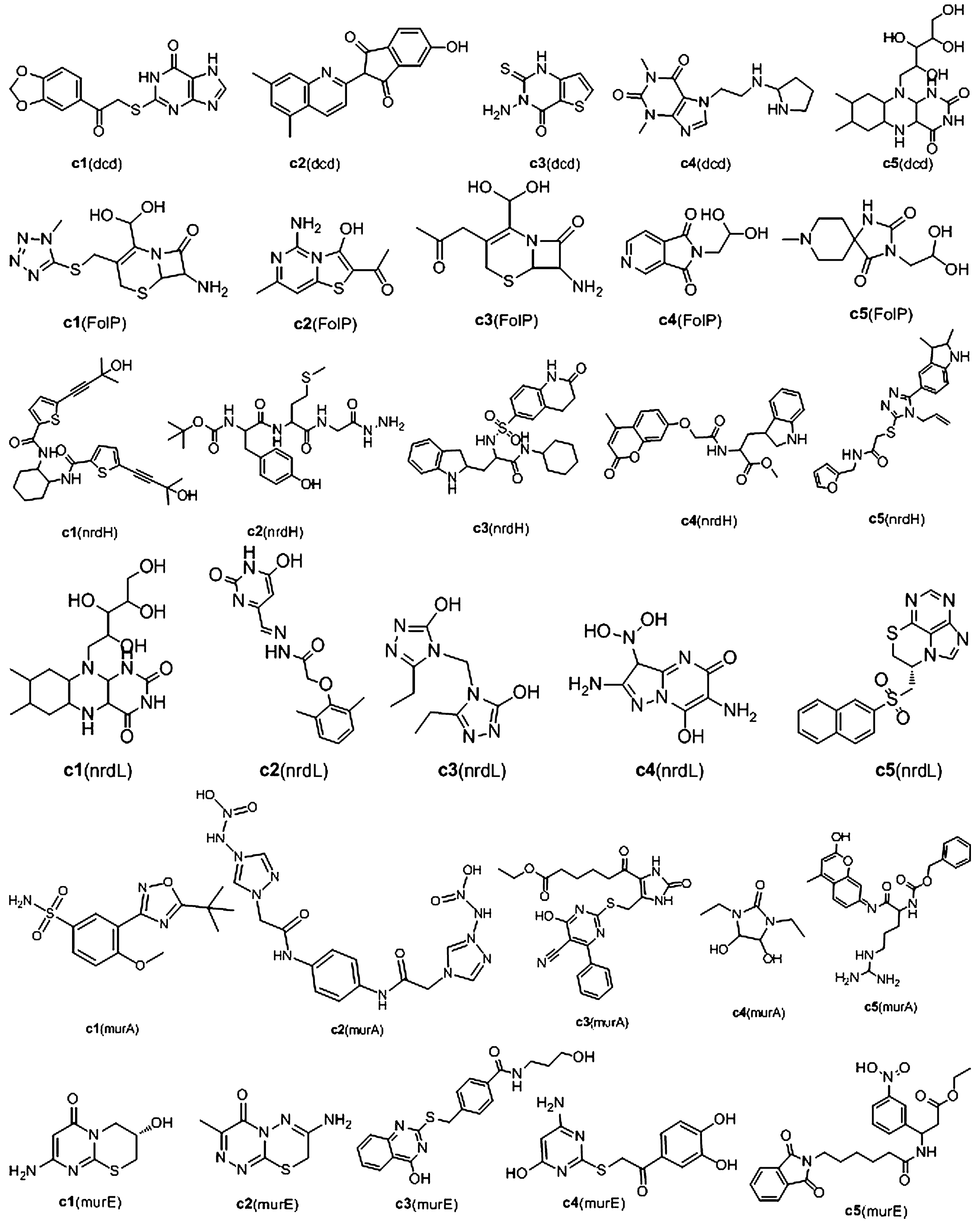


Figure 3: Chemical structures for the top five compounds predicted by GLIDE.

has been identified as a target in *A. hydrophila* (41). This protein is localized to the membrane and therefore it is also suitable for vaccine development.

Two other targets include putative amino acid carrier protein (*Cp1002*_1332, sodium transport) and large-conductance mechanosensitive channel protein (*Cp1002*_0665⁄mscL, transport, and membrane biogenesis), which are novel targets. These two targets are highly conserved in *Corynebacterium* and are membrane localized. Therefore, they were shown to be potential targets in developing both anti-*Cp* drugs and vaccines for all five hosts.

Cytochrome *C* biogenesis protein (resB) is non-homologous to all five hosts. Moreover, resB is an essential gene in *B. pseudomallei*, and it is involved in cytochrome C biogenesis. It also plays the role of an essential cofactor in oxidoreduction process (31). In this study, we also found resB as a potential novel target to inhibit the oxidoreduction process of *Cp*. resB is a PSE-localized protein; therefore, it may be potential to vaccine development. However, in *M. tuberculosis*, it is not found to be a potential target.

Bacterial two-component and secretion systems are unique pathways to bacteria and are critical for growth and survival of the organism in extreme conditions. Preprotein translocase subunit (secE) of the bacterial secretion system has been demonstrated as target in *Escherichia Coli* (47) and *N. gonorrhoeae* (33). We also found secE to be a potential target for all *Cp* strains but not in *M. tuberculosis* or *C. diphtheriae* (data not shown). Owing to its membrane localization, it may have potential for anti-*Cp* vaccine development.

From host-pathogen common pathways, several proteins were identified as potential drug and vaccine targets in *Cp*. Cytoplasmic enzymes, homoserine dehydrogenase (thrA) and homoserine kinase (thrB), which are involved in glycine, homoserine, threonine, and lysine metabolism pathways, were selected for *Cp*. Both these targets were non-homologous to all five hosts and have previously been identified as targets in *Mycobacterium* (42,48). But thrB is not found to be a target in *C. diphtheriae*.

From the histidine metabolism pathway, imidazole glycerol-phosphate dehydratase (hisB) and ATP phosphoribosyl transferase (hisG) were selected as targets in *Cp*. Both of these enzymes are identified as potential drug targets in many bacteria including *Mycobacterium* and *Pseudomonas* (42,44,49), and nitrobenzothiazole is used as an inhibitor for *M. tuberculosis* hisG (50); however, as *Cp* hisG was a partial horse homolog, it may not be a good target in developing anti-*Cp* drug for wide-ranging number of hosts.

Membrane enzyme phosphoribose diphosphate (ubiA) is involved in glycan biosynthesis and metabolism and, in *Cp*, we identified ubiA as potential target. ubiA is a reported target in *M. tuberculosis* (51), and the disruption of ubiA resulted in a complete loss of cell wall arabinan and death of *C. glutamicum* (52). We have also found that ubiA can also be targeted in *C. diphtheriae*. Being a membrane enzyme, it may also serve the dual purpose.

Cytoplasmic enzyme 1-deoxy-D-xylulose 5-phosphate reductoisomerase (dxr) is essential in the methylerythritol phosphate (MEP) pathway (53). The MEP pathway is extensively targeted in *M. tuberculosis* (53,54), and dxr is a promising target for *Mycobacterium* (55) and *Salmonella* (34). Fosmidomycin is an effective antibiotic that inhibits dxr (56). Our results also suggest that dxr is a potential target in *Cp* for goat, sheep, and bovine, but because of its partial sequence homology with horse, further analysis is required to explore its potentiality in multiple hosts.

Biotin is essential for the growth of various bacteria including *Sinorhizobium meliloti* (57); therefore, biotin biosynthesis pathways are important for bacterial survival and growth. Three cytoplasmic enzymes [biotin synthase family transferase (*Cp1002*_0903), biotin synthase (bioB), and dethiobiotin synthetase (bioD1)] were identified from biotin biosynthesis pathways for *Cp*. However, *Cp1002*_0903 is not found in both *Mycobacterium* and *C. diphtheriae*. bioB was previously reported as an essential gene as well as drug target in *S. typhi* (34) and *A. hydrophila* (41) that are human pathogens. Owing to a PAI-related protein, bioD1 is a good target against *Cp* considering sheep, goat, and bovine but not for horse as it has partial sequence to horse.

From thiamin biosynthesis pathway, thiamine monophosphate kinase (thiL), a cytoplasmic enzyme, is considered. Thus, thiL is a reported target in *M. leprae* (44) and as per our analysis, thiL is a broad-spectrum target (present in all six pathogen considered), which can also be used to develop drug for broad host range (all five hosts in this study).

Cell redox homeostasis is an essential survival mechanism for any intracellular pathogen like *Cp*. Among the several key cytoplasmic enzymes in this pathway, glutaredoxin-like protein (NrdH) was identified as an essential enzyme as well as drug target for *Cp*. NrdH is a novel redoxin in *E. coli* having thioredoxin-like activity (58) and was also found to be a good target in *Mycobacterium*(43). However, in our analysis, this was not found to be a target in *C. diphtheriae*.

FolP is an identified target in *A. hydrophila* (41) and *N. gonorrhoeae* (33). The enzyme catalyzes a condensation reaction yielding dihydropteroate, an intermediary metabolite, that is subsequently converted to tetrahydrofolic acid and is essential for the syntheses of purine, thymidylate, glycine, methionine, pantothenic acid, and *N*-formylmethionyl-tRNA. FolP was found to be a cytoplasmic enzyme and was conserved in *Corynebacterium*. Owing to the fact that the enzyme was found to be essential in the tetrahydrofolate biosynthesis pathway and to be a non-host homolog of *Cp* as applicable to other CMN species and all five hosts, it demonstrated high potentiality as an attractive target, possibly targeted by sulfonamide antibiotics.

Precorrin-4 C11-methyltransferase (cobM) is a crucial enzyme in adenosylcobalamin biosynthesis II, siroheme biosynthesis, and porphyrin and chlorophyll metabolism pathways for all bacteria. It is a potential target in *M. tuberculosis* (42) and *A. hydrophila* (41). In this analysis, we found that cobM is a universal target for all pathogens considered here.

Deoxycytidine triphosphate deaminase (dcd) is an important enzyme in the dUTP and pyrimidine deoxyribonucleotides de novo biosynthesis process. dcd was recently identified as a drug target in *Mycobacterium* (42). Here, we found dcd to be an essential gene in *Cp* and also for *C. diphtheriae* that is non-homologous to all five hosts and also associated with PAI, making dcd an attractive target in all studied pathogens.

Six novel targets consisting of three (rplA, rpmB, and H+ anti-porter subunit-c) from host-pathogen's common pathways and rest three (amino acid career protein, mscL, and resB) from pathogen's unique pathways have been identified. As a result, none except rpmB was found to be universal target because they are not present in *M. tuberculosis*. However, considering *Cp*, all unique pathway-related three targets may be used for developing anti-*Cp* drug as well as vaccine.

Five important broad-spectrum targets (murA, murE, folP, nrdL, and dcd), and *Cp*- and *M. tuberculosis*-specific nrdH were modeled and subjected to virtual screening to identify new molecular agents specific to these targets. We selected these targets because of their potentiality to be targets in other CMN group of human pathogens too, although they are not novel targets for *Cp*. Total 30 compounds, five for each target, have been identified, and one compound [c5(dcd)/c1(nrdL)] was found to be useful in targeting both dcd and nrdL. There is no specific drug available till date to treat *Cp* infection. Therefore, identified compounds can be tested for their efficacy to attain the corresponding targets toward the development of anti-*Cp* and anti-CMN drugs.

Conclusion

In this study, we identified several drug and vaccine targets that are common to four *Cp* strains (*Cp1002*, *Cp*C*231*, *CpFrc41,* and *CpI-19*). Twenty targets were found common to CMN group of pathogens including *Cp* with respect to a broad range of hosts (goat, sheep, bovine, horse, and human). It was also found that some targets can be used for all host ranges, and some are host specific. In general, the peptidoglycan biosynthesis pathway was most important for targeting, followed by ABC-type transport system. Glycan biosynthesis–related ubiA, biotin synthesis pathway enzymes bioB and thiL, cell redox homeostasis regulator NrdH, tetrahydrofolic acid biosynthesis–related folP, and dUTP- and pyrimidine deoxyribonucleotides biosynthesis–related dcd were found to be attractive targets in *Cp* with respect to all considered hosts. We also identified six novel targets that are not reported in any other bacteria, which can be used for broad host range. We also identified potential compounds for our six selected targets using virtual screening. All these targets and identified candidate lead compounds require experimental validation and consideration that the pathogen remains protected inside abscesses, thus proper delivery methods need to be developed. Several targets were found to be strain specific and some were specific to hosts. We have not considered most of the hypothetical proteins because of their strain specificity. These strain- and host-specific targets can be further explored. Currently, we are analyzing hypothetical proteins to enrich the target list. Also, we are adopting fold-level homology modeling and simulation methods for these identified targets and validating to develop broad-spectrum novel drugs and vaccines against CMN group of pathogens for a broad range of hosts.

Acknowledgments

D.B. conceived the method, designed the experiment, coordinated the entire work, and wrote the paper; D.B., N.J., and S.T. annotated and analyzed all genomes for target identification; V.D'A., A.A., A.R.S., L.C.G., S.C.S., A.M. A.S., and V.A. assembled and annotated newly sequenced genomes; D.B., A.B., and L.L. performed virtual screening; A.N.M., A.K., and V.A. monitored and provided technical guided throughout the work.

The work was mostly carried out at IIOAB; D.B., S.T., and N.J. thank IIOAB and its all collaborating Labs; S.T., N.J. had DBT's student fellowship; S.T., N.J., and A.K. acknowledge DBT's Bioinformatics facilities at DAVV; A.N.M. acknowledges the funding of DBT-BIF sanction No.BT/BI/25-001/2006 to ANM; For V.D'A., A.A., A.R.S., L.C.G., S.C.S., A.M. A.S., and V.A., this work was a part of the Rede Paraense de Genômica e Proteômica supported by Fundação de Amparo a Pesquisa do Estado do Pará (FAPESPA). V.A., A.S., and A.M., were supported by Conselho Nacional de Desenvolvimento Científico e Tecnolœgico (CNPq) and also had financial support from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Fundação de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG).

References

1. Bayan N., Houssin C., Chami M., Leblon G. (2003) Mycomembrane and S-layer: two important structures of *Corynebacterium glutamicum* cell envelope with promising biotechnology applications. J Biotechnol;104:55–67.
2. Hard G.C. (1969) Electron microscopic examination of *Corynebacterium ovis*. J Bacteriol;97:1480–1485.
3. Songer J.G., Beckenbach K., Marshall M.M., Olson G.B., Kelley L. (1988) Biochemical and genetic characterization of Corynebacterium pseudotuberculosis. Am J Vet Res;49:223–226.
4. Hall V., Collins M.D., Hutson R.A., Lawson P.A., Falsen E., Duerden B.I. (2003) *Corynebacterium atypicum* sp. nov., from a human clinical source, does not contain corynomycolic acids. Int J Syst Evol Microbiol;53:1065–1068.
5. Hard G.C. (1975) Comparative toxic effect of the surface lipid of *Corynebacterium ovis* on peritoneal macrophages. Infect Immun;12:1439–1449.
6. Dorella F.A., Pacheco L.G.C., Oliveira S.C., Miyoshi A., Azevedo V. (2006) Corynebacterium pseudotuberculosis: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. Vet Res;37:201–218.
7. Williamson L.H. (2001) Caseous lymphadenitis in small ruminants. Vet Clin North Am Food Anim Pract;17:359–371.
8. Merchant I.A., Packer R.A. (1967) The genus corynebacterium. In: Merchant I.A., Packer R.A., editors. Veterinary Bacteriology and Virology. Iowa: The Iowa State University Press; p. 425–440.

9. Brown C.C., Olander H.J., Alves S.F. (1987) Synergistic hemolysis-inhibition titers associated with caseous lymphadenitis in a slaughterhouse survey of goats and sheep in Northeastern Brazil. Can J Vet Res;51:46–49.
10. Collett M.G., Bath G.F., Cameron C.M. (1994) Corynebacterium pseudotuberculosis infections. In: Coetzer J.A.W., Thomson G.R., Tustin R.C., Kriek N.P.J., editors. Infectious Diseases of Livestock with Special Reference to Southern Africa. Cape Town: Oxford University Press; p. 1387–1395.
11. Paton M., Walker S., Rose I., Watt G. (2003) Prevalence of caseous lymphadenitis and usage of caseous lymphadenitis vaccines in sheep flocks. Aust Vet J;81:91–95.
12. Unanian M., Silva A.F., Pant K. (1985) Abscesses and caseous lymphadenitis in goats in tropical semi-arid north-east Brazil. Tropic Anim Health Prod, 17:57–62.
13. Yeruham I., Elad D., Van-Ham M., Shpigel N.Y., Perl S. (1997) *Corynebacterium pseudotuberculosis* infection in Israeli cattle: clinical and epidemiological studies. Vet Rec;140:423–427.
14. Ben Saïd M.S., Ben Maitigue H., Benzarti M. *et al.* (2002) Epidemiological and clinical studies of ovine caseous lymphadenitis. Arch Inst Pasteur Tunis;79:51–57.
15. Binns S.H., Bailey M., Green L.E. (2002) Postal survey of ovine caseous lymphadenitis in the United Kingdom between 1990 and 1999. Vet Rec;150:263–268.
16. Connor K.M., Quirie M.M., Baird G., Donachie W. (2000) Characterization of United Kingdom isolates of *Corynebacterium pseudotuberculosis* using pulsed-field gel electrophoresis. J Clin Microbiol;38:2633–2637.
17. Paton M., Rose I., Hart R. *et al*. (1994) New infection with *Corynebacterium pseudotuberculosis* reduces wool production. Aust Vet J;71:47–49.
18. Arsenault J., Girard C., Dubreuil P. *et al*. (2003) Prevalence of and carcass condemnation from maedi-visna, paratuberculosis and caseous lymphadenitis in culled sheep from Quebec, Canada. Prev Vet Med;59:67–81.
19. Ayers J.L. (1977) Caseous lymphadenitis in goats and sheep: a review of diagnosis, pathogenesis, and immunity. J Am Vet Med Assoc;171:1251–1254.
20. Peel M.M., Palmer G.G., Stacpoole A.M., Kerr T.G. (1997) Human lymphadenitis due to *Corynebacterium pseudotuberculosis*: report of ten cases from Australia and review. Clin Infect Dis;24:185–191.
21. Paton M. (1993) Control of cheesy gland in sheep. West Aust J Agric;34:31–37.
22. Menzies P.I., Hwang Y.T., Prescott J.F. (2004) Comparison of an interferon-gamma to a phospholipase D enzyme-linked immunosorbent assay for diagnosis of *Corynebacterium pseudotuberculosis* infection in experimentally infected goats. Vet Microbiol;100:129–137.
23. Fontaine M.C., Baird G., Connor K.M., Rudge K., Sales J. , Donachie W. (2006) Vaccination confers significant protection of sheep against infection with a virulent United Kingdom strain of *Corynebacterium pseudotuberculosis*. Vaccine;24:5986–5996.
24. Piontkowski M.D., Shivvers D.W. (1998) Evaluation of a commercially available vaccine against *Corynebacterium pseudotuberculosis* for use in sheep. J Am Vet Med Assoc;212:1765–1768.
25. Garg D.N., Nain S.P.S., Chandiramani N.K. (1985) Isolation and characterization of *Corynebacterium ovis* from sheep and goats. Indian Vet J;62:805–808.
26. Stanford K., Brogden K.A., McClelland L.A., Kozub G.C., Audibert F. (1998) The incidence of caseous lymphadenitis in Alberta sheep and assessment of impact by vaccination with commercial and experimental vaccines. Can J Vet Res;62:38–43.
27. D'Afonseca V., Prosdocimi F., Dorella F.A. *et al.* (2010) Survey of genome organization and gene content of *Corynebacterium pseudotuberculosis*. Microbiol Res;165:312–320.
28. Silva A., Schneider M.P., Cerdeira L. *et al.* (2010) Complete genome sequence of *Corynebacterium pseudotuberculosis* I-19, strain isolated from Israel Bovine mastitis. J Bacteriol; ;193:323–4.
29. Sakharkar K.R., Sakharkar M.K., Chow V.T.K. (2004) A novel genomics approach for the identification of drug targets in pathogens, with special reference to *Pseudomonas Aeruginosa*. In silico Biol;4:355–360.
30. Dutta A., Singh S.K., Ghosh P. *et al.* (2006) In silico identification of potential therapeutic targets in the human pathogen *Helicobacter pylori*. In silico Biol;6:43–47.
31. Chong C.E., Lim B.S., Nathan S., Mohamed R. (2006) In silico analysis of *Burkholderia pseudomallei* genome sequence for potential drug targets. In silico Biol;6:341–346.
32. Asif S.M., Asad A., Faizan A. *et al.* (2009) Dataset of potential targets for *Mycobacterium tuberculosis* H37Rv through comparative genome analysis. Bioinformation;4:245–248.
33. Barh D., Kumar A. (2009) In silico identification of candidate drug and vaccine targets from various pathways in *Neisseria gonorrhoeae*. In silico Biol;9:225–231.
34. Rathi B., Sarangi A.N., Trivedi N. (2009) Genome subtraction for novel target definition in *Salmonella typhi*. Bioinformation;4:143–150.
35. Gish W., States D.J. (1993) Identification of protein coding regions by database similarity search. Nat Genet;3:266–272.
36. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. (1990) Basic local alignment search tool. J Mol Biol;215:403–410.
37. Zhang R., Lin Y. (2009) DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. Nucleic Acids Res;37:D455–D458.
38. Kanehisa M., Goto S. (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res;28:27–30.
39. Jones D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol;292:195–202.
40. Irwin J.J., Shoichet B.K. (2005) ZINC – a free database of commercially available compounds for virtual screening. J Chem Inf Model;45:177–182.
41. Sharma V., Gupta P., Dixit A. (2008) In silico identification of putative drug targets from different metabolic pathways of *Aeromonas hydrophila*. In silico Biol;8:331–338.
42. Anishetty S., Pulimi M., Pennathur G. (2005) Potential drug targets in *Mycobacterium tuberculosis* through metabolic pathway analysis. Comput Biol Chem;29:368–378.
43. Leiting W.U., Jianping X.I.E. (2010) Comparative genomics analysis of Mycobacterium NrdH-redoxins. Microb Pathog;48:97–102.

44. Shanmugam A., Natarajan J. (2010) Computational genome analyses of metabolic enzymes in *Mycobacterium leprae* for drug target identification. Bioinformation;4:392–395.
45. Chhabra G., Sharma P., Anant A. *et al.* (2010) Identification and modeling of a drug target for *Clostridium perfringens* SM101. Bioinformation;4:278–289.
46. Barh D., Misra A.N. (2009) Scientific commons: epitope design from transporter targets in *N. gonorrhoeae*.
47. Driessen A.J.M., Haril U.F., Wickner W. (2003) The enzymology of protein translocation across the *Escherichia coli* plasma membrane.
48. Hasan S., Daugelat S., Rao P.S.S., Schreiber M. (2006) Prioritizing genomic drug targets in pathogens: application to *Mycobacterium tuberculosis*. PLoS Comp Biol;2:e61.
49. Perumal D., Lim C.S., Sakharkar K.R., Sakharkar M.K. (2007) Differential genome analyses of metabolic enzymes in *Pseudomonas aeruginosa* for drug target identification. In silico Biol;7:453–465.
50. Cho Y., Ioerger T.R., Sacchettini J.C. (2008) Discovery of novel nitrobenzothiazole inhibitors for *Mycobacterium tuberculosis* ATP phosphoribosyl transferase (HisG) through virtual screening. J Med Chem;51:5984–5992.
51. Huang H., Berg S., Spencer J.S. *et al.* (2008) Identification of amino acids and domains required for catalytic activity of DPPR synthase, a cell wall biosynthetic enzyme of *Mycobacterium tuberculosis*. Microbiology (Reading, England);154:736–743.
52. Alderwick L.J., Radmacher E., Seidel M. *et al.* (2005) Deletion of Cg-emb in corynebacterianeae leads to a novel truncated cell wall arabinogalactan, whereas inactivation of Cg-ubiA results in an arabinan-deficient mutant with a cell wall galactan core. J Biol Chem;280:32362–32371.
53. Ershov I.V. (2007) 2-C-methylerythritol phosphate pathway of isoprenoid biosynthesis as a target in identifying of new antibiotics, herbicides, and immunomodulators (Review). Prikl Biokhim Mikrobiol;43:133–157.
54. Eoh H., Brennan P.J., Crick D.C. (2009) The *Mycobacterium tuberculosis* MEP (2C-methyl-d-erythritol 4-phosphate) pathway as a new drug target. Tuberculosis (Edinburgh, Scotland);89:1–11.
55. Brown A.C., Parish T. (2008) Dxr is essential in *Mycobacterium tuberculosis* and fosmidomycin resistance is due to a lack of uptake. BMC Microbiol;8:78.
56. Shigi Y. (1989) Inhibition of bacterial isoprenoid synthesis by fosmidomycin, a phosphonic acid-containing antibiotic. J Antimicrobial Chemother;24:131–145.
57. Watson R.J., Heys R., Martin T., Savard M. (2001) Sinorhizobium meliloti cells require biotin and either cobalt or methionine for growth. Appl Environ Microbiol;67:3767–3770.
58. Jordan A., Aslund F., Pontis E., Reichard P., Holmgren A. (1997) Characterization of *Escherichia coli* NrdH. A glutaredoxin-like protein with a thioredoxin-like activity profile. J Biol Chem;272:18044–18050.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Selected 38 common targets in CMN group of pathogens including *Cp*.

Table S2. Comparative 3D modeling data.

Table S3. Hits properties of top five compounds for each selected proteins.

Table S4. Lists the protein residue IDs that are in contact with at least one of the top five compounds.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

II.II TRANSCRIPTÔMICA

II.II.1 Differential transcriptional profile of *Corynebacterium pseudotuberculosis* in response to abiotic stresses.
Pinto AC, de Sá PH, Ramos RT, Barbosa S, Barbosa HP, Ribeiro AC, Silva WM, Rocha FS, Santana MP, de Paula Castro TL, Miyoshi A, Schneider MP, Silva A, **Azevedo V**.
BMC Genomics. 2014 Jan 9;15:14. doi: 10.1186/1471-2164-15-14.

Neste trabalho, utilizando a tecnologia de nova geraçao, RNAseq, os autores observaram o perfil de expressao genica de Corynebacterium pseudotuberculosis em diferentes condiçoes de estresse, simulando as condiçoes desfavorareis encontradas no hospedeiro, como acidez, osmolaridade e alta temperatura. Na avaliaçao de unidade formadora de colonia, a bacteria se mostrou altamente resistente aos estresses, demonstrando uma resposta favoravel à sobrevivencia a estas condiçoes. Os resultados mostraram um catalogo genico que podem ser explorados em estudos relacionados ao desenvolvimento de vacina, diagnostico, ou antimicrobianos. Entre os genes que mais se destacaram, como alto valor de expressao, em relaçao ao controle, foram proteinas hipoteticas, demonstrando a necessidade de estudos aprofundados da bacteria. Alem destas, os genes expressos apresentaram-se envolvidos em processo de infecçao, de adesao, regulaçao, e virulencia. Este foi o primeiro trabalho de RNAseq no Brasil, e utilizou-se a plataforma SOLiD.

BMC Genomics

RESEARCH ARTICLE Open Access

Differential transcriptional profile of *Corynebacterium pseudotuberculosis* in response to abiotic stresses

Anne Cybelle Pinto[1], Pablo Henrique Caracciolo Gomes de Sá[2], Rommel T J Ramos[2], Silvanira Barbosa[2], Hivana P Melo Barbosa[2], Adriana Carneiro Ribeiro[2], Wanderson Marques Silva[1], Flávia Souza Rocha[1], Mariana Passos Santana[1], Thiago Luiz de Paula Castro[1], Anderson Miyoshi[1], Maria P C Schneider[2], Artur Silva[2] and Vasco Azevedo[1*]

Abstract

Background: The completion of whole-genome sequencing for *Corynebacterium pseudotuberculosis* strain 1002 has contributed to major advances in research aimed at understanding the biology of this microorganism. This bacterium causes significant loss to goat and sheep farmers because it is the causal agent of the infectious disease caseous lymphadenitis, which may lead to outcomes ranging from skin injury to animal death. In the current study, we simulated the conditions experienced by the bacteria during host infection. By sequencing transcripts using the SOLiD^{TM} 3 Plus platform, we identified new targets expected to potentiate the survival and replication of the pathogen in adverse environments. These results may also identify possible candidates useful for the development of vaccines, diagnostic kits or therapies aimed at the reduction of losses in agribusiness.

Results: Under the 3 simulated conditions (acid, osmotic and thermal shock stresses), 474 differentially expressed genes exhibiting at least a 2-fold change in expression levels were identified. Important genes to the infection process were induced, such as those involved in virulence, defence against oxidative stress, adhesion and regulation, and many genes encoded hypothetical proteins, indicating that further investigation of the bacterium is necessary. The data will contribute to a better understanding of the biology of *C. pseudotuberculosis* and to studies investigating strategies to control the disease.

Conclusions: Despite the veterinary importance of *C. pseudotuberculosis*, the bacterium is poorly characterised; therefore, effective treatments for caseous lymphadenitis have been difficult to establish. Through the use of RNAseq, these results provide a better biological understanding of this bacterium, shed light on the most likely survival mechanisms used by this microorganism in adverse environments and identify candidates that may help reduce or even eradicate the problems caused by this disease.

Keywords: Differential gene expression, Transcripts, RNAseq, SOLID™, Stress, *C. pseudotuberculosis*

Background

Corynebacterium pseudotuberculosis is a Gram-positive pathogenic bacterium belonging to the class Actinobacteria, which is a member of the *Corynebacterium*, *Mycobacterium*, *Nocardia* and *Rhodococcus* genera (the CMNR group). The CMNR group shares several common characteristics, including (i) the organisation of the cell wall, which is mainly composed of peptidoglycan, arabinogalactan and mycolic acids and (ii) the high G + C content of the genome (47-74%) [1].

The bacterium causes caseous lymphadenitis disease, which affects small ruminants and large animals (such as horses and cattle) worldwide and can infect humans [1]. Therefore, there is an urgent need to control the disease through the development of effective vaccines, therapies and diagnostic kits.

Because *C. pseudotuberculosis* is a facultative intracellular microorganism found preferentially in macrophages in the host, during the infection process, the bacterium

* Correspondence: vasco@icb.ufmg.br
[1]Department of General Biology, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Av. Antônio Carlos, Belo Horizonte 31.270-901, Brazil
Full list of author information is available at the end of the article

is exposed to a number of environmental changes that are far from ideal [2].

After phagocytosis, the phagosome quickly becomes acidic (pH ~ 4.6-5.0) [3] negatively affecting the metabolism and damaging macromolecules in the invading cell. In addition, other intracellular stresses negatively affect the microorganism, including oxidative thermal shock and nitrosative, surface, osmotic and starvation stresses; however, the bacterium manages to escape and persist in the environment [2].

To survive in this environment, the pathogen must mount an immediate and adequate, protective response that is reflected initially by transcriptional changes in specific sets of genes [4]. In this context, sigma factors, which coordinate the expression of these genes under different types of stresses are important [2]; these factors include *sigS* in *E.coli,* which is involved in trehalose synthesis during osmotic stress, with trehalose serving as an important osmoprotectant in this type of stress [4]. In *Mycobacterium tuberculosis*, RT-PCR (real-time polymerase chain reaction) was used to demonstrate the transcriptional profile of 10 sigma factors during the exponential growth phase. The role of these sigma factors was analysed under different conditions of stress, and a number of the factors demonstrated increased expression in response to one kind of stimulus, whereas others responded to more than one stimulus. The resistance to different environmental stresses is associated with the ability of the pathogenic bacteria to survive in the host, and different sigma factors play fundamental roles in the survival of the pathogen. For example, in *M. smegmatis, sigB* is involved in the response to oxidative stress, and *sigE* was shown to play a role in the regulation of genes involved in the response to acid stress, thermal shock and sodium dodecyl sulphate (SDS) exposure [5,6].

The functions of sigma factors include involvement in the adaptation to stress, the interaction of the bacterium with the extracellular medium and in a number of cases, with bacterial virulence [7].

For *C. pseudotuberculosis* 1002, there is still no information regarding the role of sigma factors in the regulation of genes involved in bacterial survival throughout the infection process. In addition, few virulence determinants contributing to bacterial survival have been identified. Therefore, investigations related to the control of caseous lymphadenitis have been difficult [1]. To date, the virulence determinants most studied in *C. pseudotuberculosis* infection include the following: the PLD (phospholipase D) protein, an exoprotein that is also considered leukotoxic, contributing to the formation of lesions and the destruction of caprine macrophages during infection [8];the *fagABC* operon and the *fagD* gene, which play a role in the virulence of the bacterium and have been identified as genes involved in iron acquisition [9]; the high concentration of cell wall lipids, which renders the microorganism resistance to digestion by cellular enzymes and allows it to persist as a facultative intracellular parasite [10] and CP40, identified as an immunogenic protein that exhibits proteolytic activity as a serine protease [11].

The availability of the *C. pseudotuberculosis* strain 1002 genome (access number CP001809) has allowed the further investigation and characterisation of the microorganism, which is poorly characterised despite its importance to agribusiness. Therefore, to generate additional information, the transcriptional profile of *C. pseudotuberculosis* was analysed using cDNA sequencing with SOLiD™ 3 plus next-generation technology (Life TechnologiesTM, CA). Using this technology, we investigated the molecular characteristics of the microorganism that allow it to persist in the host and the mechanisms the bacteria use to escape the host immune response. Additionally, in an attempt to contribute to the elimination of caseous lymphadenitis in caprine and ovine populations, we investigated a large number of targets related to *C. pseudotuberculosis* virulence by simulating the specific conditions tolerated by the bacterium when invading the host (i.e., acidic, osmotic and high temperature stresses).

Results and discussion

The sequencing of cDNA is an attractive technology for the investigation of gene expression in prokaryotic organisms because it provides a high level of coverage and high sensitivity for the detection of transcripts at considerably lower costs compared to traditional methods [12]. Currently, sequencing technology is considered the gold standard for the analysis of gene expression levels [13]. Therefore, because the microorganism can survive environmental changes in the host during infection, we analysed the transcriptional profile of *C. pseudotuberculosis* strain 1002 using RNAseq technology. Acid, osmotic and thermal shock stresses were used to identify genes involved in bacterial tolerance of these unfavourable environments.

Stress-generating agents were applied to the cells at an optical density (OD) of 0.2 ($A_{600nm} = 0.2$), and the cell viability analysis demonstrated a reduction in replication of approximately 27% under thermal stress, 34% under acid stress and 23% under osmotic stress (Figure 1). A lack of growth or reduced growth is normal during periods of environmental change as the organism attempts to adapt to and physiologically adjust to the new environment [14].

The cDNA samples were sequenced using the SOLiD™ 3 Plus platform, which allowed the analysis of the gene expression profile of the microorganism in the early exponential phase. Using the Bioscope programme,

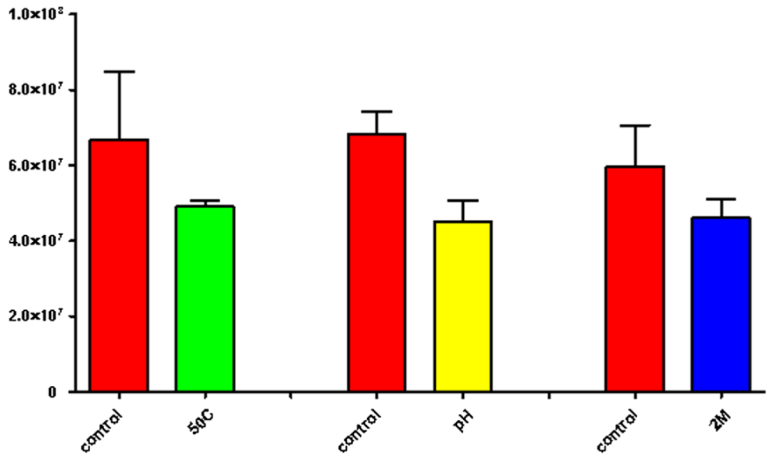


Figure 1 Number of viable cells under each condition. The control conditions correspond to approximately 6.7 x 10^7 cells mL^{-1} of C. *pseudotuberculosis* strain 1002, with the exception of the osmotic stress control, which corresponds to 6.0 x 10^7 cells mL^{-1}. Red, control conditions. Green, thermal shock stress. Yellow, acid stress. Blue, osmotic stress. Figure taken from [16].

unique readings were mapped on the genome, and gene expression was quantified based on the RPKM (reads per kilobase of coding sequence per million mapped) [15]. Ribosomal transcripts were filtered using Bioscope, and the total number of readings obtained before and after application of the filter is shown in Table 1.

Genomic coverage was inferred using the data generated in Bioscope, which represents the level of expression in the RNAseq experiments. The osmotic stress produced the largest number of uniquely mapped transcripts throughout the entire genome, followed by the thermal stress, acid stress and control conditions (Table 2).

The automatic annotation of the genome using the Fgenes software (www.softberry.com), followed by manual curation of the *C. pseudotuberculosis* strain 1002 genome, identified 2,090 coding regions. From the cDNA sequencing performed using SOLiD^{TM} at the beginning of the exponential phase, 2,055 transcripts active in control were identified (equivalent to 98.32% of the transcribed genome) and 35 genes (1.67%) were considered non-transcripts, exhibiting a RPKM value of 0. Under the different stress conditions, 2,065 (98.80%) transcripts were produced under osmotic stress, 2,063 (98.70%) transcripts under thermal stress and 2,064 (97.76%) under acid stress.

According to the DEGseq analysis software, of the 2,065 transcripts produced following the osmotic stress, 889 (43.05%) were considered differentially expressed compared to the control (*p-value* <0.001) [16] (Figure 2), 565 of these genes were induced and 324 were repressed. In the thermal stress experiment, 543 (26.32%) transcripts were considered differentially expressed, of which 374 were induced and 169 were repressed. In the acid stress, 811 (39.30%) transcripts were considered differentially expressed, of which 519 were induced and 292 were repressed.

Among the differentially expressed genes, the genes exhibiting a 2-fold change in expression (at least 2x relative to the control) were selected for analysis. The fold-change values were calculated based on the RPKM value between the stress and the control, in which a value greater than one indicated induced gene expression and a value less than one indicated repressed gene expression.

In a previous study reported by our group [17], sequencing of transcripts from *C. pseudotuberculosis* strain Cp31 (biovar equi) was performed using an Ion Torrent

Table 1 Number of total readings obtained during sequencing

	pH	2 M	50°C	Control
Gross Data	17,393,077	18,783,810	21,622,844	25,235,478
Filtered Ribosomal Transcripts	9,738,772	9,564,434	9,971,878	9,270,342

Gross data include readings of all transcripts. Filtered ribosomal transcripts refer to the number of readings of transcripts not considered in the analysis. pH- acid stress; 2 M- osmotic stress, 50°C -thermal stress and control -no stress.

Table 2 Number of unique readings mapped in the genome of *C. pseudotuberculosis* 1002 and coverage of the transcripts in the genome

	2 M	50°C	pH	Control
Uniquely mapped readings	2,016,131	1,633,118	1,764,047	1,650,975
Genome coverage	43x	39x	37x	35x

2 M- osmotic stress; 50°C– thermal stress; pH- acid stress and control-no stress.

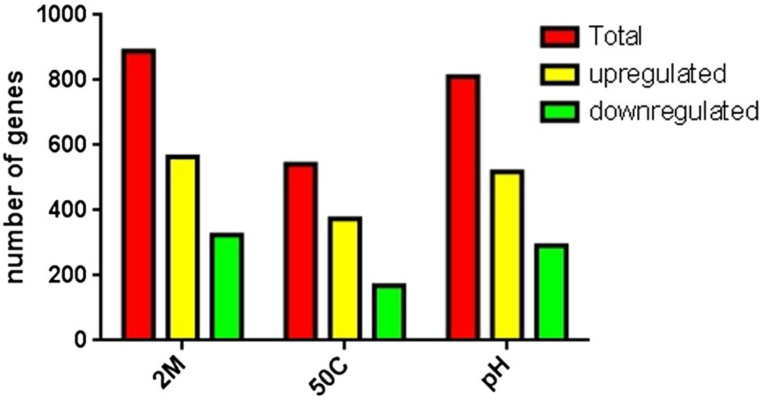


Figure 2 Number of genes differentially expressed relative to the control. Red, genes transcribed under each condition (sum of yellow and green). Yellow, genes induced relative to the control. Green, genes repressed relative to the control. 2 M, osmotic stress; 50°C, thermal stress and pH, acid stress.

platform (Life Technologies). In this study, a comparative analysis between 2 rRNA depletion methodologies was performed, and the transcripts were submitted to *ab initio* assembly. Both transcriptomes were then submitted to gene ontology analysis according to biological processes and molecular functions. Data were obtained only under physiological conditions using brain heart infusion (BHI) media. The authors observed that few transcripts represented genes involved in pathogenicity or the cell adhesion process and concluded that contact with the host may influence the induction of these transcripts.

The present study simulated some of the environmental conditions that the pathogen faces in the host. The cell adhesion process was among those most represented in the conditions described below, and these levels showed greater than a 2-fold change. This result indicated that there were a greater number of transcripts representing genes that participate in the cell adhesion process when compared to the physiological condition. Together, these findings demonstrate that contact with the host most likely influences the transcription of genes essential for bacterial survival.

Genes induced in the biological processes of the osmotic stress stimulon

The Blast2GO programme was used to identify the biological processes most abundant in the cells under osmotic stress. However, for a more detailed analysis of the processes (Figure 3), the CoreStImulon (CSI) programme [18] was used, which identified the genes present in each of the processes determined in the Blast2GO programme (see Additional file 1: Figure S1).

The majority of the genes induced under osmotic stress were part of the oxidoreduction process, represented by 4 genes (see Additional file 1: Figure S1). Normal aerobic metabolism induces the production of active oxygen molecules, which are increased following exposure to certain environments [19]. The results demonstrated that the growth of the bacterium was reduced but was not interrupted under these conditions; therefore, the bacterium survived in the environment. This observation was confirmed by the biosynthetic process, which was comprised of 3 genes (see Additional file 1: Figure S1), indicating that the bacterium was able to invest energy into the replication process and survived in the unfavourable environment.

The adhesion process, comprising 3 genes, was prominent under osmotic stress. Adhesion is essential for the initiation of the infectious process because the bacterium-host interaction establishes a pathogenic relationship. The Cp1002_0988 gene, encoding a hypothetical protein, is located within 1 of the pathogenicity islands of *C. pseudotuberculosis* 1002, indicating its importance in the development of the disease. Additionally, this gene exhibited a 6.8-fold change in expression levels compared with the control (see Additional file 2: Table S1). The Cp1002_1764 and Cp1002_1765 genes, also identified as encoding hypothetical proteins, exhibited 2.7-fold and 4-fold changes in expression, respectively, compared with the control.

Pathogenic bacteria have developed highly sophisticated signal transduction systems that control the coordinated expression of a number of virulence determinants in response to environmental stresses, and changes in osmolarity contribute to the expression of the genes [20]; therefore, further studies to identify the proteins encoded by these genes and to evaluate the true contribution of these proteins will be necessary.

Genes induced in the biological processes of the acid stress stimulon

Under acid stress conditions, the induction of genes involved in the processes of cellular adhesion and

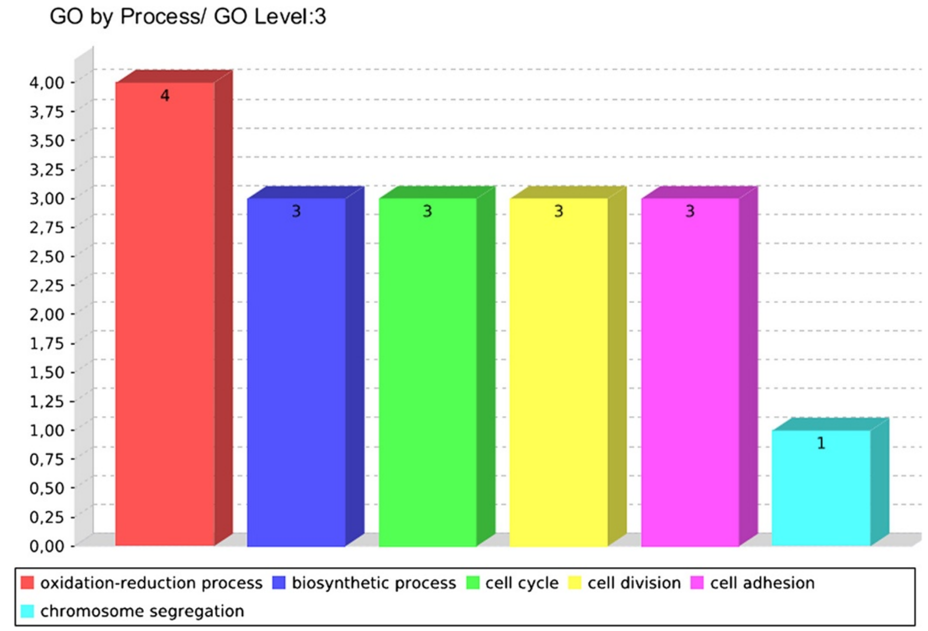


Figure 3 Biological processes most evident among the genes induced in the osmotic stimulon. Figure obtained with the CSI program.

oxidoreduction in response to stress were of paramount importance because they are associated with virulence (Figure 4) (see Additional file 3: Figure S2). The adhesion processes were composed of hypothetical proteins, 1 of which was characterised as a secreted protein containing an LPxTG domain, which may be an important vaccine candidate. Through analysing the genes that made up the cellular oxidoreduction process, we observed the presence of genes with functions that appeared essential for the persistence of the bacterium in this environment.

The Cp1002_2043 gene exhibited a 7.8-fold change in expression (see Additional file 4: Table S2). This gene, which may encode the Dps protein, was linked to the

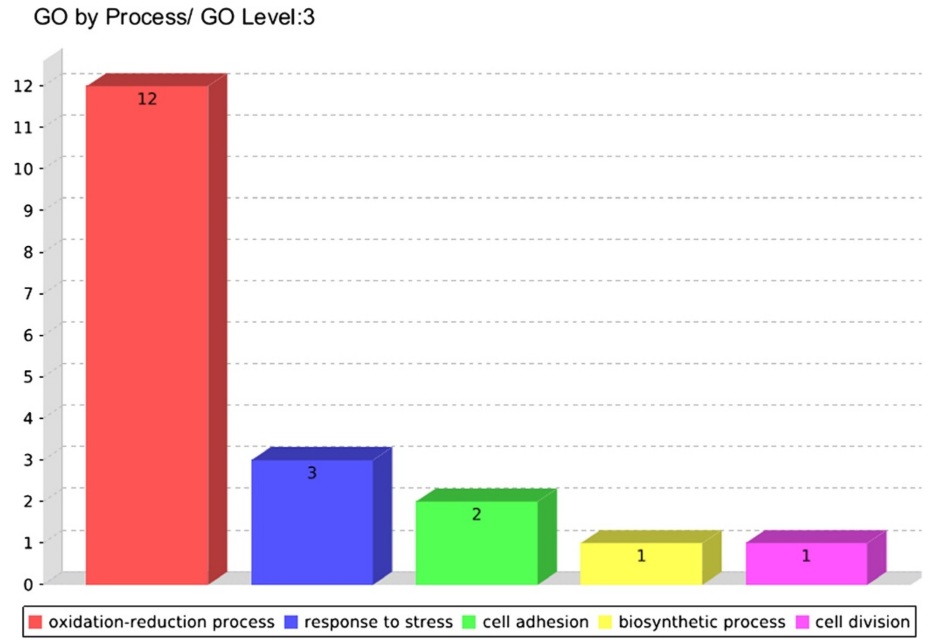


Figure 4 Biological processes most evident among the genes induced in the acid stimulon. Figure obtained using the CSI program.

processes of stress response and cellular oxidoreduction (see Additional file 3: Figure S2). This protein protects the organism against oxidative stress because it stores iron in a bioavailable form, reducing the possibility of the production of reactive oxygen molecules. Under acidic conditions, the production of molecules that produce reactive oxygen species is increased. Reports have demonstrated that in cells in which the pH is decreased, the ratio of HOO^- to O_2^- increases [19], increasing the chance of producing hydrogen peroxide (H_2O_2). Therefore, the increased number of genes constituting the cellular oxidoreduction process is justified. Another important gene in the oxidoreduction process was Cp1002_0173, which exhibited a 4-fold change in expression and is presumed to encode a catalase that plays a role in reducing the concentration of H_2O_2 in the cell.

The mechanisms used by the bacterium to resist the damage caused by reactive oxygen species are essential for survival within the macrophage [21]. Therefore, these proteins might promote the survival of the bacterium in the acid medium starting at the early exponential growth phase.

The Cp1002_1192 *msrB* (methionine sulphoxide peptide reductase) gene, which was 1 of the genes present in the oxidoreduction process, exhibited a high (16-fold) change in expression. Because the MsrA protein may be required to maintain the role of adhesins, the high increase in *msrB* expression suggested that the protein contributes to the survival of the pathogen in the host, the resistance to oxidative stress in vitro and the adhesion capability of eukaryotic cells [22]. The methionine (Met) residues in proteins exposed at the cell surface are thought to be involved in capturing reactive oxygen species, and a complex comprised of MsrA and MsrB reduces the oxidised Met residues, removing the reactive oxygen species [23]. Met is the amino acid most sensitive to reactive oxygen species, and the oxidation of the Met residue in a protein alters the protein structure (or prevents translation), drastically affecting the function [24]. There is evidence that only the MsrB domain is present in *C. pseudotuberculosis* 1002. However, the loss of the *msrA-msrB* domain or *msrB* alone in *Helicobacter pylori* resulted in a reduction in virulence in mouse models, likely because of the oxidation of important proteins [25]. Therefore, the process of cellular oxidoreduction may involve genes that contribute predominantly to the maintenance and persistence of the bacterium in media harmful to the cell.

Genes induced in the biological processes of the thermal shock stimulon

We analysed the genes involved in the oxidoreduction process under thermal stress. The Cp1002_1785 (*betA*) gene is among the genes that exhibited the highest fold-change in expression values (Figure 5) (see Additional file 5: Figure S3), and the gene may encode choline dehydrogenase. The induced gene exhibited a 5-fold change in expression (See Additional file 6: Table S3) compared to the control. The protein belongs to the oxidoreductase family, which catalyses the oxidation of choline to glycine betaine via the intermediate betaine aldehyde. The protein promotes increased tolerance to hypersalinity and freezing and contributes to the osmotic balance of the cell under stress. The osmoprotectors not only play a role in osmotic balance but also act as effective stabilisers of enzymatic function, providing protection against salinity, high temperatures, freezing, thawing and even dryness [26]. Under conditions of high salinity, the osmoprotectors, together with the transport system, function as virulence factors in certain pathogenic bacteria [20]. Understanding how these elements operate under this condition of stress and their relationship with pathogenesis will be important for future studies.

The adhesion process comprises the Cp1002_1765 gene, which may encode a secreted protein. Cp1002_1765 exhibited a 2-fold change in expression compared with the control, and it is an important candidate for studies related to controlling the disease caseous lymphadenitis.

Among the genes involved in the process of the response to stress, the Cp1002_1895 gene, which may encode a heat shock regulatory protein (HspR), exhibited the largest fold-change in expression (4x relative to the control). This protein acts as a negative regulator of the expression of genes encoding chaperones and proteases in different bacteria under physiological conditions. The heat shock proteins play a key role in cellular metabolism under all growth conditions, monitoring the folding, assembly and translocation of cellular proteins [27]. In *M. tuberculosis*, HspR represses the operon formed by the *dnaK-grpE-dnaJ-hspR* genes through interaction at the HAIR (HspR-associated inverted repeats) region located in the 5′UTR region of the genes. In *C. pseudotuberculosis*, the *hspR* gene is located in the reverse strand, below the *dnaJ, GrpE* and *dnaK* genes, indicative of their regulation by HspR.

It has been suggested that the HspR protein acts as a repressor of other genes linked to virulence/pathogenicity [27]. A study in *M. tuberculosis* demonstrated that the partial disruption of heat-shock regulation influences virulence because the bacterium loses the ability to establish a chronic infection [28]. Mutation of *hspR* in the organism produced increased expression of the DnaK chaperone, which is highly antigenic, resulting in an enhanced immune response in the host. Furthermore, it was demonstrated that DnaK acts as co-repressor for

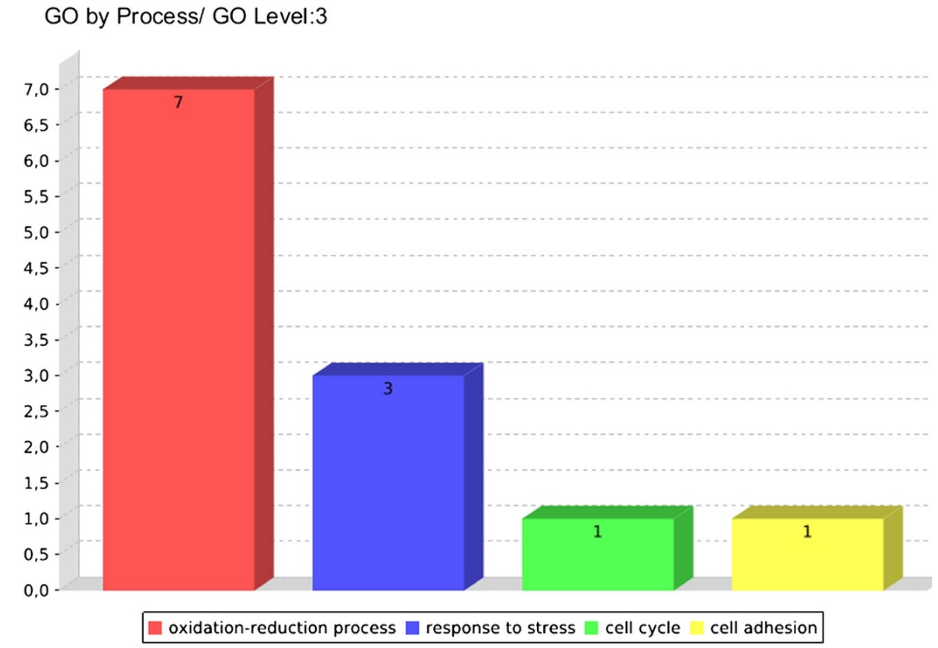


Figure 5 Biological processes most evident among the genes induced in the thermal stimulon. Figure obtained using the CSI program.

HspR; the activity of HspR is dependent on DnaK. A proposed mechanism for the regulation of *hspR* expression is that under conditions of heat shock, the operon is induced, leading to the increased synthesis of DnaK and HspR. When the concentration of these proteins reaches a critical level, they bind to the promoter of the operon, resulting in repression [29]. To investigate whether this scenario occurred in *C. pseudotuberculosis*, we analysed the gene encoding DnaK and demonstrated that the expression of the gene was induced 2.5-fold relative to the control, which is consistent with the proposed mechanism (see Additional file 6: Table S3).

Distribution of genes among the simulated conditions

From the Venn diagram (Figure 6) it was possible to observe the distribution of genes per condition, demonstrating 29 genes active in all 3 simulated environments. The analysis of the genes demonstrated that a variation in the fold-change in expression, ranging from 2- to 43-times the control, was observed among the stresses. Of these genes, 48% encoded hypothetical proteins and demonstrated a higher fold-change in expression (see Additional file 7: Table S4) under all conditions. The data demonstrating the high number of unidentified genes reflects the lack of information on *C. pseudotuberculosis*, despite its importance in agribusiness. In the host, the bacterium suffers various stresses simultaneously; therefore, identifying the proteins and their role in the cell is essential for a better understanding of the molecular mechanisms of the bacterium. Presumably, the proteins contribute strongly to the survival and persistence of the bacterium in unfavourable environments; therefore, these proteins are potential candidates for the development of vaccines, diagnostic kits or therapies for caseous lymphadenitis.

Among the shared repressed genes were those encoding proteins related mostly to energy metabolism, sugar transport, amino acids that contribute greatly to the maintenance and replication of the organism in the environment [30], such as *argJ* [31], *nanK* [32], *opp* [33] (see Additional file 8: Table S5) and hypothetical proteins. These data are consistent with the colony-forming unit experiments, which demonstrated a decrease in replication in the stressful environments (Figure 1). A reduction in growth is a survival strategy and occurs in essentially all stressful situations because the organism's ability to perceive the environment and to modulate the response-controlling mechanisms of resistance, metabolism and other processes [34] which are increased, specific and suitable for the new conditions, is essential.

Expression of genes encoding sigma factors under simulated conditions

In their natural environment, or in the host during the infection process, the bacterium is exposed to disturbances that require a fast and adaptive response to ensure the survival of the pathogen. Therefore, it is necessary for the bacterium to change the pattern of expression of the genes encoding proteins that directly combat the deleterious nature of the stress [35].

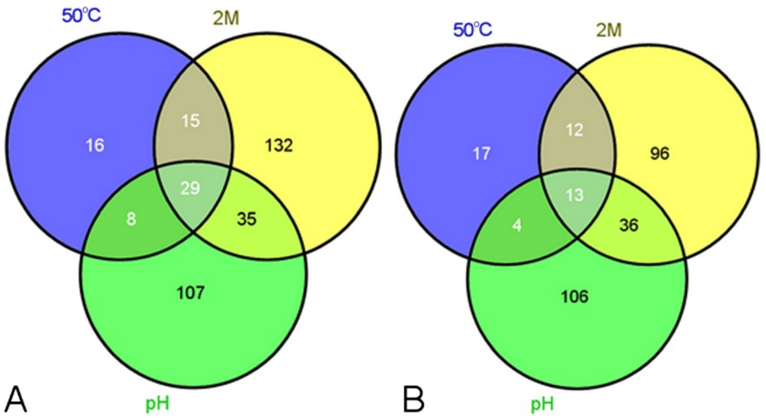


Figure 6 Venn diagram of the conditions tested. Distribution of the genes indicating differential expression (**A**-induction and **B**-repression) with at least a 2-fold change relative to the control. 50°C, thermal shock stress; 2 M, osmotic stress and pH, acid stress.

Certain sigma factors play an important and fundamental role in regulating the expression of virulence genes [36]; therefore, it is important to analyse these genes individually.

In the *C. pseudotuberculosis* strain 1002 genome, 8 genes encoding sigma factors were identified, which included the essential sigma factor SigA, non-essential and alternative SigB and 6 alternative factors belonging to the group of extracytoplasmic factors SigC, SigD, SigE, SigH, SigK and SigM, which are dispensable and induced frequently in response to specific conditions, for example, in response to stress.

Under the osmotic stress conditions, the DEGseq programme demonstrated that only the genes encoding sigma factors A and M were induced (were within the established cutoff of a fold-change of at least 2) (see Additional file 9: Table S6). Under acidic conditions, the genes encoding sigma factors B, E and H were differentially expressed and under thermal stress conditions, the genes encoding sigma factors A, D and H were considered differentially expressed; however, the expression of these genes were below the cutoff for the analysis (Figure 7).

It is possible that *sigA* encodes sigma factor RpoD, alternatively named sigma 70, which promotes the binding of RNA polymerase to specific sites by activating the transcription of most genes essential for exponential growth in *Escherichia coli* (*E. coli*) [37]. In *C. pseudotuberculosis* 1002, the gene contains the 4 domains conserved in the sigma 70 family. Sporadically, σA can act as an alternative sigma factor that is specifically required for the expression of virulence genes. An investigation of the role of *sigA* in *E. coli* in which the *rpoD* gene was induced (from a lack of amino acids and thermal shock), suggested that the protein was involved in the mechanism of recovery from stress [38]. *Streptomyces spp.* contains several homologues of the principal factor that are not essential for replication under normal conditions but appear to play a role under certain growth conditions [39].

Because *sigA* demonstrated changes in expression between the control and the stress conditions in *C. pseudotuberculosis* 1002, it may play a role as an alternative sigma factor, regula [40] ting genes involved in the maintenance of the bacterium.

The gene encoding sigma M was considered differentially expressed only under osmotic stress. This gene may not be expressed preferentially in the early exponential phase. Microarray analysis of *C. glutamicum* in the exponential phase ($OD_{610nm} = 0.2$ to 0.3) demonstrated that the disruption of *sigM* did not affect the level of transcription of genes induced by thermal shock, which implies that this sigma factor is not involved in the regulation of gene expression in response to this stress [40].

In *M. tuberculosis*, expression of *sigM* occurred in the stationary phase and only under conditions of thermal stress [41]. In *C. glutamicum*, experiments revealed that the deletion of *sigM* caused a reduction in the number of viable cells under conditions of heat shock, cold shock and disulphide (an oxidative stress subtype) stress in the exponential growth phase ($OD_{600nm} = 0.7$). Furthermore, experiments using real-time PCR demonstrated that the transcription of *sigM* increased significantly after the application of the stresses. These results suggest that sigma factor M is involved in the stress response [42], albeit with a strong indication that the involvement occurs at the late growth phase.

Because there was an increase in the expression of *sigM* under osmotic stress in *C. pseudotuberculosis* 1002, a study of the regulon will be required to identify the

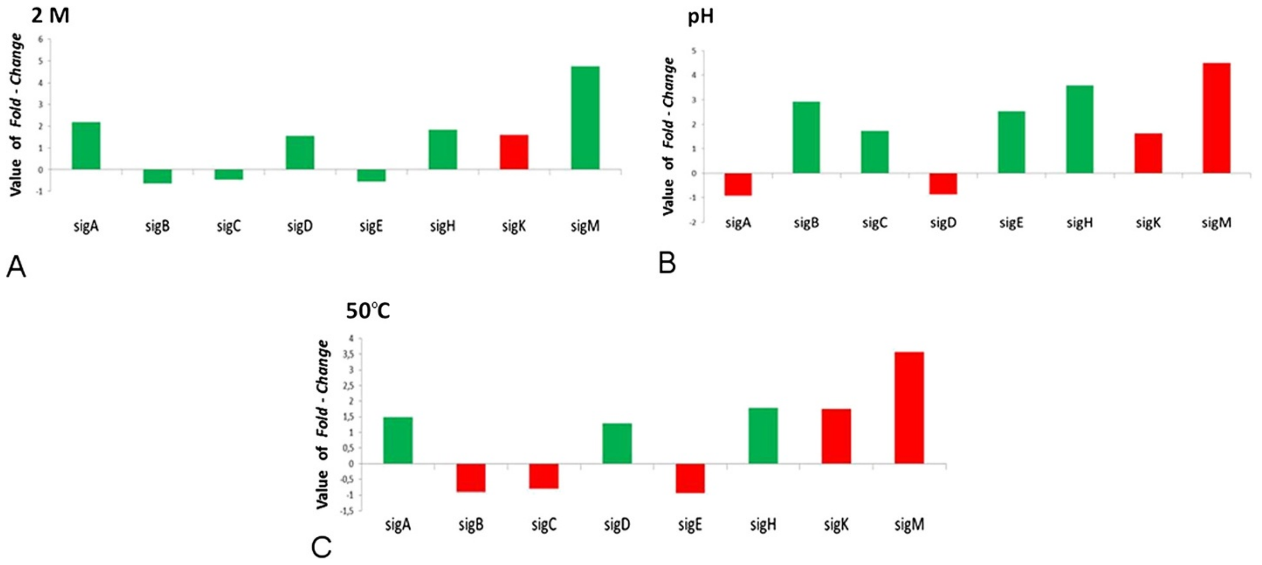


Figure 7 Expression of genes encoding sigma factors under conditions of simulated stresses. The fold-change value was based on the ratio between the stress RPKM and the control RPKM. Green columns indicate the differentially expressed genes, independent of the fold-change cutoff established for the analysis, which was at least 2x. Red columns represent the genes not differentially expressed (*p-value* >0.001), established using the DEGseq programme. **a,** Osmotic stress. **b,** Acid stress. **c,** Thermal stress.

genes encoding proteins that assist in the persistence of the bacterium in this hostile environment.

Sigma B is a very common sigma factor in the stress response. The role of the protein in the response to acidity tolerance was established in *Bacillus subtilis*, *Brevibacterium flavum*, *Listeria monocytogenes* and *Staphylococcus aureus* [43-45]. In *B. subtilis*, *SigB* regulates the majority of stress responses, thereby contributing to the transcription of more than100 genes [46]. Another study demonstrated a reformulation of transcription during the infection process and the relevance of sigma B in controlling the expression of virulence genes important for adaptation to the intestinal environment in *L. monocytogenes* [47]. In *L. monocytogenes*, σB contributed to cell survival under different stress conditions [48], and the absence of the factor reduced the ability of the species to invade epithelial cells [49]. It is known that σB contributes to virulence in several Gram-positive pathogens [2], and these observations underscore the need to study the factor separately in *C. pseudotuberculosis* and to identify the regulon that contributes to the survival and escape of the bacterium from the host immune system.

The *sigE* gene, encoding the presumed factor SigE, a member of the σ70 subfamily exhibiting extracytoplasmic function, regulates functions related to perception and response to changes in the periplasm and in the extracytoplasmic environment. In *Haemophilus influenzae,* the expression of *rpoE* increased 102-fold after phagocytosis by macrophages, and the survival of *Haemophilus influenzae* containing a mutated *rpoE* was reduced compared to wild type [50]. According to a study in *Vibrio cholerae* [51], *rpoE* mutants attenuate virulence, and the ability of the bacteria to colonise the mouse intestines is reduced.

According to a study in *C. pseudotuberculosis* 1002 [52], compared with the wild type strain, the mutant 1002 strain is more sensitive to agents that generate nitrosative, acid (pH 5.5) and surface stresses, which indicates the important role SigE plays in the persistence of the bacterium in hostile environments.

The investigation of a mutant *sigE* in *M. tuberculosis* H37Rv [53] demonstrated that the mutant was more sensitive to various environmental stresses, such as thermal shock, SDS and oxidative agents but not to acidity. In a report on *M. Smegmatis* [6], the mutant *sigE* was more sensitive to acidity, hydrogen peroxide, SDS and thermal shock than the wild type. In the *M. tuberculosis* H37Rv study [53], the authors determined that *sigE* is required for the stress response and for the ability to grow and survive inside macrophages. Furthermore, it was suggested that *sigE* influences the level of *sigB* in the cell; however, the expression of *sigB* is not completely influenced by *sigE* because the level of *sigB* was reduced in the mutant strain, whereas the mRNA levels of other sigma factors were not affected.

In *C. pseudotuberculosis* 1002 under acidic conditions, *sigE* was induced and the expression of both *sigE* and *sigB* was significant, whereas under conditions where the mRNA level of *sigE* was not considered significant, the *sigB* expression was also not significant. These sigma factors may influence each other at this exponential phase; however, studies that are more specific will be necessary to verify this hypothesis.

The *sigH* gene, which encodes the extracytoplasmic sigma factor H, was induced (within the established cut-off value) only under the acidic conditions. In *M. tuberculosis,* the *sigH* factor is involved in the response to different stresses and it has been suggested that it regulates the expression of genes involved in the intracellular survival of the microorganism. Additionally, proteins that are part of the *sigH* facto regulon may interact with the host's immune system, modulating its response [54]. This observation suggests that at the beginning of the *C. pseudotuberculosis* 1002 replication process, the expression of the *sigH* gene and its regulon are required for the bacterium to persist in an acidic environment.

In *C. glutamicum*, a mutation in the *sigH* gene blocked the transcription of *sigM*, and the identification of the *sigH* promoter upstream of *sigM* implied that the factor is under the direct transcriptional control of *sigH* [42]. It was determined that in *C. pseudotuberculosis* 1002, when *sigH* expression increased under conditions of stress, *sigM* expression also increased, although the *sigM* transcripts did not exhibit significant induction during this growth phase. These results suggest that the *sigH* gene responded to an environment similar to that encountered by the bacterium in the host, highlighting the importance of identifying the regulon and the genes responsible for the persistence of the bacterium in the environment.

Identification of non-coding RNA

The prediction using the RFAM programme identified 5 non-coding RNAs (ncRNAs) in the genome of *C. pseudotuberculosis* strain 1002 (Table 3), including riboswitches (thiamine pyrophosphate [TPP] yybP-ykoY), ncRNA tmRNA (SsrA) and ncRNA mraW.

In most cases, the ncRNAs in bacteria play a role in regulating cellular response during environmental change, which is beneficial to the organism, which needs to adapt quickly and efficiently to these changes and may be impaired by the cell because of the energy cost involved in the expression of a large number of genes. Compared to the synthesis of regulatory proteins, less energy is needed for the synthesis of a small RNA (sRNA) [55]. Among the ncRNAs detected in the genome, riboswitches (elements that control expression in response to various metabolites and can function as sensors or binding sites for a receptor protein that senses a change in the cellular environment) were identified through the analysis of active transcripts. Furthermore, other elements that may be responsible for the synthesis of peptidoglycan, which promotes the expression of genes important under stress conditions [56,57], were identified.

Table 3 Identification *in silico* of ncRNA in the genome of *Corynebacterium pseudotuberculosis* strain 1002 using the RFAM programme

ID	Score	Strand	ncRNA
RF00023	136.03	+	tmRNA.1
RF00059	57.42	+	TPP.1
RF01747	54.15	-	msiK.1
RF00080	49.16	-	yybP-ykoY.1
RF01746	48.99	-	mraW.1

Coverage of the transcripts was analysed in the predicted element with the highest score. Figure 8 shows the coverage under the control condition and under the stress conditions for the ncRNA tmRNA. This element exhibits dual properties, acting as both messenger RNA and carrier RNA in the rescue of stalled ribosomes [58]. Under acidic and thermal shock conditions, coverage of this ncRNA was slightly superior compared to the control condition. Therefore, it is possible that in the unfavourable environment, because of the reduction in replication, this element is required to attempt the rescue of stalled ribosomes to avoid a loss in protein synthesis, allowing the bacterium to continue to replicate in the environment, albeit at a slower pace. In *Mycoplasma pneumoniae*, *tmRNA* proved essential for growth and in *Salmonella typhimurium*, this element is necessary for survival inside macrophages. It is believed that under conditions of stress, the cells become more sensitive to the activity of *tmRNA*, suggesting the importance of the ability of cells to adapt and survive in different environments [59].

Conclusion

Next-generation technology allowed the identification of a large number of genes presumed to be required by *C. pseudotuberculosis* 1002 for survival in unfavourable environments, such as acidity, thermal shock and osmotic stress. A number of these genes encode hypothetical proteins, which highlight the need for further investigation of the microorganism. Among the most relevant biological processes identified under all simulated conditions were the processes of adhesion, stress response and oxidoreduction. In these processes, genes involved in the virulence of the organism were affected and should be investigated in more detail to identify the roles they play in the cell, especially inside the host. Furthermore, it is believed that the identification of these genes may contribute to the development of vaccines that are more effective, diagnostic kits and therapies for caseous lymphadenitis.

The expression of sigma factors varied under the different conditions and at the beginning of the exponential phase, these important factors involved in the regulation

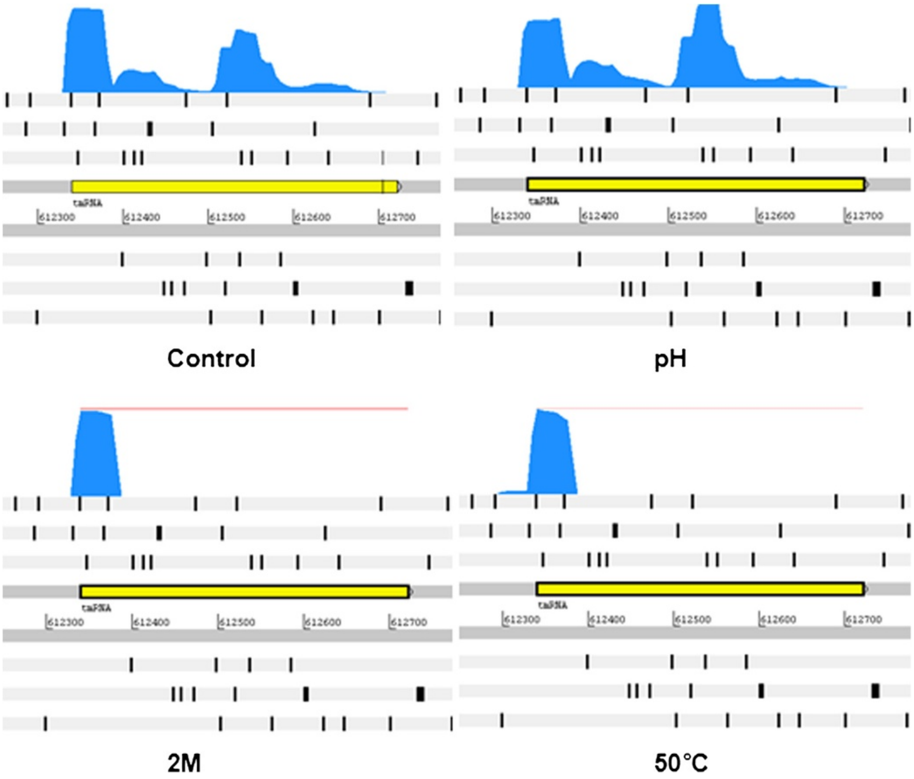


Figure 8 Coverage of ncRNA transcripts predicted by RFAM. Control; pH, acid medium; 2 M, osmotic medium and 50°C, thermal shock. Figure obtained in Artemis using the file .BAM, generated in the Bioscope programme.

of genes required for the maintenance of microorganisms under different environmental conditions were observed. Understanding the regulon of the genes would clarify the biology of the organism.

The predicted ncRNAs were identified by the coverage of the transcripts and these factors presumably contribute to the regulation of genes related to the persistence of the bacterium in harmful environments. Further studies will be required to confirm the role of the ncRNAs in escaping the host immune system and contributing to the survival of the bacterium in hostile environments.

The data regarding the expression of induced or repressed genes do not necessarily indicate protein translation; therefore, future experiments will focus on understanding the biology of the transcripts and their products.

Methods

Culture conditions: obtaining bacterial cells

Corynebacterium pseudotuberculosis strain 1002 was grown in petri dishes containing BHI media (broth composed of (g/L): calf-brain infusion 200.00, beef-heart infusion 250.00, proteose peptone 10.00, dextrose 2.00, sodium chloride 5.00, di-sodium phosphate 2.50 pH 7.4 ± 0.2 at 25°C) at room temperature (RT). One colony was used to prepare the pre-inoculum in 20 mL of BHI media supplemented with 0.05% Tween 80. The culture was grown overnight at 37°C in a shaker at 160 rpm. One millilitre of this pre-inoculum was used to prepare the inoculum in an Erlenmeyer flask containing 100 mL fresh BHI, and this culture was incubated at 37°C at 160 rpm. This preparation was monitored until the beginning of the exponential growth phase ($A_{600} = 0.2$), which was reached approximately 2.5 hours after the initial inoculation (see Additional file 10: Figure S4).

Application of stresses

After the culture reached the beginning of the exponential growth phase, the inoculum was divided into 4 50-mL Falcon tubes (1 for each condition), each containing a final volume of 20 mL, and these tubes were then centrifuged for 3 minutes at 8,000 rpm at RT. The pellet was resuspended in fresh BHI specific to each condition. For the acid stress condition, the media was supplemented with hydrochloric acid (which the pH changed to 5). Osmotic stress was achieved with 2 M NaCl, and thermal stress was induced by resuspending the pellet in BHI medium pre-heated to 50°C. In the control condition, bacterial pellets were resuspended in BHI medium at a physiological condition. After the addition of culture media, the tubes were kept in a shaker at 37°C and 160 rpm for 15 minutes, with the exception of the thermal stress sample that was subjected to a temperature of 50°C. An aliquot of each condition was used for decimal dilutions from 10^{-1} to 10^{-6}, from which 10^{-4} to 10^{-6} bacteria were seeded in BHI agar, and petri dishes were kept at 37ºC for 48 hours for viability analysis and colony counting (this step was performed in duplicate). The remaining sample was subjected to centrifugation at RT for 3 minutes at 8,000 rpm, and the pellet was resuspended in 2 ml of RNAlater, according to the manufacturer's instructions.

RNA extraction

The bacteria suspended in RNAlater® buffer were subjected to total RNA extraction using the ChargeSwitch® total RNA cell kit (Invitrogen, USA) in accordance with the manufacturer's recommendations, including the following adaptations: after the addition of the lysis buffer (Invitrogen), the material was transferred to 2-mL tubes partially filled with 1-mm diameter glass microbeads (Bertin Technologies). The cells were lysed mechanically using a Prescellys 24 homogeniser, set at 6,500 rpm, for 2 cycles (15 seconds per cycle) with an interval of 30 seconds between the cycles. The samples were centrifuged for 1 minute, and the supernatant transferred to fresh 2-ml tubes and incubated in a dry bath at 60°C for 15 minutes (represents the complete original protocol). DNase was added to eliminate the residual genomic DNA. The elution of the total RNA from the magnetic beads was performed using 100 μL of milli-Q RNase-free water. The amount of total RNA was assessed using a Qubit® 2.0 fluorometer (Invitrogen).

mRNA enrichment through rRNA depletion

To enrich the mRNA, rRNA from each total RNA sample was removed using the Ribominus™ Transcriptome Isolation kit for yeast and bacteria (Invitrogen, USA), in accordance with the manufacturer's recommendations. The rRNA-depleted RNA was used for cDNA synthesis using the SOLiD™ Total RNA-Seq kit in accordance with the standard protocol recommended by the manufacturer, and the material was quantified in a Qubit® 2.0 fluorometer (Invitrogen).

Sequencing in SOLiD™

The depleted RNA was fragmented using RNase III in preparation for amplification of the cDNA library, which was produced by reverse transcription from adapters attached to the ends of the RNA molecules, in accordance with the SOLiD™ Total RNA-Seq kit protocol (Life Technologies™, CA). Next, 6% denaturing polyacrylamide gel electrophoresis was performed and fragments of appropriate sizes (150 to 250 bases) were cut from the gel for cDNA amplification using PCR. Following recommended protocols, the cDNA was purified and the sizes were confirmed using 2% agarose electrophoresis. The PCR amplification in emulsion was performed using primers complementary to the adapters, in accordance with the Applied Biosystems SOLiD™ 3 Plus System Templated Bead Preparation Guide. After amplification, the microspheres were deposited onto slides for sequencing in accordance with the manufacturer's recommendations. The SOLiD™ 3 Plus system was used to sequence the 50-nucleotide RNA reads.

Analysis *in silico*

After obtaining the reading files using the SOLiD™ technique, the data were loaded into Bioscope version 1.2.1-5 programme (Life Technologies™, CA), and the gene expression data were obtained and quantified as reads per kilobase of coding sequence per million reads (RPKM) [15]. The DEGseq programme [16] was used to identify differentially expressed genes. The programme uses the output file from Bioscope and the RPKM values as the input file. For differentially expressed genes, a cutoff value of p <0.001 was defined. The data from each stimulon were analysed using Blast2GO (http://www.blast2go.com), and the results were exported to CoreStImulon programme [18] to search more rapidly for the genes comprising each biological process defined by Blast2GO.

To predict the non-coding RNA, a similarity search was performed in the RFAM database [60] using the script rfam_scan-1.0.3.pl (http://rfam.sanger.ac.uk/), and sRNAs smaller than 70 bp were discarded to minimise the number of false positives.

Transcriptomic coverage for all sRNAs annotated by RFAM was confirmed by manual curation in the program Artemis using the file BAM.

Additional files

Additional file 1: Figure S1 Report of the biological process for the osmotic medium. The file contains the genes induced in the biological processes in the osmotic medium stimulon, which exhibited fold-change values equal to or greater than 2x relative to the control. (PDF 26 kb)

Additional file 2: Table S1 Values of RPKM and fold-change of genes differentially expressed in the osmotic medium. The table contains the differentially expressed genes in the osmotic medium and their respective RPKM and fold-change values. The column marked TRUE indicates that the genes were considered differentially expressed (*p-value* <0.001), and FALSE indicates that the genes were not considered differentially expressed (*p-value* >0.001). (XLS 287 kb)

Additional file 3: Figure S2 Report on the biological process for the acid medium. The file contains genes induced from the biological processes in the acid medium stimulon, which exhibited fold-change values equal to or greater than 2x relative to the control. (PDF 19 kb)

Additional file 4: Table S2 Values of RPKM and fold-change of genes differentially expressed in acid medium. The table contains the differentially expressed genes in the acid medium and their respective

RPKM and fold-change values. The column marked TRUE indicates the genes were considered differentially expressed (*p-value* <0.001), and FALSE indicates that the genes were not considered differentially expressed (*p-value* >0.001). (XLS 285 kb)

Additional file 5: Figure S3 Report on the biological process under thermal shock. The file contains genes induced in the biological processes in the thermal shock stimulon, which exhibited fold-change values equal to or greater than 2x relative to the control. (PDF 18 kb)

Additional file 6: Table S3 Values of RPKM and fold-change of genes differentially expressed under thermal shock. The table contains the differentially expressed genes under thermal shock and their respective RPKM and fold-change values. The column marked TRUE indicates that the genes were considered differentially expressed (*p-value* <0.001), and FALSE indicates that the genes were not considered differentially expressed (*p-value* > 0.001). (XLS 806 kb)

Additional file 7: Table S4 Values of RPKM and fold-change of genes considered induced under the 3 conditions. The table contains the differentially expressed genes induced under the 3 conditions simultaneously and their respective fold-change values and product names. The table contains information for each stress separated by colour. (XLS 28 kb)

Additional file 8: Table S5 Values of RPKM and fold-change of genes considered repressed under the 3 conditions. The table contains the differentially expressed genes repressed under the 3 conditions simultaneously and their respective fold-change values and product names. The table contains information for each stress separated by colour. (XLS 20 kb)

Additional file 9: Table S6 Values of RPKM and fold-change of genes encoding sigma factors. The table contains the RPKM and fold-change values of the genes encoding sigma factors and the genes differentially or not differentially expressed are indicated. The asterisk indicates genes that were not considered differentially expressed. (XLS 22 kb)

Additional file 10: Figure S4 Growth Curve. Plot showing the growth curve of *C. pseudotuberculosis* strain 1002 under the control condition and measured at an optical density of 600 nm. Triangles indicate the density values at each hour. The arrow shows the time-point when the stresses were induced. (OD_{600nm} = 0.2) – indicates the beginning of the exponential phase. (PDF 73 kb)

Abbreviations

RNA-Seq: high-throughput sequencing of cDNA libraries; PLD: Phospholipase D; RPKM: Reads per kilobase of coding sequence per million mapped; CSI: CoreStImulon; H_2O_2: Hydrogen peroxide; msrB: peptide methionine sulphoxide reductase; Met: Methionine; cDNA: complementary DNA synthesised from RNA; rRNA: ribosomal RNA; tRNA: transfer RNA; Sig and σ: Sigma factor; spp: species; ncRNA: non-coding RNA; sRNA: small RNA; BHI: Brain heart infusion broth.

Competing interests

The authors declare there are no competing interests.

Authors' contributions

ACP, WMS and FSR performed the bacterial growth and stress application experiments. VA, AS, MPCS and AM offered support for the sequencing of the transcripts, preparation of the reagents and the analysis tools. VA coordinated and directed the research. ACP, SB and HPMB performed the sequencing experiments in SOLiD^{TM}. ACP analysed the data. PHCGS, RTJR and ACR offered support in bioinformatics. ACP, MPS and TLPC wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the Genome and Proteome Network of the State of Pará [Rede Paraense de Genômica e Proteômica] for supporting the development of this work and the Research Support Foundation of the State of Pará [Fundação de Amparo a Pesquisa do Estado do Pará] for funding the study. The authors also thank the National Council for Scientific and Technological Development [Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq] and the Coordination for Improvement of Higher Education Personnel [Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CAPES] for the financial support provided to the authors.

Author details

[1]Department of General Biology, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Av. Antônio Carlos, Belo Horizonte 31.270-901, Brazil. [2]Genome and Proteome Network of the State of Pará, Universidade Federal do Pará, R.Augusto Corrêa, Belém 66.075-110, Brazil.

Received: 4 July 2013 Accepted: 13 December 2013
Published: 9 January 2014

References

1. Dorella FA, *et al*: **Review article Corynebacterium pseudotuberculosis : microbiology , biochemical properties , pathogenesis and molecular studies of virulence.** *Vet Res* 2006, **37:**201–218.
2. Kazmierczak MJ, Wiedmann M, Boor KJ: **Alternative Sigma Factors and Their Roles in Bacterial Virulence.** *Society* 2005, **69:**527–543.
3. Jin Y, Tian Y, Zhang W, Jang SH, Jen AK, Meldrum DR: **Tracking bacterial infection of macrophages using a novel red-emission pH sensor.** *Anal Bioanal Chem* 2010, **398**(3)**:**1375–1384.
4. Ramos JL, Gallegos MT, Marqués S, Ramos-González MI, Espinosa-Urgel M, Segura a: **Responses of Gram-negative bacteria to certain environmental stressors.** *Curr Opin Microbiol* 2001, **4:**166–171.
5. Manganelli R, Dubnau E, Tyagi S, Kramer FR, Smith I: **Differential expression of 10 sigma factor genes in Mycobacterium tuberculosis.** *Mol Microbiol* 1999, **31**(2)**:**715–724.
6. Wu Q-L, Kong D, Lam K, Husson RN: **A mycobacterial extracytoplasmic function sigma factor involved in survival following stress.** *J Bacteriol* 1997, **179**(9)**:**2922–2929.
7. Missiakas D, Raina S: **The extracytoplasmic function sigma factors: role and regulation.** *Mol Microbiol* 1998, **28**(6)**:**1059–1066.
8. Tashjian J, Campbell S: **Interaction between caprine macrophages and Corynebacterium pseudotuberculosis: An electron microscopic study.** *Am J Vet Res* 1983, **44**(4)**:**690–693.
9. Billington SJ, Esmay PA, Songer JG, Jost BH: **Identification and role in virulence of putative iron acquisition genes from Corynebacterium pseudotuberculosis.** *FEMS Microbiol Lett* 2002, **208:**41–45.
10. Ca M, Gyles CL: **Characterization of strains of corynebacterium pseudotuberculosis.** *Can J Comp Med* 1982, **46:**206–208.
11. Wilson MJ, Brandon MR, Walker J: **Molecular and biochemical characterization of a protective 40-kilodalton antigen from Corynebacterium pseudotuberculosis.** *Infect Immun* 1995, **63:**206–211.
12. Sorek R, Cossart P: **Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity.** *Nat Rev Genet* 2010, **11:**9–16.
13. Sendler E, Johnson GD, Krawetz SA: **Local and global factors affecting RNA sequencing analysis.** *Anal Biochem* 2011, **419**(2)**:**317–322.
14. Jozefczuk S, Klie S, Catchpole G, Szymanski J, Cuadros-Inostroza A, Steinhauser D, Selbig J, Willmitzer L: **Metabolomic and transcriptomic stress response of Escherichia coli.** *Mol Syst Biol* 2010, **6:**364.
15. Mortazavi A, Williams BA, Mccue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5:**7–8.
16. Wang L, Feng Z, Wang X, Wang X, Zhang X: **DEGseq: an R package for identifying differentially expressed genes from RNA-seq data.** *Bioinformatics* 2010, **26**(1)**:**136–138.
17. Castro TL: **Ion Torrent-based transcriptional assessment of a Corynebacterium pseudotuberculosis equi strain reveals denaturing high-performance liquid chromatography a promising rRNA depletion method.** *Microb Biotechnol* 2013, **6**(2)**:**1–10.
18. Pinto AC, Ramos RT, Silva WM, Rocha FS, Barbosa S, Miyoshi A, Schneider MP, Silva A, Azevedo V: **The core stimulon of Corynebacterium pseudotuberculosis strain 1002 identified using ab initio methodologies.** *Integr Biol* 2012, **4**(7)**:**789–794.
19. Farr S, Kogoma T: **Oxidative stress responses in Escherichia coli and Salmonella typhimurium.** *Microbiol Rev* 1991, **55**(4)**:**561–585.
20. Sleator RD, Hill C: **Bacterial osmoadaptation: the role of osmolytes in bacterial stress and virulence.** *FEMS Microbiol Rev* 2002, **26**(1)**:**49–71.
21. Ratledge C, Dover LG: **Iron metabolism in pathogenic bacteria.** *Annu Rev Microbiol* 2000, **54:**881–941.

22. Sasindran SJ, Saikolappan S, Dhandayuthapani S: **Methionine sulfoxide reductases and virulence of bacterial pathogens.** *Future Microbiol* 2007, **2:**619–630.
23. Delaye L, Becerra A, Orgel L, Lazcano A: **Molecular Evolution of Peptide Methionine Sulfoxide Reductases (MsrA and MsrB): On the Early Development of a Mechanism That Protects Against Oxidative Damage.** *J Mol Evol* 2007, **64**(1):15–32.
24. Dhandayuthapani S, Jagannath C, Nino C, Saikolappan S, Sasindran SJ: **Methionine sulfoxide reductase B (MsrB) of Mycobacterium smegmatis plays a limited role in resisting oxidative stress.** *Tuberculosis (Edinb)* 2009, **89**(Suppl 1)**:**S26–32.
25. Alamuri P, Maier RJ: **Methionine sulphoxide reductase is an important antioxidant enzyme in the gastric pathogen Helicobacter pylori.** *Mol Microbiol* 2004, **53:**1397–1406.
26. Rajan LA, Joseph TC, Thampuran N, James R: **Functional Characterization and Sequence Analysis of Choline Dehydrogenase from Escherichia coli.** *Genet Eng Biotechnol J* 2010.
27. Das Gupta T, Bandyopadhyay B, Das Gupta SK: **Modulation of DNA-binding activity of Mycobacterium tuberculosis HspR by chaperones.** *Microbiology* 2008, **154:**484–490.
28. Stewart GR, Wernisch L, Stabler R, Mangan JA, Hinds J, Laing KG, Young DB, Butcher PD: **Dissection of the heat-shock response in Mycobacterium tuberculosis using mutants and microarrays.** *Microbiology* 2002, **148** (10)**:**3129–3138.
29. Bandyopadhyay B, Gupta TD, Roy D, Gupta SKD: **DnaK Dependence of the Mycobacterial Stress-Responsive Regulator HspR Is Mediated through Its Hydrophobic C-Terminal Tail.** *J Bacteriol* 2012, **194**(17)**:**4688–4697.
30. Isabella VM, Clark VL: **Deep sequencing-based analysis of the anaerobic stimulon in Neisseria gonorrhoeae.** *BMC Genomics* 2011, **12:**51.
31. Weerasinghe JP, Dong T, Schertzberg MR, Kirchhof MG, Sun Y, Schellhorn HE: **Stationary phase expression of the arginine biosynthetic operon argCBH in Escherichia coli.** *BMC Microbiol* 2006, **6:**14.
32. Weihofen W, Berger M, Chen H, Saenger W, Hinderlich S: **Structures of human N-Acetylglucosamine kinase in two complexes with N-Acetylglucosamine and with ADP/glucose: insights into substrate specificity and regulation.** *J Mol Biol* 2006, **364:**388–399.
33. Goodell EW, Higgins CF: **Uptake of cell wall peptides by Salmonella typhimurium and Escherichia coli.** *J Bacteriol* 1987, **169**(8)**:**3861–3865.
34. Rohde KH, Veiga DF, Caldwell S, Balázsi G, Russell DG: **Linking the transcriptional profiles and the physiological states of Mycobacterium tuberculosis during an extended intracellular infection.** *PLoS Pathog* 2012, **8**(6)**:**e1002769.
35. Hecker M, Völker U: **General stress response of Bacillus subtilis and other bacteria.** *Adv Microb Physiol* 2001, **44:**35–91.
36. Gomez M, Doukhan L, Nair G, Smith I: **sigA is an essential gene in Mycobacterium smegmatis.** *Mol Microbiol* 1998, **29:**617–628.
37. Helmann JD, Chamberlin MJ: **Structure and function of bacterial sigma factors.** *Annu Rev Biochem* 1988, **57**(1)**:**839–872.
38. Taylor WE, Straus DB, Grossman AD, Burton ZF, Gross CA, Burgess RR: **Transcription from a heat-inducible promoter causes heat shock regulation of the sigma subunit of E. coli RNA polymerase.** *Cell* 1984, **38**(2)**:**371–381.
39. Buttner M, Chater K, Bibb M: **Cloning, disruption, and transcriptional analysis of three RNA polymerase sigma factor genes of Streptomyces coelicolor A3 (2).** *J Bacteriol* 1990, **172**(6)**:**3367–3378.
40. Ehira S, Teramoto H, Inui M, Yukawa H: **Regulation of Corynebacterium glutamicum heat shock response by the extracytoplasmic-function sigma factor SigH and transcriptional regulators HspR and HrcA.** *J Bacteriol* 2009, **191:**2964–2972.
41. Agarwal N, Woolwine SC, Tyagi S, Bishai WR: **Characterization of the Mycobacterium tuberculosis sigma factor SigM by assessment of virulence and identification of SigM-dependent genes.** *Infect Immun* 2007, **75**(1)**:**452–461.
42. Nakunst D, Larisch C, Hüser AT, Tauch A, Pühler A, Kalinowski J: **The extracytoplasmic function-type sigma factor SigM of Corynebacterium glutamicum ATCC 13032 is involved in transcription of disulfide stress-related genes.** *J Bacteriol* 2007, **189:**4696–4707.
43. Kovács T, Hargitai A, Kovács KL, Mécs I: **pH-dependent activation of the alternative transcriptional factor σB in Bacillus subtilis.** *FEMS Microbiol Lett* 1998, **165**(2)**:**323–328.
44. Bischoff M, Entenza J, Giachino P: **Influence of a Functional sigB Operon on the Global Regulators sar and agr inStaphylococcus aureus.** *J Bacteriol* 2001, **183**(17)**:**5171–5179.
45. Ferreira A, Sue D, O'Byrne CP, Boor KJ: **Role of Listeria monocytogenes σB in survival of lethal acidic conditions and in the acquired acid tolerance response.** *Appl Environ Microbiol* 2003, **69**(5)**:**2692–2698.
46. Vijay K, Brody MS, Fredlund E, Price CW: **A PP2C phosphatase containing a PAS domain is required to convey signals of energy stress to the σB transcription factor of Bacillus subtilis.** *Mol Microbiol* 2000, **35**(1)**:**180–188.
47. Toledo-Arana A, Dussurget O, Nikitas G, Sesto N, Guet-Revillet H, Balestrino D, Loh E, Gripenland J, Tiensuu T, Vaitkevicius K, *et al*: **The Listeria transcriptional landscape from saprophytism to virulence.** *Nature* 2009, **459:**950–956.
48. Ferreira A, O'Byrne CP, Boor KJ: **Role of sigB in Heat, Ethanol, Acid, and Oxidative Stress Resistance and during Carbon Starvation in Listeria monocytogenes.** *Appl Environ Microbiol* 2001, **67**(10)**:**4454–4457.
49. Garner M, Njaa B, Wiedmann M, Boor K: **Sigma B contributes to Listeria monocytogenes gastrointestinal infection but not to systemic spread in the guinea pig infection model.** *Infect Immun* 2006, **74**(2)**:**876–886.
50. Craig JE, Nobbs A, High NJ: **The Extracytoplasmic Sigma Factor, sigE, Is Required for Intracellular Survival of Nontypeable Haemophilus influenzae in J774 Macrophages.** *Infect Immun* 2002, **70**(2)**:**708–715.
51. Kovacikova G, Skorupski K: **The alternative sigma factor σE plays an important role in intestinal survival and virulence in Vibrio cholerae.** *Infect Immun* 2002, **70**(10)**:**5355–5362.
52. Pacheco LG, Castro TL, Carvalho RD, Moraes PM, Dorella FA, Carvalho NB, Slade SE, Scrivens JH, Feelisch M, Meyer R, *et al*: **A Role for Sigma Factor sigma(E) in Corynebacterium pseudotuberculosis Resistance to Nitric Oxide/Peroxide Stress.** *Front Microbiol* 2012, **3:**126.
53. Manganelli R, Voskuil MI, Schoolnik GK, Smith I: **The Mycobacterium tuberculosis ECF sigma factor sigmaE: role in global gene expression and survival in macrophages.** *Mol Microbiol* 2001, **41:**423–437.
54. Dutta NK, Mehra S, Martinez AN, Alvarez X, Renner NA, Morici LA, Pahar B, Maclean AG, Lackner AA, Kaushal D: **The stress-response factor SigH modulates the interaction between Mycobacterium tuberculosis and host phagocytes.** *PloS One* 2012, **7**(1)**:**e28958.
55. Mikulík K, Palečková P, Felsberg J, Bobek J, Zídková J, Halada P: **SsrA genes of streptomycetes and association of proteins to the tmRNA during development and cellular differentiation.** *Proteomics* 2008, **8**(7)**:**1429–1441.
56. Mehta NB, *et al*: **Riboswitches: classification, function and insilico approach.** *Int J* 2010, **1:**409–420.
57. Weinberg Z, Wang JX, Bogue J, Yang J, Corbino K, Moy RH, Breaker RR: **Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes.** *Genome Biol* 2010, **11**(3)**:**R31.
58. Bessho Y, Shibata R, Sekine S-i, Murayama K, Higashijima K, Hori-Takemoto C, Shirouzu M, Kuramitsu S, Yokoyama S: **Structural basis for functional mimicry of long-variable-arm tRNA by transfer-messenger RNA.** *Proc Natl Acad Sci* 2007, **104**(20)**:**8293–8298.
59. Wassarman KM: **Small RNAs in bacteria: diverse regulators of gene expression in response to environmental changes.** *Cell* 2002, **109**(2)**:**141–144.
60. Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S, Finn RD, Nawrocki EP, Kolbe DL, Eddy SR: **Rfam: Wikipedia, clans and the "decimal" release.** *Nucleic Acids Res* 2011, **39**(suppl 1)**:**D141–D145.

doi:10.1186/1471-2164-15-14
Cite this article as: Pinto *et al.*: Differential transcriptional profile of *Corynebacterium pseudotuberculosis* in response to abiotic stresses. *BMC Genomics* 2014 **15**:14.

II.II.2 The core stimulon of *Corynebacterium pseudotuberculosis* strain 1002 identified using ab initio methodologies.
Pinto AC, Ramos RT, Silva WM, Rocha FS, Barbosa S, Miyoshi A, Schneider MP, Silva A, **Azevedo V**.
Integr Biol (Camb). 2012 Jul;4(7):789-94. doi: 10.1039/c2ib00092j. Epub 2012 May 28.

Baseados nos dados de RNAseq da *C.pseudotuberculosis*, foi possível desenvolver um software o qual permitiu visualizar os transcritos compartilhados entre as condiçoes estudadas (acidez, osmolaridade, termico) e integrar as informaçoes aos dados obtidos a partir do BLAST E BLAST2GO. Assim, foi criado o CoreStimulon, CSI, que permite distinguir os genes que compoe o processo biologico, a funçao e localizaçao celular, identificando e quantificando quais genes compoe cada processo.

Integrative Biology

Dynamic Article Links

Cite this: *Integr. Biol.*, 2012, **4**, 789–794

www.rsc.org/ibiology

PAPER

The core stimulon of *Corynebacterium pseudotuberculosis* strain 1002 identified using *ab initio* methodologies†‡

Anne Cybelle Pinto,§*[a] Rommel T. J. Ramos,§*[b] Wanderson Marques Silva,[c] Flávia Souza Rocha,[c] Silvanira Barbosa,[b] Anderson Miyoshi,[c] Maria P. C. Schneider,[b] Artur Silva[b] and Vasco Azevedo*[c]

Received 25th August 2011, Accepted 22nd April 2012

DOI: 10.1039/c2ib00092j

Corynebacterium pseudotuberculosis is a bacterium which causes diseases such as caseous lymphadenitis in small ruminants, resulting in large-scale economic losses for agribusiness worldwide. Consequently, this bacterium including its transcriptional profile analysis has been the focus of various studies. Identification of the transcripts that appear under conditions that simulate the environment encountered by this bacterial species in the host is of great importance in discovering new targets for the production of more efficient vaccines. We sequenced the cDNA of *Corynebacterium pseudotuberculosis* strain 1002, using the SOLiD V3 system, under the following conditions: osmotic stress (2 M), acidity (pH), heat shock (50 °C) and control condition (*N*). To identify the transcripts shared among the stimulons and integrate this information with the results from BLAST and BLAST2GO, we developed the software CoreStImulon (CSI) which allows the user to individually distinguish the genes in terms of their participation in biological processes, their function and cellular location. In the biosynthetic processes, eleven genes represented in the core stimulon and twenty genes in the control were observed. This validates the hypothesis that the organisms strategy for surviving in a hostile environment is through growth reduction. The oxidation reduction process, response to stress process, and cell adhesion are controlled by genes that contribute to bacterial cell maintenance under stress conditions; these could be involved in their pathogenicity. The methodology for identification of transcripts obtained by *ab initio* assembly and shared among the stimulons permitted candidates selection for vaccine studies. CSI is available at https://sourceforge.net/projects/corestimulon/.

Introduction

Corynebacterium pseudotuberculosis (*C. pseudotuberculosis*) is a Gram-positive pathogenic bacterium which belongs to the CMNR group, and causes diseases such as caseous lymphadenitis, which mainly affects goats and sheep, resulting in economic losses to goat and sheep farmers. Unfortunately, few vaccine targets have been identified, which makes development of proper control measures difficult.[1] Using the new sequencing technologies, it has become possible to obtain complete genomes of various strains of this species[2] facilitating further studies concerning this bacterium.[3] After the genome structure became known, discovery of genes that are differentially expressed under

Insight, innovation, integration

The Core Stimulon allowed identification of expressed genes under different conditions based on *ab initio* assemblies and the integration with its Blast2GO results grouped by biological process, function and cell component which enabled a selection of possible vaccine targets, some experimentally characterized and others that will be studied in the future. Furthermore, new transcripts were identified through Core Stimulon database and they will be incorporated into the genome annotation previously deposited.

[a] *Department of Microbiology, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Av. Antônio Carlos, Belo Horizonte, 31.270-901, Brazil. E-mail: acybelle@gmail.com; Tel: +55 (31) 34092873*
[b] *Genome and Proteome Network of the State of Pará, Universidade Federal do Pará, R.Augusto Corrêa, Belém, 66.075-110, Brazil. E-mail: rommelramos@ufpa.br*
[c] *Department of General Biology, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Av. Antônio Carlos, Belo Horizonte, 31.270-901, Brazil. E-mail: vasco@icb.ufmg.br; Tel: +55 (31) 34092610*

† Published as part of an iBiology themed issue entitled "Computational Integrative Biology". Guest Editor: Prof. Jan Baumbach.

‡ Electronic supplementary information (ESI) available. See DOI: 10.1039/c2ib00092j

§ These authors contributed equally to this work.

Table 1 Number of transcribed hits obtained using the Blast search in relation to CDS of strain 1002

Condition	Hits transcribed *versus* CDS (strain 1002)
2 M	299 (42 153 bp)
50 °C	200 (26 901 bp)
N	95 (12 897 bp)
pH	380 (53 118 bp)
Total	974 (135 069 bp)

Quantity of transcripts and DNA base pairs, by condition, without significant hits in relation to the CDS databank of strain 1002, using blastx.

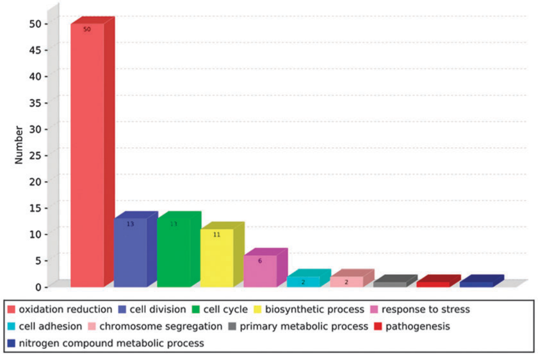


Fig. 1 Frequencies of genes that are part of the Core Stimulon, divided by the process (level 3) of the Blast2GO.

conditions that simulate the environment encountered by this microorganism during the infectious process will help identify new targets. In response to stress, the bacteria can activate the transcription of genes/operons, the products of which have important roles in bacterial cell maintenance in response to environmental challenges, or repress expression of transcripts that are not necessary at certain times.[4] The group of genes or operons that are differentially expressed in response to an environmental challenge is denominated stimulon. With reference to the concept of a core genome, which is a set of normally obtained orthologous genes in bacterial genome studies,[5] we use the term core stimulon, to determine the set of transcripts that are shared (but not orthologs) between stimulons, in order to study different environmental conditions simulated *in vitro* in a species: *C. pseudotuberculosis* 1002.

Response to stress normally involves a combination of a specific response, to minimize deleterious effects, and a general response to repress genes involved in ribosomal translation and biogenesis.[6] These cause the cessation or reduction in growth that is generally observed in response to stress, this being an important strategy for physiological cellular to a novel environmental condition.[7]

In this study, we used SOLiD™ technology to investigate the transcriptional profile of *C. pseudotuberculosis* 1002 under conditions that simulate the host environment: BHI supplemented with sodium chloride (2 M), with hydrochloric acid (pH), to simulate the acidity of macrophages, and heat shock (50 °C), along with BHI medium used as a control (*N*). To facilitate integration of the transcriptome data and identify which transcripts are shared among the stimulons (core stimulon) and their Gene Ontology (GO) classification, we developed the software CoreStImulon (CSI).

Results and discussion

The transcripts under control (*N*), acid stress (pH), thermal stress (50 °C) and osmotic stress (2 M) conditions, obtained using NGS sequencing, and assembled using an *ab initio* approach were run against the CDS (coding sequence) of the *C. pseudotuberculosis* 1002 genome using Blast (blastx), to identify transcripts but without considering expression levels, however only the presence of the transcript, achieved through *ab initio* assembly, is already an indication that there was expression in each condition and detection by CoreStimulon. One hundred genes that were expressed under all conditions, except for the control condition, were selected for analysis (ESI 1‡). For each condition, we identified the transcripts that did not give hits with alignments greater than 40 bp and *e*-value less than or equal to 1×10^{-5} (Table 1), giving a total of 974 sequences that may represent new coding regions not specified in advance in genome annotation.

Based on the Blast2GO results, the CSI program allowed us to identify the genes present in the various biological process, molecular function and cell location categories. The genes that are part of the core stimulon were analyzed individually and the classification of each of them was visualized in the report (ESI 1‡). Also, we were able to visualize the number of genes present in GO terms for the process (Fig. 1), followed by identification of all of the genes present in each term (ESI 1‡).

Through the incorporated Blast2GO results of the level 3 biological process into CSI, it was observed that the core stimulon genes are mainly involved in the processes of oxidative reduction, cell division, cell cycle, and biosynthetic process (Fig. 1), which allowed us to infer that they are activated when the cell encounters a hostile environment. In this way it was found that strain 1002 cells reduce but do not stop growth, as also reported by Schell[7] who indicated that this is a survival strategy.

Proof of this fact was obtained through experimental "colony forming unit" data, which indicated an average of 28% reduction in bacterial growth when subjected to the various types of stress as illustrated in Fig. 2 besides *in silico* confirmation, in which the biosynthetic process of the condition control

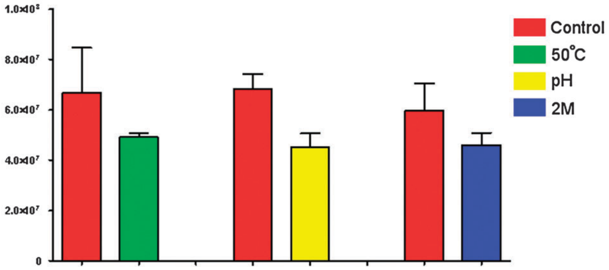


Fig. 2 Number of viable cells under each condition. Control condition amounts, on average, 6.70×10^{7} mL^{-1} cells of the strain 1002, except for control of osmotic which is equivalent to 6.0×10^{7} mL^{-1}. red, condition control; green, thermal stress; yellow, acidic stress; blue, osmotic stress.

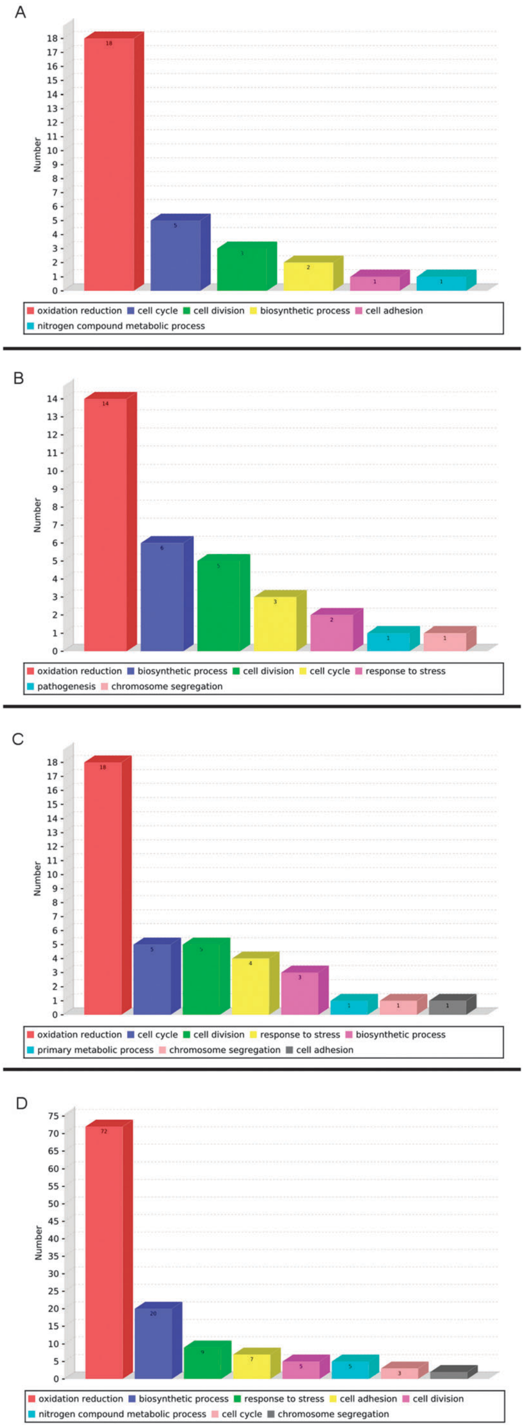


Fig. 3 Distribution of the transcripts according to GO terms and conditions of occurrence. A, osmotic stress; B, thermal stress; C, acid stress; D, control (except the core stimulon genes).

(Fig. 3D) takes the second position, presenting twenty genes involved in this activity, while in the core stimulon eleven genes are candidates for this role, and its process takes the fourth place (Fig. 1).

The genes involved in the processes of stress response, cell adhesion, and pathogenesis were few, though they have great importance for maintaining the bacterium under stress conditions. Through these results, it can be inferred that the adaptive response is already activated at the beginning of the exponential growth phase under all stress conditions, since a few specific genes are activated.

The redox process is among those most activated in the core stimulon, possibly because it is found in all aerobic organisms, functioning as a defense against toxic oxidants that result from incomplete reduction of oxygen.[8] A total amount of fifty genes were candidate genes for this activity in the core stimulon; in the control process 72 genes were found to be involved which was not detected in the core. Analyzing the list of genes which composes each biological process of level 3, which are associated with the terms GO, it can be observed that in the core stimulon there probably existed genes with accentuated expression when the amount of reactive oxygen species exceeds the baseline, by exposure to certain environments, such as catalase, which plays the role of eliminating excess hydrogen peroxide which tends to increase their production in hostile environments.[9] Under the control condition, the amount of genes increases, possibly because all transcribed GO term genomes were considered an end, however the genes which are part of this process (ESI 2‡) are probably related to a response to the active oxygen molecules produced as an inevitable by-product of normal aerobic metabolism (Fig. 3A–C).

In the core stimulon, the process of stress response is composed of six genes, among them is the gene that encodes the putative Cp1002_0660 "two component system response regulator", which is also part of the process of pathogenesis (ESI 1‡). The regulator of a two-component system, together with the sensor (histidine kinase), is of utmost importance for the adaptation of the bacterium in different media, for regulating the expression of genes that contribute to the survival and growth in a hostile environment.[10] Under the control condition, a larger number of the genes composes this process, and while analyzing them (ESI 1‡), it was observed that they are arguably encoding of proteins involved in a general response to stress.

Another gene which is part of this process is the *dps* gene (Cp1002_2043) (ESI 1‡), which encodes the supposed protein to DNA protection during starvation. These proteins are capable of storing iron in the bioavailable form and protect cells against oxidative stress.[11] The hydroxyl radical is able to induce breaks in DNA, lipid peroxidation and degradation of biomolecules. The protection provided by Dps protein occurs in two ways: either by connecting to the Fe^{2+} ion preventing the formation of toxic hydroxyl radicals catalyzed by the Fenton reaction or by connecting to the DNA protecting oxidative radicals.[12] And as you can see the process of oxidoreduction occurs intensively among the stresses, and the Dps protein seems to play an essential role in the survival of bacteria in the middle, seeking to reduce the harmful effects caused by the process that proved to be the most abundant in the cell.

In the cellular adhesion, the gene Cp1002_1765 which codes the supposed "secreted protein" (ESI 1‡) promotes interaction of the bacteria with the host cell, allowing it to establish a pathogenic relationship. This adhesion involves surface organelles, including pili or fimbriae.

The pili of Gram-positive bacteria have not been studied extensively, though their importance for establishing the infectious process is known.[13] Under the control condition, seven genes were candidates for this function, however, only two genes were found in the core stimulon, which makes it an important candidate for studies of vaccine targets, since it is positively regulated under stress conditions.

The other gene involved in the core stimulon adhesion is described as "ABC type iron uptake system substrate-binding protein" (ESI 1‡). Iron is essential and required by bacteria for survival and for a successful infectious process. Ability to acquire iron is probably the principal determinant for discovering whether a microorganism is capable of maintaining itself within the host.[14]

Without this ability, the microorganism would be incapable of growing and would be eliminated by the host defense system or would die due to a lack of nutrients.

Acquisition of iron is recognized to be one of fundamental steps for the development of any pathogen in a host.[15] Consequently, the iron transport system that was induced under all of the various stress conditions in *C. pseudotuberculosis* 1002 confirms one of the capabilities of this bacterium, allowing it to survive within the host.

The program CSI allowed us to detect the core stimulon, where genes shared by the stimulons under stress conditions, and that are involved in the adaptive response, are expressed. This allows rapid adjustments within a hostile ambient, guaranteeing survival and an infectious process. The program facilitates the search for possible vaccine candidates, since it reduces the number of genes that need to be analyzed, selecting only genes that are expressed under certain conditions.

Using this program, we were able to identify the genes that are involved in the processes, functions and cellular components through incorporation of Blast2GO results, which facilitated identification of the targets therefore giving the user category options to choose from in analyzing the results (in this work, we opted for the biological process). The advantage of CSI compared to Blast2GO is the simplification of the process of identifying each gene that comprises the categories of Blast2GO, which features an extensive list of the genes and their respective categories (biological processes, cellular components and functions), complicating the search by a specific gene.

Materials and methods

Sample data

The bacterium *Corynebacterium pseudotuberculosis* strain 1002 (CP001809.1), isolated from infected goats in Brazil,[3] was used as the model organism for this research.

Cultivation of the bacteria and obtaining bacterial cells

C. pseudotuberculosis 1002 was grown for 24 hours in Brain Heart Infusion broth (BHI: Oxoid, Hampshire, UK) media at 37 °C, 160 rpm, diluted 1 : 100. Then the bacteria were transferred to new BHI media, diluted 1 : 100, grown at 37 °C, at 160 rpm, until reaching the beginning of the exponential growth phase (A_{600} = 0.2). After attaining the optical density, the inoculation was distributed among four 50 mL Falcon tubes, each containing a final volume of 15 mL, centrifuged for three minutes at 8000 rpm, at room temperature and each pellet resuspended in new BHI medium modified according to each stress condition. To apply the acid stress, the BHI medium was supplemented with hydrochloric acid, until it reached a pH of 5; osmotic stress was induced by the addition of NaCl to the medium, until it reached 2 M; and thermal stress was applied by transferring the tube to a shaker maintained at 50 °C.

All of the tubes, including the control resuspended in BHI medium, were maintained at 37 °C, except for the thermal stress trial, and agitated at 160 rpm for 15 minutes. Decimal dilutions were then made from 10^{-1} to 10^{-6}; the 10^{-4}–10^{-6} dilutions were seeded onto BHI 1.5% bacteriological agar plates, at 37 °C to determine viability and to make colony counts. The rest of the inoculant was centrifuged for three minutes at 8000 rpm at room temperature and the pellet resuspended in 2 mL of RNAlater, following the manufacturer's instructions.

Obtaining RNA and sequencing with SOLiD™

The bacteria stabilized with RNAlater® were subjected to total RNA extraction with the kit ChargeSwitch® Total RNA Cell (Invitrogen), following the manufacturer's recommendations, with the following adaptations: after addition of lysing buffer (Invitrogen), the material was transferred to 2 mL tubes partially filled with 1 mm glass microspheres (Bertin Technologies). The cells were mechanically lysed using a Prescellys 24 homogenizer configured to agitate tubes at 6500 rpm for two 15 second cycles, with an interval of 30 seconds between the cycles.

The samples were then centrifuged for one minute and the supernatant transferred to new 2 mL tubes, which were then incubated in a dry bath at 600 °C for 15 minutes, following the original protocol. DNase was added to eliminate the residual genomic DNA. Elution of the total RNA on the magnetic beads was done with 100 µL of Milli-Q RNase-free water. The quantity of total RNA was measured using a Qubit® 2.0 Fluorometer (Invitrogen). In order to enrich the mRNA, the rRNA of each sample of total RNA was removed using the Invitrogen Ribominus™ Transcriptome Isolation (Yeast and Bacteria) kit, following manufacturer's recommendations. The RNA depleted of rRNA was used to synthesize the cDNA with the standard protocol of the SOLiD™ Total RNA-Seq kit, according to recommendations, and the material was quantified in the Qubit® 2.0 Fluorometer (Invitrogen).

The depleted RNA was fragmented using RNase III, to prepare amplification of the cDNA library produced through reverse transcription, using adapters bound to the extremities of RNA molecules, following the SOLiD™ Total RNA-Seq Kit (Applied Biosystems, CA) protocol. After this step, the material was run through a 6% denaturing polyacrylamide gel and the appropriate size fragments (150–250 bases) were cut for cDNA amplification by PCR. Immediately afterward, continuing the protocol, the reaction was purified, confirmed by electrophoresis in 2% agarose gels. This was followed by emulsion PCR amplification, using primers complementary to the adapters based on recommendations from the Applied Biosystems SOLiD™ 3 Plus System Templated Bead Preparation Guide. After amplification, the microspheres were deposited onto sequencing slides based on manufacturer's recommendations. The system used to sequence the 50 nt RNA reads was SOLiDTM 3 plus.

Data treatment

The transcripts obtained from sequencing with SOLiD V3 were submitted to a quality filter with the software Quality Assessment[16] which removed the reads with read quality lower than Phred 15, and sequencing errors were corrected with SAET (SOLiD™ Accuracy Enhancement Tool).

Assembling the transcripts

The reads obtained from the cDNA sequencing were assembled with the software Velvet[17] with variation in the parameters for all conditions (Table 2), totaling 60 assemblies for each condition.

Four multifasta files were generated, containing the contigs obtained from all of the assemblies for each condition. These files were submitted to Simplifier (Rommel Ramos, unpublished data) to remove redundant sequences, reducing the effort required to curate them.

Identification of the transcripts

The transcripts generated by Velvet were used to run Blastx on the CDS of the *C. pseudotuberculosis* 1002 genome. The CDS found under all conditions, except normal, with alignments greater than 40 bp and with *e*-value lower than or equal to 1×10^{-5} were identified, constituting the core stimulon.

Blast2GO

To analyze the gene products in the functional categories, the multifasta files containing the transcripts (NT) were submitted to Blast2GO, running the following steps: Blast, GO mapping and annotation, and then exportation of the sequence table through the menu, menu File > Export.

CoreStImulon tool

CoreStImulon (CSI) receives as entry files the CDS of the reference genome and the transcripts obtained from the *ab initio* assembly, besides the Blast (blastx) results of the transcript of each condition in relation to the CDS; it generates analytic reports of the core stimulon, consisting of overall graphs and graphs specific for each component, function and biological process using the GO level, when results of Blast2GO (optional) are included. This software was implemented with the JAVA programming language using Swing (http://java.sun.com/), and reports generated using the software IReport (http://java.sun.com/docs/books/tutorial/uiswing/). CSI is available at https://sourceforge.net/projects/corestimulon/.

Table 2 Ranges of parameters used to produce the transcript

Condition	K-mer (range)	Cov. cutoff (range)	Expec. Cov. (range)
pH	29–35 (2)	2–10 (2)	5–45 (20)
2 M	29–35 (2)	2–10 (2)	5–45 (20)
50 °C	29–35 (2)	2–10 (2)	5–45 (20)
Control	29–35 (2)	2–10 (2)	5–45 (20)

Ranges of parameters used to produce the transcripts for the conditions pH, 2 M, 50 °C and control.

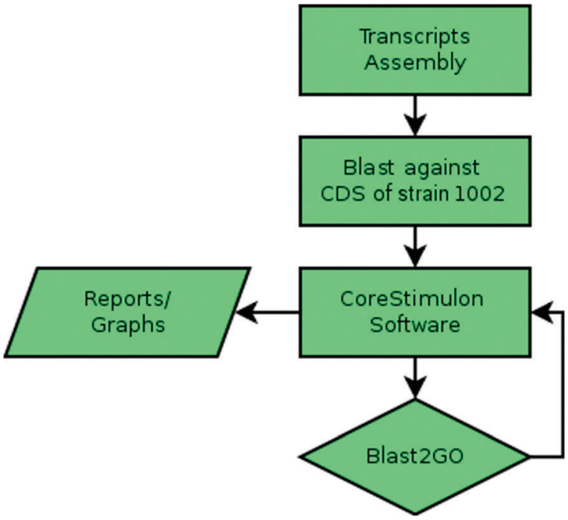


Fig. 4 Identification pipeline of the Core Stimulon.

Pipeline

The steps followed to identify and characterize transcripts shared among the different stress conditions are presented in Fig. 4.

Conclusions

CoreStImulon (CSI) allowed us to detect the core stimulon, group of exclusive common genes of stress conditions involved in the adaptation, responsible for rapid response in hostile environments. Thereby, the CSI helped us to identify possible vaccine candidates by reducing the amount of genes to be analyzed, thereby enabling data integration with Blast2GO and simplification of the search for genes associated with specific GO terms.

Acknowledgements

This work was part of the Rede Paraense de Genômica e Proteômica supported by Fundação de Amparo a Pesquisa do Estado do Pará. V.A., A.S., M.P.C.S., F.S.R., and A.C.P. were supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). R.T.J.R. and W.M.S. acknowledge support from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). We also thank Sheila Gordo and Cibelle Moreira for constructive criticism of the manuscript.

Notes and references

1 F. A. Dorella, L. G. C. Pacheco, S. C. Oliveira, A. Miyoshi and V. Azevedo, *Vet. Res.*, 2006, **37**, 201.
2 (*a*) J. C. Ruiz, V. D'Afonseca, A. Silva, A. Ali, A. C. Pinto, A. R. Santos, A. a. M. C. Rocha, D. O. Lopes, F. a. Dorella, L. G. C. Pacheco, M. P. Costa, M. Z. Turk, N. Seyffert, P. M. R. O. Moraes, S. C. Soares, S. S. Almeida, T. L. P. Castro, V. a. C. Abreu, E. Trost, J. Baumbach, A. Tauch, M. P. C. Schneider, J. McCulloch, L. T. Cerdeira, R. T. J. Ramos, A. Zerlotini,

A. Dominitini, D. M. Resende, E. M. Coser, L. M. Oliveira, A. L. Pedrosa, C. U. Vieira, C. T. Guimarães, D. C. Bartholomeu, D. M. Oliveira, F. R. Santos, É. M. Rabelo, F. P. Lobo, G. R. Franco, A. F. Costa, I. M. Castro, S. R. C. Dias, J. a. Ferro, J. M. Ortega, L. V. Paiva, L. R. Goulart, J. F. Almeida, M. I. T. Ferro, N. P. Carneiro, P. R. K. Falcão, P. Grynberg, S. M. R. Teixeira, S. Brommonschenkel, S. C. Oliveira, R. Meyer, R. J. Moore, A. Miyoshi, G. C. Oliveira and V. Azevedo, *PLoS One*, 2011, **6**, e18551; (*b*) A. Silva, M. P. C. Schneider, L. Cerdeira, M. S. Barbosa, R. T. J. Ramos, A. R. Carneiro, R. Santos, M. Lima, V. D. Afonseca, S. S. Almeida, A. R. Santos, S. C. Soares, A. C. Pinto, A. Ali, F. A. Dorella, F. Rocha, V. Augusto, C. D. Abreu, E. Trost, A. Tauch, N. Shpigel, A. Miyoshi and V. Azevedo, *J. Bacteriol.*, 2011, **193**, 323–324; (*c*) E. Trost, L. Ott, J. Schneider, J. Schröder, A. Goesmann, P. Husemann, J. Stoye, A. Dorella, F. S. Rocha, S. D. C. Soares, D. Afonseca, A. Miyoshi, J. Ruiz, A. Silva, A. Burkovski, N. Guiso and O. F. Join-lambert, *Genome*, 2010, **11**, 728.

3 (*a*) L. G. C. Pacheco, S. E. Slade, N. Seyffert, A. R. Santos, T. L. Castro, W. M. Silva, A. V. Santos, S. G. Santos, L. M. Farias, M. A. Carvalho, A. M. Pimenta, R. Meyer, A. Silva, J. H. Scrivens, S. C. Oliveira, A. Miyoshi, C. G. Dowson and V. Azevedo, *BMC Microbiol.*, 2011, **11**(1), 12; (*b*) D. Barh, N. Jain, S. Tiwari, B. P. Parida, V. D'Afonseca, L. Li, A. Ali, A. R. Santos, L. C. Guimarães, S. de Castro Soares, A. Miyoshi, A. Bhattacharjee, A. N. Misra, A. Silva, A. Kumar and V. Azevedo, *Chem. Biol. Drug Des.*, 2011, **78**(1), 73.

4 J. L. Ramos, M. T. Gallegos, S. Marqués, M. I. Ramos-González, M. Espinosa-Urgel and A. Segura, *Curr. Opin. Microbiol.*, 2001, **4**, 166–171.

5 R. Hengge-Aronis, *J. Mol. Microbiol. Biotechnol.*, 2002, **4**, 341–346.

6 S. Jozefczuk, S. Klie, G. Catchpole, J. Szymanski, A. Cuadros-Inostroza, D. Steinhauser, J. Selbig and L. Willmitzer, *Mol. Syst. Biol.*, 2010, **6**, 364.

7 M. A. Schell, *Organization*, 1993, **47**, 597–626.

8 N. A. Buchmeier, C. J. Lipps, M. Y. So and F. Heffron, *Mol. Microbiol.*, 1993, **7**, 933–936.

9 R. G. Lloyd, *J. Bacteriol.*, 1991, **173**, 5414–5418.

10 A. M. Stock, V. L. Robinson and P. N. Goudreau, *Annu. Rev. Biochem.*, 2000, **69**, 183–215.

11 G. E. Soto and S. J. Hultgren, *J. Bacteriol.*, 1999, **181**, 1059–1071.

12 S. Tokishita and T. Mizuno, *Mol. Microbiol.*, 1994, **13**, 435–444.

13 D. a. Los and N. Murata, *Biochim. Biophys. Acta*, 2004, **1666**, 142–157.

14 M. Sritharan, *Indian J. Med. Microbiol.*, 2006, **24**, 163–164.

15 C. Ratledge and L. G. Dover, *Annu. Rev. Microbiol.*, 2000, **54**, 881–941.

16 R. T. J. Ramos, A. R. Carneiro, Ribeiro, J. Baumbach, V. Azevedo, M. P. C. Schneider and A. Silva, *BMC Res. Notes*, 2011, **4**, 130.

17 D. R. Zerbino and E. Birney, *Genome Res.*, 2008, **18**, 821–829.

II.III PROTEÔMICA

II.III.1 Conserved host-pathogen PPIs. Globally conserved inter-species bacterial PPIs based conserved host-pathogen interactome derived novel target in *C. pseudotuberculosis*, *C. diphtheriae*, *M. tuberculosis*, *C. ulcerans*, *Y. pestis*, and *E. coli* targeted by Piper betel compounds.

Barh D, Gupta K, Jain N, Khatri G, León-Sicairos N, Canizalez-Roman A, Tiwari S, Verma A, Rahangdale S, Shah Hassan S, dos Santos AR, Ali A, Guimarães LC, Thiago Jucá Ramos R, Devarapalli P, Barve N, Bakhtiar M, Kumavath R, Ghosh P, Miyoshi A, Silva A, Kumar A, Misra AN, Blum K, Baumbach J, **Azevedo V**.

Integr Biol (Camb). 2013 Mar;5(3):495-509. doi: 10.1039/c2ib20206a.

Integrative Biology

RSCPublishing

PAPER

View Article Online
View Journal

Conserved host–pathogen PPIs†

Cite this: DOI: 10.1039/c2ib20206a

Debmalya Barh,‡*[ab] Krishnakant Gupta,[ac] Neha Jain,[a] Gourav Khatri,[ac] Nidia León-Sicairos,[d] Adrian Canizalez-Roman,[d] Sandeep Tiwari,[a] Ankit Verma,[ac] Sachin Rahangdale,[ac] Syed Shah Hassan,[e] Anderson Rodrigues dos Santos,[e] Amjad Ali,[e] Luis Carlos Guimarães,[e] Rommel Thiago Jucá Ramos,[f] Pratap Devarapalli,[g] Neha Barve,[ac] Marriam Bakhtiar,[e] Ranjith Kumavath,[g] Preetam Ghosh,[ah] Anderson Miyoshi,[e] Artur Silva,[f] Anil Kumar,[c] Amarendra Narayan Misra,[bi] Kenneth Blum,[ajkl] Jan Baumbach[m] and Vasco Azevedo‡[e]

Received 27th August 2012,
Accepted 5th November 2012

DOI: 10.1039/c2ib20206a

www.rsc.org/ibiology

Although attempts have been made to unveil protein–protein and host–pathogen interactions based on molecular insights of important biological events and pathogenesis in various organisms, these efforts have not yet been reported in *Corynebacterium pseudotuberculosis* (*Cp*), the causative agent of Caseous Lymphadenitis (CLA). In this study, we used computational approaches to develop common conserved intra-species protein–protein interaction (PPI) networks first time for four *Cp* strains (*Cp FRC41*, *Cp 316*, *Cp 3/99-5*, and *Cp P54B96*) followed by development of a common conserved inter-species bacterial PPI using conserved proteins in multiple pathogens (*Y. pestis, M. tuberculosis, C. diphtheriae, C. ulcerans*, *E. coli*, and all four *Cp* strains) and *E. Coli* based experimentally validated PPI data. Furthermore, the interacting proteins in the common conserved inter-species bacterial PPI were used to generate a conserved host–pathogen interaction (HP-PPI) network considering human, goat, sheep, bovine, and horse as hosts. The HP-PPI network was validated, and acetate kinase (Ack) was identified as a novel broad spectrum target. Ceftiofur, penicillin, and two natural compounds derived from *Piper betel* were predicted to inhibit Ack activity. One of these *Piper betel* compounds found to inhibit *E. coli O157:H7* growth similar to penicillin. The target specificity of these *betel* compounds, their effects on other studied pathogens, and other *in silico* results are currently being validated and the results are promising.

[a] *Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, West Bengal-721172, India. E-mail: dr.barh@gmail.com; Fax: +91-944 955 0032; Tel: +91-944 955 0032*

[b] *Department of Biosciences and Biotechnology, School of Biotechnology, Fakir Mohan University, Jnan Bigyan Vihar, Balasore, Orissa, India*

[c] *School of Biotechnology, Devi Ahilya University, Khandwa Road Campus, Indore, MP, India*

[d] *Unidad de investigacion, Facultad de Medicina, Universidad Autónoma de Sinaloa. Cedros y Sauces, Fraccionamiento Fresnos, Culiacán Sinaloa 80246, México*

[e] *Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil*

[f] *Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém, PA, Brazil*

[g] *Department of Genomic Science, School of Biological Sciences, Riverside Transit Campus, Central University of Kerala, Kasaragod, India*

[h] *Department of Computer Science and Center for the Study of Biological Complexity, Virginia Commonwealth University, 401 West Main Street, Room E4234, P.O. Box 843019, Richmond, Virginia 23284-3019, USA*

[i] *Center for Life Sciences, School of Natural Sciences, Central University of Jharkhand, Ranchi, Jharkhand State, India*

[j] *University of Florida, College of Medicine, Gainesville, Florida, USA*

[k] *Global Integrated Services Unit University of Vermont Center for Clinical & Translational Science, College of Medicine, Burlington, VT, USA*

[l] *Dominion Diagnostics LLC, North Kingstown, Rhode Island, USA*

[m] *Computational Biology Group, Department of Mathematics and Computer Science, University of Southern Denmark, Campusvej 55, DK-5230 Odense, Denmark*

† Electronic supplementary information (ESI) available: Supplementary Tables 1–8. See DOI: 10.1039/c2ib20206a

‡ DB and VA conceived the idea; DB designed the study, collected and analyzed primary data to finalize the protocol, coordinated and leaded the entire project, and wrote the manuscript. DB, KG, NJ, GK, ST, AV, and SR performed all *in silico* analyses; SSH, ARS, AA, LCG, and ATJR performed *Cp* genome annotation and cross checked all other analyses; PD, RK, MB, NB, and PG cross checked all analyses; NLS and ACR conducted microbial experiments with betel compounds; AK, KB, ANM, AM, PG, JB, and VA provided technical consultations and reviewed the manuscript. All authors have read and approved the final manuscript.

Insight, innovation, integration

Here, for the first time we represent the intra-species PPIs in *C. pseudotuberculosis* (*Cp*). Further, a novel method was used to develop common conserved inter-species bacterial PPIs for *C. pseudotuberculosis*, *Y. pestis*, *M. tuberculosis*, *C. diphtheriae*, *C. ulcerans*, *C. glutamicum*, and *E. coli* (pathogenic, nonpathogenic, closed and distant taxa) to identify the conserved common essential PPIs in these bacteria. This inter-species bacterial PPI was then used to make conserved common host–pathogen interactions. Using network analysis strategies and subtraction genomics approaches, from this conserved common host–pathogen interactions; Ack was identified as a key target for all these bacteria. Virtual screening shows Penicillin and Ceftiofur can inhibit Ack. However, *Piper betel* derived Piperdardine and Dehydropipernonaline are predicted to have similar or superior effects compared to Penicillin and Ceftiofur on Ack. Piperdardine inhibits *E. coli O157:H7* growth similar to penicillin and can also work on other pathogens in a similar way.

Introduction

Protein–protein interactions (PPIs) are crucial events in several biological processes. PPI-based decoding of the functionality of uncharacterized proteins can reveal unknown molecular mechanisms behind important biological events within a cell or at the system level.[1,2] Therefore, PPIs of an entire proteome or between a set of proteins in a pathogen and its corresponding host can be useful in identifying precise molecular mechanisms of host–pathogen interactions, thereby leading to the development of effective drug targets against the pathogen.[3–5] Initial computational approaches for the prediction of PPIs were based on the structural context of proteins. However, in the post-genomic era, the focus has shifted, and sequence information is now used.[6,7] The availability of genomic and proteomic data and the advent of yeast two-hybrid, affinity purification, mass spectrometry, and other high-throughput techniques have tremendously enriched the field. Recently, a number of computational approaches have also been developed to facilitate the prediction and study of these ubiquitous interactions. A number of *in silico* approaches were recently reviewed that highlight the use of genomic, structural, and biological contexts of proteins and genes in complete genomes for PPI predictions and determination of the functional relationship among them.[8] Using these approaches, the development of highly reliable PPIs in several organisms including yeast[9] and human[10] are close to completion. However, false-positive interactions are a concern.[11,12] Similarly, sequence-based computational methods including gene neighborhood,[13] phylogenetic profiles,[14] gene fusion,[15] co-evolution,[16] and domain interactions,[17] along with several newly developed methods, have been used to generate genome-/proteome-wide interactions in a number of organisms including, *M. tuberculosis*[18] and *E. coli*.[19] Genomic sequences are used as the primary data sources in these prediction techniques, which assume that evolutionary co-inherited gene pairs have a functional association.[20,21] Similarly, amino acid (AA) sequence-based PPIs identify interacting protein pairs that have specific AA residues due to their co-evolution or binding to one another.[22] Yeats *et al.* have catalogued the commonly occurring domains for PPIs.[23–25] However, in general, a PPI denotes the binding of proteins to other proteins.

Concurrently, *in silico* host–pathogen interactions have been reported in many organisms, including *Plasmodium*,[26,27] *M. tuberculosis*,[28] and *Streptococcus*.[29] Combined computational and yeast two-hybrid based approaches have been recently published for *B. anthracis*, *F. tularensis*, and *Y. pestis* PPIs.[5] Although, it gives only 20% positive interactions and therefore produces a high degree of false-negative interaction,[30] the yeast two-hybrid method and related high-throughput and computational interaction data have been analyzed to identify targets in many pathogens.

Although extensive studies have been conducted for host–pathogen interactions and target identification in *M. tuberculosis*[31–34] and *Corynebacterium diphtheriae*,[35–38] another member of the *Corynebacterium, Mycobacterium, Nocardia,* and *Rhodococcus* (CMNR) group of pathogens, *C. pseudotuberculosis*, remains uninvestigated with respect to both its PPI and host–pathogen interactions. *C. pseudotuberculosis* causes Caseous Lymphadenitis (CLA) or "cheesy gland" in small ruminants worldwide, which can result in a significant economic loss.

CLA is characterized by the formation of external or internal abscesses, chronic limb infections (lymphangitis) and lymphadenitis.[39,40] It also infects visceral organs such as the liver, spleen, kidneys and lungs.[41] Although the bacterium rarely infects humans, there are reports of human lymphadenitis, and clinical strains have been isolated.[42] Other important pathogens in the CMNR group, *M. tuberculosis* and *C. diphtheriae*, cause tuberculosis and diphtheria, respectively. According to the WHO, approximately 1.7 million people died from tuberculosis in 2009 and 50,000 died from diphtheria in 2004. *Yersinia pestis* causes plague and poses a threat for use in bioterrorism.[43] Most of its isolates are derived from *Y. pseudotuberculosis*,[44] and lymphadenitis or lymphadenopathy caused by *Cp* is one of the symptoms of a *Y. pestis*[45–47] and *M. tuberculosis*[48,49] infection.

Here, for the first time, using a combination of comparative, functional, and phylogenomics approaches, supported by published, experimentally validated data we report (a) a probable conserved PPIs in the *Cp* proteome. (b) Further, we created proteome-wide common conserved PPIs for a number of pathogenic and non-pathogenic bacteria (*C. pseudotuberculosis*, *C. diphtheriae, C. ulcerans, M. tuberculosis*, *Y. pestis*, and *E. coli*). (c) Thereafter, the proteins involved in this common conserved intra-species bacterial PPIs were used to generate host–pathogen interactions considering human, goat, sheep, and horse as hosts. This host–pathogen PPI was based on experimentally validated published host–pathogen interactions data. (d) By analyzing the host–pathogen interaction networks, we identified common conserved targets in these pathogens.

(e) Finally, we use the identified targets to develop broad spectrum drugs from an existing antibiotic regime and phytochemicals derived from *Piper betel*.

Materials and methods

Selection of highly identical conserved proteins in *Cp*, other bacteria, and hosts

Selection of conserved genes for intra-species *Cp* PPI. In this work, we aimed to develop PPIs based on sequences. Therefore, highly identical common conserved proteins of Cp were selected using comparative genomics/proteomics approaches using the BLAST tool.[50] As there is no report on *Cp* PPIs so far, first we approached to develop intra-species common conserved PPIs for four *Cp* strains (strain FRC41, 316, 3/99-5, and P54B96) that were isolated from four different hosts and recently sequenced. The strain *FRC41* (biovar *ovis*) was isolated from a human; strain *316* (biovar *equi*) was isolated from a horse; strain *3/99-5* (biovar *ovis*) was isolated from a sheep; and strain *P54B96* (biovar *ovis*) was isolated from an antelope. Highly conserved and common proteins of these four strains were selected using BLASTp cut off values: $E = 0.0001$ and $\geq 80\%$ identity. Such BLAST parameters were used to select identical sequences from different strains of a species.[51]

Selection of conserved genes for inter-species bacterial PPI. Next, the highly identical common conserved genes across a wide range of pathogenic and non-pathogenic bacteria from the same and distant taxa (*E. coli*, *Y. pestis, M. tuberculosis, C. diphtheriae, C. ulcerans*, *C. glutamicum,* and all four *Cp* strains) were selected using the BLAST option available in the Prokaryotic Sequence homology Analysis (PSAT) Tool.[52] The PSAT tool was selected because it compares gene neighborhoods, gene clusters, homologs, and orthologs among multiple bacterial genomes in a single run. It also accounts information of gene context including weak alignment scores therefore provides better sensitivity compared to other available comparative analysis methods. To get the homolog list we used *Y. pestis* genome as reference and compared with *M. tuberculosis*, *C. diphtheriae, C. glutamicum, and E. coli.* The BLAST score thresholds were set to: $E = 0.01$, bit score ≥ 100, identity $\geq 35\%$ that was used in our previous report to identify homologs essential genes.[51] The common homolog genes in these bacteria were selected and further tested for their presence in *C. ulcerans* and pool of conserved common genes of four Cp strains, and other selected bacterial strains (Table S1, ESI†) using NCBI BLASTp with same parameters. Finally, the common conserved genes that are present in all these selected bacteria were collected and the common conserved *E. coli* K12 genes were used in further analysis as most of the required experimentally validated data are available for this species. The list of bacteria used in this analysis is represented in Table S1 (ESI†).

Selection of conserved genes in hosts. A range of hosts (human, goat, sheep, bovine, and horse) were selected based on the commonality of the pathogenesis from the selected pathogenic organisms. The conserved genes in these hosts were identified using the general NCBI BLASTp program (cut off values: $E = 0.01$, bit score ≥ 100, identity $\geq 35\%$). In all cases, the name of the protein or the functionality was matched during the selection.

Classification and functional annotations of common conserved bacterial proteins

The common conserved inter-species bacterial proteins were functionally classified as per the Clusters of Orthologous Groups classifications (COGs).[53] *E. coli* genes were subjected to the COGNITOR BLAST (using default parameters) to group the proteins under each COG functional classifications. Each class of COG consists of evolutionary conserved (at least 3 distant lineages) individual protein or groups of paralogs having similar cellular function under 18 classes. Therefore, the COG database and its classification are very useful in comparative, evolutionary, and phylogenetic analysis of new genome or gene to assign their biological functions.[54] Additionally, the proteins were annotated for their functionality using the NCBI and UniProt[55] databases. Pathogenicity islands (PAIs) encode various virulence factors including type III secretion system proteins of a bacterium that are required for infection. Hence, to check the virulence of the common bacterial proteins, each protein was tested with the help of the BLASTp option at the Pathogenicity Island Database (PAIDB) server.[56] The PIDB contains all reported PAIs from 497 pathogenic bacterial strains. The database also contains more than 310 predicted PAIs from 118 prokaryots. To map the pathway involvements of these conserved proteins, we used the KEGG pathway database.[57]

Generation of intra- and inter-species bacterial PPI, validation, and analysis

The bacterial PPIs were developed and analyzed using VisANT 3.0.[58,59] VisANT is an integrative platform for developing PPIs and network prediction, construction, editing, analysis, and visualization. It develops biological interactions based on data derived from 102 methods (computational and both high- and low- throughput experimental methods). The tool can integrate and mine KEGG[57] pathways in biological interactions and multi-scale analysis and visualization of multiple pathways can also be done.

Intra-species PPI of four Cp. The *Cp* genome is not available in VisANT. Therefore, a combination of genomic context-based methods including comparative and phylogenetic profiling,[14] gene or domain fusion,[15] and gene neighborhood methods[13] were used to develop the intra-species PPIs for the conserved *C. pseudotuberculosis* proteins of the selected four *Cp* strains. The resultant PPIs along with KEGG pathways were incorporated in the VisANT for network analysis and *in silico* validation of the intra-species *Cp* PPIs.

Inter-species bacterial PPI. The common conserved inter-species bacterial PPIs for *Y. pestis, E. coli, M. tuberculosis, C. glutamicum, C. diphtheriae,* and *C. ulcerans* and all four *Cp* strains were developed using VisANT. We used common conserved *E. coli K12* proteins to develop this PPI as multiple experimentally validated data for *E. coli* PPIs are available in VisANT. Additionally, the VisANT generated *E. coli* based conserved PPIs were evaluated using anti-tag co-immunoprecipitation-based binding PPIs from *E. coli*.[60] Next, the COG-based classification was applied to construct interacting protein hubs (a group of proteins under a common COG). Further, KEGG pathways were incorporated into the PPI network and analyzed in VisANT for identification of correlations among the interacting individual proteins, hubs, connecting nodes, and pathways to determine if the selected common conserved proteins and their PPIs are involved in bacterial essential metabolic process as well as in pathogenesis. This is with the agreement that as we have taken common conserved proteins of multiple pathogenic and non-pathogenic bacteria from same and different taxa; the proteins and their resultant interspecies PPIs must be involved in bacterial essential metabolic as well as pathogenic pathways.

Host–pathogen protein–protein interactions (HP-PPIs)

Cp infects a broad range of hosts, commonly goat, sheep, and horse,[51] and in rare cases, human.[42] However, the other pathogens investigated in this analysis do affect humans. With the exception of the human host, the genomes of the other hosts (goat, sheep and horse) have not been fully characterized. It is presumed that the goat, sheep and horse genomes have protein products similar to those of human, as they are higher mammals.[51] Several symptoms are shared between a *Cp, Y. pestis,*[45–47] and *Mycobacterium*[48,49] infection. *Mycobacterium* also falls under the same bacterial group of *Cp* (the CMNR group of pathogens). Therefore, we used our identified common conserved proteins (that interact with each other and forms common conserved PPIs) in our previous analysis step (inter-species PPIs) to generate a common conserved host–pathogen interaction that will be common to all the selected pathogens and hosts.

Although several computational approaches based HP-PPIs have been reported over time for a number of pathogens,[26,28,61,62] instead of using computational methods, we made our HP-PPIs based on published experimentally validated host–pathogen protein–protein binding data. To achieve the HP-PPIs; yeast two-hybrid assay based *Y. pestis-human* PPIs,[5,63] liquid chromatography-tandem mass spectrometry based surface-affinity profiling data for *S. gallolyticus-human* PPIs,[64] and protein microarray based *streptococcu*-human PPIs[29] were extracted from corresponding published literatures. Although the yeast two-hybrid screens generate significant degree of false negatives interactions,[65] we had no other option to generate the host pathogen PPIs because of unavailability of any other high throughput experimental data.

In addition to these literature based data, 7180 experimentally validated host–pathogen protein binding interactions for 21 pathogens with the human proteins from the PathoSystems Resource Integration Center (Patric) database[66] and 24 253 PPIs between 58 hosts and 416 pathogen species from HPIDB database[67] were downloaded to enrich our interaction data. While the Patric contains interactions of bacterial proteins with only human; the HPIDB provides PPIs data for multiple hosts (including human, mouse, rat, and bovine, chicken *etc.*).

Next, the identified common conserved bacterial proteins those interact with each other in intra-species bacterial PPIs were manually correlated with human interacting counterparts based on the collected experimentally validated host–pathogen interaction data. In some cases, the correlation was difficult as the interacting partner protein from the bacteria was from species that is not considered in our analysis. Therefore, we used comparative genomics BLAST to identify if the interacting bacterial partner is a homologue to any of our selected common conserved bacterial proteins and if there is a $>35\%$ identity, we considered the interaction for our purpose.

Towards validating and determining the significance of the HP-PPIs

To identify and evaluate the significance of the host–pathogen interactions involved in the host response to the pathogenesis and the key bacterial proteins involved in the pathogenesis, we performed two analyses of the HP-PPIs. First, we performed gene set enrichment and enriched functional clustering based on Gene Ontology using the well known tool: Database for Annotation Visualization and Integrated Discovery (DAVID Vs6.7)[68] for the host proteins in the HP-PPIs. Further, we used ToppGene[69] for candidate gene prioritization, identification of network key nodes, and centrality analysis of the interacting host proteins by mapping their involvement in host pathways affected due to infection. ToppGene is a platform for gene set enrichment, functional annotations, and protein interactions network based candidate gene prioritization. It also provides information about relative importance of a candidate gene in a PPI network. For ToppGene analysis, the training sets for the respective biological processes were collected from data available at the Molecular signature Database (MsigDB).[70] The key biological processes were selected that are modulated within the host such as TLR signaling and inflammatory pathways, immunity, cytoskeleton reorganization, phagocytosis, and apoptosis in response to infection of *Y. pestis, E. coli, M. tuberculosis* and several other pathogenic bacteria as described in manually curated PHIDIAS host–pathogen interactions database.[71] Finally, the interacting pathway-specific key host proteins were selected based on the ToppGene analysis.

The key bacterial proteins in the HP-PPIs that are involved in the pathogenesis were identified based on the functionality analysis. The functional annotation was done using the NCBI, UniProt,[55] and KEGG databases.[57] Additionally, the sub-cellular localization of the proteins were determined using CELLO[72] and ''Effective''[73] tools. While CELLO identifies extracellular, outer membrane, inner membrane, periplasmic, and cytoplasmic proteins; the ''Effective'' specifically predicts bacterial secreted proteins. The virulence was checked using PAIDB database.[56]

Identification of targets from the host–pathogen PPIs and virtual screening

From the host–pathogen interaction network, the interacting essential non-host homolog bacterial proteins were identified as probable targets based on the method and criteria as described by Barh *et al.*, 2011.[51] Briefly, the interacting essential bacterial proteins were selected based on Database of Essential Genes (DEG)[74] BLASTp (cut off values: $E = 0.01$, bit score ≥ 100,

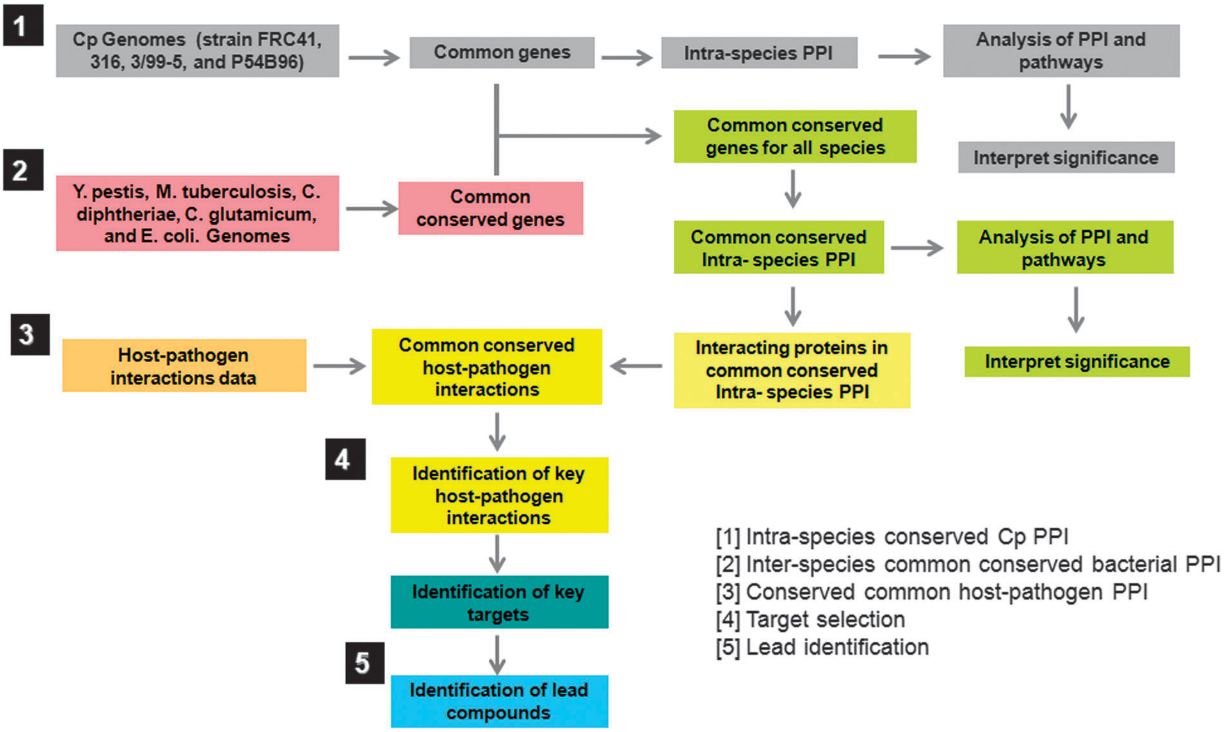

1
Cp Genomes (strain FRC41, 316, 3/99-5, and P54B96)
Common genes
Intra-species PPI
Analysis of PPI and pathways
Interpret significance
Common conserved genes for all species
2
Y. pestis, M. tuberculosis, C. diphtheriae, C. glutamicum, and E. coli. Genomes
Common conserved genes
Common conserved Intra- species PPI
Analysis of PPI and pathways
3
Host-pathogen interactions data
Common conserved host-pathogen interactions
Interacting proteins in common conserved Intra- species PPI
Interpret significance
4
Identification of key host-pathogen interactions
Identification of key targets
5
Identification of lead compounds
[1] Intra-species conserved Cp PPI
[2] Inter-species common conserved bacterial PPI
[3] Conserved common host-pathogen PPI
[4] Target selection
[5] Lead identification

Fig. 1 Simple flow diagram of the overall strategy used to develop intra-species *Cp* PPI, inter-species bacterial PPI, host–pathogen interaction PPI, and identification of targets from the host–pathogen interactions.

identity ≥ 35%). Further, the non-host essential bacterial homologs were identified by subjecting essential proteins in NCBI BLASTp program against human, mouse, sheep, horse, and bovine proteomes. Finally, the bacterial essential non-host homolog core and PAI associated proteins having ≤100 KDa molecular weight and are involved in bacteria's multiple unique essential metabolic pathways were selected as putative targets.

The bacterial targets were modeled using the Phyre 2[75] and Swiss model servers[76] and validated using the SAVS server Vs.4. (http://services.mbi.ucla.edu/SAVES/). A ligand library was developed with 30 well known antibiotics used against the selected pathogens and effective drugs for *Cp*.[77] In India, a *Cp* infection is rare in areas where the cattle feed on betel vine leaves and stalks. Therefore, 120 compounds derived from betel vine were also used to enrich the ligand library and for testing these betel compounds on the identified targets. The catalytic pockets within the target proteins were determined using Molegro Virtual Docker.[78] The docking was performed using GOLD software[79] and the five best ligands based on their GOLD score. The overall strategy is represented in Fig. 1.

Growth inhibitory effect of *Piper betel* compounds: preliminary validation

The best lead compounds from *Piper betel* were tested for their individual growth inhibition efficacy against the pathogenic *E. coli* O157:H7. The bacteria were cultured in Mueller Hinton (MH) broth (Sigma-Aldrich Co. LLC) at 37 °C for 6 hours to reach the log phase. Then, cells were harvested by centrifugation and 10^7 CFU mL^{-1} cells were resuspended in tubes containing MH broth and 10, 100 μM or 1, 10, and 100 mM concentrations of the *Piper betel* compounds. Treatment with 100 μg ml^{-1} of ampicillin was used as control. Cultures were then incubated at 37 °C for 2 hours in a shaker. The number of colony-forming units (CFUs) was counted each 30 min interval by obtaining the CFU/ml from serial 10-fold dilutions prepared in MH agar (Sigma-Aldrich Co. LLC).

Results

Bacterial protein–protein interactions

Common conserved intra-strain PPI in Cp. We identified 1783 genes common to our 4 *Cp* selected strains. Using the computational approaches, we found 4186 conserved interactions common to these *Cp* strains. We found total 874 proteins are involved in these interactions. The number of predicted PPIs based on phylogenetic profile, domain fusion, and gene neighborhood methods are 2392, 2388, and 245, respectively. To analyze the pathways falling in these conserved interactions, we fed the PPIs and *Cp FRC41* metabolic pathways (obtained from KEGG) into VisANT. Upon analysis, we found that 68 pathways can be mapped in this intra-strain PPI of the Cp. These pathways include various metabolisms, two component systems, ABC transporters, and bacterial secretion systems among others that are important for bacterial survival and pathogenesis. Therefore, our selected conserved common proteins and the developed intra-strain PPI of *Cp* will be useful to explain the biology and pathogenesis of the bacteria if further analyzed.

Although this PPI of *Cp* is very preliminary of its kind, we are reporting it because there is no report so far on *Cp* PPI. As our main aims are to develop conserved common inter-species bacterial PPIs and use the same to develop conserved common host–pathogen interactions to finally identify conserved common broad spectrum target; we did not analyze the intra-strain PPI of *Cp* in detail.

Inter-species common conserved bacterial PPIs. To generate the common conserved inter-species bacterial PPIs, first we identified common conserved proteins in *Y. pestis CO92, E. coli K-12 DH10B, E. coli O157:H7, M. tuberculosi H37Rvs, C. diphtheriae,* and *C. glutamicum R* using PAST server. Seventy eight proteins were found to be conserved in all these species. Further, we checked if all these proteins are conserved in other virulent and non-virulent strains of various strains of these bacteria and *Cp* strains *i.e.* from closed and distance taxa. To achieve this we used amino acid sequences of these 78 *Y. pestis CO92* proteins and performed comparative BLASTp in NCBI server against proteomes of *E. coli str. K-12 substr. MG1655*, *C. glutamicum ATCC 13032 Kitasato*, *C. urealyticum DSM 7109*, *M. tuberculosis CDC1551*, *M. ulcerans Agy99*, and four of our *Cp* strains (*FRC41, 316, 3/99-5, and P54B96*). We found all these 75 proteins are conserved in all these selected species and strains (Table S2, ESI†).

As various experimental PPI data are available for *E. coli str. K-12*, we selected conserved 75 proteins of this species to make the common conserved inter-species PPIs using VisANT. In VisANT, these 75 proteins form a PPI network with 1674 interactions involving 666 interacting nodes where 1210, 755, and 281 interactions are based on the tandem affinity purification, inferred by authors, and anti tag co-immunoprecipitation methods, respectively. There are interactions based on computational and other experimental methods such as cross-linking studies among others. Twenty seven total pathways were mapped in this PPI (Table S3a, ESI†). However, while we did internal interactions among these 75 proteins, we found only 142 interactions involving 23 pathways (Table S3b, ESI†). These 75 interacting proteins fall under 14 COGs (Fig. 2) and with the exception of 3 proteins, all other proteins were found to be virulent as per the PAIDB – BLASTp analysis (Table S2, ESI†).

We selected pathogenic and non-pathogenic organisms from the same and distant taxa and their conserved genes to make the inter-species PPIs. Therefore, the resultant PPIs are common and conserved in all the bacterial species considered and the PPIs should involve pathways that are essential for bacterial survival as well as for pathogenesis. To check this, KEGG pathways were incorporated in the PPIs using VisANT's ''expand pathways'' option and the interactions along with the pathways were analyzed. The analysis showed that the interacting networks were well linked and fit with various pathways that are well known for their involvements in bacterial survival and virulence such as various metabolism, two-component

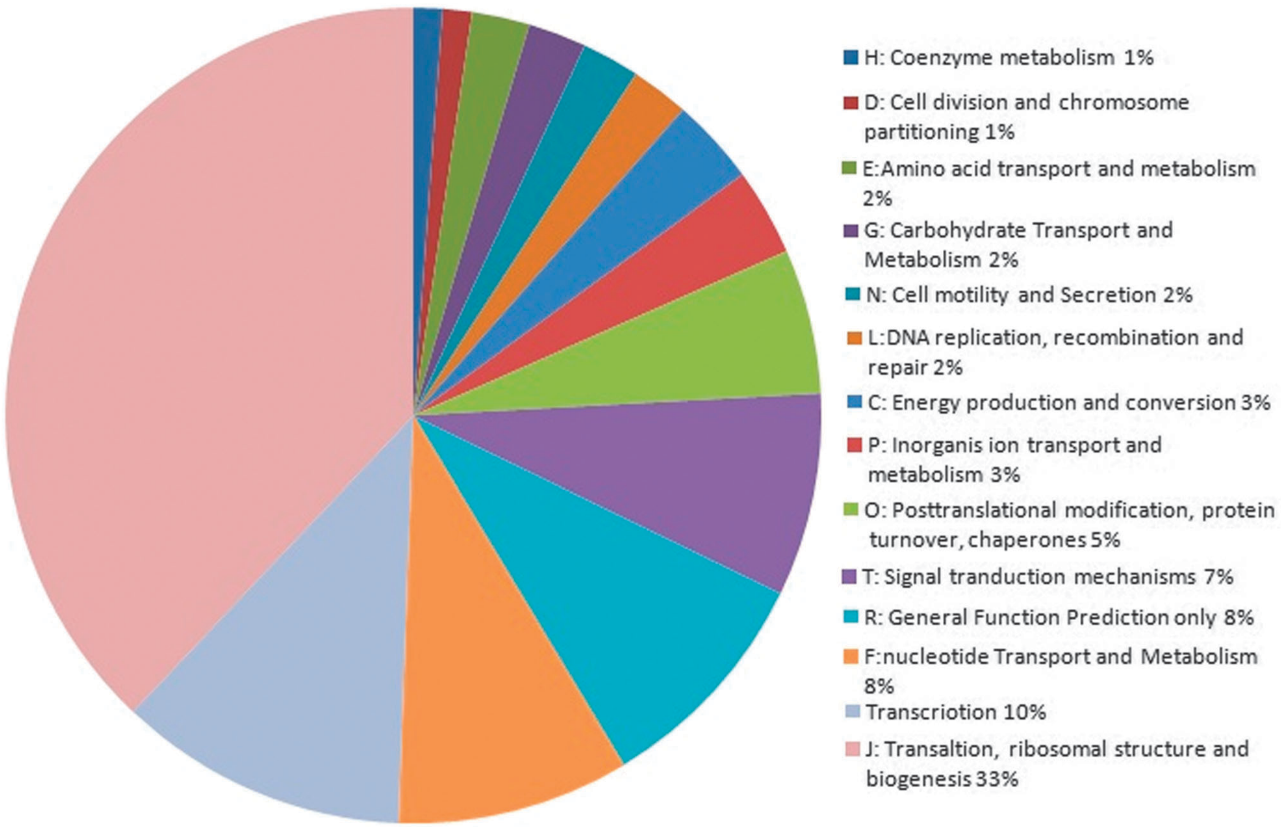


Fig. 2 Clusters of Orthologous Groups (COG) classifications of common conserved proteins of four *C. pseudotuberculosis* strains, *Y. pestis, M. tuberculosis, C. glutamicum, C. diphtheriae, C. ulcerans*, and *E. coli*.

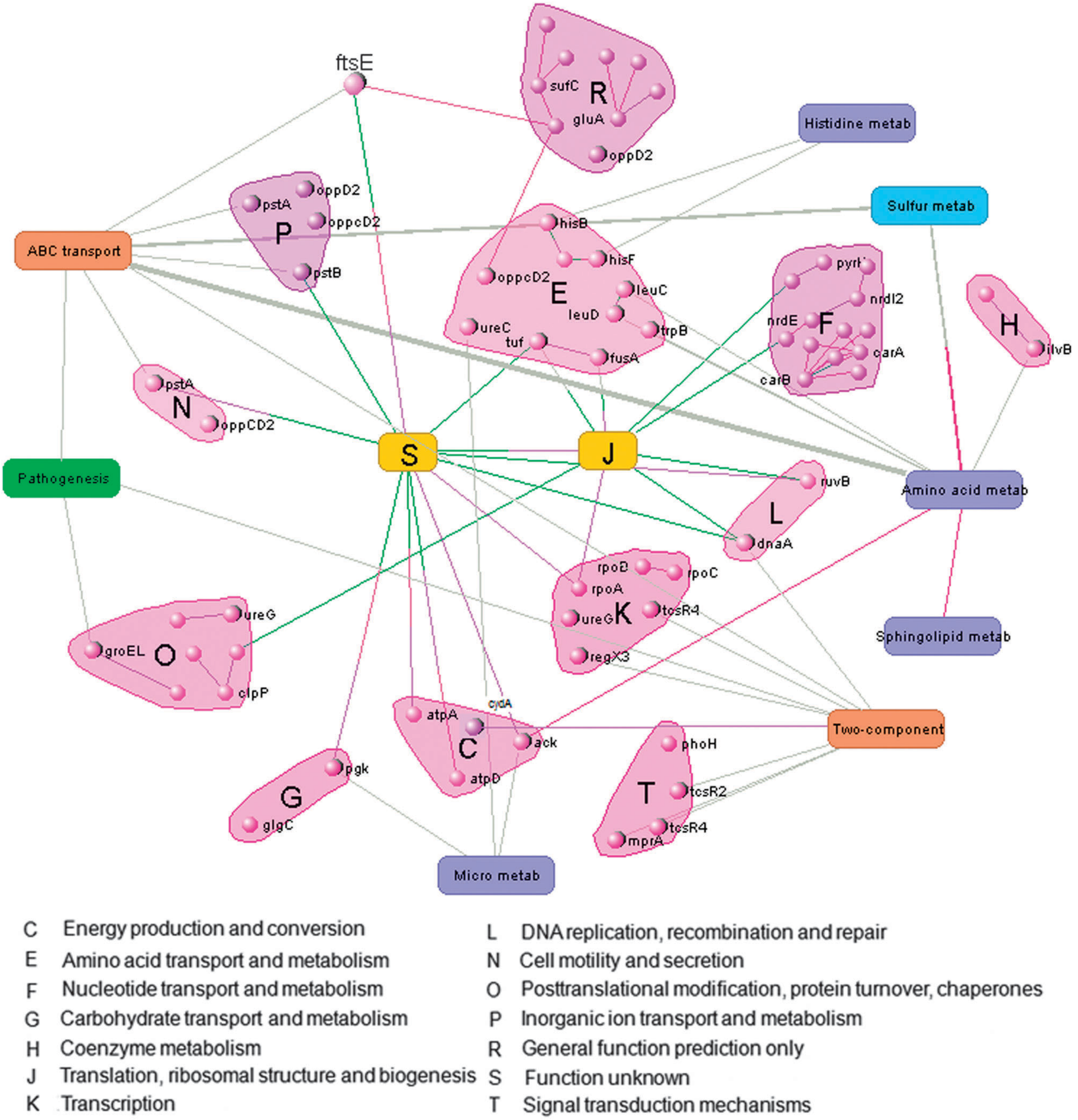

ftsE
sufC
R
gluA
oppD2
Histidine metab
oppD2
pstA
oppcD2
P
pstB
ABC transport
hisB
Sulfur metab
hisF
oppcD2
E
leuC
leuD
ureC
tuf
trpB
fusA
pyrH
nrdI2
nrdE
F
carA
carB
H
ilvB
pstA
N
oppCD2
S
J
Pathogenesis
ruvB
Amino acid metab
L
dnaA
rpoB
rpoC
rpoA
ureG
K
tcsR4
regX3
Sphingolipid metab
ureG
groEL
O
clpP
cydA
atpA
C
ack
atpD
Two-component
phoH
T
tcsR2
tcsR4
mprA
pgk
G
glgC
Micro metab
C Energy production and conversion
E Amino acid transport and metabolism
F Nucleotide transport and metabolism
G Carbohydrate transport and metabolism
H Coenzyme metabolism
J Translation, ribosomal structure and biogenesis
K Transcription
L DNA replication, recombination and repair
N Cell motility and secretion
O Posttranslational modification, protein turnover, chaperones
P Inorganic ion transport and metabolism
R General function prediction only
S Function unknown
T Signal transduction mechanisms

Fig. 3 The conserved common PPIs with COG classifications of *Cp FRC41*, *Cp 316*, *Cp 3/99-5*, *Cp P54B96, Y. pestis, M. tuberculosis, C. gluticum, C. diptherae, C. ulcerans*, and *E. coli*. Important bacterial pathways involving these proteins and the relationship of these proteins and pathways are also shown. The relationships (edgs) between hubs and individual proteins are determined using VisANT.

system,[80] ABC transporter,[81,82] redox signaling,[83] and sphingolipid metabolism[84,85]-like pathways (Fig. 3), supporting the accuracy and significance of our PPIs.

Host–pathogen protein–protein interactions (HP-PPIs)

To make the HP-PPIs, we used the conserved bacterial proteins that interact with at least another protein of the bacteria in the inter-species bacterial PPI. Using the procedure described in the methods and such conserved interacting proteins, we identified 14 bacterial proteins that interact with 122 host proteins. Functional annotations of these bacterial proteins revealed that eight are cytoplasmic enzymes and five are membrane localized. All these 14 proteins were predicted to be involved in virulence as per the PAIDB and the DEG based analysis showed; all these proteins are encoded by essential genes. Further, the functional annotation of these 14 proteins revealed that, they are involved in bacterial various essential metabolic pathways as well as pathogenicity-related pathways

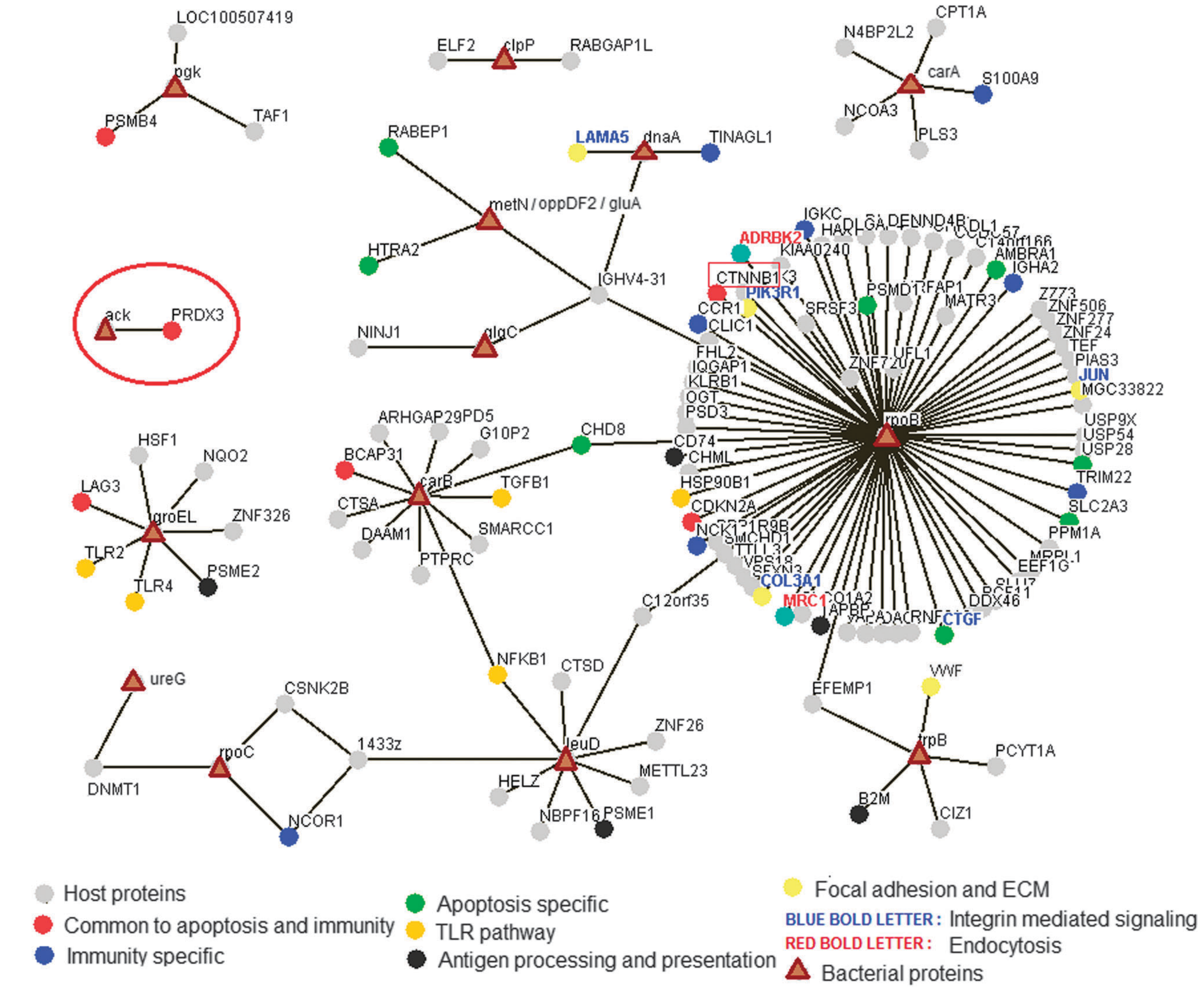


Fig. 4 Common conserved host–pathogen interaction network of multiple pathogens (four *C. pseudotuberculosis* strains, *Y. pestis, M. tuberculosis, C. glutamicum, C. diphtheriae, C. ulcerans*, and *E. coli*) and their usual hosts.

such as two-component systems (dnaA) and ABC transporters (gluA) (Table S4, ESI†). All 122 interacting host proteins were found in well-known bacterial infection associated host pathways such as integrin-mediated signaling, endocytosis, TLR signaling, immunity, apoptosis, inflammation, and redox signaling[71] (Table S5, ESI†). The ToppGene-based gene set enrichment analysis ranked CTNB1 and PIK3R1 at positions one and four, respectively. Both proteins interact with rpoB and are involved in immunity, apoptosis, and cell matrix adhesion (Table S6, ESI†). The bacterial proteins rpoB, carA, carB, leuD, groEL and their host interacting partners IGHV4-31, NFKB1, CHD8, and C12orf35, respectively were the key nodes in the host–pathogen protein–protein interaction network based on the degree of interactions and centrality analysis (Fig. 4).

Drug target and lead selection

From the host–pathogen protein–protein interaction network, we identified common conserved bacterial targets using subtractive genomics as described by Barh *et al.*, 2011.[51] The 14 identified genes were essential for the selected group of pathogens, and the cytoplasmic Acetate kinase (Ack) [EC = 2.7.2.1, Mass = 43.3 KDa] involved in the metabolism of taurine, hypotaurine, pyruvate, propanoate, and methane metabolism is the only non-host homolog satisfying most of the criteria of an ideal target for being (a) an essential non-host homolog enzyme for multiple organisms, (b) core gene of the organisms, (c) involvement in organisms' multiple unique and essential pathways, (d) PAI-related enzyme, and (e) less than 100 KDa molecular weight[51] (Table S4, ESI†). This common conserved target binds to host PRDX3 in yeast two hybrid assay (Fig. 4). PRDX3 is involved in the immune system, apoptosis, cell proliferation, and redox signaling-like pathways. Therefore, interaction of Ack-PRDX3 affects all these biological processes in the host, supporting a mechanism of bacterial infection.

Four active sites were found in the modeled Ack using the Molegro Virtual Docker (Table S7, ESI†). The GOLD fitness score and MVD analysis of the docking showed that of the group of 30 selected antibiotics, ceftiofur and penicillin, commonly used to treat *Cp*, Diphtheria, Tuberculosis, and *Y. pestis* infections, were probably effective against Ack (Fig. 5 and Table S8, ESI†). Additionally, piperdardine and dehydropipernonaline derived from *Piper betel* were also predicted to be effective and possibly had a similar or superior inhibitory activity against the target as compared to penicillin and ceftiofur (Table S8, ESI†).

Piperdardine inhibits *E. coli O157:H7* growth

Viable cells were counted during the culture in MH media containing the compounds in order to investigate their growth-inhibiting effect on *E. coli* O157:H7. We observed that addition of 100.0 μM of piperdardine or their higher concentration dramatically decrease in the CFU counts, similar to bacteria treated with ampicillin (Fig. 6).

Discussion

PPIs derived information along with a molecular basis for host–pathogen interactions are important in finding effective targets against a pathogen. Computational or high-throughput approaches based on the development of genome- or proteome-wide PPI networks have been applied to various organisms,[9,10,18,19,26–29] allowing for the extraction of important information for specific biological processes. Predicted host–pathogen PPIs have been reported for *HIV*,[86,87] *Dengue virus*,[88] *Mycobacterium*, *apicomplexa*, *kinetoplastida*,[28] and *P. falciparum*.[26,27] Experimentally validated interactions and their implementations in drug or vaccine development against the various pathogens have also been reported for group-B *streptococcus*,[29] *Corynebacterium diphtheriae*,[36,89] *M. tuberculosis*,[31–33] *Yersinia pestis*,[90] and *Yersinia pseudotuberculosis*.[91] However, these experiments were conducted for a small fraction of pathogenic proteins. Recently, yeast-two hybrid-based proteome-wide

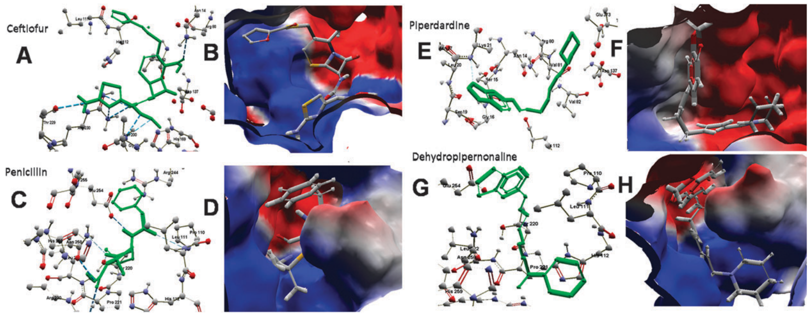


Fig. 5 Docking of Ack with ceftiofur (A–B), penicillin (C–D), piperdardine (E–F), and dehydropipernonaline (G–H).

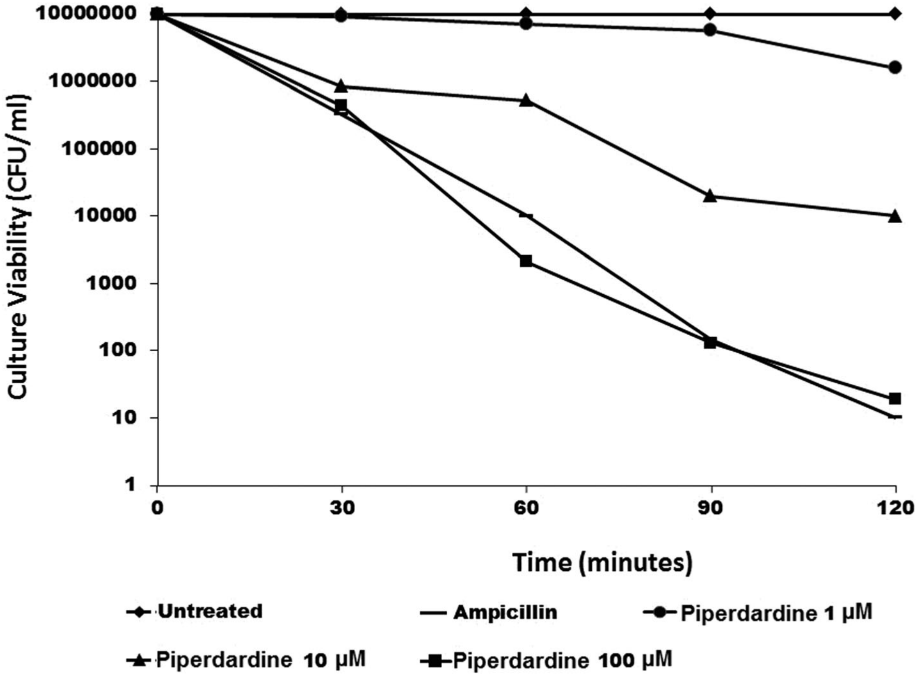


Fig. 6 Inhibitory effects of Piperdardine on the growth of *Escherichia coli* 0157:H7 as compared to ampicillin.

host–pathogen protein–protein binding interactions were reported for *B. anthracis*, *F. tularensis*, and *Y. pestis*,[5] and a number of novel interactions were documented for these pathogens.

In this report, for the first time we represent 4186 common conserved intra-species PPIs for four *Cp* strains (*Cp FRC41*, *Cp 316*, *Cp 3/99-5*, and *Cp P54B96)* using phylogenetic profile, domain fusion, and gene neighborhood methods. In *Cp*, we found 874 proteins are involved in these interactions. The recently reported experimental PPI data on *M. tuberculosis H37Rv*, another CMNR group of pathogens, revealed ~8000 novel interactions.[92] When we compared our intra-species PPIs of *Cp* with these *M. tuberculosis* data, we found half of the number of *M. tuberculosis* interactions in *Cp.* This difference may be due to the larger genome size of *M. tuberculosis* (4062 genes, almost double that of the *Cp* genome), the methods applied, and the phylogenetically conserved proteins in *Cp*. The sixty eight pathways mapped in the *Cp* PPI belong to both the bacterial essential metabolic and virulence pathways. Therefore, our developed *Cp* PPI will be significant in explaining biology and pathology of *Cp* upon further analysis.

While we developed the inter-species PPIs using common phylogenetically conserved proteins of different groups of organisms (pathogenic, non-pathogenic, and same and distance taxa), including four *Cp* strains*, Y. pestis, M. tuberculosis, C. glutamicum, C. diphtheriae, C. ulcerans*, and *E. coli*, we only observed 75 common interacting proteins that constituted a network of 142 interactions among each other and 1674 interactions involving 666 proteins to form the PPI network; however, important essential metabolic pathways and virulence related pathway can be mapped in these networks supporting the usefulness of the PPI in describing the common physiological process and virulence of these selected pathogens. It is also profound from these results that, the species-specific global PPIs exhibit a large number of interactions. However, number of interactions in conserved PPIs across a distantly related species of similar pathogenesis is reduced drastically, although essential and important pathogenesis-related proteins and pathways were found in the network.

Human-based host–pathogen interactions have been reported for a number of individual pathogens.[5,29,63,64] However, a common conserved HP-PPI for a number of pathogens and hosts have not been reported. Here, for the first time we used common conserved proteins from a broad spectrum of hosts (human, goat, sheep, and horse) to study the interactions. Additionally, for the first time, we have extended the strategy to generate conserved and common host–pathogen interactions for a group of pathogens using inter-species common conserved interacting proteins of *Cp, Y. pestis, M. tuberculosis, C. diphtheriae, C. ulcerans*, and *E. coli* with a mode of pathogenesis common to these selected hosts. This strategy helped to gain insight into common conserved host–pathogen interactions across a wide range of organisms and to identify broad spectrum targets in a single analysis. The PAI-related proteins are thought to be involved in pathogenesis.[93] Our results support this finding, and we found that the 14 identified conserved pathogen proteins involved in host–pathogen interactions were located in PAIs. These proteins are also involved in essential metabolic and virulence pathways. Similarly, GSEA, candidate gene prioritization, key nodes, and centrality analysis of the interacting host proteins revealed that they are involved in most of the infection-related signaling pathways,[71] supporting the rationality of the developed host–pathogen interaction networks.

Based on the strategy of target identification,[51] Ack was selected as a broad spectrum target from the host–pathogen interaction network. Ack is essential to *E. coli*,[94] *M. genitalium*,[95] and *M. pulmonis*[96] and is predicted to be a target in *S. aureus*.[97] The HP-PPI showed that Ack interacted with Peroxiredoxin 3 (PRDX3) from the host. PRDX3 is a peroxidase and is involved in the NF-kappaB cascade, cell proliferation, apoptosis, and redox signaling. Redox-sensitive proteins in pathogens make them resistant to oxidative stress and antibiotics,[98] and manipulation of the redox state can be an important strategy for the management of Tuberculosis.[99] Ack, our identified target, is a kinase that interacts with the redox protein PRDX3 of the host. We hypothesized that the binding of Ack to PRDX3 modulates PRDX3 activity, thereby disrupting the redox signaling and immune system of the host. This interaction may help in SOD-mediated fibrocyte activation and scar or abscess formation[100] in lymphadenitis, the common symptom of *Cp* and *Y. Pestis* infections. It may also be a vital mechanism for drug resistance in these pathogens, disrupting the host redox system.

However, to interfere mitochondrial functions during pathogenesis, a bacterial protein needs to reach and bind to mitochondrial protein of the host.[101] Bacteria that possess type III and type IV secretion system like injection machinery can directly inject bacterial proteins into the host cell cytoplasm during infection process.[102,103] As per the "Effective"[73] prediction, Ack of *M. tuberculosis H37Rv* is a type III secreted protein and according to *Couto et al.* (2012), Ack is probably secreted or localizes to bacterial surface during *M. mycoides* infection in cattle and plays a role in immunogenic responses in the host.[104] Therefore, it might be possible that Ack is injected into host cell through bacterial secretion system during infection and upon resealed into the host cytoplasm it interacts with mitochomdrial PRDX3. However, it should be proved experimentally and this is one of the future scopes of this research.

Virtual screening showed that ceftiofur and penicillin could be effective antibiotics against the selected pathogens considering the target Ack. The natural products piperdardine and dehydropipernonaline from *Piper betel* had shown a similar or superior effect on Ack as per our *in silico* analysis. Until now, no experimental data were available that tested the efficacy of compounds targeted to Ack, and validation is thereby necessary using conventional antibiotics and our identified *Piper betel* compounds. The leaf extract of *Piper betel* has proven to be useful as an antimicrobial,[105,106] antioxidant,[107] anti-inflammatory,[108] and immunomodulator.[109] However, the specific compounds in the plant that produce these properties are yet to be determined. In our preliminary validation, we observed that, 100.0 μM of piperdardine inhibits *E. coli* O157:H7 growth similar to penicillin. Therefore, it is presumed that these compounds may also be effective against other pathogens considered in this work. We are currently testing the bactericidal effects of these betel compounds against *C. pseudotuberculosis*, *C. diphtheriae*, *M. tuberculosis*, *C. ulcerans*, and *Y. pestis* and their target specificity to Ack. The results are highly promising.

Conclusion

This study demonstrates intra-species PPI for *Cp* and illustrates the potential and importance of inter-species bacterial protein–protein and host–pathogen interactions in broad spectrum target identification. We report the conserved intra-species PPIs of *Cp* and a common conserved host pathogen-interaction network for *Y. pestis, M. tuberculosis, C. diphtheriae, C. ulcerans*, *E. coli*, and four *Cp* strains. Ack was identified as a broad spectrum target for all these pathogens considering human, goat, sheep, and horse as hosts. Ceftiofur, penicillin and two natural compounds derived from *Piper betel*, piperdardine and dehydropipernonaline, were predicted to be effective against Ack activity. Validation shows piperdardine is a highly effective antibacterial agent. The *in silico* approaches used in this work were supposed to be effective in developing and analyzing inter-species global bacterial PPIs as well as host–pathogen interactions to identify drug targets.

Competing interests

The authors declare that they have no competing interests.

Financial disclosure

This work was carried out without any financial support or grant.

References

1 R. Sharan, I. Ulitsky and R. Shamir, Network-based prediction of protein function, *Mol. Syst. Biol.*, 2007, **3**, 88.

2 E. D. Levy and J. B. Pereira-Leal, Evolution and dynamics of protein interactions and networks, *Curr. Opin. Struct. Biol.*, 2008, **18**, 349–357.

3 F. Hormozdiari, R. Salari, V. Bafna and S. C. Sahinalp, Protein–protein interaction network evaluation for identifying potential drug targets, *J. Comput. Biol.*, 2010, **17**, 669–684.

4 Y. Y. Wang, J. C. Nacher and X. M. Zhao, Predicting drug targets based on protein domains, *Mol. BioSyst.*, 2012, **8**, 1528–1534.

5 M. D. Dyer, C. Neff, M. Dufford, C. G. Rivera, D. Shattuck, J. Bassaganya-Riera, T. M. Murali and B. W. Sobral, The human-bacterial pathogen protein interaction networks of Bacillus anthracis, Francisella tularensis, and Yersinia pestis, *PLoS One*, 2010, **5**, e12089.

6 J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li and H. Jiang, Predicting protein–protein interactions based only on sequences information, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**(11), 4337–41.

7 X. W. Zhao, Z. Q. Ma and M. H. Yin, Predicting protein–protein interactions by combing various sequence-derived features into the general form of Chou's Pseudo amino acid composition, *Protein Pept. Lett.*, 2011, **19**(5), 492–500.

8 L. Skrabanek, H. K. Saini, G. D. Bader and A. J. Enright, Computational prediction of protein–protein interactions, *Mol. Biotechnol.*, 2008, **38**, 1–17.

9 H. Yu, P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J. F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A. S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A. L. Barabasi, J. Tavernier, D. E. Hill and M. Vidal, High-quality binary protein interaction map of the yeast interactome network, *Science*, 2008, **322**, 104–110.

10 U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksoz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach and E. E. Wanker, A human protein–protein interaction network: a resource for annotating the proteome, *Cell*, 2005, **122**, 957–968.

11 I. Ispolatov, A. Yuryev, I. Mazo and S. Maslov, Binding properties and evolution of homodimers in protein–protein interaction networks, *Nucleic Acids Res.*, 2005, **33**, 3629–3635.

12 M. P. Stumpf, T. Thorne, S. E. de, R. Stewart, H. J. An, M. Lappe and C. Wiuf, Estimating the size of the human interactome, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 6959–6964.

13 T. Dandekar, B. Snel, M. Huynen and P. Bork, Conservation of gene order: a fingerprint of proteins thatphysically interact, *Trends Biochem. Sci.*, 1998, **23**, 324–328.

14 M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg and T. O. Yeates, Assigning protein functions by comparative genome analysis: protein phylogenetic profiles, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**, 4285–4288.

15 A. J. Enright, I. Iliopoulos, N. C. Kyrpides and C. A. Ouzounis, Protein interaction maps for complete genomes based on gene fusion events, *Nature*, 1999, **402**, 86–90.

16 C. S. Goh, A. A. Bogan, M. Joachimiak, D. Walther and F. E. Cohen, Co-evolution of proteins with their interaction partners, *J. Mol. Biol.*, 2000, **299**, 283–293.

17 M. Singhal and H. Resat, A domain-based approach to predict protein–protein interactions, *BMC Bioinf.*, 2007, **8**, 199.

18 K. Raman and N. Chandra, Mycobacterium tuberculosis interactome analysis unravels potential pathways to drug resistance, *BMC Microbiol.*, 2008, **8**, 234.

19 M. Rashid, S. Ramasamy and G. P. Raghava, A simple approach for predicting protein–protein interactions, *Curr. Protein Pept. Sci.*, 2010, **11**(7), 589–600.

20 P. M. Bowers, S. J. Cokus, D. Eisenberg and T. O. Yeates, Use of logic relationships to decipher protein network organization, *Science*, 2004, **306**, 2246–2249.

21 D. Barker and M. Pagel, Predicting functional gene links from phylogenetic-statistical analyses of whole genomes, *PLoS Comput. Biol.*, 2005, **1**, e3.

22 N. Tuncbag, G. Kar, O. Keskin, A. Gursoy and R. Nussinov, A survey of available tools and web servers for analysis of protein–protein interactions and interfaces, *Briefings Bioinf.*, 2009, **10**, 217–232.

23 C. Yeats, J. Lees, P. Carter, I. Sillitoe and C. Orengo, The Gene3D Web Services: a platform for identifying, annotating and comparing structural domains in protein sequences, *Nucleic Acids Res.*, 2011, **39**, W546–W550.

24 S. Hunter, P. Jones, A. Mitchell, R. Apweiler, T. K. Attwood, A. Bateman, T. Bernard, D. Binns, P. Bork, S. Burge, C. E. de, P. Coggill, M. Corbett, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, M. Fraser, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, C. McMenamin, H. Mi, P. Mutowo-Muellenet, N. Mulder, D. Natale, C. Orengo, S. Pesseat, M. Punta, A. F. Quinn, C. Rivoire, A. Sangrador-Vegas, J. D. Selengut, C. J. Sigrist, M. Scheremetjew, J. Tate, M. Thimmajanarthanan, P. D. Thomas, C. H. Wu, C. Yeats and S. Y. Yong, InterPro in 2011: new developments in the family and domain prediction database, *Nucleic Acids Res.*, 2012, **40**, D306–D312.

25 J. Lees, C. Yeats, J. Perkins, I. Sillitoe, R. Rentzsch, B. H. Dessailly and C. Orengo, Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis, *Nucleic Acids Res.*, 2012, **40**, D465–D471.

26 M. D. Dyer, T. M. Murali and B. W. Sobral, Computational prediction of host–pathogen protein–protein interactions, *Bioinformatics*, 2007, **23**, i159–i166.

27 S. Wuchty, Computational prediction of host-parasite protein interactions between P. falciparum and H. sapiens, *PLoS One*, 2011, **6**, e26960.

28 F. P. Davis, D. T. Barkan, N. Eswar, J. H. McKerrow and A. Sali, Host pathogen protein interactions predicted by comparative modeling, *Protein Sci.*, 2007, **16**, 2585–2596.

29 I. Margarit, S. Bonacci, G. Pietrocola, S. Rindi, C. Ghezzo, M. Bombaci, V. Nardi-Dei, R. Grifantini, P. Speziale and G. Grandi, Capturing host–pathogen interactions by protein microarrays: identification of novel streptococcal proteins binding to human fibronectin, fibrinogen, and C4BP, *FASEB J.*, 2009, **23**, 3100–3112.

30 H. Huang, B. M. Jedynak and J. S. Bader, Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps, *PLoS Comput. Biol.*, 2007, **3**(11), e214.

31 S. Gagneux, K. DeRiemer, T. Van, M. Kato-Maeda, B. C. de Jong, S. Narayanan, M. Nicol, S. Niemann, K. Kremer, M. C. Gutierrez, M. Hilty, P. C. Hopewell and P. M. Small, Variable host–pathogen compatibility in Mycobacterium tuberculosis, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 2869–2873.

32 D. P. Cifuentes, M. Ocampo, H. Curtidor, M. Vanegas, M. Forero, M. E. Patarroyo and M. A. Patarroyo, Mycobacterium tuberculosis Rv0679c protein sequences involved in host-cell infection: potential TB vaccine candidate antigen, *BMC Microbiol.*, 2010, **10**, 109.

33 K. Raman, A. G. Bhat and N. Chandra, A systems perspective of host–pathogen interactions: predicting disease outcome in tuberculosis, *Mol. BioSyst.*, 2010, **6**, 516–530.

34 Y. Wang, T. Cui, C. Zhang, M. Yang, Y. Huang, W. Li, L. Zhang, C. Gao, Y. He, Y. Li, F. Huang, J. Zeng, C. Huang, Q. Yang, Y. Tian, C. Zhao, H. Chen, H. Zhang and Z. G. He, Global protein–protein interaction network in the human pathogen Mycobacterium tuberculosis H37Rv, *J. Proteome Res.*, 2010, **9**, 6665–6677.

35 V. Kolodkina, T. Denisevich and L. Titov, Identification of Corynebacterium diphtheriae gene involved in adherence to epithelial cells, *Infect., Genet. Evol.*, 2011, **11**, 518–521.

36 L. Ott, M. Holler, R. G. Gerlach, M. Hensel, J. Rheinlaender, T. E. Schaffer and A. Burkovski, Corynebacterium diphtheriae invasion-associated protein (DIP1281) is involved in cell surface organization, adhesion and internalization in epithelial cells, *BMC Microbiol.*, 2010, **10**, 2.

37 L. Ott, M. Holler, J. Rheinlaender, T. E. Schaffer, M. Hensel and A. Burkovski, Strain-specific differences in pili formation and the interaction of Corynebacterium diphtheriae with host cells, *BMC Microbiol.*, 2010, **10**, 257.

38 E. Trost, J. Blom, S. S. de Castro, I. H. Huang, A. Al-Dilaimi, J. Schroder, S. Jaenicke, F. A. Dorella, F. S. Rocha, A. Miyoshi, V. Azevedo, M. P. Schneider, A. Silva, T. C. Camello, P. S. Sabbadini, C. S. Santos, L. S. Santos, R. Hirata, Jr., A. L. Mattos-Guaraldi, A. Efstratiou, M. P. Schmitt, H. Ton-That and A. Tauch, Pan-genomics of Corynebacterium diphtheriae: Insights into the genomic diversity of pathogenic isolates from cases of classical diphtheria, endocarditis and pneumonia, *J. Bacteriol.*, 2012, **194**(12), 3199–3215.

39 L. H. Williamson, Caseous lymphadenitis in small ruminants, *Vet. Clin. North Am.: Food Anim Pract.*, 2001, **17**, 359–371, vii.

40 M. Aleman, S. J. Spier, W. D. Wilson and M. Doherr, Corynebacterium pseudotuberculosis infection in horses: 538 cases (1982–1993), *J. Am. Vet. Med. Assoc.*, 1996, **209**, 804–809.

41 R. G. Batey, Pathogenesis of caseous lymphadenitis in sheep and goats, *Aust. Vet. J.*, 1986, **63**, 269–272.

42 E. Trost, L. Ott, J. Schneider, J. Schroder, S. Jaenicke, A. Goesmann, P. Husemann, J. Stoye, F. A. Dorella, F. S. Rocha, S. C. Soares, V. D'Afonseca, A. Miyoshi, J. Ruiz, A. Silva, V. Azevedo, A. Burkovski, N. Guiso, O. F. Join-Lambert, S. Kayal and A. Tauch, The complete genome sequence of Corynebacterium pseudotuberculosis FRC41 isolated from a 12-year-old girl with necrotizing lymphadenitis reveals insights into gene-regulatory

networks contributing to virulence, *BMC Genomics*, 2010, **11**, 728.

43 G. J. Annas, Bioterror and ''bioart'' – a plague o' both your houses, *N. Engl. J. Med.*, 2006, **354**, 2715–2720.

44 M. Drancourt, Plague in the genomic area, *Clin. Microbiol. Infect.*, 2012, **18**, 224–230.

45 T. Karttunen, K. Nevasaari, O. Rasanen, P. J. Taskinen and M. Alavaikko, Immunoblastic lymphadenopathy with a high serum Yersinia enterocolitica titer. A case report, *Cancer*, 1983, **52**, 2281–2284.

46 S. J. Nesbitt, L. O. Neville, F. R. Scott and D. M. Flynn, Yersinia pseudotuberculosis in a 3 year old and rapid response to cefotaxime, *J. R. Soc. Med.*, 1994, **87**, 418–419.

47 J. E. Comer, D. E. Sturdevant, A. B. Carmody, K. Virtaneva, D. Gardner, D. Long, R. Rosenke, S. F. Porcella and B. J. Hinnebusch, Transcriptomic and innate immune responses to Yersinia pestis in the lymph node during bubonic plague, *Infect. Immun.*, 2010, **78**, 5086–5098.

48 P. R. Mohapatra and A. K. Janmeja, Tuberculous lymphadenitis, *J. Assoc. Physicians India*, 2009, **57**, 585–590.

49 J. Knox, G. Lane, J. S. Wong, P. G. Trevan and H. Karunajeewa, Diagnosis of Tuberculous Lymphadenitis Using Fine Needle Aspiration Biopsy, *Int. Med. J.*, 2012, DOI: 10.1111/j.1445-5994.2012.02748.x.

50 S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, 1997, **25**, 3389–3402.

51 D. Barh, N. Jain, S. Tiwari, B. P. Parida, V. D'Afonseca, L. Li, A. Ali, A. R. Santos, L. C. Guimaraes, S. S. de Castro, A. Miyoshi, A. Bhattacharjee, A. N. Misra, A. Silva, A. Kumar and V. Azevedo, A novel comparative genomics analysis for common drug and vaccine targets in Corynebacterium pseudotuberculosis and other CMN group of human pathogens, *Chem. Biol. Drug Des.*, 2011, **78**, 73–84.

52 C. Fong, L. Rohmer, M. Radey, M. Wasnick and M. J. Brittnacher, PSAT: a web tool to compare genomic neighborhoods of multiple prokaryotic genomes, *BMC Bioinf.*, 2008, **9**, 170.

53 R. L. Tatusov, D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova and E. V. Koonin, The COG database: new developments in phylogenetic classification of proteins from complete genomes, *Nucleic Acids Res.*, 2001, **29**, 22–28.

54 M. Kaufmann, The Role of the COG Database in Comparative and Functional Genomics, *Curr. Bioinf.*, 2006, **1**, 291–300.

55 M. Magrane and U. Consortium, *UniProt Knowledgebase: a hub of integrated protein data. Database*, Oxford, 2011, bar009.

56 S. H. Yoon, Y. K. Park, S. Lee, D. Choi, T. K. Oh, C. G. Hur and J. F. Kim, Towards pathogenomics: a web-based resource for pathogenicity islands, *Nucleic Acids Res.*, 2007, **35**, D395–D400.

57 M. Kanehisa and S. Goto, KEGG: kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.*, 2000, **28**, 27–30.

58 Z. Hu, J. Mellor, J. Wu and C. DeLisi, VisANT: an online visualization and analysis tool for biological interaction data, *BMC Bioinf.*, 2004, **5**, 17.

59 Z. Hu, D. M. Ng, T. Yamada, C. Chen, S. Kawashima, J. Mellor, B. Linghu, M. Kanehisa, J. M. Stuart and C. DeLisi, VisANT 3.0: new modules for pathway visualization, editing, prediction and construction, *Nucleic Acids Res.*, 2007, W625–W632.

60 G. Butland, J. M. Peregrin-Alvarez, J. Li, W. Yang, X. Yang, V. Canadien, A. Starostine, D. Richards, B. Beattie, N. Krogan, M. Davey, J. Parkinson, J. Greenblatt and A. Emili, Interaction network containing conserved and essential protein complexes in Escherichia coli, *Nature*, 2005, **433**, 531–537.

61 N. Tyagi, O. Krishnadev and N. Srinivasan, Prediction of protein–protein interactions between Helicobacter pylori and a human host, *Mol BioSyst.*, 2009, **5**(12), 1630–1635.

62 O. Krishnadev and N. Srinivasan, Prediction of protein–protein interactions between human host and a pathogen and its application to three pathogenic bacteria, *Int. J. Biol. Macromol.*, 2011, **48**(4), 613–619.

63 H. Yang, Y. Ke, J. Wang, Y. Tan, S. K. Myeni, D. Li, Q. Shi, Y. Yan, H. Chen, Z. Guo, Y. Yuan, X. Yang, R. Yang and Z. Du, Insight into bacterial virulence mechanisms against host immune response *via* the Yersinia pestis-human protein–protein interaction network, *Infect. Immun.*, 2011, **79**(11), 4413–4424.

64 A. Boleij, C. M. Laarakkers, J. Gloerich, D. W. Swinkels and H. Tjalsma, Surface-affinity profiling to identify host–pathogen interactions, *Infect. Immun.*, 2011, **79**(12), 4777–4783.

65 T. Stellberger, R. Häuser, A. Baiker, V. R. Pothineni, J. Haas and P. Uetz, Improving the yeast two-hybrid system with permutated fusions proteins: the Varicella Zoster Virus interactome, *Proteome Sci.*, 2010, **8**, 8.

66 J. J. Gillespie, A. R. Wattam, S. A. Cammer, J. L. Gabbard, M. P. Shukla, O. Dalay, T. Driscoll, D. Hix, S. P. Mane, C. Mao, E. K. Nordberg, M. Scott, J. R. Schulman, E. E. Snyder, D. E. Sullivan, C. Wang, A. Warren, K. P. Williams, T. Xue, H. S. Yoo, C. Zhang, Y. Zhang, R. Will, R. W. Kenyon and B. W. Sobral, PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species, *Infect. Immun.*, 2011, **79**, 4286–4298.

67 R. Kumar and B. Nanduri, HPIDB – a unified resource for host–pathogen interactions, *BMC Bioinf.*, 2010, **11**(Suppl 6), S16.

68 D. W. Huang, B. T. Sherman and R. A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nat. Protoc.*, 2009, **4**(1), 44–57.

69 B. E. A. B. J. A. Chen, ToppGene Suite for gene list enrichment analysis and candidate gene prioritization, *Nucleic Acids Res.*, 2009, 37.

70 A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander and J. P. Mesirov, Gene set

enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 15545–15550.

71 Z. Xiang, Y. Tian and Y. He, PHIDIAS: a pathogen-host interaction data integration and analysis system, *Genome Biol.*, 2007, **8**(7), R150.

72 C. S. Yu, C. J. Lin and J. K. Hwang, Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on *n*-peptide compositions, *Protein Sci.*, 2004, **13**, 1402–1406.

73 M. A. Jehl, R. Arnold and T. Rattei, Effective–a database of predicted secreted bacterial proteins, *Nucleic Acids Res.*, 2011, D591–D595.

74 R. Zhang, H. Y. Ou and C. T. Zhang, DEG: a database of essential genes, *Nucleic Acids Res.*, 2004, **1**, D271–D272.

75 L. A. Kelley and M. J. Sternberg, Protein structure prediction on the Web: a case study using the Phyre server, *Nat. Protoc.*, 2009, **4**, 363–371.

76 K. Arnold, L. Bordoli, J. Kopp and T. Schwede, The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling, *Bioinformatics*, 2006, **22**, 195–201.

77 S. J. Spier, Corynebacterium pseudotuberculosis infection in horses: An emerging disease associated with climate change?, *Equine vet. Educ*, 2008, **20**, 37–39.

78 R. Thomsen and M. H. Christensen, MolDock: a new technique for high-accuracy molecular docking, *J. Med. Chem.*, 2006, **49**, 3315–3321.

79 M. L. Verdonk, J. C. Cole, M. J. Hartshorn, C. W. Murray and R. D. Taylor, Improved protein-ligand docking using GOLD, *Proteins*, 2003, **52**, 609–623.

80 T. Tobe, The roles of two-component systems in virulence of pathogenic Escherichia coli and Shigella spp, *Adv. Exp. Med. Biol.*, 2008, **631**, 189–99.

81 J. S. Klein and O. Lewinson, Bacterial ATP-driven transporters of transition metals: physiological roles, mechanisms of action and roles in bacterial virulence, *Metallomics*, 2011, **3**(11), 1098–1108.

82 V. G. Lewis, M. P. Ween and C. A. McDevitt, The role of ATP-binding cassette transporters in bacterial pathogenicity, *Protoplasma*, 2012, **249**(4), 919–942.

83 A. Trivedi, N. Singh, S. A. Bhat, P. Gupta and A. Kumar, Redox biology of tuberculosis pathogenesis, *Adv. Microbiol. Physiol.*, 2012, **60**, 263–324.

84 L. J. Heung, C. Luberto and M. Del Poeta, Role of sphingolipids in microbial pathogenesis, *Infect. Immun.*, 2006, **74**(1), 28–39.

85 D. An, C. Na, J. Bielawski, Y. A. Hannun and D. L. Kasper, Membrane sphingolipids as essential molecular signals for Bacteroides survival in the intestine, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**(Suppl 1), 4666–4671.

86 P. Evans, W. Dampier, L. Ungar and A. Tozeren, Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs, *BMC Med. Genomics*, 2009, **2**, 27.

87 O. Tastan, Y. Qi, J. G. Carbonell and J. Klein-Seetharaman, Prediction of interactions between HIV-1 and human proteins by information integration, *Pac. Symp. Biocomput.*, 2009, 516–527.

88 J. M. Doolittle and S. M. Gomez, Mapping protein interactions between Dengue virus and its human and insect hosts, *PLoS Neglected Trop. Dis.*, 2011, **5**, e954.

89 V. Kolodkina, T. Denisevich and L. Titov, Identification of Corynebacterium diphtheriae gene involved in adherence to epithelial cells, *Infect. Genet. Evol.*, 2011, **11**, 518–521.

90 B. Li and R. Yang, Interaction between Yersinia pestis and the host immune system, *Infect. Immun.*, 2008, **76**, 1804–1811.

91 C. G. Zhang, A. D. Gonzales, M. W. Choi, B. A. Chromy, J. P. Fitch and S. L. McCutchen-Maloney, Subcellular proteomic analysis of host–pathogen interactions using human monocytes exposed to Yersinia pestis and Yersinia pseudotuberculosis, *Proteomics*, 2005, **5**, 1877–1888.

92 Y. Wang, T. Cui, C. Zhang, M. Yang, Y. Huang, W. Li, L. Zhang, C. Gao, Y. He, Y. Li, F. Huang, J. Zeng, C. Huang, Q. Yang, Y. Tian, C. Zhao, H. Chen, H. Zhang and Z. G. He, Global protein–protein interaction network in the human pathogen Mycobacterium tuberculosis H37Rv, *J. Proteome Res.*, 2010, **9**, 6665–6677.

93 H. Schmidt and M. Hensel, Pathogenicity islands in bacterial pathogenesis, *Clin. Microbiol. Rev.*, 2004, **17**, 14–56.

94 S. Y. Gerdes, M. D. Scholle, J. W. Campbell, G. Balazsi, E. Ravasz, M. D. Daugherty, A. L. Somera, N. C. Kyrpides, I. Anderson, M. S. Gelfand, A. Bhattacharya, V. Kapatral, M. D'Souza, M. V. Baev, Y. Grechkin, F. Mseeh, M. Y. Fonstein, R. Overbeek, A. L. Barabasi, Z. N. Oltvai and A. L. Osterman, Experimental determination and system level analysis of essential genes in Escherichia coli MG1655, *J. Bacteriol.*, 2003, **185**, 5673–5684.

95 J. I. Glass, N. Assad-Garcia, N. Alperovich, S. Yooseph, M. R. Lewis, M. Maruf, C. A. Hutchison, III, H. O. Smith and J. C. Venter, Essential genes of a minimal bacterium, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 425–430.

96 C. T. French, P. Lao, A. E. Loraine, B. T. Matthews, H. Yu and K. Dybvig, Large-scale transposon mutagenesis of Mycoplasma pulmonis, *Mol. Microbiol.*, 2008, **69**, 67–76.

97 V. K. Morya, V. Dewaker, S. D. Mecarty and R. Singh, In silico Analysis Metabolic Pathways for Identification of Putative Drug Targets for Staphylococcus aureus, *J. Comput. Sci. Syst. Biol.*, 2010, **3**, 062–069.

98 P. R. Chen, P. Brugarolas and C. He, Redox signaling in human pathogens, *Antioxid. Redox Signaling*, 2011, **14**, 1107–1118.

99 A. Kumar, A. Farhana, L. Guidry, V. Saini, M. Hondalus and A. J. Steyn, Redox homeostasis in mycobacteria: the key to tuberculosis control?, *Expert Rev. Mol. Med.*, 2011, **13**, e39.

100 M. C. Vozenin-Brotons, V. Sivan, N. Gault, C. Renard, C. Geffrotin, S. Delanian, J. L. Lefaix and M. Martin, Antifibrotic action of Cu/Zn SOD is mediated by TGF-beta1 repression and phenotypic reversion of myofibroblasts, *Free Radical Biol. Med.*, 2001, **30**, 30–42.

101 V. Kozjak-Pavlovic, K. Ross and T. Rudel, Import of bacterial pathogenicity factors into mitochondria, *Curr. Opin. Microbiol.*, 2008, **11**(1), 9–14.

102 G. R. Cornelis, The type III secretion injectisome, *Nat. Rev. Microbiol.*, 2006, **4**(11), 811–25.

103 S. Backert and T. F. Meyer, Type IV secretion systems and their effectors in bacterial pathogenesis, *Curr. Opin. Microbiol.*, 2006, **9**(2), 207–217.

104 M. S. R. Couto, C. S. Klein, D. Voss-Rech and H. Terenzi, Extracellular Proteins of Mycoplasma synoviae, *ISRN Vet. Sci.*, 2012, **2012**, 6.

105 R. Nair and S. Chanda, Antimicrobial Activity of Terminalia catappa, Manilkara zapota and Piper betel Leaf Extract, *Indian J. Pharm. Sci.*, 2008, **70**, 390–393.

106 I. Ali, F. G. Khan, K. A. Suri, B. D. Gupta, N. K. Satti, P. Dutt, F. Afrin, G. N. Qazi and I. A. Khan, *In vitro* antifungal activity of hydroxychavicol isolated from Piper betle L, *Ann. Clin. Microbiol. Antimicrob.*, 2010, **9**, 7.

107 N. Dasgupta and B. De, Antioxidantactivity of PiperbetleL. leafextract *in vitro*, *Food Chem.*, 2004, **88**, 219–224.

108 S. Ganguly, S. Mula, S. Chattopadhyay and M. Chatterjee, An ethanol extract of Piper betle Linn. mediates its anti-inflammatory activity *via* down-regulation of nitric oxide, *J. Pharm. Pharmacol.*, 2007, **59**, 711–718.

109 D. G. Kanjwani, T. P. Marathe, S. V. Chiplunkar and S. S. Sathaye, Evaluation of immunomodulatory activity of methanolic extract of Piper betel, *Scand. J. Immunol.*, 2008, **67**, 589–593.

II.III.2 The *Corynebacterium pseudotuberculosis* in silico predicted pan-exoproteome.
Santos AR, Carneiro A, Gala-García A, Pinto A, Barh D, Barbosa E, Aburjaile F, Dorella F, Rocha F, Guimarães L, Zurita-Turk M, Ramos R, Almeida S, Soares S, Pereira U, Abreu VC, Silva A, Miyoshi A, **Azevedo V**.
BMC Genomics. 2012;13 Suppl 5:S6. doi: 10.1186/1471-2164-13-S5-S6. Epub 2012 Oct 19.

Utilizando as cinco primeiras linhagens de *C. pseudotuberculosis* sequenciadas e inteiramente montadas, sendo quatro feitas pelo meu grupo de pesquisa, foram feitas predições por software a respeito da localização subcelular dos proteomas. Essas predições resultaram em um conjunto de proteínas preditas como exportadas presentes nas cinco linhagens, além de outras presentes em subconjuntos menores que cinco. As proteínas preditas como exportadas nas cinco linhagens foram catalogadas e referenciadas por um identificador único denominado panlocus. Atualmente esse é o padrão de anotação de genomas de banco de dados de genomas do NCBI

BMC Genomics

RESEARCH Open Access

The *Corynebacterium pseudotuberculosis in silico* predicted pan-exoproteome

Anderson R Santos[1], Adriana Carneiro[2], Alfonso Gala-García[1], Anne Pinto[1], Debmalya Barh[3], Eudes Barbosa[1], Flávia Aburjaile[1], Fernanda Dorella[1], Flávia Rocha[1], Luis Guimarães[1], Meritxell Zurita-Turk[1], Rommel Ramos[2], Sintia Almeida[1], Siomar Soares[1], Ulisses Pereira[1], Vinícius C Abreu[1], Artur Silva[2], Anderson Miyoshi[1], Vasco Azevedo[1*]

From X-meeting 2011 - International Conference on the Brazilian Association for Bioinformatics and Computational Biology
Florianópolis, Brazil. 12-15 October 2011

Abstract

Background: Pan-genomic studies aim, for instance, at defining the core, dispensable and unique genes within a species. A pan-genomics study for vaccine design tries to assess the best candidates for a vaccine against a specific pathogen. In this context, rather than studying genes predicted to be exported in a single genome, with pan-genomics it is possible to study genes present in different strains within the same species, such as virulence factors. The target organism of this pan-genomic work here presented is *Corynebacterium pseudotuberculosis*, the etiologic agent of caseous lymphadenitis (CLA) in goat and sheep, which causes significant economic losses in those herds around the world. Currently, only a few antigens against CLA are known as being the basis of commercial and still ineffective vaccines. In this regard, the here presented work analyses, *in silico*, five *C. pseudotuberculosis* genomes and gathers data to predict common exported proteins in all five genomes. These candidates were also compared to two recent *C. pseudotuberculosis in vitro* exoproteome results.

Results: The complete genome of five *C. pseudotuberculosis* strains (1002, C231, I19, FRC41 and PAT10) were submitted to pan-genomics analysis, yielding 306, 59 and 12 gene sets, respectively, representing the core, dispensable and unique *in silico* predicted exported pan-genomes. These sets bear 150 genes classified as secreted (SEC) and 227 as potentially surface exposed (PSE). Our findings suggest that the main *C. pseudotuberculosis in vitro* exoproteome could be greater, appended by a fraction of the 35 proteins formerly predicted as making part of the variant *in vitro* exoproteome. These genomes were manually curated for correct methionine initiation and redeposited with a total of 1885 homogenized genes.

Conclusions: The *in silico* prediction of exported proteins has allowed to define a list of putative vaccine candidate genes present in all five complete *C. pseudotuberculosis* genomes. Moreover, it has also been possible to define the *in silico* predicted dispensable and unique *C. pseudotuberculosis* exported proteins. These results provide *in silico* evidence to further guide experiments in the areas of vaccines, diagnosis and drugs. The work here presented is the first whole *C. pseudotuberculosis in silico* predicted pan-exoproteome completed till today.

* Correspondence: vasco@icb.ufmg.br
[1]Molecular and Celular Genetics Laboratory, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
Full list of author information is available at the end of the article

Background

Reverse Vaccinology (RV) [1] analyses the genome sequence of a pathogen, which is an expected coded sequence for all the possible expressed genes in the pathogen's life cycle. All Open Reading Frames (ORF's) derived from the genome sequence can be evaluated using a computer program to determine their ability as vaccine candidates, giving special attention to exported proteins, as these are essential in host-pathogen interactions. Examples of such interactions include: (i) adherence to host cells, (ii) invasion of the cell to which there is compliance, (iii) damage to host tissues, (iv) environmental stresses resistance from the defense machinery of the cell being infected, and (v) mechanisms for subversion of host immune response [2-5].

Regarding exported proteins, these can distinguish between those that are exported to the cell wall, and after cleaved, release the mature portion into the extracellular milieu, which are referred to as secreted proteins (SEC), and those proteins exported to the cell wall which, even after cleaved, do not release the mature portion to the extracellular milieu, due to one or more hydrophobic motifs causing anchoring to the cell wall, and which are referred to as potentially surface exposed proteins (PSE). Different PSE subcategories exist according to the presence of a carboxy (C) or amino (N) terminal portion anchored to the cell wall, lipoproteins (E), end terminal loops (L), retention signals-like such as LGxTG, LysM, GW, Choline binding and PG binding (R), in combination or not with other PSE subcategories [6].

The term 'Reverse' from RV can be explained by the reverse genetics (RG) technique. Before the dawn of genomic, there were attempts to discover the responsible genes from a phenotype, reversing the research path of Crick's Central Dogma [7] (DNA → RNA → Protein) discovery. Holding the likely gene sequence, several techniques can be used to identify gene sequence modifications responsible for changes in the organism's phenotype. Crick's Central Dogma principle is also used for RV, as this technique searches within a gene sequence for possible proteins that could act as antigens capable of stimulating an immune response in a host organism [8].

The concept of RV was adapted to fit a new reality of widespread availability of genomic data [9]. With this technique, instead of searching for targets in a single strain or subspecies of an organism, it is now possible to simultaneously research in dozen of genomes, exploring potential joint antigens or exclusive ones to multiple genomes [10]. The availability of a large number of genomes to implement RV has lead to the emergence of the pan-genomics reverse vaccinology concept [11], which can also apply to the concepts of core, extended (dispensable) and character (unique) genomes. While the core genome is composed of exported genes (genes that transcribe for exported proteins) that are common to these multiple strains and could represent candidates for a vaccine, the dispensable genome consists of genes that are absent in at least one of the strains of the studied species and the unique genome consists of genes that are specific to only a particular a strain [10]. From the standpoint of vaccines, the core genome represents to be a good candidate to compose a vaccine that is suitable for all studied strains. In this regard, the first step to enable any pan-genomic reverse vaccinology study is to predict the core genome, along this work denominated *in silico* predicted pan-exoproteome (ISPPE). The model organism here analyzed *(C. pseudotuberculosis)* is a Gram-positive (GRAM+) bacterium, intracellular facultative parasite that affects small ruminants causing a chronic infectious pyogranulomatous disease characterized by the formation of abscesses in lymph nodes [12]. This pathogen infects mainly goats and sheep causing caseous lymphadenitis, but can also infect a huge variety of hosts throughout the world such as camels, horses, cattle, buffaloes, llamas, alpacas and, more rarely, humans [13-18], causing different diseases with different degrees of severity in each of them [12,19].

Results and discussion

In silico exoproteome prediction schema

As shown in our proposed prediction schema (Figure 1), the software SurfG+ (Surface Gram positive), specially configured for GRAM+ bacteria, is responsible for most of the sub-cellular classifications, which vary between cytoplasmic (CYT), membrane (MEM), SEC and PSE (Figure 2). SurfG+ was configured for GRAM+ bacteria. Figure 1 represents the prediction schema using SurfG+ and three additional software, TatP 1.0 [20], SecretomeP 2.0 [21] and NclassG+ [22], which are specialized in non-classical secretion prediction. SurfG+ incorporates SignalP 3.0 predictor, responsible for identification of classical putative secreted proteins or exported proteins by the SEC pathway [23].

The results obtained after running SurfG+, TapP, SecretomeP and NClassG+ have gave rise to two gene data sets labeled as SEC and PSE, which correspond to the *C. pseudotuberculosis* ISPPE. These ISPPE data sets are composed of putative proteins present fivefold (5x), fourfold (4x), threefold (3x), twofold (2x) or onefold (1x), where fivefold means that a gene was predicted in all five strains, four fold meaning that a gene was predicted in four strains, and so on. A gene fold was obtained by reciprocal blast results, as described in the methods section. Since not all predicted genes are named, it was necessary to create a pan genome identifier, here denominated pan *locus*, to nominate each unique gene fold. The pan *locus* is unique within a pan genome and is shared by all homologous genes. For example, when a putative exported protein

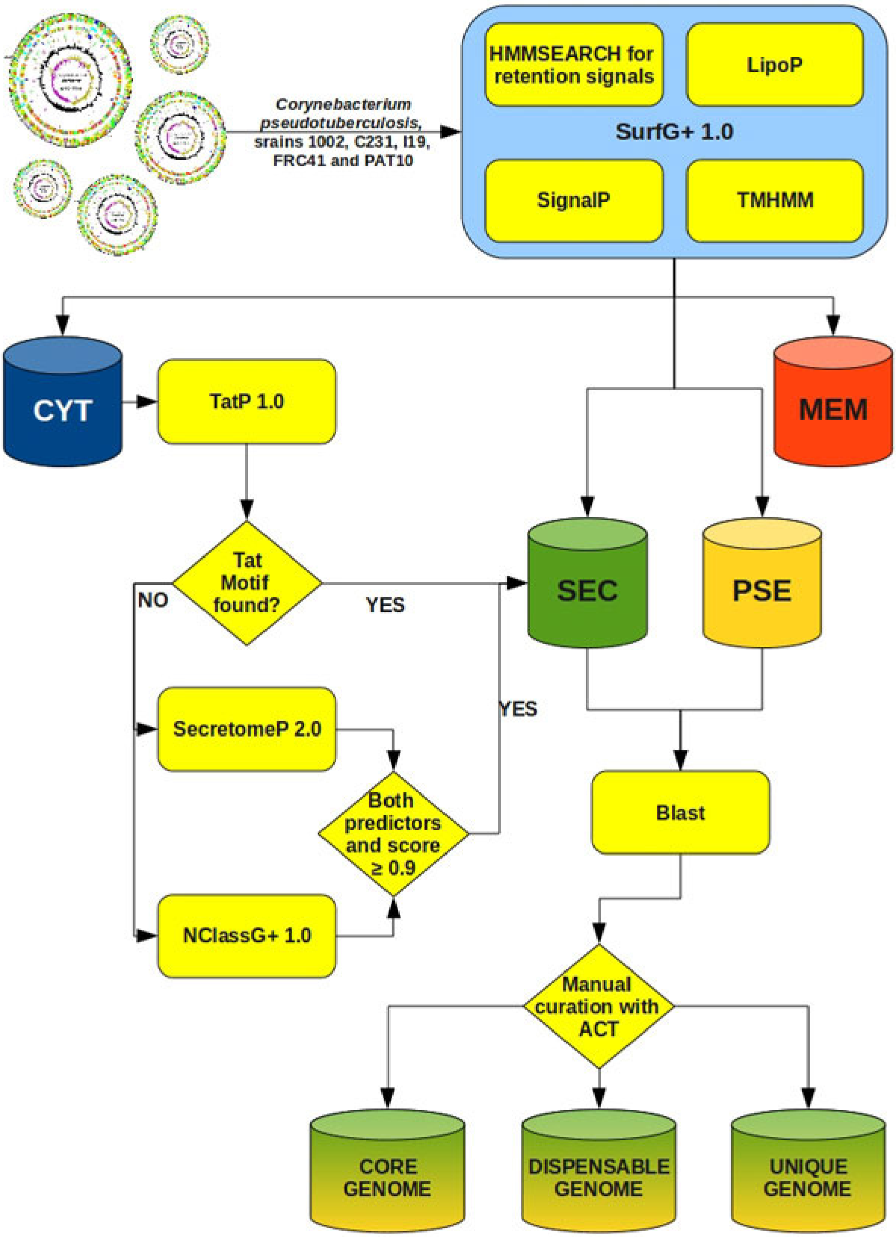


Figure 1 ***C. pseudotuberculosis*** **pan genomic prediction schema**. Software used, identified sub-cellular compartments and flow scheme to create the final pan genomic data sets.

was found within the five strains, each gene copy received the same pan *locus* to facilitate further data processing and identification. Following, it was necessary to confirm these results by systematical manual curation of each gene using the ACT tool from the Artemis software package [24]. Once completed this manual curation, it was possible to answer several questions regarding the correctness of each blast result and, as a consequence, it was possible to identify, for instance, that a gene formerly classified as 1x was indeed a 5x, as the other four gene copies were created starting beyond the signal peptide motif. After initial methionine correction, and also taking into account homologous genes, a new prediction step indicated all remaining putative proteins to be exported, composing the core ISPPE. However, gene's start positions incorporating a less probable signal peptide motif were also observed. In general, genes formerly predicted as Nx proved to be correct by manual curation as the remaining (5-N)x genes were predicted as cytoplasmic, PSE or pseudogenes. These results are particularly interesting because they compose

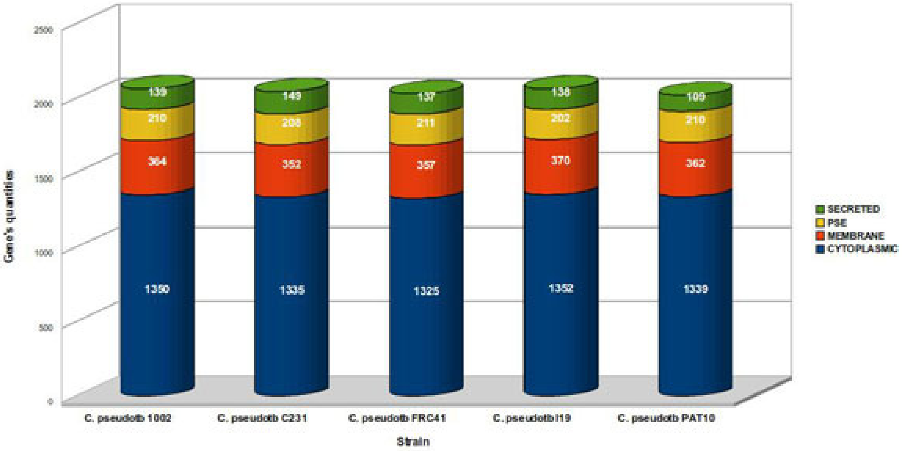


Figure 2 Predicted gene quantities by sub-cellular compartment from full *C. pseudotuberculosis* genomes. Classification of more than 10,000 distinct genes from the five different *C. pseudotuberculosis* strains in the four sub-cellular categories: cytoplasmic (CYT), membrane (MEM), potentially surface exposed (PSE) and secreted (SEC). Predictions were made using the schema presented in Figure 1.

the dispensable and unique ISPPE data sets. These genome annotation corrections, as a consequence of these analyses, were incorporated into the official annotation of the five *C. pseudotuberculosis* strains deposited at GenBank in August, 2011. This genomes are also available in the additional file 1, as EMBL files.

Classical and non-classical secreted putative proteins

Figure 3 exhibits the *in silico* predicted pan secretome results for *C. pseudotuberculosis*, which comprise 150 genes, out of 377 from the whole ISPPE, representing 750 *locus_tags* in the five studied *C. pseudotuberculosis* strains. However, despite representing 750 *locus_tags*, not all were predicted as secreted. If at least one gene copy, within a specific pan *locus*, was not predicted as secreted, it still received the same pan *locus* but was not classified as part of the predicted core secretome. There are 122 genes composing the predicted core secretome (5x), followed by 25 genes constituting the predicted dispensable secretome (4x, 3x and 2x) and just 3 genes as the predicted unique secretome (1x). These results were obtained applying the prediction schema from Figure 1; however, different contributions were obtained from different predictors, as shown in Figure 4.

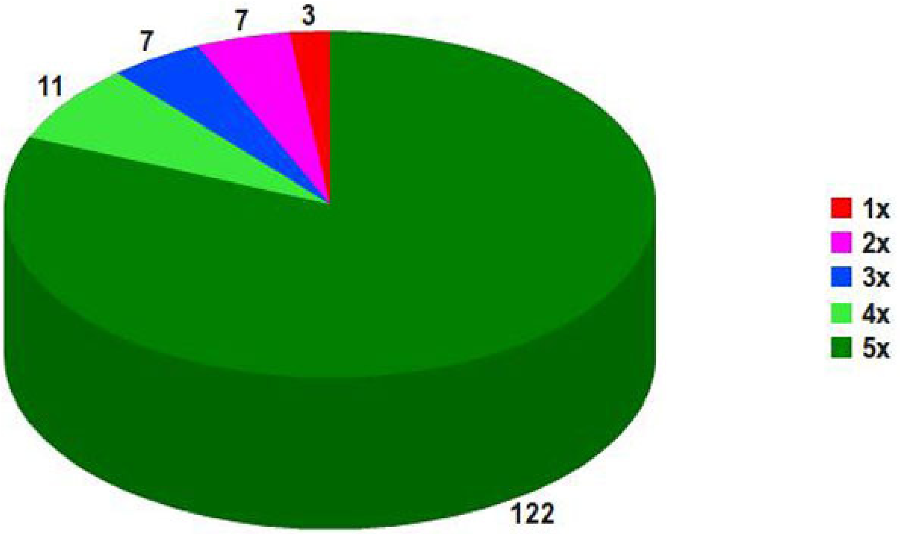


Figure 3 Predicted *C. pseudotuberculosis* pan secretome. Predictions for 150 genes from strains 1002, C231, I19, FRC41 and PAT10 made by SurfG+ 1.0, TatP 1.0 Server and SecretomeP 2.0 Server.

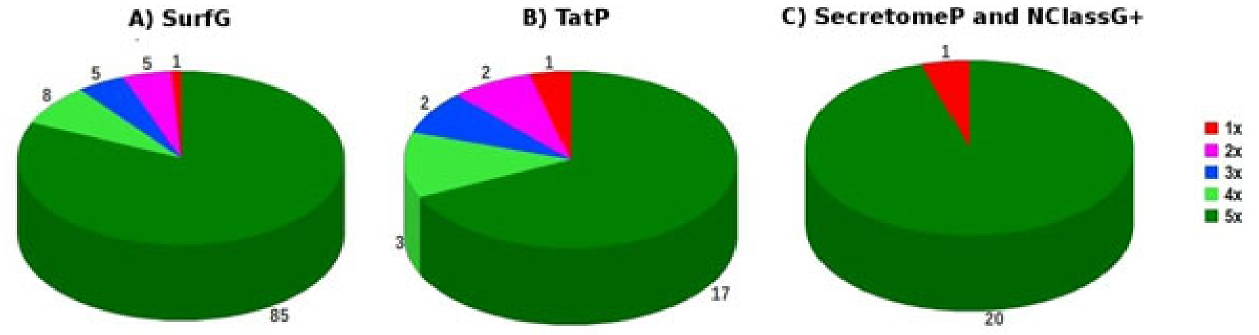


Figure 4 Predicted *C. pseudotuberculosis* pan secretome by predictor software. Predicted secreted genes coverage in the predicted pan secretome of the five bacterial strains separated by predictor software SurfG+, TatP and SecretomeP.

SurfG+ predicted 104 genes, corresponding 85, 18 and 1 to the predicted core, dispensable and unique secretome respectively. On the other hand, TatP predicted 25 genes, of which 17, 7 and 1 corresponded to the predicted core, dispensable and unique secretome respectively. Finally, SecretomeP and NClassG+ predicted 21 genes, corresponding 20 and 1 to the predicted core and unique secretome respectively. It can be easily observed that the main predicted portion is originated by SurfG+, as it predicts putative proteins possibly secreted by the SEC pathway. A considerable portion of genes (~31%), only within the predicted core secretome, comes from non-classical secretion predictors that cannot be ignored when the subject is about vaccine candidates.

The dispensable and unique *C. pseudotuberculosis* predicted secretomes contain ~8%, or 58 *locus_tags*, not predicted as secreted. Putative proteins predicted as CYT, PSE and putative frame shifts (pseudogenes) account for 22, 24 and 10 *locus_tags* respectively. In the dispensable and unique *C. pseudotuberculosis in silico* predicted secretomes, the numbers of genes identified as membrane integral or absent in a genome are insignificant. Nevertheless, the manual curation step ensured no annotation errors in these predictions, making it possible to claim the hypothesis that these differences could be due to environment adaptations. A table containing the complete list of *C. pseudotuberculosis* secreted proteins is available in the additional file 2.

Potentially surface exposed (PSE) putative proteins

The SurfG+ software was calibrated by the cell wall thickness for each *C. pseudotuberculosis* strain. Figure 5 shows 184 genes, out of 377 from the whole ISPPE, comprising the predicted core surfaceome (5x), 34 genes composing the predicted dispensable surfaceome (4x, 3x and 2x) and just 9 genes as predicted unique surfaceome (1x). These 227 genes account for 1135 *locus_tags* in all five strains. In this set, homologous genes within a pan *locus* do not ever share the same sub-cellular prediction. Genes predicted as MEM, CYT, SEC and putative pseudogenes account for 29, 23, 20 and 17 distinct *locus_tags*, respectively. Genes predicted as MEM (~3%) compose the second major group. This could be explained by the fact that membrane proteins already contain hydrophobic extension and could be more susceptible to expose or occult parts of a protein to the extracellular milieu. However, the same reasoning does not suit to explain the third major group of *locus_tags* with surfaceome pan *locus* that correspond to proteins predicted as secreted ones. These 20 *locus_tags* that were predicted as secreted, but also received surfaceome pan *locus*, raise a question; do these fit SEC or PSE labels? There exist no simple paths to estimate their sub-cellular compartment by software, since some *locus_tags* were predicted as PSE receiving surfaceome pan *locus* and other were predicted as SEC and also received secretome pan *locus*. Ten pan *locus* (plcppse193, plcppse194, plcppse205, plcppse218, plcppse226, plcpsec096, plcpsec097, plcpsec098, plcpsec100, plcpsec101) faces this question, as some genes appear in both the predicted secretome and surfaceome.

The PSE subcategories show predominance of genes, as presented in Figure 6. Most of the 1045 genes predicted as PSE are cell wall anchored outward C-terminal (~40%) (≥ 50 AA long), followed by lipoproteins (~24%), outward loops (~11%) (≥ 100 AA long) and outward N-terminal (~17%) (≥ 50 AA long), whereas genes containing retention signals (PSE R) account only for ~8%.

The PSE results of all strains were analyzed considering that a significant cell wall thickness difference between strain I19 and the other ones was observed (~34 nm versus ~24 nm). Despite the significant cell wall thickness difference, a small difference was predicted in the genome, which accounts for a decrease in the number of PSE and an increase in the number of MEM genes in *C. pseudotuberculosis* strain I19. A table containing the complete list of *C. pseudotuberculosis* PSE proteins is available in the additional file 3.

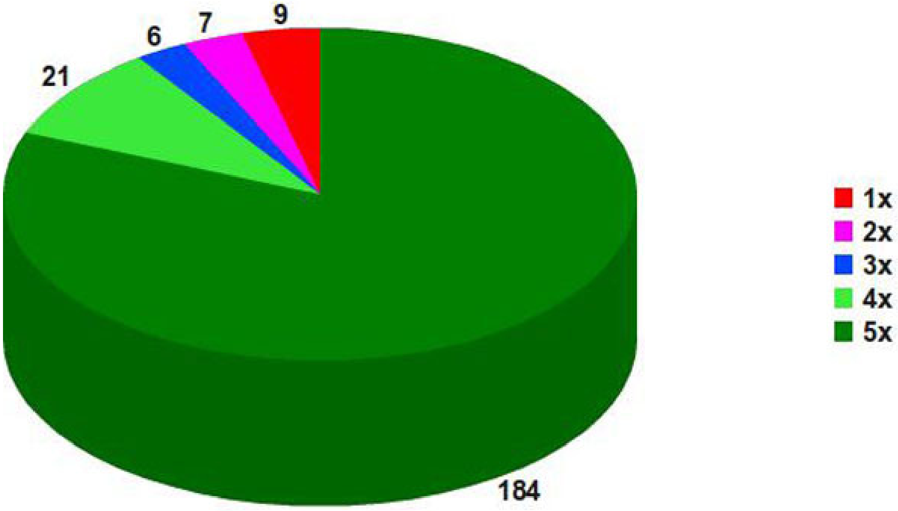


Figure 5 Predicted *C. pseudotuberculosis* pan surfaceome. Pan surfaceome predictions for 227 genes from strains 1002, C231, I19, FRC41 and PAT10, performed by SurfG+ 1.0.

Revised *in vitro* exoproteome results

The 104 observed genes in both TPP/LC-MS^E [25] and 2-DE-MALDI-TOF/TOF, (Silva WM, Seyffert N, Castro TLP, Santos AV, Pacheco LGC, Santos AR, Ciprandi A, Zurita-Turk M, Dorella FA, Andrade HM, Pimenta AMC, Silva A, Miyoshi A, Azevedo V, unpublished observations) experiments were compared with the ISPPE results here presented. This comparison, explained in the methods section, brought novel insights into the *in vitro* exoproteome and showed the possibility of having additional genes in the main *C. pseudotuberculosis in vitro* exoproteome. In Table 1 are listed all 35 proteins of the variant *in vitro* exoproteome (strains 1002 and C231), that correspond to ~23% of the total amount. These proteins were found to be highly conserved in the five compared *C. pseudotuberculosis* strains and comprise the core ISPPE. Moreover, it was verified that three proteins (ADL20466, ADL20097 e ADL19973), previously classified as belonging to the variant *in vitro* exoproteome of strains 1002 [25], did actually belong to the main *in vitro* exoproteome. These findings

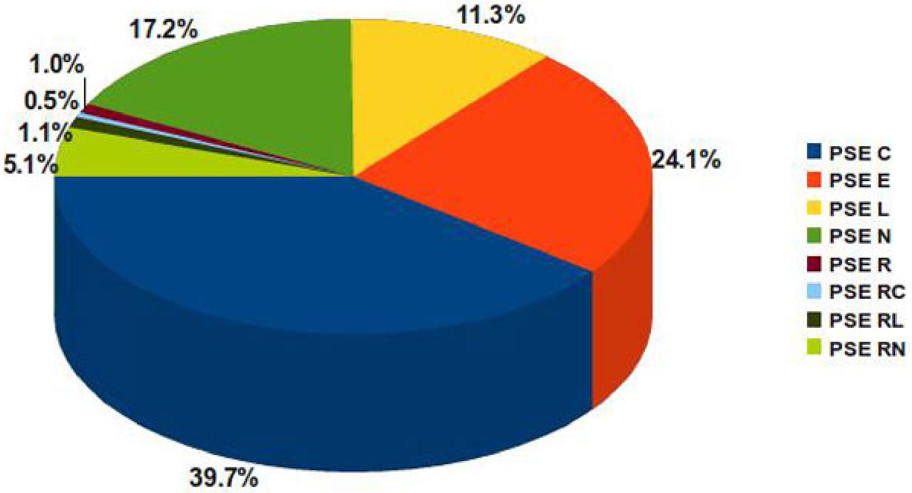


Figure 6 Predicted *C. pseudotuberculosis* pan surfaceome by PSE subcategories. PSE categories are distributed in outward C-terminal or N-terminal portion greater than or equal 50 AA. Outward N or C terminal greater than 100 AA are classified as L. Lipogenes identified by LipoP are classified as E and retention signals identified by HMMSEARCH profiles are classified as R. These labels can also be conjugated to create other PSE subcategories.

Table 1 Core *C. pseudotuberculosis in silico* predicted pan-exoproteome found in the variant *in vitro* exoproteome

Protein identifier	*locus_tag*	Gene name	Product	Predicted local sub-cellular	GenBank organism identifier
ADL19972	Cp1002_0064		Hypothetical protein	PSE E	CP001809
ADL20140	Cp1002_0237	*slpA*	Surface layer protein A	SEC	CP001809
ADL20222	Cp1002_0320		Hypothetical protein	PSE N	CP001809
ADL20288	Cp1002_0388		L,D-transpeptidase catalytic domain, region YkuD	SEC	CP001809
ADL20391	Cp1002_0497	*malE*	Maltose/maltodextrin transport system substrate-binding protein	PSE E	CP001809
ADL20455	Cp1002_0562	*sprT*	Trypsin	PSE C	CP001809
ADL20477	Cp1002_0584	*cynT*	Carbonic anhydrase	PSE E	CP001809
ADL20508	Cp1002_0615		Hypothetical protein	SEC	CP001809
ADL20574	Cp1002_0681	*rpfB*	Resuscitation-promoting factor RpfB	SEC	CP001809
ADL20656	Cp1002_0766		Hypothetical protein	SEC	CP001809
ADL21028	Cp1002_1144	*yceG*	Amino deoxychorismate lyase	SEC	CP001809
ADL21239	Cp1002_1362		Hypothetical protein	PSE E	CP001809
ADL21302	Cp1002_1425	*ctaC*	Cytochrome c oxidase subunit II	PSE C	CP001809
ADL21537	Cp1002_1669		Hypothetical protein	SEC	CP001809
ADL21667	Cp1002_1802	*lipY*	Secretory lipase	SEC	CP001809
ADL09524	CpC231_0025	*pld*	Phospholipase D	SEC	CP001829
ADL09532	CpC231_0033	*pbpA*	Penicillin-binding protein A	SEC	CP001829
ADL09691	CpC231_0196		Hypothetical protein	SEC	CP001829
ADL09697	CpC231_0203	*pbpB*	Penicillin binding protein transpeptidase	SEC	CP001829
ADL09852	CpC231_0360	*oppA1*	Oligopeptide-binding protein oppA	PSE E	CP001829
ADL09871	CpC231_0379		Hypothetical protein	SEC	CP001829
ADL09872	CpC231_0380	*malE*	Maltotriose-binding protein	PSE E	CP001829
ADL09990	CpC231_0503	*lytR*	Transcriptional regulator lytR	PSE C	CP001829
ADL10248	CpC231_0766		Hypothetical protein	SEC	CP001829
ADL10460	CpC231_0982	*ciuA*	Iron ABC transporter substrate-binding	PSE E	CP001829
ADL10489	CpC231_1012	*yceI*	Protein yceI	SEC	CP001829
ADL10626	CpC231_1150		Zinc metallopeptidase	PSE C	CP001829
ADL10663	CpC231_1187		Lipoprotein	PSE E	CP001829
ADL10880	CpC231_1409	*pknL*	Serine/threonine protein kinase	PSE N	CP001829
ADL11196	CpC231_1737		Corynomycolyl transferase	SEC	CP001829
ADL11213	CpC231_1756		Hypothetical protein	SEC	CP001829
ADL11326	CpC231_1871		Hypothetical protein	PSE N	CP001829
ADL11338	CpC231_1885		Membrane protein	SEC	CP001829
ADL11339	CpC231_1886		Hypothetical protein	SEC	CP001829
ADL11410	CpC231_1959	*glpQ*	Glycerophosphoryl diester phosphodiesterase	PSE E	CP001829

The 35 proteins listed in this table were not found in the experimental main *in vitro* exoproteome [47; 48] but were found in the *in silico* predicted pan-exoproteome of all five *C. pseudotuberculosis* strains.

give raise to the possibility that more proteins of the variant *in vitro* exoproteome indeed make part of the main *in vitro* exoproteome.

This comparison also served as a rebuttal argument against some specific genes. The Cp1002_0369 gene, classified under the plcpsec100 pan *locus* as a pseudogene, was identified by the *in vitro* exoproteome experiment. Interestingly, this gene copy also suits the plcppse226 pan *locus*. Both pan *locus* make part of previous related genes that already showed difficulties to be classified, by software, into any potential sub-cellular compartment, as some genes within the pan *locus* fit both SEC and PSE labels. The *in silico* predictions enforces that there are at least three secreted proteins, inspite of the other two gene copies being predicted as having PSE and CYT labels.

Furthermore, the genes plcppse180, plcppse192, plcpsec077, plcpsec095 and plcpsec099 also had both genes found in the main *in vitro* exoproteome of strains 1002 and C231, but were not classified in the ISPPE. The plcppse180 pan *locus* holds a putative pseudogene (CpPAT10_0459), and is therefore not present in the *in silico* predicted core surfaceome. Other genes were predicted as cytoplasmic. It is possible that these genes were wrongly assembled since there is evidence that at least two homologous genes, from strains 1002 and C231, are exported to the extracellular milieu.

Core *C. pseudotuberculosis* ISPPE candidates homologous to *Mtb*

Within the core *C. pseudotuberculosis* ISPPE, homologous genes to those of the previously studied *Mycobacterium tuberculosis* H37Rv (*Mtb*) were observed. In this work we present some of these homologous genes featuring at least 90% protein alignment and 50% identity within this alignment. These cut-offs were obtained during the search for *C. pseudotuberculosis* homologous genes in the *Mtb* genome.

The core *C. pseudotuberculosis* ISPPE, that accounts for ~81% of the total, is composed of 306 genes or 1,530 distinct *locus_tags*, being ~40% predicted as SEC and ~60% predicted as PSE proteins, of which 20 genes present high similarity to *Mtb*'s genes (Table 2); however, not all of these *Mtb* genes have known functions.

In this regard, here we only discuss some of these *Mtb*'s genes with experimental evidence. The plcppse174 pan *locus* shows 51% protein identity with Rv3915 (YP_178027.1), a gene named *cwlM* that was the first autolysin gene identified and cloned from *Mtb*. This finding offers a new drug target class that could alter the permeability of the mycobacterium cell wall and enhance the effectiveness of treatments for tuberculosis [26]. Applying principles of *in vivo* expression technology (IVET), it was possible to identify upregulated genes from *Mtb* in an *in vitro* simulation of anaerobic persistence condition. The upregulated genes under hypoxic condition (dissolved oxygen <1%) include Rv0050 (*ponA1*), a penicillin binding protein that has 52% protein identity to the plcppse165 pan *locus* and 90% alignment extension [27]. The plcpsec122 pan *locus* shows ~58% protein identity with Rv2752c (NP_217268.1), a unique bi-functional *Mtb* gene that owns both β-lactamase and RNase activities. Both activities are lost upon deletion of the 100 AA long C-terminal 100 tail, which contains an additional loop when compared to the RNase J of *Bacillus subtilis* [28]. As it can be observed, the plcppse080 pan *locus* appears twice in Table 2, as it is homologous to both NADH dehydrogenase gene copies of *Mtb*, *ndh* (NP_216370.1) and *ndhA* (NP_214906.1), with ~57% protein identity. In *Mtb*, energy generation is mainly performed by type II dehydrogenases *ndh* and *ndhA*, being both, as such, essential genes [29].

The plcpsec113 pan *locus* is homologous to the *glmU* gene (NP_215534.1), holding ~59% protein identity and more than 90% alignment extension. This gene is essential in *Mtb*, being required for optimal bacterial growth, and has been selected as a possible drug target for structural and functional investigation [30]. *GlmU* is a bifunctional acetyltransferase/uridyltransferase that catalyses the formation of UDP-GlcNAc from GlcN-1-P. UDP-GlcNAc is the substrate for two important biosynthetic pathways: lipopolysaccharide and peptidoglycan synthesis. Due to its important roles, *glmU* had its conformational structure solved [30]. The plcpsec113 pan *locus* for *C. pseudotuberculosis* is an interesting putative drug candidate since it is predicted to be secreted, part of the core ISPPE and is able to infer its conformational structure by homology modeling using *Mtb glmU*.

Several genes involved in mannoglycoconjugate biosynthesis have shown to be involved in virulence, due to their central role in biosynthesis of major surface-associated glycoconjugates. Within these genes, the *Mtb* gene *manB* (Rv3264c) is defined as a GDP-mannose pyrophosphorylase (GDPMP) and disruption of its activity leads to decrease of surface-associated mannosylated lipoglycans. For GDPMP, this decrease correspond directly to reduced virulence in both BALB/c mice and cultured human macrophages [31]. The *Mtb manB* gene holds 69% protein identity to the plcpsec110 pan *locus* and more than 90% alignment extension, making plcpsec110 a considerable putative drug target.

Mycolic acids and multimethyl-branched fatty acids are found uniquely in the cell envelope and are essential for survival, virulence and antibiotic resistance of *Mtb*. Acyl-CoA carboxylases (ACCases) commit acyl-CoAs to the biosynthesis of these unique fatty acids. Previous studies indicate that AccD5 is important for cell envelope lipid biosynthesis and its disruption leads to pathogen death [32]. The *Mtb* gene *accD5* (NP_217797.1) had its structure determined and also shows ~74% protein identity to the plcppse045 pan *locus* in more than 90% alignment extension, making it also a promising candidate for further vaccine candidate evaluations.

Moreover, it was demonstrated that *Mtb* can use heme as an iron source, suggesting that *Mtb* contains a yet-unknown heme acquisition system [33]. We found that the *C. pseudotuberculosis* plcpsec076 pan *locus* holds ~52% protein identity to the *Mtb* gene *hemE* (NP_217194.1) and more than 90% alignment size, therefore also representing an interesting drug target for *C. pseudotuberculosis.*

Candidates filtering

The here presented results provide a plethora of putative vaccine candidates never seen before for

Table 2 Core *C. pseudotuberculosi s in silic o* predicted pan-exoproteome homologous to *Mtb*'s proteins

Corynebacterium pseudotberculosis				*Mycobacterium tuberculosis*				
pan *locus*	Reference genome *locus_tag*	ORF size	% of amino acid alignment's identity	ORF size	*locus_tag*	Gene name	protein ID	Annotated product
plcpsec106	cpfrc_00104	488	69.10	461	Rv3790		NP_218307.1	oxidoreductase
plcpsec076	cpfrc_00276	371	51.56	357	Rv2678c	*hemE*	NP_217194.1	uroporphyrinogen decarboxylase
plcppse023	cpfrc_00283	535	52.51	529	Rv0528		NP_215042.1	transmembrane protein
plcppse045	cpfrc_00491	543	73.72	548	Rv3280	*accD5*	NP_217797.1	propionyl-CoA carboxylase beta chain
plcpsec110	cpfrc_00506	362	69.03	359	Rv3264c	*manB*	YP_177951.1	D-alpha-D-mannose-1-phosphate guanylyltransferase MANB
plcpsec111	cpfrc_00508	151	51.45	139	Rv3259		NP_217776.1	hypothetical protein
plcpsec113	cpfrc_00705	487	58.67	495	Rv1018c	*glmU*	NP_215534.1	bifunctional N-acetylglucosamine-1-phosphate uridyltransferase/glucosamine-1-phosphate acetyltransferase
plcpsec115	cpfrc_00945	64	63.33	64	Rv1642	*rpmI*	NP_216158.1	50S ribosomal protein L35
plcppse080	cpfrc_01015	452	57.08	470	Rv0392c	*ndhA*	NP_214906.1	membrane NADH dehydrogenase
plcppse080	cpfrc_01015	452	58.10	463	Rv1854c	*ndh*	NP_216370.1	NADH dehydrogenase
plcpsec041	cpfrc_01074	403	62.96	381	Rv1488		NP_216004.1	hypothetical protein
plcpsec119	cpfrc_01121	504	53.71	457	Rv1407	*fmu*	NP_215923.1	Fmu protein (SUN protein)
plcppse085	cpfrc_01126	417	55.58	418	Rv1391	*dfp*	NP_215907.1	bifunctional phosphopantothenoylcysteine decarboxylase/phosphopantothenate synthase
plcpsec138	cpfrc_01214	79	68.42	82	Rv2708c		NP_217224.1	hypothetical protein
plcpsec122	cpfrc_01267	683	57.76	558	Rv2752c		NP_217268.1	hypothetical protein
plcpsec124	cpfrc_01393	239	57.83	250	Rv2149c	*yfiH*	NP_216665.1	hypothetical protein
plcppse104	cpfrc_01424	412	50.38	429	Rv2195	*qcrA*	NP_216711.1	Rieske iron-sulfur protein QcrA
plcpsec128	cpfrc_01757	313	59.42	322	Rv3579c		NP_218096.1	tRNA/rRNA methyltransferase
plcppse131	cpfrc_01798	480	62.21	491	Rv2443	*dctA*	NP_216959.1	C4-dicarboxylate-transport transmembrane protein DctA
plcppse165	cpfrc_02038	721	52.00	678	Rv0050	*ponA1*	YP_177687.1	bifunctional penicillin-binding protein 1A/1B
plcppse174	cpfrc_02102	393	51.41	406	Rv3915	*cwlM*	YP_178027.1	hydrolase

Related *C. pseudotuberculosis*'s proteins containing at least 50% amino acid identity and 90% alignment size to the *Mtb* H37Rv's proteins.

C. pseudotuberculosis. However, genes predicted as MEM and CYT account respectively for 18% and 65% of the *in silico* predicted pan genome. Despite the 227 surfaceome and 150 secretome genes here presented, these only represents ~16% of the *C. pseudotuberculosis in silico* predicted pan genome. Most of the genes remain inaccessible for the current *in silico* prediction techniques and it is possible that these neglected genes could also be good candidates against *C. pseudotuberculosis*. These findings raise the need for more elaborated and driven software or prediction schemas capable of uncovering these major genome neglected portions. Using the prediction schema here presented, it was possible to include more than ~2% of non-classic secreted putative proteins that compose putative vaccine candidates. However, this low income amount of vaccine candidates is due to the optional parameter selected in our prediction schema, the non-classic secreted score greater than or equal 0.90. If using the default parameter from the software secretomeP and NClassG+, this income would be increased up to ~6% and the final income of putative vaccine candidates would be ~20%, using a couple of motifs predictors as depicted in Figure 1. The current reverse vaccinology software allows obtaining a number of candidates closer to 20% of the *C. pseudotuberculosis* genome. These considerations raise a question: supposing that novel software for unexplored secretion pathways come into scenario, what is the genome's percentage that could be selected as putative vaccine candidates? Supposing that this percentage reaches 40%, how could the problem of choosing between almost one thousand putative vaccine candidates to be used for the next vaccine production stage for *C. pseudotuberculosis* be solved? This dilemma could be solved by using further software prediction just like those addressing epitopes MHC class I and II allele affinity [34]; however, this could be just a part of the solution. There are chances of solving this dilemma by means of broader vaccine projects, which would take into account particular variables for each target organism in order to minimise research efforts and the number of possible vaccine candidates [35].

In silico versus non-in silico

It is broadly known that *in silico* genome investigations could give evidence about the genome's function and structure. It is also known that such *in silico* investigations could only be proved or denied by non-*in silico* experiments. Therefore, such reasonable thinking is not a single-hand avenue. Non-*in silico* experiments could be improved by means of more comprehensive or specific approaches with the objective of getting a closer answer to the reality for biological questions. The fact is that *in silico* analyses cannot vary when executed over and over again and no matter how many folds are run. We know that exactly 122 genes will be always predicted as having classical exportation motifs; on the other hand, we cannot expect the same behavior from non-*in silico* analysis. Some real proteins could be or not be found in an *in vitro* or *in vivo* exoproteome result, due to an uncountable number of factors [21]. Therefore, we suggest that the core *C. pseudotuberculosis* ISPPE could be composed of a larger number of predicted genes, but such confirmation could only be affirmed with additional non-*in silico* exoproteome experiments.

Conclusions

The *in silico* pan-exoproteome prediction methodology applied to the pathogen *C. pseudotuberculosis* helps to raise new insights into putative vaccine candidates against CLA. Additional investigations of the *in vitro* exoproteome of two strains of *C. pseudotuberculosis*, 1002 and C231, showed evidence that the major part of the variant *in vitro* exoproteome is contained in the core ISPPE. A simultaneous curation of the *in silico* predicted core secretome and surfaceome within the five *C. pseudotuberculosis* strains also contributed to homogenize the genome annotations and it was possible to fix the most probable putative methionine proteins. Moreover, putative miss assembled genes, formerly classified as pseudogenes by *in silico* analyses, were also revised. The efforts to create a *C. pseudotuberculosis* ISSPE catalogue proved to be necessary and computationally viable to ensure a uniform set of putative vaccine candidates free of annotation errors.

Methods

Genomes

The analyzed *C. pseudotuberculosis* genomes were obtained from the GenBank according to the following accession numbers: EMBL: CP001809 (strain 1002), EMBL: CP001809 (strain C231), EMBL: CP002251 (strain I19), EMBL: CP002924 (strain PAT10) and EMBL: NC_014329 (strain FRC41).

Prediction schema

Predicted genes from all five *C. pseudotuberculosis* strain genomes were exported as amino acid fasta files using the Artemis software. These fasta files were passed as parameters to SurfG+ 1.0 (Figure 1), and lists of genes predicted as CYT, SEC, PSE and MEM were created by this software. Genes formerly predicted as CYT by SurfG+ were then submitted to the TapP 1.0 predictor; when a Tat motif was found, the putative protein was automatically classified as SEC, otherwise, another prediction round would took place using two other non-classic secretion predictors, SecretomeP 2.0 and NclassG+ 1.0. With a positive prediction from both software and a prediction score greater than or equal to 0.90, the genes were automatically classified as SEC. The SEC and PSE data sets were finally submitted to a reciprocal blastp processing and posterior filtering, giving rise to the fivefold categories according to folds occurring in each strain: 5x, 4x, 3x, 2x and 1x. The results were then manually curated using the ACT software and strain 1002, the first to be sequenced and annotated. The strain 1002 was disposed, in ACT software, in the middle of two pairs of the other two genome strains, facilitating to exhibit differences among all of them.

SurfG+ 1.0

Sub-cellular localization prediction of *C. pseudotuberculosis* putative proteins was made by *in silico* analysis using the SurfG+ 1.0 software [6]. SurfG+ is a pipeline for protein sub-cellular prediction that incorporates common software, such as SignalP, LipoP and TMHMM to search for motifs. It also creates novel HMMSEARCH profiles to predict cell wall retention signals. SurfG+ starts searching, in the following order for: retention signals, lipoproteins, SEC pathway export motifs and transmembrane motifs. If none of these motifs are found in a protein sequence, it is then characterized as CYT. A novel possible characterization introduced by SurfG+ is its ability to better distinguish between MEM and PSE, by informing an expected cell wall thickness in amino acids. Using the literature or an electronic microscopy it is possible to estimate a reasonable cell wall thickness value for prokaryotic organisms. By means of this last option, *C. pseudotuberculosis* genes were classified into four different sub-cellular locations: CYT, MEM, PSE, or SEC.

TatP 1.0 Server

Twin-arginine signal peptide motifs were predicted using the on line server hosted by http://www.cbs.dtu.dk/services/TatP/[20]. Only putative proteins formerly classified as CYT by SurfG+ were submitted to the TatP analyses. There were no intersections between SignalP and TatP predictions.

SecretomeP 2.0 and NClassG+ 1.0

Non-classical secreted putative proteins were predicted using the online server hosted by http://www.cbs.dtu.dk/services/SecretomeP/[21]. NClassG+ [22], a second non-classical secreted protein predictor, was also used; however, the predictions were directly performed contacting the software authors. This double check prediction ensured greater accuracy. Only those genes formerly classified as CYT by SurfG+ and without the twin-arginine signal peptide motifs were submitted to a non-classical secreted analysis. Despite the significant scores of SecretomeP and Nclass+, ranging between 0.5 and 1.0, only those genes with a score greater than or equal to 0.9 were selected, in order to ensure a minimal false positive in future wet lab experiments, the focus of our research group.

Pan genome

To predict the *C. pseudotuberculosis* pan genome, reciprocal blastp results were used. All the putative proteins predicted as SEC were put apart in a single amino acid fasta file to make a reciprocal blast. A similar file was also created for the proteins predicted as PSE. To avoid homologous mismatches, the blastp results obtained using the PAM70 substitution matrix and the 10^{-6} e-value were manually filtered. In this regard, the first step was to establish the alignment size and identity percentages of cut-offs, being 89.58 and 50.00%, respectively, for SEC putative proteins, whereas for PSE putative proteins, these cut-offs were 88.16 and 48.80%, respectively. Identity percentages closer to 50% are explained by frame shifts not annotated until this work. All the putative proteins from the five strains (query) with alignment size and identity percentages higher than these cut-offs had no more than one group of blast hits (subject) against the others strains. Moreover, within each of these blast hits groups, there was a blast hit from the query protein against it self as subject. The results were manually curated using the ACT software, from the Artemis package [24], using the strain 1002 as reference strain for the other two strains. This ACT view was composed by strains C231-1002-I19 and FRC41-1002-I19. Each putative protein predicted as SEC and PSE was compared against their other four homologues for correct initial methionine, frame shifts and finally annotating the correct sub-cellular location.

Revised *in vitro* exoproteome results

In lists 1 and 2 of the annex are both gene *locus* present in the *C. pseudotuberculosis* ISPPE, together with the quantity of homologous genes present in the all five genomes. These results were inserted in a relational database, denominated *C. pseudotuberculosis* Data Base (CpDB) [36], in a specific table called ‘exopred’. The list of the *in vitro* exoproteome proteins was also inserted to the CpDB into a table called ‘exo’ that discriminates the identification of each protein regarding GenBank (protein id), as well as in which strains it is found. To make a relationship between the ‘exopred’ and ‘exo’ tables, a third table of the CpDB, called ‘gene’, which contains all the functional annotation of the genomes of *C. pseudotuberculosis*, was created. The CpDB is the repository of the pan genome of *C. pseudotuberculosis*, harbouring the genomes since their initial genomic prediction, deposited in the GenBank, as well as the annotation corrections for future deposits. For this last purpose, the CpDB stores the identification of each protein according to the GenBank. In this way, it is possible to make a link between the three tables in the form of a clause of JOIN of the SQL: “*... WHERE gene.locus_tag = exopred.locus_tag AND gene.protein_id = exo.protein_id AND exopred.pangenome_coverage = ‘5x’ ...*”. This clause returns the registries of the CpDB whose *locus_tag* in the gene table is equal to the *locus_tag* of the explored table, being this same gene in the protein_id field in the exo table with prediction of belonging to all five genomes. Other conditions can also be included, such as for example, restraining the results to specific genes of a *C. pseudotuberculosis* strain or simultaneously present in the exoproteome of specific strains.

Additional material

Additional file 1: *C. pseudotuberculosis* genomes. The five *C. pseudotuberculosis* genomes here checked, as EMBL files.

Additional file 2: Predicted *C. pseudotuberculosis* pan secretome. List of the 150 genes for 750 *locus_tags* from the five *C. pseudotuberculosis* strains.

Additional file 3: Predicted *C. pseudotuberculosis* pan surfaceome. List of the 227 genes for 1135 locus_tags from the five *C. pseudotuberculosis* strains.

Acknowledgements

We would like to thank Daniel Restrepo-Montoya, Camilo Pino, Luis F Nino, Manuel E Patarroyo and Manuel A Patarroyo for their support with NGClass+ predictions. We would also like to thank to LISI, ALGOS-UN and FIDIC bioinformatics staff which were funded supported by Asociación Investigación Solidaria SADAR, Caja Navarra (CAN) (Navarra, Spain) and the Spanish Agency for International Development Cooperation (AECID).
Funding: CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brazil.
This article has been published as part of *BMC Genomics* Volume 13 Supplement 5, 2012: Proceedings of the International Conference of the Brazilian Association for Bioinformatics and Computational Biology (X-meeting 2011). The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/13/S5.

Author details

[1]Molecular and Celular Genetics Laboratory, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil. [2]DNA Polimorfism Laboratory, Universidade Federal do Pará, Campus do Guamá - Belém, PA, Brazil. [3]Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology, Nonakuri, Purba Medinipur, West Bengal, India.

Authors' contributions

VA encouraged the research, BMC application, provided references, applied biological knowledge and gave final approval of the version to be published. ARS conducted all the software analyses, manually corrected annotation errors in the five genomes, developed the prediction schema and wrote the paper. EB contributed to the manually curation of all pseudogenes from all bacterial strains. MZT made substantial contributions to the design and interpretation of the manuscript. UP, AG, FD, FS, AC, AP, DB, FF, LG, RR, SA, SS, VCA, AS and AM have given final approval of the version to be published.

Competing interests

The authors declare that they have no competing interests.

Published: 19 October 2012

References

1. Rappuoli R: **Reverse vaccinology.** *Curr Opin Microbiol* 2000, **3**:445-450.
2. Sibbald MJJB, van Dij JML: **Secretome Mapping in Gram-Positive Pathogens. In Karl Wooldridge (ed.), Bacterial Secreted Protein: Secretory Mechanisms and Role in Pathogenesis.** *Caister Academic Press* 2009, 193-225.
3. Simeone R, Bottai D, Brosch R: **ESX/type VII secretion systems and their role in host-pathogen interaction.** *Curr Opin Microbiol* 2009, **12**:4-10.
4. Stavrinides J, McCann HC, Guttman DS: **Host-pathogen interplay and the evolution of bacterial effectors.** *Cell Microbiol* 2008, **10**:285-292.
5. Bhavsar AP, Guttman JA, Finlay BB: **Manipulation of host-cell pathways by bacterial pathogens.** *Nature* 2007, **449**:827-834.
6. Barinov A, Loux V, Hammani A, Nicolas P, Langella P, Ehrlich D, Maguin E, van de Guchte M: **Prediction of surface exposed proteins in Streptococcus pyogenes, with a potential application to other Gram-positive bacteria.** *Proteomics* 2009, **9**:61-73.
7. Strasser BJ: **A world in one dimension: Linus Pauling, Francis Crick and the central dogma of molecular biology.** *Hist Philos Life Sci* 2006, **28**:491-512.
8. Rappuoli R: **IS15 Developing vaccines in the era of genomics and toll receptors.** *Immunology* 2005, **116**:1.
9. Rinaudo CD, Telford JL, Rappuoli R, Seib KL: **Vaccinology in the genome era.** *J Clin Invest* 2009, **119**:2515-2525.
10. Lapierre P, Gogarten JP: **Estimating the size of the bacterial pan-genome.** *Trends Genet* 2009, **25**:107-110.
11. Bambini S, Rappuoli R: **The use of genomics in microbial vaccine development.** *Drug Discov Today* 2009, **14**:252-260.
12. Dorella FA, Pacheco LG, Seyffert N, Portela RW, Meyer R, Miyoshi A, Azevedo V: **Antigens of Corynebacterium pseudotuberculosis and prospects for vaccine development.** *Expert Rev Vaccines* 2009, **8**:205-213.
13. Afzal M, Sakir M, Hussain MM: **Corynebacterium pseudotuberculosis infection and lymphadenitis (taloa or mala) in the camel.** *Trop Anim Health Prod* 1996, **28**:158-162.
14. Aleman M, Spier SJ, Wilson WD, Doherr M: **Corynebacterium pseudotuberculosis infection in horses: 538 cases (1982-1993).** *J Am Vet Med Assoc* 1996, **209**:804-809.
15. Silva A, Schneider MPC, Cerdeira L, Barbosa MS, Ramos RTJ, Carneiro AR, Santos R, Lima M, D'Afonseca V, Almeida SS, Santos AR, Soares SC, Pinto AC, Ali A, Dorella FA, Rocha F, de Abreu VAC, Trost E, Tauch A, Shpigel N, Miyoshi A, Azevedo V: **Complete genome sequence of Corynebacterium pseudotuberculosis I19, a strain isolated from a cow in Israel with bovine mastitis.** *J Bacteriol* 2011, **193**:323-324.
16. Selim SA: **Oedematous skin disease of buffalo in Egypt.** *J Vet Med B Infect Dis Vet Public Health* 2001, **48**:241-258.
17. Peel MM, Palmer GG, Stacpoole AM, Kerr TG: **Human lymphadenitis due to Corynebacterium pseudotuberculosis: report of ten cases from Australia and review.** *Clin Infect Dis* 1997, **24**:185-191.
18. Trost E, Ott L, Schneider J, Schröder J, Jaenicke S, Goesmann A, Husemann P, Stoye J, Dorella FA, Rocha FS, Soares SDC, D'Afonseca V, Miyoshi A, Ruiz J, Silva A, Azevedo V, Burkovski A, Guiso N, Join-Lambert OF, Kayal S, Tauch A: **The complete genome sequence of Corynebacterium pseudotuberculosis FRC41 isolated from a 12-year-old girl with necrotizing lymphadenitis reveals insights into gene-regulatory networks contributing to virulence.** *BMC Genomics* 2010, **11**:728.
19. Ruiz JC, D'Afonseca V, Silva A, Ali A, Pinto AC, Santos AR, Rocha AAMC, Lopes DO, Dorella FA, Pacheco LGC, Costa MP, Turk MZ, Seyffert N, Moraes PMRO, Soares SC, Almeida SS, Castro TLP, Abreu VAC, Trost E, Baumbach J, Tauch A, Schneider MPC, McCulloch J, Cerdeira LT, Ramos RTJ, Zerlotini A, Dominitini A, Resende DM, Coser EM, Oliveira LM, Pedrosa AL, Vieira CU, Guimarães CT, Bartholomeu DC, Oliveira DM, Santos FR, Rabelo EM, Lobo FP, Franco GR, Costa AF, Castro IM, Dias SRC, Ferro JA, Ortega JM, Paiva LV, Goulart LR, Almeida JF, Ferro MIT, Carneiro NP, Falcão PRK, Grynberg P, Teixeira SMR, Brommonschenkel S, Oliveira SC, Meyer R, Moore RJ, Miyoshi A, Oliveira GC, Azevedo V: **Evidence for Reductive Genome Evolution and Lateral Acquisition of Virulence Functions in Two Corynebacterium pseudotuberculosis Strains.** *PLoS ONE* 2011, **6**:e18551.
20. Bendtsen JD, Nielsen H, Widdick D, Palmer T, Brunak S: **Prediction of twin-arginine signal peptides.** *BMC Bioinformatics* 2005, **6**:167.
21. Bendtsen JD, Kiemer L, Fausbøll A, Brunak S: **Non-classical protein secretion in bacteria.** *BMC Microbiol* 2005, **5**:58.
22. Restrepo-Montoya D, Pino C, Nino LF, Patarroyo ME, Patarroyo MA: **NClassG+: A classifier for non-classically secreted Gram-positive bacterial proteins.** *BMC Bioinformatics* 2011, **12**:21.
23. Petersen TN, Brunak S, von Heijne G, Nielsen H: **SignalP 4.0: discriminating signal peptides from transmembrane regions.** *Nat Methods* 2011, **8**:785-786.
24. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation.** *Bioinformatics* 2000, **16**:944-945.
25. Pacheco LGC, Slade SE, Seyffert N, Santos AR, Castro TLP, Silva WM, Santos AV, Santos SG, Farias LM, Carvalho MAR, Pimenta AMC, Meyer R, Silva A, Scrivens JH, Oliveira SC, Miyoshi A, Dowson CG, Azevedo V: **A combined approach for comparative exoproteome analysis of Corynebacterium pseudotuberculosis.** *BMC Microbiol* 2011, **11**:12.
26. Deng LL, Humphries DE, Arbeit RD, Carlton LE, Smole SC, Carroll JD: **Identification of a novel peptidoglycan hydrolase CwlM in Mycobacterium tuberculosis.** *Biochim Biophys Acta* 2005, **1747**:57-66.
27. Saxena A, Srivastava V, Srivastava R, Srivastava BS: **Identification of genes of Mycobacterium tuberculosis upregulated during anaerobic persistence by fluorescence and kanamycin resistance selection.** *Tuberculosis (Edinb)* 2008, **88**:518-525.
28. Sun L, Zhang L, Zhang H, He Z: **Characterization of a Bifunctional β-Lactamase/Ribonuclease and Its Interaction with a Chaperone-Like Protein in the Pathogen Mycobacterium tuberculosis H37Rv.** *Biochemistry (Moscow)* 2011, **76**:350-358.
29. Velmurugan K, Chen B, Miller JL, Azogue S, Gurses S, Hsu T, Glickman M, Jacobs WRJ, Porcelli SA, Briken V: **Mycobacterium tuberculosis nuoG is a virulence gene that inhibits apoptosis of infected host cells.** *PLoS Pathog* 2007, **3**:e110.
30. Zhang Z, Bulloch EMM, Bunker RD, Baker EN, Squire CJ: **Structure and function of GlmU from Mycobacterium tuberculosis.** *Acta Crystallogr D Biol Crystallogr* 2009, **65**:275-283.
31. McCarthy TR, Torrelles JB, MacFarlane AS, Katawczik M, Kutzbach B, Desjardin LE, Clegg S, Goldberg JB, Schlesinger LS: **Overexpression of Mycobacterium tuberculosis manB, a phosphomannomutase that increases phosphatidylinositol mannoside biosynthesis in Mycobacterium smegmatis and mycobacterial association with human macrophages.** *Mol Microbiol* 2005, **58**:774-790.
32. Lin T, Melgar MM, Kurth D, Swamidass SJ, Purdon J, Tseng T, Gago G, Baldi P, Gramajo H, Tsai S: **Structure-based inhibitor design of AccD5, an essential acyl-CoA carboxylase carboxyltransferase domain of Mycobacterium tuberculosis.** *Proc Natl Acad Sci USA* 2006, **103**:3072-3077.
33. Jones CM, Niederweis M: **Mycobacterium tuberculosis can utilize Heme as an iron source.** *J Bacteriol* 2011, **193**:1767-1770.
34. Davies MN, Flower DR: **Harnessing bioinformatics to discover new vaccines.** *Drug Discov Today* 2007, **12**:389-395.
35. Santos A, Ali A, Barbosa E, Silva A, Miyoshi A, Barh D, Azevedo V: **The reverse vaccinology - A contextual overview.** *IIOABJ* 2011, **2**:8-15.
36. Azevedo V, Santos AR, Soares S, Ali A, Pinto A, Magalhaes A, Barbosa E, Ramos R, Cerdeira L, Carneiro A, Abreu V, Almeida S, Schneider P, Silva A, Miyoshi A: **Automated functional annotation.** In *Bioinformatics - trends and methodologies. Volume 1.* InTechOpen;Mahdavi MA 2011:722.

doi:10.1186/1471-2164-13-S5-S6
Cite this article as: Santos *et al.*: The *Corynebacterium pseudotuberculosis in silico* predicted pan-exoproteome. *BMC Genomics* 2012 **13**(Suppl 5):S6.

II.III.3 A combined approach for comparative exoproteome analysis of *Corynebacterium pseudotuberculosis*.

Pacheco LG, Slade SE, Seyffert N, Santos AR, Castro TL, Silva WM, Santos AV, Santos SG, Farias LM, Carvalho MA, Pimenta AM, Meyer R, Silva A, Scrivens JH, Oliveira SC, Miyoshi A, Dowson CG, **Azevedo V**.

BMC Microbiol. 2011 Jan 17;11(1):12. doi: 10.1186/1471-2180-11-12.

Pacheco *et al. BMC Microbiology* 2011, **11**:12
http://www.biomedcentral.com/1471-2180/11/12

BMC Microbiology

RESEARCH ARTICLE Open Access

A combined approach for comparative exoproteome analysis of *Corynebacterium pseudotuberculosis*

Luis GC Pacheco[1,2,3], Susan E Slade[4], Núbia Seyffert[2], Anderson R Santos[2], Thiago LP Castro[2], Wanderson M Silva[2], Agenor V Santos[1], Simone G Santos[5], Luiz M Farias[5], Maria AR Carvalho[5], Adriano MC Pimenta[1], Roberto Meyer[3], Artur Silva[6], James H Scrivens[4], Sérgio C Oliveira[1], Anderson Miyoshi[2], Christopher G Dowson[4], Vasco Azevedo[2*]

Abstract

Background: Bacterial exported proteins represent key components of the host-pathogen interplay. Hence, we sought to implement a combined approach for characterizing the entire exoproteome of the pathogenic bacterium *Corynebacterium pseudotuberculosis*, the etiological agent of caseous lymphadenitis (CLA) in sheep and goats.

Results: An optimized protocol of three-phase partitioning (TPP) was used to obtain the *C. pseudotuberculosis* exoproteins, and a newly introduced method of data-independent MS acquisition (LC-MS^E) was employed for protein identification and label-free quantification. Additionally, the recently developed tool SurfG+ was used for *in silico* prediction of sub-cellular localization of the identified proteins. In total, 93 different extracellular proteins of *C. pseudotuberculosis* were identified with high confidence by this strategy; 44 proteins were commonly identified in two different strains, isolated from distinct hosts, then composing a core *C. pseudotuberculosis* exoproteome. Analysis with the SurfG+ tool showed that more than 75% (70/93) of the identified proteins could be predicted as containing signals for active exportation. Moreover, evidence could be found for probable non-classical export of most of the remaining proteins.

Conclusions: Comparative analyses of the exoproteomes of two *C. pseudotuberculosis* strains, in addition to comparison with other experimentally determined corynebacterial exoproteomes, were helpful to gain novel insights into the contribution of the exported proteins in the virulence of this bacterium. The results presented here compose the most comprehensive coverage of the exoproteome of a corynebacterial species so far.

Background

Corynebacterium pseudotuberculosis is a facultative intracellular pathogen that belongs to the so-called CMN (*Corynebacterium-Mycobacterium-Nocardia*) group, a distinct subgroup of the *Actinobacteria* that also includes other highly important bacterial pathogens, such as *Corynebacterium diphtheriae* and *Mycobacterium tuberculosis*. The most distinctive feature of these Gram-positive bacteria is the unique composition of the cell envelope, characterized by the presence of long chain fatty acids, known as mycolic acids, on the surface of the cell [1,2].

The main recognizable disease caused by *C. pseudotuberculosis* is caseous lymphadenitis (CLA) in sheep and goats, though this bacterium can also infect several other hosts, including humans [1,3]. Typical manifestations of CLA in small ruminants include formation of abscesses in superficial and internal lymph nodes, and in visceral organs [3]. Despite the important economic losses caused by this disease to sheep and goat husbandry worldwide, no effective treatment exists, and the efficacy of the currently available vaccines and diagnostic methods is still controversial [4].

The search for *C. pseudotuberculosis* molecular determinants that contribute to CLA pathogenesis lead to the

* Correspondence: vasco@icb.ufmg.br
[2]Department of General Biology, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Av. Antônio Carlos, Belo Horizonte, 31.270-901, Brazil
Full list of author information is available at the end of the article

recognition of two exported proteins as the major virulence-associated factors of this bacterium known to date: a secreted phospholipase D (PLD) [5]; and an ABC-type transporter component of an iron uptake system (FagB) [6]. In fact, one might expect that the majority of the virulence determinants of *C. pseudotuberculosis* would be present in the exoproteome, *i.e.* the entire set of bacterial proteins found in the extracellular milieu [7]. This is because exported proteins participate in essential steps of the host-pathogen interplay, including: (i) adhesion to host cells; (ii) invasion; (iii) damage to host tissues; (iv) resistance to environmental stresses during infection; and (iv) subversion of the host's immune response mechanisms [8-10].

In two previous attempts to characterize the *C. pseudotuberculosis* exoproteome, our group optimized a protocol of salting out of proteins using sulfate and butanol, known as three-phase partitioning (TPP), for isolation of the extracellular proteins of this bacterium [11], and generated a library of *C. pseudotuberculosis* mutant strains possessing transposon insertions in genes coding for probable exported proteins [12]. In the former study, we were able to determine the optimal conditions for obtaining the best recovery of immunoreactive extracellular proteins of *C. pseudotuberculosis* [11]. The second study in turn, enabled us to identify various previously uncharacterized *C. pseudotuberculosis* exported proteins, being that at least two of them are apparently involved in virulence [12]. Now, the very recent conclusion of the *C. pseudotuberculosis* Genome Project by our group, associated to the current availability of high-throughput proteomic technologies, permitted us to perform a much more comprehensive analysis of this bacterium's exoproteome.

In this study, we sought to implement a combined approach for comparative exoproteome analysis of different *C. pseudotuberculosis* strains. The strategy included: (i) the previously optimized TPP protocol for isolation of the extracellular proteins [11]; (ii) a newly introduced method of data-independent LC-MS acquisition (LC-MS^E) for protein identification and quantification [13,14]; and (iii) the recently developed tool SurfG+ for *in silico* prediction of protein sub-cellular localization in Gram-positive bacteria [15]. We believe that the experimental approach used is very suitable for profiling bacterial exoproteomes, as it shown to be easily applicable to different strains with very good reproducibility. This is an advantage over what is commonly observed for proteomic approaches based on two-dimensional (2D) gel electrophoresis, where there is more variability, but is apparently the method of choice for most of the bacterial exoproteome studies published recently [16-20]. Furthermore, the LC-MS^E method provides high subproteome coverage, due to enhanced sensitivity, and allows for label-free analysis of differentially expressed proteins [14]; this latter possibility enables the detection of variations in the exoproteomes of different strains that could be missed by simply profiling the exoproteins, and meets the growing interest in performing physiological proteomic studies of bacteria [21,22].

We were able to identify 93 different *C. pseudotuberculosis* extracellular proteins with high confidence by analyzing the exoproteomes of two strains isolated from different hosts that presented distinct virulence phenotypes under laboratory conditions [23,24]. Most of the identified proteins were predicted *in silico* to have an extracytoplasmic localization. To the best of our knowledge, these results compose the largest inventory of experimentally confirmed exoproteins of a single corynebacterial species to date. Importantly, the comparative exoproteome analyses permitted us to speculate on the probable contributions of different *C. pseudotuberculosis* extracellular proteins to the virulence of this bacterium.

Results and Discussion

Exoproteome analysis of *Corynebacterium pseudotuberculosis*

The extracellular proteins of two *C. pseudotuberculosis* strains, one isolated from a goat (strain 1002) the other from a sheep (strain C231), cultivated in a chemically-defined medium, were extracted/concentrated by the TPP technique. The trypsinized protein samples were then submitted to LC-MS^E analysis.

Seventy soluble extracellular proteins of the 1002 strain could be confidentially identified by this methodology, whereas the number of proteins identified in the exoproteome of the C231 strain was sixty-seven. Altogether, 93 different *C. pseudotuberculosis* exoproteins were identified in this study (Figure 1). These findings agree with the results of previous experiments by our group, in which we have used a 2D-PAGE based strategy for a preliminary appraisal of the *C. pseudotuberculosis* exoproteome (additional file 1). Eighty protein spots, mostly concentrated in the pI range between 3.0 and 6.0, could be reproducibly visible in the 2D gels generated from TPP-extracted extracellular proteins of the 1002 strain (additional file 1). The fact that we have found 70 proteins in the exoproteome of this strain with high confidence when using the LC-MS^E method (Figure 1) indicates that this novel methodology allowed us to identify virtually the complete set of extracellular proteins that are commonly observed in the gel based methodologies (additional file 1). Moreover, the expected existence of protein isoforms among the eighty protein spots observed in the 2D gels, and the identification by LC-MS^E of many proteins out of the pI range 3.0-6.0, suggests that the latter methodology is much more suitable for obtaining a comprehensive coverage of

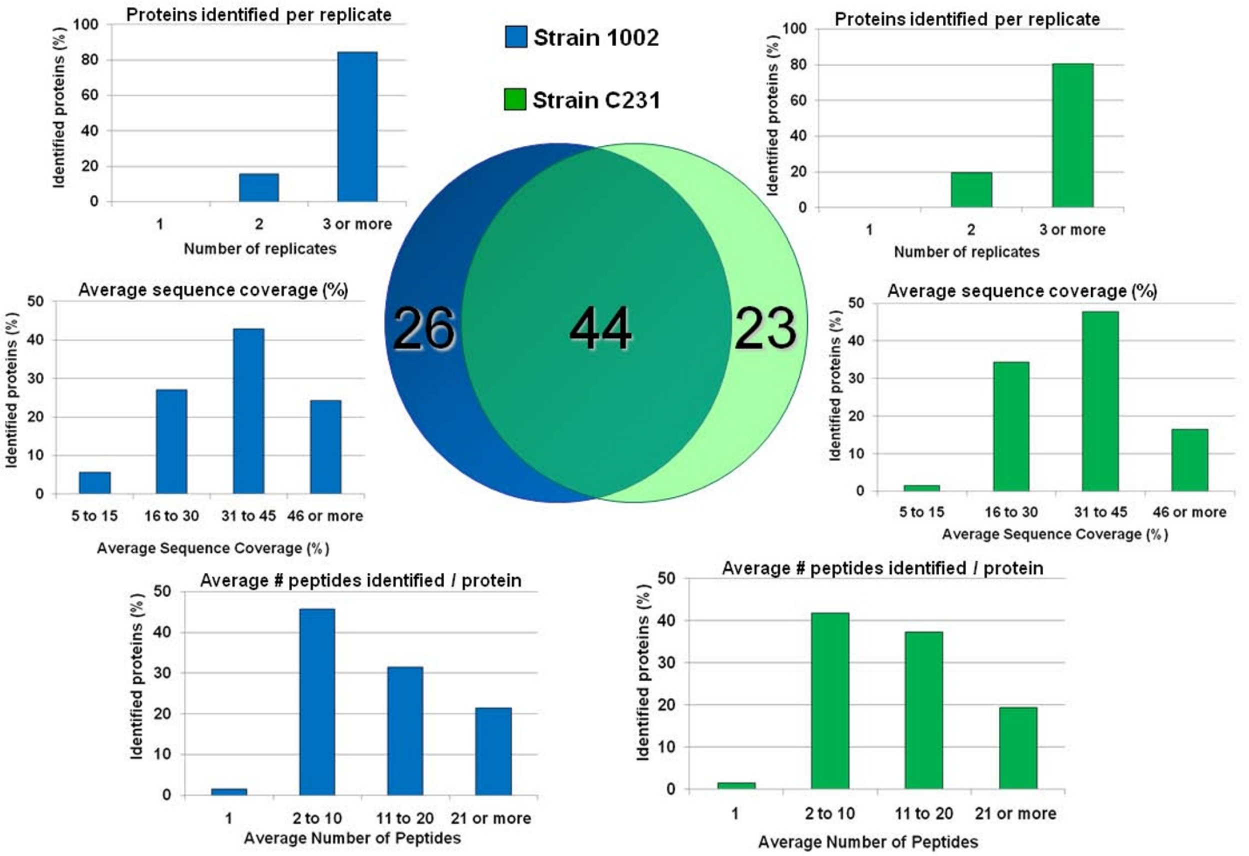


Figure 1 Analysis of the extracellular proteins of two different *C. pseudotuberculosis* strains allowed for identification of the core and variant exoproteomes. TPP-extracted extracellular proteins of the strains 1002 and C231 of *C. pseudotuberculosis* were submitted to LC-MS^E analysis. The Venn-diagram shows the numbers of commonly identified and variant exoproteins between the strains. The number of replicates in which a given protein was observed, the average peptides identified per protein, and the average sequence coverage of the proteins in each exoproteome studied, are shown as frequency distributions for comparison purposes.

the bacterial exoproteome. Noteworthy, is the use of LC-MS^E for exoproteome profiling which required (i) much less time and labor than the gel based proteomic strategy, and (ii) much less protein sample necessary for each experimental replicate, with only 0.5 μg per replicate used in the LC-MS^E compared to 150 μg for the 2D gels [refer to Patel *et al*. [25] for a comprehensive comparison on these proteomic strategies].

The performance of the combined methodology used in the present study (TPP/LC-MS^E) for mapping the *C. pseudotuberculosis* exoproteome was very similar for both strains analyzed, as can be seen by the average numbers of peptides observed per protein in the two proteomes (16.5 and 15.0) and by the average sequence coverage of the proteins identified (37.5% and 35.0%) (Figure 1). Consistent with this, the majority of the proteins detected in each extracellular proteome were shared by the goat and sheep isolates; this permitted us to define a core *C. pseudotuberculosis* exoproteome composed of 44 proteins out of the 93 different extracellular proteins identified. Additional files 2, 3 and 4 list all the proteins identified in the exoproteomes of the two *C. pseudotuberculosis* strains, along with molecular weights, isoelectric points, main orthologs, predicted sub-cellular localizations, number of peptides experimentally observed, and sequence coverage.

Searches of similarity against publicly available protein databases using the Blast-p tool [26] showed that ortholog proteins can be found in the pathogenic *Corynebacterium diphtheriae* for most of the identified *C. pseudotuberculosis* exoproteins (additional files 2, 3 and 4), as would be expected due to the close phylogenetic relationship of these species [27]. Nevertheless, no significant orthologs could be found for six proteins of the *C. pseudotuberculosis* exoproteome, even when using the position-specific iterated BLAST (PSI-BLAST) algorithm [28], namely the proteins [GenBank:ADL09626], [GenBank:ADL21925], [GenBank:ADL11253], [GenBank:ADL20222], [GenBank:ADL09871], and [GenBank:ADL21537] (additional files 2, 3 and 4). With the exception of [GenBank:ADL11253], all these proteins were predicted by different tools as being truly exported proteins. This means they are the only five exoproteins identified in this study which are probably unique for *C. pseudotuberculosis*.

Prediction of sub-cellular localization of the identified proteins

Most of the proteins identified in the exoproteomes of the two *C. pseudotuberculosis* strains were also predicted to have a probable extracytoplasmic localization after *in silico* analysis of the sequences of these proteins with different bioinformatics tools, thereby corroborating our *in vitro* findings (Figure 2, additional file 5). It is important to note here that we are considering the exoproteome as the entire set of proteins released by the bacteria into the extracellular milieu. That means we are looking to: (i) proteins possessing classical signals for active exportation by the different known mechanisms, which are directly secreted into the cell supernatant or that remain exposed in the bacterial cell surface and are eventually released in the growth medium [7]; and (ii) proteins exported by non-classical pathways, without recognizable signal peptides [29]. Besides, one might also expect to observe in the extracellular proteome a small number of proteins primarily known to have cytoplasmic localization; although some of these proteins are believed to be originated from cell lysis or leakage, like in the extreme situation reported by Mastronunzio *et al*. [19], a growing body of evidence suggests that moonlighting proteins (in this case, cytoplasmic proteins that assume diverse functions in the extracellular space) may be commonly found in the bacterial exoproteomes [29-32].

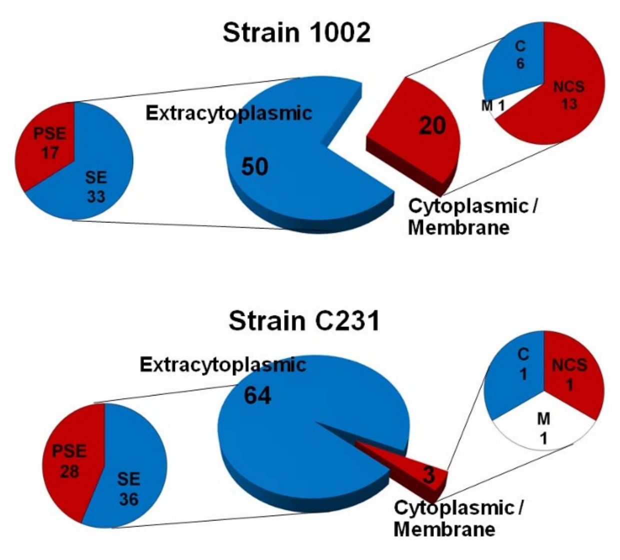


Figure 2 Most of the identified *C. pseudotuberculosis* exoproteins were predicted by the SurfG+ program as having an extracytoplasmic localization. The proteins identified in the exoproteomes of each *C. pseudotuberculosis* strain were analyzed by SurfG+ and attributed a probable final sub-cellular localization. Proteins classified as having a cytoplasmic localization were further analyzed with the SecretomeP tool for prediction of non-classical (leaderless) secretion. Besides, literature evidence for exportation by non-classical pathways was also used to re-classify the cytoplasmic proteins (see text for details). SE = secreted; PSE = potentially surface exposed; C = cytoplasmic; M = membrane; NCS = non-classically secreted.

By using the recently developed tool SurfG+ we were able to classify the identified *C. pseudotuberculosis* proteins into four different categories: (i) secreted, (ii) potentially surface exposed (PSE), (iii) membrane and (iv) cytoplasmic (Figure 2, additional files 2, 3 and 4). Basically, this software brings together the predictions of global protein localizations performed by a series of well-known algorithms, and innovates by allowing for an accurate prediction of PSE proteins [15]. This possibility of classification provides us with valuable information on the proteins identified, as bacterial surface exposed proteins are believed to play important roles in the host-pathogen interactions during infection and many of these proteins have been shown to be highly protective when used in vaccine preparations [33,34].

From a total of 93 different *C. pseudotuberculosis* proteins identified in this study, 75% (70) could be predicted as containing signals for active exportation (secretion or surface exposition) following SurfG+ analysis (Figure 2). Taken together, these proteins represent roughly 50% of all predicted secreted proteins in the recently sequenced genome of *C. pseudotuberculosis*, and around 15% of all predicted PSE proteins of this bacterium (A.R. Santos, pers. comm.).

The concordance of our *in vitro* identification of exoproteins with the *in silico* predictions of protein exportation is higher than what has normally been observed in recent exoproteome analyses of different bacteria [17-19,35,36]. For comparison, Hansmeier *et al*. [17] reported that exportation signals could be predicted in only 42 (50%) out of 85 different proteins identified in the extracellular and cell surface proteomes of *Corynebacterium diphtheriae*. The authors of this study are not the only to speculate on a probably important contribution of cross-contamination of the protein sample during preparation procedures for the observation of high numbers of proteins not predicted as having extracellular location in the bacterial exoproteomes [17,31]. We believe that the proportionally higher identification of proteins possessing exportation signals in the present study could have happened due to a series of different factors, including: (i) our methodology for isolation of the bacterial extracellular proteins might have extracted less “contaminant” cytoplasmic proteins than did other methodologies reported in previous studies; (ii) the combined strategy used by SurfG+ to predict protein sub-cellular localization might have performed better in the identification of exported proteins than happened with other strategies, sometimes based in only one prediction tool; (iii) the fact that we have included in the final exoproteome lists only proteins identified with high confidence, in at least two experimental replicates, reduced significantly the possibilities of false-positive identifications that might account for some of the unexpected proteins; and finally (iv) the lower proportion of proteins primarily regarded as cytoplasmic might be actually a typical characteristic of the *C. pseudotuberculosis* exoproteome.

Non-classically secreted proteins

Intriguingly, a much higher proportion (29.0%) of the exoproteome of the 1002 strain of *C. pseudotuberculosis* was composed by proteins predicted by SurfG+ as not having an extracytoplasmic location, when compared to only 4.5% in the exoproteome of the strain C231 (Figure 2). The possibility of these proteins being non-classically secreted has been evaluated using the SecretomeP algorithm [29]. We have also reviewed the literature for evidence of other bacterial exoproteomes that could support the extracellular localization found for these proteins in our study.

High SecP scores (above 0.5) could be predicted for 5 of the 19 proteins in the exoproteome of the 1002 strain considered by SurfG+ as having a cytoplasmic location (additional files 2 and 3); this could be an indicative that they are actually being secreted by non-classical mechanisms [29]. Nonetheless, 2 of these 5 proteins ([GenBank:ADL09626] and [GenBank:ADL20555]) were also detected in the exoproteome of the C231 strain, in which they were predicted by SurfG+ as possessing an extracytoplasmic location (additional file 2). A comparative analysis of the sequences encoding these proteins in the genomes of the two *C. pseudotuberculosis* strains showed that the disparate results were generated due to the existence of nonsense mutations in the genome sequence of the 1002 strain, which impaired the identification of signal peptides for the two proteins at the time of SurfG+ analysis (data not shown). We believe that it is unlikely that these differences represent true polymorphisms, as the proteins were identified in the extracellular proteome, indicating the real existence of exportation signals. This indeed demonstrates the obvious vulnerability of the prediction tools to the proper annotation of the bacterial genomes. On the other hand, the assignment of high SecP scores to these two proteins, even though they are not believed to be secreted by non-classical mechanisms, would be totally expected, as the SecretomeP is a predictor based on a neural network trained to identify general features of extracellular proteins; this means the prediction tool will attribute SecP scores higher than 0.5 to most of the secreted proteins, regardless the route of export [29].

We have found reports in the literature that strongly support the extracellular localization observed for 8 of the 14 remaining proteins considered as non-secretory by SurfG+ and SecretomeP in the exoproteome of the 1002 strain, and without any detectable signal peptide (additional files 2 and 3, Figure 2). Among these proteins there are the elongation factors Tu and Ts [16,33, 35,37-39]; the glycolytic enzymes triosephosphate isomerase, phosphoglycerate kinase and phosphoglycerate mutase [16-20,37-40]; the chaperonin GroES [16-18, 20,39]; a putative peptidyl prolyl cis trans isomerase [17,18,35,37,41]; and a hydroperoxide reductase enzyme [17,35,39].

Proteins primarily regarded as cytoplasmic have consistently been identified in the exoproteomes of different bacterial species, and moonlighting roles in the extracellular environment have already been demonstrated for some of them [31,32], including evasion of host's immune system [42], adhesion to host cells [43,44], folding of extracytoplasmic proteins [41,45], and interaction between microorganisms [40,46]. Noteworthy, specific evidences for active secretion of such cytoplasmic proteins have been demonstrated for only a few examples to date, and demonstration of an extracellular function is still missing for many of these proteins [30,31].

The variant exoproteome may account for differential virulence of the two *C. pseudotuberculosis* strains

A considerable number (49/93) of the extracellular proteins identified in this work was observed in only one of the two strains studied, then composing a variant experimental *C. pseudotuberculosis* exoproteome (additional files 3 and 4). Highly variant exoproteomes have also been reported recently for other Gram+ bacterial pathogens [20,36,39,47-49], and such a variation may be considered an important factor leading to the observable phenotypic dissimilarities and ultimately to differential virulence of the various strains [50,51]. Hecker *et al.* [36] reported on how the composition of the exoproteome can vary extremely within a single species, *Staphylococcus aureus*, being that only 7 out of 63 identified extracellular proteins were found in all the twenty-five clinical isolates studied.

One of the most intriguing results in the present study was the detection of the phospholipase D (PLD) protein only in the extracellular proteome of the strain C231 (additional file 4). As the regulation of PLD expression was demonstrated to be complex and highly affected by multiple environmental factors [52], we sought to detect this protein in the culture supernatant of the

C. pseudotuberculosis 1002 strain grown in a rich medium (brain-heart infusion broth) instead of only chemically-defined medium (CDM), but these attempts were also unfruitful (data not shown). Besides, we were not able to detect secretion of PLD following total exoproteome analysis of the 1002 strain grown under specific stress generating conditions (Pacheco *et al.*, unpublished). The results strongly indicate that this protein is actually not being secreted by the 1002 strain in culture.

PLD is an exotoxin considered as the major virulence factor of *C. pseudotuberculosis* [5,52]. It possesses sphingomyelinase activity that contributes to endothelial permeability and then to spreading of the bacteria within the host [5]. Mutation of the *pld* gene in *C. pseudotuberculosis* rendered strains no longer capable of causing caseous lymphadenitis (CLA) in sheep and goats; the potential of these strains to be used as live attenuated vaccines was already evaluated [53-55]. Similarly, the strain 1002 of *C. pseudotuberculosis* was already tested as a possible live attenuated vaccine against CLA due to its natural low virulent status, and administration of this bacterium to goats did not cause lesions formation [23,56]. The molecular mechanisms leading to the low virulence of the 1002 strain however remain undetermined so far. We believe that non-secretion of PLD might be one of the main factors responsible for the lowered virulence of the strain. Importantly, we currently cannot affirm that the 1002 strain does not produce this protein while infecting a mammalian host. Besides, this strain still retains the capability of causing localized abscesses and disease in susceptible mice (Pacheco *et al.*, unpublished results).

Other proteins believed to be associated with the virulence of *C. pseudotuberculosis* were also identified exclusively in the exoproteome of the C231 strain, namely FagD and Cp40 (Table 1). The former protein is a component of an iron uptake system, whose coding sequences are clustered immediately downstream of the *pld* gene in the *C. pseudotuberculosis* genome [6]. The latter protein is a secreted serine protease shown to be protective against CLA when used to vaccinate sheep [57].

Strikingly, one variant protein of the *C. pseudotuberculosis* exoproteome, a conserved hypothetical exported protein with a cutinase domain [GenBank:ADL10384], has its coding sequence present in the genome of the C231 strain but absent from the genome of the 1002 strain (additional file 6). The genomic structure of the gene's surroundings is indicative of a region prone to recombination events, such as horizontal gene transfer [58]. In fact, it seems that gene gain and loss are frequent events leading to variations observed in the bacterial exoproteomes [39,59].

Variation of the core exoproteome: differential expression analysis of the common proteins by LC-MS^E

In addition to identifying qualitative variations in the exoproteomes of the two *C. pseudotuberculosis* strains, we were also able to detect relative differences in expression of the proteins common to the two proteomes through label-free protein quantification by the LC-MS^E method. Relative protein quantification by this method can be obtained with basis on the accurate precursor ion mass and electrospray intensity data, acquired during the low energy scan step of the alternating scan mode of MS acquisition [14]. Importantly, this quantitative attribute of the technique opens up new possibilities of utilization, as grows the interest on the so-called physiological proteomics [21].

Thirty-four out of 44 proteins commonly identified in the exoproteomes of the strains 1002 and C231 of *C. pseudotuberculosis* were considered by the PLGS quantification algorithm as having significantly variable expression (score > 250; 95% CI) (Figure 3, additional files 2 and 7). If we further filter these results for the proteins presenting differential expression higher than 2-fold between the strains, we end up with only four proteins up-regulated in the 1002 strain and sixteen in the C231 strain (Figure 3).

Among the group of proteins not presenting considerable variations in expression between the two *C. pseudotuberculosis* strains, proteins probably participating in basic bacterial physiological processes could be easily identified, as would be expected, including cell shape maintenance and cell division (penicillin binding protein, transglycosylases, peptidases, PGRP amidase) [60]; and iron uptake and utilization (HmuT) [61] (Figure 3, additional file 2). In this sense, one might also speculate that the hypothetical proteins identified as non variant in the two strains may have functions associated to the general physiology of *C. pseudotuberculosis*, when grown in minimal medium.

The most up-regulated proteins were observed in the extracellular proteome of the C231 strain, including two cell envelope-associated proteins [62], namely the major secreted (mycoloyltransferase) protein PS1 (10-fold up-regulated), and the S-layer protein A (8-fold up-regulation) (Figure 3). This may be indicative of differences on cell envelope-related activities in the two *C. pseudotuberculosis* strains, such as nutrient acquisition, protein export, adherence and interaction with the host [63]. Dumas *et al.* [49] compared the exoproteomes of *Listeria monocytogenes* strains of different virulence groups, and found that altered expression (up- or down-regulation) of a protein related to the bacterial cell wall could be a marker of specific virulence phenotypes. Additionally, surface associated proteins have been shown to undergo phase and antigenic variation in some bacterial

Table 1 Formerly and newly identified[‡] exported proteins that may be associated with the virulence phenotype of *Corynebacterium pseudotuberculosis* strains

Protein Description[a]	GenBank Accession	Identified in the exoproteome of the strain[b]:		Orhologs found in other Corynebacteria[c]:		References
		1002	C231	Pathogenic	Non-pathogenic	
Phospholipase D (PLD)	ADL09524.1	No	Yes	Yes	No	[54]
Iron siderophore binding protein (FagD)	ADL09528.1	No	Yes	Yes	Yes	[6]
Serine proteinase precursor (CP40)	ADL11339.1	No	Yes	No	No	[57]
Putative iron transport system binding (secreted) protein	ADL10460.1	No	Yes	Yes	No	[12]
Glycerophosphoryl diester phosphodiesterase	ADL11410.1	No	Yes	Yes	No	This work. [72]
Putative surface-anchored membrane protein	ADL20074.1	Yes	Yes	Yes	No	This work.
Putative hydrolase (lysozyme-like)	ADL20788.1	Yes	Yes	Yes	No	This work.
Putative secreted protein	ADL21714.1	Yes	Yes	Yes	No	This work.
Putative sugar-binding secreted protein	ADL09872.1	No	Yes	Yes	No	This work.

[‡] The inclusion criteria followed three main requisites: (i) experimental detection of the proteins in the exoproteomes of the pathogenic *C. diphtheriae* and *C. jeikeium*; (ii) non-detection of the proteins in the exoproteomes of the non-pathogenic *C. glutamicum* and *C. efficiens*; and (iii) *in silico* detection of ortholog proteins in pathogenic, but not in non-pathogenic, corynebacteria through search of similarity against public protein repositories.

[a] This protein list is not meant to be all-inclusive. Rather, it wants to give an overview of the exported proteins identified in this study for which it was possible to speculate on a probable involvement in *C. pseudotuberculosis* virulence after comparative proteomic analyses.

[b] Proteins identified in this study by TPP/LC-MS^E.

[c] Searches of similarity against publicly available protein databases using Blast-p.

pathogens, and ultimately affect the infectivity potential of different strains [50].

Comparative analyses of corynebacterial exoproteomes

Recent studies attempted to characterize the extracellular proteomes of other pathogenic (*C. diphtheriae* and *C. jeikeium*) and non-pathogenic (*C. glutamicum* and *C. efficiens*) corynebacterial species [17,37,64,65]. All these studies used 2D-PAGE to resolve the extracellular proteins of the different corynebacteria, and PMF by MALDI-TOF-MS was the method of choice in most of them for protein identification [17,37,64,65]. Figure 4 shows the numbers of proteins identified in the exoproteomes of all strains studied, in comparison to the numbers obtained in the present study for *C. pseudotuberculosis*. Despite one study with the strain R of *C. glutamicum*, which reports identification of only two secreted proteins [65], all the corynebacterial strains had somehow similar numbers of extracellular proteins identified, ranging from forty-seven in *C. jeikeium* K411 to seventy-four in *C. diphtheriae* C7s(-)$^{tox-}$. Importantly, the fact that we have identified in this study 93 different exoproteins of *C. pseudotuberculosis*, through the analysis of two different strains, means that our dataset represents the most comprehensive exoproteome analysis of a corynebacterial species so far.

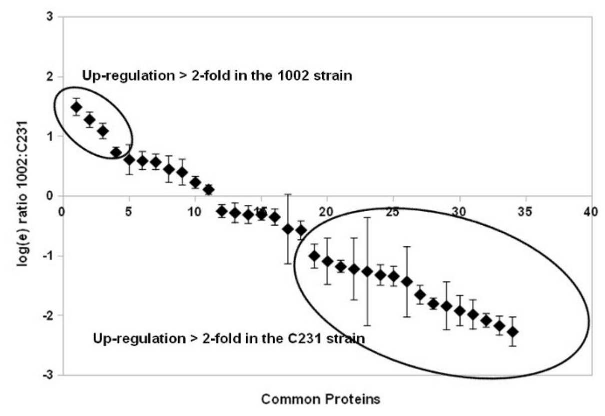


Figure 3 Differential expression of the proteins composing the core *C. pseudotuberculosis* exoproteome, evaluated by label-free relative quantification using LC-MS^E. Results are shown as natural log scale of the relative quantifications (1002:C231) for each protein. Only proteins that were given a variation score higher than 250 by PLGS quantification algorithm are presented. Proteins regulated more than 2-fold in each strain are indicated. Protein identification numbers correspond to additional files 2 and 7: Tables S1 and S4.

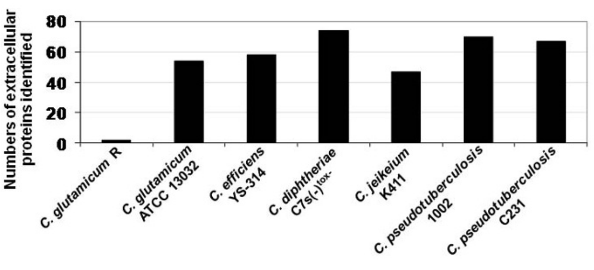


Figure 4 Comparative analysis of corynebacterial exoproteomes. Numbers of extracellular proteins identified in previous corynebacterial exoproteome analyses [17,37,69,70] in comparison to those identified in this study with the two strains of *C. pseudotuberculosis*.

Regardless the different methodologies employed to characterize the exoproteomes of the various corynebacteria, we sought to identify extracellular proteins commonly identified in most of the studies, taking the catalogue of *C. pseudotuberculosis* exoproteins generated in this work as the comparison dataset. Besides corroborating our findings, the objective here was to identify extracellular proteins that could be associated exclusively to pathogenic corynebacterial species.

In total, 34 proteins identified in the exoproteome of the strain 1002 of *C. pseudotuberculosis* were found to be present in the experimentally determined extracellular proteomes of other corynebacteria, whereas the number of common corynebacterial exoproteins in the C231 strain was 32 (Figure 5). Only 6 proteins were consistently identified in all the corynebacterial exoproteomes, including pathogenic and non-pathogenic species: (i) S-layer protein A [62]; (ii) resuscitation-promoting factor RpfB [66]; (iii) cytochrome c oxidase subunit II [67]; (iv) a putative esterase; (v) a NLP/P60 family protein (putative cell wall-associated hydrolase) [68]; and (vi) a trehalose corynomycolyl transferase (Figure 5, additional file 8). Interestingly, three of these six proteins are predicted to be regulated by the same transcription factor [GenBank:ADL09702], a member of the cAMP receptor protein (Crp) family of transcription regulators which are found controlling a diversity of physiological functions in various bacteria [69].

Twelve proteins of the exoproteome of the 1002 strain and fifteen of the C231 strain were also detected experimentally only in the exoproteomes of other pathogenic corynebacteria, namely *C. diphtheriae* and *C. jeikeium* (Figure 5). Altogether, this represents 19 different *C. pseudotuberculosis* proteins (additional file 8). A search of similarity using the sequences of these proteins against publicly available databases, believed to contain the predicted proteomes of all corynebacteria with completely sequenced genomes, showed that 6 of these

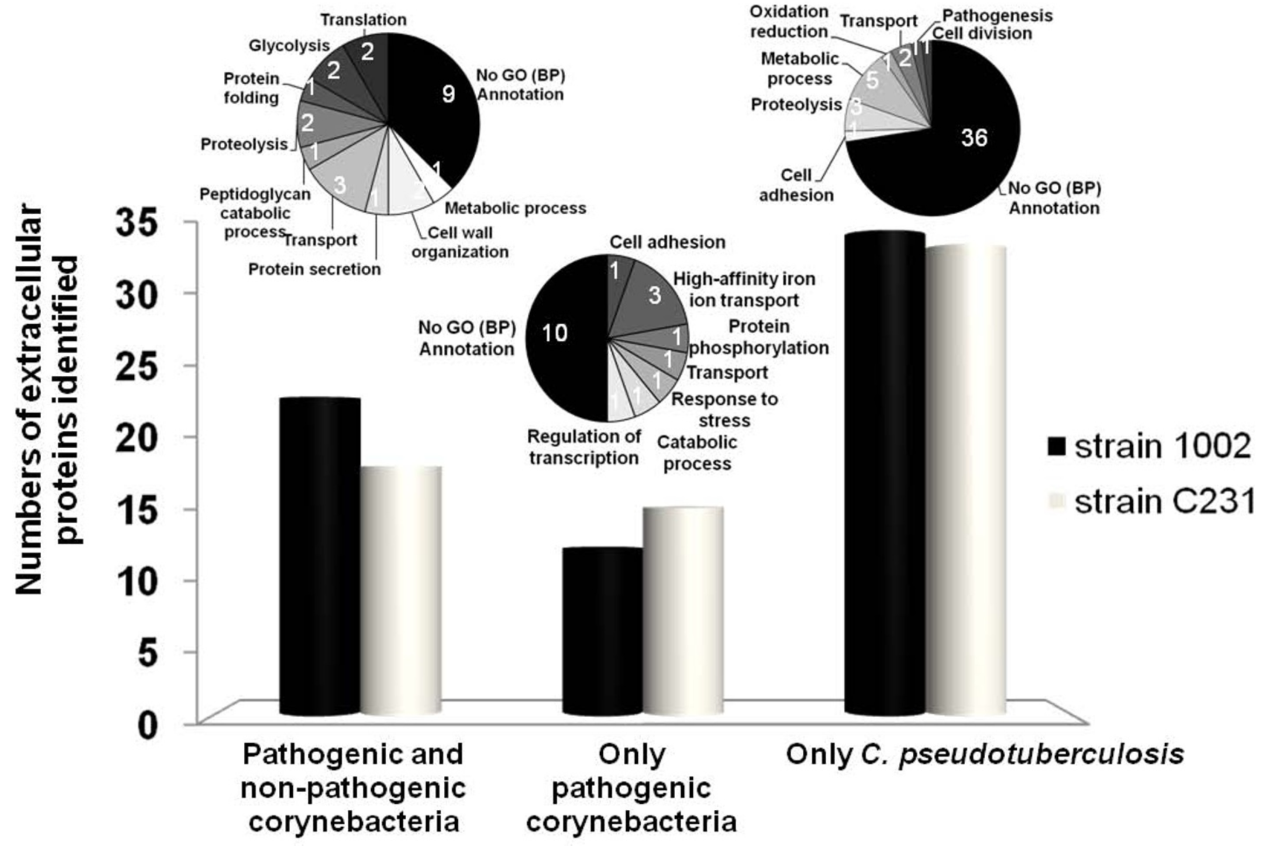


Figure 5 Distribution of orthologous proteins of the *C. pseudotuberculosis* experimental exoproteins throughout other experimentally confirmed corynebacterial exoproteomes. Pathogenic species: *C. diphtheriae* C7s(-)$^{tox-}$ and *C. jeikeium* K411 [17,69]; non-pathogenic species: *C. glutamicum* ATCC13032 and *C. efficiens* YS-314 [37,70]. Pie charts show Gene Ontology (GO) functional annotations for the 93 different *C. pseudotuberculosis* exoproteins identified (24 commonly identified in pathogenic and non-pathogenic corynebacteria; 19 commonly identified only in pathogenic corynebacteria; and 50 only identified in *C. pseudotuberculosis*). Annotations were obtained following analyses with the Blast2GO tool [84], used through the web application available at http://www.blast2go.org/start_blast2go.

19 proteins are apparently absent from non-pathogenic corynebacterial species (Table 1). Moreover, 5 of these proteins are predicted to be part of regulatory networks already shown to be involved in virulence functions, including those regulated by the diphtheria toxin repressor (DtxR)-like protein [70] and the cAMP-binding transcription regulator GlxR [71].

Two proteins presented orthologs highly distributed in various bacterial pathogens: (i) a putative iron transport system binding (secreted) protein [GenBank:ADL10460]; and (ii) a putative glycerophosphoryl diester phosphodiesterase [GenBank:ADL11410]. Interestingly, an ortholog of this latter protein was included recently in a list of seventeen proteins found to be very common in pathogenic bacteria and absent or very uncommon in non-pathogens, representing then probable virulence-associated factors [72]. In fact, reports in the literature can be found that associate orthologs of the two aforementioned proteins with virulence phenotypes [73,74]. Noteworthy, both proteins were detected in this study only in the exoproteome of the C231 strain of *C. pseudotuberculosis*, the more virulent one.

Conclusions

There seems to be a growing interest in profiling the exoproteomes of bacterial pathogens, due to the distinguished roles played by exported proteins on host-pathogen interactions [10]. Classical proteomic profiling strategies, normally involving two-dimensional (2D) gel electrophoresis, have been extensively used for this purpose [16-20]. Nevertheless, the introduction of more high-throughput proteomic technologies brings new perspectives to the study of bacterial exoproteomes, as it makes it easier to analyze multiple phenotypically distinct strains, yielding better subproteome coverage with fewer concerns regarding technical sensitivity and reproducibility [75]. Besides, the currently available methods for label-free quantification of proteins [76] allow us to compare the “dynamic behavior” of the exoproteome across different bacterial strains, and this in turn will help us to better identify alterations of the exoproteome that may contribute to the various virulence phenotypes.

By using a high-throughput proteomic strategy, based on a recently introduced method of LC-MS acquisition (LC-MS^E) [14], we were able to perform a very comprehensive analysis of the exoproteome of an important veterinary pathogen, *Corynebacterium pseudotuberculosis*. Comparative exoproteome analysis of two strains presenting different virulence status allowed us to detect considerable variations of the core *C. pseudotuberculosis* extracellular proteome, and thereby the number of exoproteins identified increased significantly. Most importantly, it was helpful to gain new insights into the probable participation of *C. pseudotuberculosis* exported proteins, other than the well-known PLD and FagB, in the virulence of this bacterium. Several novel targets for future work on *C. pseudotuberculosis* molecular determinants of virulence can be identified from the catalogue of exoproteins generated in this study. Interestingly, around 30% of the proteins identified were predicted by the SurfG+ software [15] as being probably surface exposed in *C. pseudotuberculosis*. Such proteins may represent promising new candidates for composing a CLA vaccine more effective than the ones currently available [4], as has been demonstrated for a series of other bacterial pathogens [33,34]. Therefore, it will be critical to further study the role of this protein set in virulence and vaccine design.

Methods

Bacterial strains and culture conditions

The strains 1002 and C231 of *Corynebacterium pseudotuberculosis* were used in this study. Strain 1002 was isolated from an infected goat in Brazil and has been shown to be naturally low virulent [23,56]; strain C231 was isolated from an infected sheep in Australia, and it showed a more virulent phenotype [24]. Species confirmation was performed by biochemical and molecular methods for both strains, as described [77]. Complete genome sequences of the two strains were generated by Genome Networks in Brazil and Australia (RGMG/RPGP and CSIRO Livestock Industries), and made available for this study (unpublished results).

C. pseudotuberculosis strains were routinely maintained in Brain Heart Infusion broth (BHI: Oxoid, Hampshire, UK) or in BHI 1.5% bacteriological agar plates, at 37°C. For proteomic studies, strains were grown in a chemically defined medium (CDM) previously optimized for *C. pseudotuberculosis* cultivation [78]. The composition of the CDM was as follows: autoclaved 0.067 M phosphate buffer [$Na_2HPO_4 \cdot 7H_2O$ (12.93 g/L), KH_2PO_4 (2.55 g/L), NH_4Cl (1 g/L), $MgSO_4 \cdot 7H_2O$ (0.20 g/L), $CaCl_2$ (0.02 g/L), and 0.05% (v/v) Tween 80]; 4% (v/v) MEM Vitamins Solution 100X (Invitrogen); 1% (v/v) MEM Amino Acids Solution 50X (Invitrogen); 1% (v/v) MEM Non Essential Amino Acids Solution 100X (Invitrogen); and 1.2% (w/v) filter-sterilized glucose.

Three-phase partitioning

Extraction/concentration of the soluble supernatant proteins of *C. pseudotuberculosis* followed the TPP protocol previously optimized by our group [11], with minor modifications. Briefly, overnight cultures (*ca*. 24 hours) of the different *C. pseudotuberculosis* strains were inoculated (1:100) separately into 500 mL of pre-warmed fresh CDM and incubated at 37°C, with agitation at 100 rpm, until reach the mid-exponential growth phase ($OD_{540\ nm}$ = 0.4; LabSystems iEMS Absorbance Plate

Reader). At this point, cultures were centrifuged at room temperature (RT) for 20 min, 4000 rpm, and 400 mL of each supernatant was transferred into new sterile flaks. Following addition of 20 μL Protease Inhibitor Cocktail P8465 (Sigma-Aldrich), supernatants were filtered through 0.22 μm filters; ammonium sulphate was added to the samples at 30% (w/v) and the pH of the mixtures were set to 4.0. Then, *n*-butanol was added to each sample at an equal volume; samples were vigorously vortexed and left to rest for 1 h at RT, until the mixtures separated into three phases. The interfacial precipitate was collected in 1.5 mL microtubes, and re-suspended in 1 mL Tris 20 mM + 10 μL protease inhibitor. Finally, samples were submitted to diafiltration and buffer exchange with NH_4HCO_3 (100 mM), using 5 kDa cut-off spin columns (Millipore).

In-solution tryptic digestion of TPP-extracted proteins

Protein samples were resuspended in 1 mL of 0.1% Rapigest (Waters Corporation, Milford, MA) and concentrated using a 5 kDa cut-off spin column. The solution was heated at 80°C for 15 minutes, reduced with dithiothreitol, alkylated with iodoacetamide and digested with 1:50 (w/w) sequencing grade trypsin for 16 hours. RapiGest was hydrolysed by the addition of 2 μL of 13 M trifluoroacetic acid, filtered using a 0.22 μm spin column and each sample was typically diluted to 1 μg/μL prior to a 1:1 dilution with a 100 fmol/μL glycogen phosphorylase B standard tryptic digest to give a final protein concentration of 500 ng/μL per sample and 50 fmol/μL phosphorylase B.

LC-MS configurations for label-free analysis (LC-MS^E)

Nanoscale LC separations of tryptic peptides for qualitative and quantitative multiplexed LC-MS analysis were performed with a nanoACQUITY system (Waters Corporation) using a Symmetry C_{18} trapping column (180 μm × 20 mm 5 μm) and a BEH C_{18} analytical column (75 μm × 250 mm 1.7 μm). The composition of solvent A was 0.1% formic acid in water, and solvent B (0.1% formic acid in acetonitrile). Each sample (total digested protein 0.5 μg) was applied to the trapping column and flushed with 0.1% solvent B for 2 minutes at a flow rate of 15 μL/min. Sample elution was performed at a flow rate of 250 nL/min by increasing the organic solvent concentration from 3 to 40% B over 90 min. Three technical replicate injections of the TPP-extracted 1002 sample and four technical replicates of the TPP-extracted C231 sample were used for subsequent data analysis in this study. These were from two biological cultures of each *C. pseudotuberculosis* stain.

The precursor ion masses and associated fragment ion spectra of the tryptic peptides were mass measured with a Q-ToF Ultima Global or Synapt HDMS mass spectrometer (Waters Corporation) directly coupled to the chromatographic system. The time-of-flight analyzers of both mass spectrometers were externally calibrated using the MS/MS spectrum from [Glu^1]-Fibrinopeptide B (human - Sigma Aldrich, UK) obtained from the doubly charged peptide ion at *m/z* 785.8426. The monoisotopic mass of the doubly charged species in MS mode was also used for post-acquisition data correction. The latter was delivered at 500 fmol/μL to the mass spectrometer via a NanoLockSpray interface using the auxiliary pump of a nanoACQUITY system at a flow rate of 500 nL/min, sampled every 60 seconds.

Accurate mass data were collected in data independent mode of acquisition by alternating the energy applied to the collision cell/s between a low and elevated energy state (MS^E). The spectral acquisition scan rate was typically 0.9 s with a 0.1 s interscan delay. On the Synapt HDMS instrument in the low energy MS mode, data were collected at constant trap and transfer collision energies (CE) of 3 eV and 1 eV respectively. In elevated energy MS mode, the trap collision energy was ramped from 15 eV to 30 eV with the transfer collision energy at 10 eV. On the Ultima Global instrument a low energy of 6 eV was applied to the collision cell, increasing from 6 eV to 35 eV in elevated MS mode.

Data processing for label-free acquisitions (MS^E)

The LC-MS^E data were processed using ProteinLynx Global Server v2.4 (Waters Corporation, Milford, MA) (see additional file 9). In brief, lockmass-corrected spectra are centroided, deisotoped, and charge-state-reduced to produce a single accurately mass measured monoisotopic mass for each peptide and the associated fragment ion. The initial correlation of a precursor and a potential fragment ion is achieved by means of time alignment. The detection and correlation principles for data independent, alternate scanning LC-MS^E data have been described [14].

Database searches

All data were searched using PLGS v2.4 against a *Corynebacterium pseudotuberculosis* database (NCBI Genome Project ID: 40687 and 40875), released in November 2009, to which the glycogen phosphorylase B and trypsin sequences had been appended. The database was randomised within PLGS generating a new concatenated database consisting of the original sequences plus one additional sequence for each entry with identical composition but randomly scrambled residues. This database contained a total of 4314 entries. A fixed modification of carbamidomethyl-C was specified, and variable modifications included were acetyl N-terminus, deamidation N, deamidation Q and oxidation M. One missed trypsin cleavage site was permitted.

For the MS^E data, the time-based correlation applied in data processing is followed by a further correlation process during the database search that is based on the physicochemical properties of peptides when they undergo collision induced fragmentation. The precursor and fragment ion tolerances were determined automatically. The initial protein identification criteria used by the $Identity^E$ algorithm within PLGS for a single replicate data file, required the detection of at least three fragment ions per peptide, seven fragment ions and a minimum of one peptide per protein.

A process analogous to the Bayesian model described by Nesvizhskii *et al.* [79] was used by PLGS to assign probability values to scores of peptide and protein identifications. Two automated mechanisms determined peptide and protein threshold identification criteria providing a 95% identification confidence interval. A background search is conducted by the search algorithm creating a discriminating decoy identification distribution. The determined peptide cut-off score, typically a log value of 6.25 for the expected 95% identification probability is automatically applied to the results.

Further more stringent filtering was then applied to the database search results from each sample to improve the confidence in the protein observations and quantitative measurements. The results from each of the *individual* replicate analyses from each sample were combined and proteins were removed that were observed in only one of the replicates. Using this additional and rigorous filter the false discovery rate was further reduced to 0.2% for this study, with an average of 16.5 peptides/protein and 37.5% sequence coverage for the TPP-extracted 1002 sample and 15 peptides/protein with 35% sequence coverage for the respective C231 sample. Proteins were observed on average in 2.81 technical replicates in the 1002 sample where 3 replicate analyses were used and 3.52 for the C231 sample in which 4 replicates were included.

Protein quantification using label-free system (MS^E)

Relative quantitative analysis between samples was performed by comparing normalized peak area/intensity of each identified peptide [80]. For relative quantification, automatic normalization was applied to the data set within PLGS using the total peptide complement of each sample. The redundant, proteotypic quantitative measurements generated from the tryptic peptide identifications from each protein were used to determine an average, relative protein fold-change, with a confidence interval and a regulation probability. The confidently identified peptides to protein ratios were automatically weighted based on their identification probability. Binary comparisons were conducted to generate an average normalized intensity ratio for all matched proteins. The entire data set of differentially expressed proteins was further filtered by considering only the identified proteins that replicated in at least two technical replicates with a score > 250 and likelihood of regulation value greater than 0.95 for upregulation and lower than 0.05 for downregulation as determined by the PLGS quantification algorithm.

In silico predictions of protein sub-cellular localization

Prediction of sub-cellular localization was performed initially for the identified proteins by using the SurfG+ program v1.0, run locally in a Linux environment, as described [15] (see additional file 9). For prediction of potentially surface exposed (PSE) proteins, a cut-off value of 73 amino acids was calculated as the minimum distance from the *C. pseudotuberculosis* outermost membrane until the surface of the cell-wall, based on electron microscopy of this bacterium's cell envelope (data not shown).

The programs TatP v1.0 and SecretomeP v2.0 were used through the web applications available at http://www.cbs.dtu.dk/services/, for prediction of twin-arginine pathway-linked signal peptides and non-classical (leaderless) secretion, respectively [29,81].

Comparative analyses of multiple corynebacterial exoproteomes

A list of experimentally observed extracellular proteins of pathogenic (*C. diphtheriae* and *C. jeikeium*) and non-pathogenic (*C. glutamicum* and *C. efficiens*) corynebacteria was identified in previously published studies [17,37,64,65]. The amino acid sequences of these proteins were retrieved from public repositories of protein sequences to create a local database. This database was used in similarity searches with the Blast-p algorithm (E-value < 10^{-4}) [26], taking the group of proteins identified in the *C. pseudotuberculosis* exoproteome as the input sequences. Additionally, transitivity clustering [82] was used to identify proteins (i) commonly detected in the exoproteomes of pathogenic and non-pathogenic corynebacteria, and proteins detected in exoproteomes of (ii) only pathogenic corynebacteria or (iii) only *C. pseudotuberculosis*. A more detailed description on the transitivity clustering analysis can be found in the supplementary material (additional file 9). The amino acid sequences of the identified *C. pseudotuberculosis* exoproteins were also used in similarity searches against public databases, namely NCBI nr and Swissprot.

Transcriptional regulation of the identified exoproteins

The search for transcription factors that regulate expression of the identified corynebacterial exoproteins was performed through the CoryneRegNet database, as described previously [83].

Accession numbers

The sequences of all proteins identified in this work are accessible through GenBank and correspond to the *Corynebacterium pseudotuberculosis* Genome Projects deposited in NCBI (IDs: 40687 and 40875).

Additional material

Additional file 1: Figure S1. Comparison between the experimental (A) and virtual (B) 2-D gels of the exoproteome of the strain 1002 of *C. pseudotuberculosis*. (A) 2D-gel with 150 μg of TPP extracted extracellular proteins of the 1002 strain. Proteins were separated in the first dimension by isoelectric focusing using strips of 3.0-5.6 NL pI range (GE Healthcare). Visualization was by Colloidal Coomassie staining. (B) The virtual 2D-gel was generated with the theoretical pI and MW values of the proteins identified by LC-MS^E.

Additional file 2: Table S1. Proteins composing the core *C. pseudotuberculosis* exoproteome, identified by LC-MS^E.

Additional file 3: Table S2. Variant exoproteome of the strain 1002 of *Corynebacterium pseudotuberculosis*.

Additional file 4: Table S3. Variant exoproteome of the strain C231 of *Corynebacterium pseudotuberculosis*.

Additional file 5: Figure S2. Predictions of LPXTG motif-containing proteins, lipoproteins and Tat-pathway associated signal peptides in the exoproteomes of the strains 1002 and C231 of *C. pseudotuberculosis*.

Additional file 6: Figure S4. A conserved hypothetical exported protein present in the Genome of the strain C231 but absent from the strain 1002 of *C. pseudotuberculosis*. The two sequenced Genomes were aligned using the Artemis Comparison Tool (ACT). The arrows point to tRNA genes.

Additional file 7: Table S4. Relative expression analysis of the extracellular proteins common to the strains 1002 and C231 of *Corynebacterium pseudotuberculosis*.

Additional file 8: Figure S5. Distribution of orthologous proteins of the *C. pseudotuberculosis* experimental exoproteins throughout other experimentally confirmed exoproteomes of pathogenic corynebacteria, as determined through transitivity clustering analysis. The 19 *C. pseudotuberculosis* exoproteins only identified in the exoproteomes of other pathogenic corynebacteria are presented in the table. *Cp = C. pseudotuberculosis*; *Cd = C. diphtheriae*; *Cj = C. jeikeium*.

Additional file 9: Supplementary information on the bioinformatics tools used in this study.

List of abbreviations

CDM: chemically defined medium; CLA: caseous lymphadenitis; LC-MS: liquid chromatography - mass spectrometry; NCS: non-classically secreted; PLD: phospholipase D; PLGS: ProteinLynx Global Server; PMF: peptide mass fingerprinting; PSE: potentially surface exposed; RGMG: Minas Gerais Genome Network; RPGP: Genome and Proteome Network of the State of Pará; TPP: Three-Phase Partitioning.

Acknowledgements

We are thankful to the Minas Gerais Genome Network (RGMG) and to the Genome and Proteome Network of the State of Pará (RPGP). We thank Dr. Robert Moore (CSIRO Livestock Industries) for providing the C231 strain of *C. pseudotuberculosis*.

This work was supported by grants from the Funding Agencies CNPq (grant CNPq/MAPA/SDA) and FAPEMIG, in Brazil; and by The Medical Research Fund and Advantage West Midlands, in the UK.

Author details

[1]Department of Biochemistry and Immunology, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Av. Antônio Carlos, Belo Horizonte, 31.270-901, Brazil. [2]Department of General Biology, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Av. Antônio Carlos, Belo Horizonte, 31.270-901, Brazil. [3]Institute of Health Sciences, Universidade Federal da Bahia, Av. Reitor Miguel Calmon, Salvador, 40.110-902, Brazil. [4]School of Life Sciences, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, United Kingdom. [5]Department of Microbiology, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Av. Antônio Carlos, Belo Horizonte, 31.270-901, Brazil. [6]Genome and Proteome Network of the State of Pará, Universidade Federal do Pará, R. Augusto Corrêa, Belém, 66.075-110, Brazil.

Authors' contributions

LGCP, SES, LMF, MARC, AMCP, RM, AS, JHS, SCO, AM, CGD, and VA conceived the idea, participated in the design of the study, and critically read the manuscript. LGCP, SES, NS, TLPC, WMS, AGV, and SGS performed microbiological and/or proteomic experiments. LGCP, SES and ARS performed bioinformatical analyses. LGCP and SES wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 15 October 2010 Accepted: 17 January 2011
Published: 17 January 2011

References

1. Dorella FA, Pacheco LGC, Oliveira SC, Miyoshi A, Azevedo V: *Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. *Vet Res* 2006, **37**:201-218.
2. Ventura M, Canchaya C, Tauch A, Chandra G, Fitzgerald GF, Chater KF, van Sinderen D: Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum. *Microbiol Mol Biol Rev* 2007, **71**:495-548.
3. Baird GJ, Fontaine MC: *Corynebacterium pseudotuberculosis* and its role in ovine caseous lymphadenitis. *J Comp Pathol* 2007, **137**:179-210.
4. Dorella FA, Pacheco LG, Seyffert N, Portela RW, Meyer R, Miyoshi A, Azevedo V: Antigens of *Corynebacterium pseudotuberculosis* and prospects for vaccine development. *Expert Rev Vaccines* 2009, **8**:205-213.
5. Hodgson AL, Bird P, Nisbet IT: Cloning, nucleotide sequence, and expression in *Escherichia coli* of the phospholipase D gene from *Corynebacterium pseudotuberculosis*. *J Bacteriol* 1990, **172**:1256-1261.
6. Billington SJ, Esmay PA, Songer JG, Jost BH: Identification and role in virulence of putative iron acquisition genes from *Corynebacterium pseudotuberculosis*. *FEMS Microbiol Lett* 2002, **208**:41-45.
7. Desvaux M, Hébraud M, Talon R, Henderson IR: Secretion and subcellular localizations of bacterial proteins: a semantic awareness issue. *Trends Microbiol* 2009, **17**:139-145.
8. Bhavsar AP, Guttman JA, Finlay BB: Manipulation of host-cell pathways by bacterial pathogens. *Nature* 2007, **449**:827-834.
9. Stavrinides J, McCann HC, Guttman DS: Host-pathogen interplay and the evolution of bacterial effectors. *Cell Microbiol* 2008, **10**:285-292.
10. Sibbald MJJB, van Dij JML: Secretome Mapping in Gram-Positive Pathogens. In *Bacterial secreted protein: secretory mechanisms and role in pathogenesis* Edited by: Karl Wooldridge 2009, 193-225.
11. Paule BJA, Meyer R, Moura-Costa LF, Bahia RC, Carminati R, Regis LF, Vale VLC, Freire SM, Nascimento I, Schaer R, Azevedo V: Three-phase partitioning as an efficient method for extraction/concentration of immunoreactive excreted-secreted proteins of *Corynebacterium pseudotuberculosis*. *Protein Expr Purif* 2004, **34**:311-316.
12. Dorella FA, Estevam EM, Pacheco LGC, Guimarães CT, Lana UGP, Gomes EA, Barsante MM, Oliveira SC, Meyer R, Miyoshi A, Azevedo V: In vivo insertional mutagenesis in *Corynebacterium pseudotuberculosis*: an efficient means to identify DNA sequences encoding exported proteins. *Appl Environ Microbiol* 2006, **72**:7368-7372.

13. Silva JC, Gorenstein MV, Li G, Vissers JPC, Geromanos SJ: **Absolute quantification of proteins by LCMSE a virtue of parallel MS acquisition.** *Mol Cell Proteomics* 2006, **5**:144-156.
14. Geromanos SJ, Vissers JPC, Silva JC, Dorschel CA, Li G, Gorenstein MV, Bateman RH, Langridge JI: **The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS.** *Proteomics* 2009, **9**:1683-1695.
15. Barinov A, Loux V, Hammani A, Nicolas P, Langella P, Ehrlich D, Maguin E, van de Guchte M: **Prediction of surface exposed proteins in *Streptococcus pyogenes*, with a potential application to other Gram-positive bacteria.** *Proteomics* 2009, **9**:61-73.
16. Trost M, Wehmhöner D, Kärst U, Dieterich G, Wehland J, Jänsch L: **Comparative proteome analysis of secretory proteins from pathogenic and nonpathogenic *Listeria* species.** *Proteomics* 2005, **5**:1544-1557.
17. Hansmeier N, Chao T, Kalinowski J, Pühler A, Tauch A: **Mapping and comprehensive analysis of the extracellular and cell surface proteome of the human pathogen *Corynebacterium diphtheriae*.** *Proteomics* 2006, **6**:2465-2476.
18. Målen H, Berven FS, Fladmark KE, Wiker HG: **Comprehensive analysis of exported proteins from *Mycobacterium tuberculosis* H37Rv.** *Proteomics* 2007, **7**:1702-1718.
19. Mastronunzio JE, Huang Y, Benson DR: **Diminished exoproteome of *Frankia* spp. in culture and symbiosis.** *Appl Environ Microbiol* 2009, **75**:6721-6728.
20. Dumas E, Desvaux M, Chambon C, Hébraud M: **Insight into the core and variant exoproteomes of *Listeria monocytogenes* species by comparative subproteomic analysis.** *Proteomics* 2009, **9**:3136-3155.
21. Hecker M, Reder A, Fuchs S, Pagels M, Engelmann S: **Physiological proteomics and stress/starvation responses in *Bacillus subtilis* and *Staphylococcus aureus*.** *Res Microbiol* 2009, **160**:245-258.
22. Becher D, Hempel K, Sievers S, Zühlke D, Pané-Farré J, Otto A, Fuchs S, Albrecht D, Bernhardt J, Engelmann S, Völker U, van Dijl JM, Hecker M: **A proteomic view of an important human pathogen–towards the quantification of the entire *Staphylococcus aureus* proteome.** *PLoS One* 2009, **4**:e8176.
23. Ribeiro OC, Silva JAH, Oliveira SC, Meyer R, Fernandes GB: **Preliminary results on a living vaccince against caseous lymphadenitis.** *Pesquisa Agropecuaria Brasileira* 1991, **26**:461-465.
24. Simmons CP, Dunstan SJ, Tachedjian M, Krywult J, Hodgson AL, Strugnell RA: **Vaccine potential of attenuated mutants of *Corynebacterium pseudotuberculosis* in sheep.** *Infect Immun* 1998, **66**:474-479.
25. Patel VJ, Thalassinos K, Slade SE, Connolly JB, Crombie A, Murrell JC, Scrivens JH: **A comparison of labeling and label-free mass spectrometry-based proteomics approaches.** *J Proteome Res* 2009, **8**:3752-3759.
26. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403-410.
27. Khamis A, Raoult D, La Scola B: **rpoB gene sequencing for identification of *Corynebacterium* species.** *J Clin Microbiol* 2004, **42**:3925-3931.
28. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.
29. Bendtsen JD, Kiemer L, Fausbøll A, Brunak S: **Non-classical protein secretion in bacteria.** *BMC Microbiol* 2005, **5**:58.
30. Vanet A, Labigne A: **Evidence for specific secretion rather than autolysis in the release of some *Helicobacter pylori* proteins.** *Infect Immun* 1998, **66**:1023-1027.
31. Bendtsen JD, Wooldridge KG: **Non-Classical Secretion.** In *Bacterial secreted proteins: secretory mechanisms and role in pathogenesis* Edited by: Karl Wooldridge 2009, 225-239.
32. Jeffery CJ: **Moonlighting proteins–an update.** *Mol Biosyst* 2009, **5**:345-350.
33. Rodríguez-Ortega MJ, Norais N, Bensi G, Liberatori S, Capo S, Mora M, Scarselli M, Doro F, Ferrari G, Garaguso I, Maggi T, Neumann A, Covre A, Telford JL, Grandi G: **Characterization and identification of vaccine candidate proteins through analysis of the group A *Streptococcus* surface proteome.** *Nat Biotechnol* 2006, **24**:191-197.
34. Doro F, Liberatori S, Rodríguez-Ortega MJ, Rinaudo CD, Rosini R, Mora M, Scarselli M, Altindis E, D'Aurizio R, Stella M, Margarit I, Maione D, Telford JL, Norais N, Grandi G: **Surfome analysis as a fast track to vaccine discovery: identification of a novel protective antigen for Group B *Streptococcus* hypervirulent strain COH1.** *Mol Cell Proteomics* 2009, **8**:1728-1737.
35. Barbey C, Budin-Verneuil A, Cauchard S, Hartke A, Laugier C, Pichereau V, Petry S: **Proteomic analysis and immunogenicity of secreted proteins from *Rhodococcus equi* ATCC 33701.** *Vet Microbiol* 2009, **135**:334-345.
36. Hecker M, Becher D, Fuchs S, Engelmann S: **A proteomic view of cell physiology and virulence of *Staphylococcus aureus*.** *Int J Med Microbiol* 2010, **300**:76-87.
37. Hansmeier N, Chao T, Pühler A, Tauch A, Kalinowski J: **The cytosolic, cell surface and extracellular proteomes of the biotechnologically important soil bacterium *Corynebacterium efficiens* YS-314 in comparison to those of *Corynebacterium glutamicum* ATCC 13032.** *Proteomics* 2006, **6**:233-250.
38. Schaumburg J, Diekmann O, Hagendorff P, Bergmann S, Rohde M, Hammerschmidt S, Jänsch L, Wehland J, Kärst U: **The cell wall subproteome of *Listeria monocytogenes*.** *Proteomics* 2004, **4**:2991-3006.
39. Sibbald MJJB, Ziebandt AK, Engelmann S, Hecker M, de Jong A, Harmsen HJM, Raangs GC, Stokroos I, Arends JP, Dubois JYF, van Dijl JM: **Mapping the pathways to staphylococcal pathogenesis by comparative secretomics.** *Microbiol Mol Biol Rev* 2006, **70**:755-788.
40. Furuya H, Ikeda R: **Interaction of triosephosphate isomerase from the cell surface of Staphylococcus aureus and alpha-(1->3)-mannooligosaccharides derived from glucuronoxylomannan of *Cryptococcus neoformans*.** *Microbiology* 2009, **155**:2707-2713.
41. Söderberg MA, Cianciotto NP: **A *Legionella pneumophila* peptidyl-prolyl cis-trans isomerase present in culture supernatants is necessary for optimal growth at low temperatures.** *Appl Environ Microbiol* 2008, **74**:1634-1638.
42. Kunert A, Losse J, Gruszin C, Hühn M, Kaendler K, Mikkat S, Volke D, Hoffmann R, Jokiranta TS, Seeberger H, Moellmann U, Hellwage J, Zipfel PF: **Immune evasion of the human pathogen *Pseudomonas aeruginosa*: elongation factor Tuf is a factor H and plasminogen binding protein.** *J Immunol* 2007, **179**:2979-2988.
43. Tsugawa H, Ito H, Ohshima M, Okawa Y: **Cell adherence-promoted activity of *Plesiomonas shigelloides* groEL.** *J Med Microbiol* 2007, **56**:23-29.
44. Feng Y, Pan X, Sun W, Wang C, Zhang H, Li X, Ma Y, Shao Z, Ge J, Zheng F, Gao GF, Tang J: ***Streptococcus suis* enolase functions as a protective antigen displayed on the bacterial cell surface.** *J Infect Dis* 2009, **200**:1583-1592.
45. Pissavin C, Hugouvieux-Cotte-Pattat N: **Characterization of a periplasmic peptidyl-prolyl cis-trans isomerase in *Erwinia chrysanthemi*.** *FEMS Microbiol Lett* 1997, **157**:59-65.
46. Bergonzelli GE, Granato D, Pridmore RD, Marvin-Guy LF, Donnicola D, Corthésy-Theulaz IE: **GroEL of *Lactobacillus johnsonii* La1 (NCC 533) is cell surface associated: potential role in interactions with the host and the gastric pathogen *Helicobacter pylori*.** *Infect Immun* 2006, **74**:425-434.
47. He X, Zhuang Y, Zhang X, Li G: **Comparative proteome analysis of culture supernatant proteins of *Mycobacterium tuberculosis* H37Rv and H37Ra.** *Microbes Infect* 2003, **5**:851-856.
48. Sumby P, Whitney AR, Graviss EA, DeLeo FR, Musser JM: **Genome-wide analysis of group a streptococci reveals a mutation that modulates global phenotype and disease specificity.** *PLoS Pathog* 2006, **2**:e5.
49. Dumas E, Meunier B, Berdagué J, Chambon C, Desvaux M, Hébraud M: **Comparative analysis of extracellular and intracellular proteomes of *Listeria monocytogenes* strains reveals a correlation between protein expression and serovar.** *Appl Environ Microbiol* 2008, **74**:7399-7409.
50. van der Woude MW, Bäumler AJ: **Phase and antigenic variation in bacteria.** *Clin Microbiol Rev* 2004, **17**:581-611, table of contents.
51. Behr MA, Sherman DR: **Mycobacterial virulence and specialized secretion: same story, different ending.** *Nat Med* 2007, **13**:286-287.
52. McKean SC, Davies JK, Moore RJ: **Expression of phospholipase D the major virulence factor of *Corynebacterium pseudotuberculosis*, is regulated by multiple environmental factors and plays a role in macrophage death.** *Microbiology* 2007, **153**:2203-2211.
53. Hodgson AL, Krywult J, Corner LA, Rothel JS, Radford AJ: **Rational attenuation of *Corynebacterium pseudotuberculosis*: potential cheesy gland vaccine and live delivery vehicle.** *Infect Immun* 1992, **60**:2900-2905.
54. McNamara PJ, Bradley GA, Songer JG: **Targeted mutagenesis of the phospholipase D gene results in decreased virulence of *Corynebacterium pseudotuberculosis*.** *Mol Microbiol* 1994, **12**:921-930.
55. Moore RJ, Rothel L, Krywult J, Radford AJ, Lund K, Hodgson AL: **Foreign gene expression in *Corynebacterium pseudotuberculosis*: development of a live vaccine vector.** *Vaccine* 1999, **18**:487-497.

56. Meyer R, Carminati R, Bahia R, Vale V, Viegas S, Martinez T, Nascimento I, Schaer R, Silva J, Ribeiro M, Regis L, Paule B, Freire S: **Evaluation of the goats humoral immune response induced by the *Corynebacterium pseudotuberculosis* lyophilized live vaccine.** *J Med Biol Sci* 2002, **1**:42-48.
57. Walker J, Jackson HJ, Eggleton DG, Meeusen EN, Wilson MJ, Brandon MR: **Identification of a novel antigen from *Corynebacterium pseudotuberculosis* that protects sheep against caseous lymphadenitis.** *Infect Immun* 1994, **62**:2562-2567.
58. Koonin EV, Makarova KS, Aravind L: **Horizontal gene transfer in prokaryotes: quantification and classification.** *Annu Rev Microbiol* 2001, **55**:709-742.
59. Nogueira T, Rankin DJ, Touchon M, Taddei F, Brown SP, Rocha EPC: **Horizontal gene transfer of the secretome drives the evolution of bacterial cooperation and virulence.** *Curr Biol* 2009, **19**:1683-1691.
60. Hett EC, Rubin EJ: **Bacterial growth and cell division: a mycobacterial perspective.** *Microbiol Mol Biol Rev* 2008, **72**:126-56, table of contents.
61. Allen CE, Schmitt MP: **HtaA is an iron-regulated hemin binding protein involved in the utilization of heme iron in *Corynebacterium diphtheriae*.** *J Bacteriol* 2009, **191**:2638-2648.
62. Puech V, Chami M, Lemassu A, Lanéelle MA, Schiffler B, Gounon P, Bayan N, Benz R, Daffé M: **Structure of the cell envelope of corynebacteria: importance of the non-covalently bound lipids in the formation of the cell wall permeability barrier and fracture plane.** *Microbiology* 2001, **147**:1365-1382.
63. Jordan S, Hutchings MI, Mascher T: **Cell envelope stress response in Gram-positive bacteria.** *FEMS Microbiol Rev* 2008, **32**:107-146.
64. Hansmeier N, Chao T, Daschkey S, Müsken M, Kalinowski J, Pühler A, Tauch A: **A comprehensive proteome map of the lipid-requiring nosocomial pathogen *Corynebacterium jeikeium* K411.** *Proteomics* 2007, **7**:1076-1096.
65. Suzuki N, Watanabe K, Okibe N, Tsuchida Y, Inui M, Yukawa H: **Identification of new secreted proteins and secretion of heterologous amylase by *C. glutamicum*.** *Appl Microbiol Biotechnol* 2009, **82**:491-500.
66. Hartmann M, Barsch A, Niehaus K, Pühler A, Tauch A, Kalinowski J: **The glycosylated cell surface protein Rpf2, containing a resuscitation-promoting factor motif, is involved in intercellular communication of *Corynebacterium glutamicum*.** *Arch Microbiol* 2004, **182**:299-312.
67. Sakamoto J, Shibata T, Mine T, Miyahara R, Torigoe T, Noguchi S, Matsushita K, Sone N: **Cytochrome c oxidase contains an extra charged amino acid cluster in a new type of respiratory chain in the amino-acid-producing Gram-positive bacterium *Corynebacterium glutamicum*.** *Microbiology* 2001, **147**:2865-2871.
68. Tsuge Y, Ogino H, Teramoto H, Inui M, Yukawa H: **Deletion of cgR_1596 and cgR_2070, encoding NlpC/P60 proteins, causes a defect in cell separation in *Corynebacterium glutamicum* R.** *J Bacteriol* 2008, **190**:8204-8214.
69. Körner H, Sofia HJ, Zumft WG: **Phylogeny of the bacterial superfamily of Crp-Fnr transcription regulators: exploiting the metabolic spectrum by controlling alternative gene programs.** *FEMS Microbiol Rev* 2003, **27**:559-592.
70. Oram D, Avdalovic A, Holmes R: **Analysis of genes that encode DtxR-like transcriptional regulators in pathogenic and saprophytic corynebacterial species.** *Infect Immun* 2004, **72**:1885-1895.
71. Kohl T, Baumbach J, Jungwirth B, Puhler A, Tauch A: **The GlxR regulon of the amino acid producer *Corynebacterium glutamicum*: in silico and in vitro detection of DNA binding sites of a global transcription regulator.** *J Biotechnol* 2008, **135**:340-350.
72. Stubben CJ, Duffield ML, Cooper IA, Ford DC, Gans JD, Karlyshev AV, Lingard B, Oyston PCF, de Rochefort A, Song J, Wren BW, Titball RW, Wolinsky M: **Steps toward broad-spectrum therapeutics: discovering virulence-associated genes present in diverse human pathogens.** *BMC Genomics* 2009, **10**:501.
73. Janson H, Melhus A, Hermansson A, Forsgren A: **Protein D the glycerophosphodiester phosphodiesterase from *Haemophilus influenzae* with affinity for human immunoglobulin D influences virulence in a rat otitis model.** *Infect Immun* 1994, **62**:4848-4854.
74. Braun V: **Iron uptake mechanisms and their regulation in pathogenic bacteria.** *Int J Med Microbiol* 2001, **291**:67-79.
75. Roe MR, Griffin TJ: **Gel-free mass spectrometry-based high throughput proteomics: tools for studying biological response of proteins and proteomes.** *Proteomics* 2006, **6**:4678-4687.
76. Panchaud A, Affolter M, Moreillon P, Kussmann M: **Experimental and computational approaches to quantitative proteomics: status quo and outlook.** *J Proteomics* 2008, **71**:19-33.
77. Pacheco LGC, Pena RR, Castro TLP, Dorella FA, Bahia RC, Carminati R, Frota MNL, Oliveira SC, Meyer R, Alves FSF, Miyoshi A, Azevedo V: **Multiplex PCR assay for identification of *Corynebacterium pseudotuberculosis* from pure cultures and for rapid detection of this pathogen in clinical samples.** *J Med Microbiol* 2007, **56**:480-486.
78. Moura-Costa LF, Paule BJA, Azevedo V, Freire SM, Nascimento I, Schaer R, Regis LF, Vale VLC, Matos DP, Bahia RC, Carminati R, Meyer R: **Chemically defined synthetic medium for *Corynebacterium pseudotuberculosis* culture.** *Rev. Bras. Saúde e Produção Animal* 2002, **3**:1-9.
79. Nesvizhskii AI, Keller A, Kolker E, Aebersold R: **A statistical model for identifying proteins by tandem mass spectrometry.** *Anal Chem* 2003, **75**:4646-4658.
80. Silva JC, Denny R, Dorschel CA, Gorenstein M, Kass IJ, Li G, McKenna T, Nold MJ, Richardson K, Young P, Geromanos S: **Quantitative proteomic analysis by accurate mass retention time pairs.** *Anal Chem* 2005, **77**:2187-2200.
81. Bendtsen JD, Nielsen H, Widdick D, Palmer T, Brunak S: **Prediction of twin-arginine signal peptides.** *BMC Bioinformatics* 2005, **6**:167.
82. Wittkop T, Emig D, Lange S, Rahmann S, Albrecht M, Morris JH, Böcker S, Stoye J, Baumbach J: **Partitioning biological data with transitivity clustering.** *Nat Methods* 2010, **7**:419-420.
83. Baumbach J, Wittkop T, Kleindt CK, Tauch A: **Integrated analysis and reconstruction of microbial transcriptional gene regulatory networks using CoryneRegNet.** *Nat Protoc* 2009, **4**:992-1005.
84. Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A: **High-throughput functional annotation and data mining with the Blast2GO suite.** *Nucleic Acids Res* 2008, **36**:3420-3435.

doi:10.1186/1471-2180-11-12
Cite this article as: Pacheco *et al.*: A combined approach for comparative exoproteome analysis of *Corynebacterium pseudotuberculosis*. *BMC Microbiology* 2011 **11**:12.

II.III.4 Putative virulence factors of *Corynebacterium pseudotuberculosis* FRC41: vaccine potential and protein expression.

Santana-Jorge KT, Santos TM, Tartaglia NR, Aguiar EL, Souza RF, Mariutti RB, Eberle RJ, Arni RK, Portela RW, Meyer R, **Azevedo V**.

Microb Cell Fact. 2016 May 16;15(1):83. doi: 10.1186/s12934-016-0479-6.

Diferentes abordagens são estudadas visando o controle de linfadenite caseosa através da identificação de novas drogas ou desenvolvimento de testes diagnóstico e vacina. No entanto, ainda não foi alcançada eficácia satisfatória para o controle desta doença. O sequenciamento de genomas de *Corynebacterium pseudotuberculosis* tem possibilitado a identificação de novos fatores de virulência deste microrganismo o que contribui para o melhor entendimento de sua biologia e das relações patógeno-hospedeiro. Neste trabalho, foram selecionadas como alvo de estudo os supostos fatores de virulência: SpaC, SodC, NanH, e PknG de *C. pseudotuberculosis* FRC41. Foram realizadas as caracterizações in silico destas proteínas a partir de predições de parâmetros físico-químicos, peptídeo sinal, domínios conservados, avaliação da sua conservação em eucariotos e epítopos. SpaC, PknG e NanH apresentaram melhor potencial vacinal em relação a SodC, após analise in silico. Uma análise em cluster foi realizada para avaliar o grau de redundância entre as sequencias dos epitopos preditos. 57 cluster foram encontrados e a maioria deles (34) foram clusters únicos. Dois cluster de PknG e um de SpaC agruparam epitopos para célula B e T (MHC I and II). Estes epitopos podem potencialmente estimular uma resposta imune tanto humoral quanto celular contra *C. pseudotuberculosis.* Os quatro possíveis alvos foram expressos com sucesso em *Escherichia coli* e foi desenvolvido um protocolo de purificação para PknG. Assim, este estudo demonstrou possíveis novos alvos que podem ser utilizados nas pesquisas de patogenicidade e desenvolvimento de vacina.

Microbial Cell Factories

RESEARCH

Open Access

Putative virulence factors of *Corynebacterium pseudotuberculosis* FRC41: vaccine potential and protein expression

Karina T. O. Santana-Jorge[1†], Túlio M. Santos[1,2†], Natayme R. Tartaglia[1], Edgar L. Aguiar[1], Renata F. S. Souza[1], Ricardo B. Mariutti[3], Raphael J. Eberle[3], Raghuvir K. Arni[3], Ricardo W. Portela[4], Roberto Meyer[4] and Vasco Azevedo[1*]

Abstract

Background: *Corynebacterium pseudotuberculosis*, a facultative intracellular bacterial pathogen, is the etiological agent of caseous lymphadenitis (CLA), an infectious disease that affects sheep and goats and it is responsible for significant economic losses. The disease is characterized mainly by bacteria-induced caseous necrosis in lymphatic glands. New vaccines are needed for reliable control and management of CLA. Thus, the putative virulence factors SpaC, SodC, NanH, and PknG from *C. pseudotuberculosis* FRC41 may represent new target proteins for vaccine development and pathogenicity studies.

Results: SpaC, PknG and NanH presented better vaccine potential than SodC after in silico analyses. A total of 136 B and T cell epitopes were predicted from the four putative virulence factors. A cluster analysis was performed to evaluate the redundancy degree among the sequences of the predicted epitopes; 57 clusters were formed, most of them (34) were single clusters. Two clusters from PknG and one from SpaC grouped epitopes for B and T-cell (MHC I and II). These epitopes can thus potentially stimulate a complete immune response (humoral and cellular) against *C. pseudotuberculosis*. Several other clusters, including two from NanH, grouped B-cell epitopes with either MHC I or II epitopes. The four target proteins were expressed in *Escherichia coli*. A purification protocol was developed for PknG expression.

Conclusions: In silico analyses show that the putative virulence factors SpaC, PknG and NanH present good potential for CLA vaccine development. Target proteins were successfully expressed in *E. coli*. A protocol for PknG purification is described.

Keywords: *Corynebacterium pseudotuberculosis*, Pathogenicity and virulence, Vaccine potential, Epitope prediction, Protein expression, Protein purification

Background

Caseous lymphadenitis (CLA) is a chronic, pyogenic, contagious disease of sheep and goat that imposes considerable economic losses for farmers in many countries [1, 2]. The disease is caused by *Corynebacterium pseudotuberculosis* (*C. pseudotuberculosis*): a gram-positive pleomorphic, non-capsulated, non-motile, fimbriated, facultative intracellular bacterium, multiplying within macrophages [1]. *Corynebacterium ulcerans and C. pseudotubercu1osis* produce phospholipase D (PLD), which is unique among corynebacteria. It promotes the hydrolysis of ester bonds in sphingomyelin in mammalian cell membranes, possibly contributing to the spread of the bacteria from the initial site of infection to the secondary sites within the host. Moreover, it provokes dermonecrotic lesions; and at higher doses it is lethal to a number of different species of laboratory and domestic animals [3–5].

*Correspondence: vasco@icb.ufmg.br
†Karina T. O. Santana-Jorge and Túlio M. Santos share first-author credit
[1] Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Avenida Antonio Carlos, 6627, Pampulha, Belo Horizonte 31270-901, Brazil
Full list of author information is available at the end of the article

CLA disease is expressed in external and visceral forms, either separately or together [3–5]. External CLA lesions appear initially as abscesses that convert later on to pyogranulomas ranging in size from millimeters to centimeters. These external lesions are mostly located within superficial lymph nodes, but infrequently in subcutaneous tissues. Wool or hair over CLA lesions may be lost due to the weak dermonecrotic action of *C. pseudotuberculosis* exotoxins and the pressure atrophy of overlying skin by the lesions. Visceral lesions are not detectable clinically but express themselves according to their number, site and effect on the involved organ. Progressive weight loss, respiratory disorders and chronic recurrent ruminal tympany are the most prominent signs that may accompany visceral CLA lesions.

Identification/removal of infected animals is a key factor for success of disease control measures. Vaccination of healthy animals is another strategy broadly recommended for disease control. In fact, control of CLA depends on vaccination in most countries [2, 5–7]. Although bacterin, toxoid, combined, and live vaccines are available, the disease has persisted even after prolonged vaccination, indicating the suppressive nature of CLA vaccination [5, 7]. *C. pseudotuberculosis* infection of farmer animals can contaminate meat and milk, putting consumers at risk due to its zoonotic potential [7]. The ability of *C. pseudotuberculosis* to infect both animals and humans makes necessary the development of new vaccines for a reliable control and management of CLA once the currently available commercial vaccines are unable to fully protect susceptible animals against the disease [7, 8]. In this way, the study of other *C. pseudotuberculosis* virulence factors that might be involved in CLA pathogenesis can provide new vaccine targets.

The complete genome sequence of a *C. pseudotuberculois* strain (FRC41) isolated from a 12-year-old girl with necrotizing lymphadenitis allowed the identification of *spaC* and *nanH* as genes encoding proteins regarded as potential virulence factors [8]. SpaC is a putative adhesive pili tip protein. The pilus structure can probably make the initial contact with host cell receptors to enable additional ligand-receptor interactions and to facilitate the efficient delivery of virulence factors and intracellular invasion [9]. NanH, by its turn, is a putative extracellular neuraminidase [8]. Neuraminidases, or sialidases, belong to a class of glycosyl hydrolases that catalyze the removal of terminal sialic acid residues from a variety of glycoconjugates and can contribute to the recognition of sialic acids exposed on host cell surfaces. Most sialidase-producing microorganisms are pathogenic or commensal when in close contact with mammalian hosts. It has been also suggested that, in some types of pathogenic bacteria, sialidases function as potential virulence factors that contribute to the recognition of sialic acids exposed on the surface of the host cell [10]. A homologous counterpart of *C. pseudotuberculois* FRC41 NanH was characterized in *C. diphtheriae* KCTC3075 and shown to be a protein containing neuraminidase and trans-sialidase activities [11].

The *C. pseudotuberculosis* FRC41 genome also encodes a putative secreted copper,zinc-dependent superoxide dismutase (SodC) that is characterized by a lipobox motif and may be anchored in the cell membrane [8]. The extracellular location of this enzyme suggests that it may protect the surface of *C. pseudotuberculosis* cells against superoxide generated externally by the mammalian host cells. In *Mycobacterium tuberculosis*, SodC contributes to the resistance of this microorganism against the oxidative burst products generated by activated macrophages [12, 13]. The protective activity of Cu,Zn-SODs has been associated with virulence in other bacteria, such as *Neisseria meningitides* and *Hemophylus ducreyi* [8].

As part of important cell signaling mechanisms, eukaryotic-like serine/threonine protein kinases encountered in bacteria are a class of molecules that also deserves attention since they are part of complex signaling pathways and play a diversity of physiological roles in developmental processes, secondary metabolism, cell division, cell wall synthesis, essential processes, central metabolism, and virulence [14, 15]. *Mycobacterium tuberculosis* genome encodes 11 eukaryotic-like serine/threonine protein kinases (PknA to PknL, except for PknC). Protein kinase G (PknG) gained particular interest because it affects the intracellular traffic of *M. tuberculosis* in macrophages. Most microbes and nonpathogenic mycobacteria quickly find themselves in lysosomes, where they are killed. By contrast, *M. tuberculosis* stays within phagosomes; the bacterium releases PknG to block phagosome-lysosome fusion. Bacteria lacking *pknG* gene are rapidly transferred to lysosomes and eliminated [16, 17]. The genome of *C. pseudotuberculosis* FRC41 has a gene encoding for a putative PknG protein [8] but its function in the bacterium still needs to be investigated.

Therefore, *C. pseudotuberculosis* SpaC, NanH, SodC, and PknG proteins may play important roles in virulence and pathogenicity. In the present work, a characterization and evaluation of the vaccine potential of these proteins were performed in silico. The heterologous expression of these putative virulence factors in *Escherichia coli* is also described.

Methods

Protein sequences

The amino acid sequences of the target proteins were retrieved from NCBI GenBank: SpaC [gb| ADK29663.1], SodC [gb| ADK28404.1], NanH [gb| ADK28179.1], PknG [gb| ADK29622.1].

Homology searches

NCBI BLASTP [18] searches in UniProtKB database [19] were performed to identify homologues of the target proteins in the CMNR group of microorganisms (from *Corynebacterium*, *Mycobacterium*, *Nocardi*, and *Rhodococcus* genera): *Corynebacteriumn, taxid:1716; Mycobacterium, taxid:1763; Nocardia, taxid:1817; Rhodococcus, taxid:1827.* Likewise, BLASTP searches in UniprotKB database were performed to identify homologues of the target proteins in mammalian species of the *Ovis* (taxid: 9935), *Bos* (taxid: 9903), *Equus* (taxid: 9789), *Equus* (taxid: 35510), Mus (taxid: 10088), *Mus* (taxid: 862507) genera and in *Homo sapiens* (taxid: 9606). BLAST Genome [18] searches in *C. pseudotuberculosis* (taxid: 1719) complete genomes available at NCBI genome database were performed to identify the presence of the target protein genes in other *C. pseudotuberculosis* strains.

Primary and secondary structure analysis, subcellular localization and prediction of protective antigens

ProtParam [20] and Self-OPtimized prediction method with alignment—SOPMA [21] of expasy server were used to analyze different physiological and physicochemical properties of the target proteins. Molecular weight, theoretical pI, amino acid composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY) were calculated using the ProtParam preset parameters. Solvent accessibility, transmembrane helices, globular regions, bend region, random coil and coiled-coil regions were predicted using SOPMA default parameters. The amino acid sequences were evaluated by PSORTb 3.0.2 [22] to predict subcellular localization of the target proteins. SignalP 4.1 [23] was used to predict the presence and location of signal peptide cleavage sites in the amino acid sequences. The method incorporates a prediction of cleavage sites and a signal peptide/non-signal peptide prediction based on a combination of several artificial neural networks. VaxiJen 2.0 [24] was used for alignment-independent prediction of protective antigens. The tool was developed to allow antigen classification solely based on the physicochemical properties of proteins without the need of sequence alignment.

B-cell epitope prediction

Linear B-cell epitopes were predicted from the target protein sequences using physicochemical properties [25] estimated by in silico methods available in DNASTAR Protean program (Madison, Wisconsin). The Jameson–Wolf method [26] was used to predict the potential antigenic determinants by combining existing methods for protein structural predictions. The results appear as multiple peaks in the antigenic index plot, with each peak signifying a potential antigenic determinant. The emini surface probability method [27] was used to predict the probability that a given region lies on the surface of a protein. The Kyte–Doolittle hydropathy method [28] predicts regional hydropathy of proteins from their amino acid sequences. Hydropathy values are assigned for all amino acids and are then averaged over a user defined window. The average is plotted at the midpoint of the window. The charge density method predicts regions of positive and negative charge by summing charge over a specific range of residues. DNASTAR developed this method using the pK tables of White et al. [29]. Since charged residues tend to lie on the surfaces of proteins, this method aids in predicting surface characteristics. Several wet lab experiments revealed that the antigenic portions were situated in beta turn regions of a protein [30] for these regions the Chou and Fasman beta turn prediction method was used [31, 32]. The Karplus–Schulz flexibility method [33] predicts backbone chain flexibility. The method is useful for resolving antigenic sites, as these regions tend to be among the most flexible in a polypeptide sequence. Conserved domains in the target proteins were identified by searching NCBI's conserved domain database (CDD) [34]. The results of each method were presented in a graphical frame. The peak of the amino acid residue segment above the threshold value (we used the default) is considered as predicted B-cell epitope. User can select any physicochemical property or a combination of two or more properties for epitope prediction. [35]. We selected amino acid segments in the target protein sequences where peaks above threshold overlapped in four or more methods. B-cell epitopes located in signal peptide or conserved domains were discarded.

T-cell epitope prediction

MHC I binding prediction was performed using the immune epitope database (IEDB) MHC I binding tool [36] and consensus [37] as prediction method which combines predictions from ANN aka NetMHC (3.4), SMM and comblib methods. Mouse MHC alleles (H-2-Db, H-2-Dd, H-2-Kb, H-2-Kd, H-2-Kk, H-2-Ld) and a peptide length of nine mer were selected to make the predictions from target proteins sequences. A median percentile rank of the four predictions methods was the Consensus representative percentile rank used to select the top 1 % of peptides. A small numbered percentile rank indicates high affinity.

MHC II binding predictions for target proteins were performed using NetMHCII 2.2 server [38] to predict binding of 15 mer peptides to two mouse MHC II alleles (H-2-IAb and H-2-IAd) using artificial neuron networks. The prediction values were given in nM IC50 values, and as a %-Rank to a set of 1,000,000 random natural peptides. Strong and weak binding (SB, WB) peptides were indicated in the output. T-cell epitopes located in signal peptide or conserved domains were discarded.

Epitope clustering

Epitope clustering was performed using the IEDB Epitope cluster analysis tool [36]. Clustal omega [39] was used to group predicted B and T-cell epitopes into clusters of similarity based on multiple sequence alignment and visual inspection. Clustal omega alignments were used to double check if single-sequence clusters generated by IEDB epitope cluster analysis tool were in fact composed of unique epitopes (no pairs).

Cloning procedures

Miniprep plasmid purifications, agarose gel electrophoresis, and *E. coli* media were as described [40]. Amino acids 2–23 and amino acids 2–31 were removed from *sodC* and *nanH* ORF sequences, respectively. These regions containing signal peptide were eliminated before cloning in order to improve protein expression since they are relatively rich in hydrophobic amino acids. ORF codons of all four target proteins were replaced by *E. coli* preferential codons [41]. Optimized ORF sequences were synthesized and individually cloned into pD444-NH expression vector (T5 promoter, IPTG inducible, strong ribosome binding site, His-tag, ampicillin resistance marker, high copy origin of replication, 4027 bp size) by DNA2.0 (Menlo Park, CA). Each ORF-containing plasmid (pD444-NH;*pknG*, pD444-NH;*spaC*, pD444-NH;*sodC*, and pD444-NH;*nanH*) was transformed into BL21(DE3) *E. coli* strains according to the OverExpress™ Electrocompetent Cells kit (Lucigen, Middleton) instructions.

Protein expression in *E. coli*

Protein expression protocol was according to OverExpress™ Electrocompetent Cells kit (Lucigen, Middleton) instructions. Briefly, transformed cell cultures at OD 0.5–0.7 were induced with 1 mM IPTG for 5 h at 37 °C. SDS-PAGE of non-induced and induced cell culture samples and Coomassie blue staining was as described [42].

Purification of PknG

Bacteria transformed with pD444-NH;*pknG* was induced as described above. Cell pellet was collected by 8000 rpm centrifugation, resuspended in buffer A (10 mM NaH_2PO_4 pH7.4, 300 mM NaCl, 1 % glycerol, 5 mM imidazole), lysed on ice with ten 15-s sonication pulses using a ultrasonic processor Marconi-MA 103 (Piracicaba, São Paulo) and centrifuged at 15,000×*g* for 15 min. The supernatant containing recombinant proteins was purified under native conditions using 1 mL of immobilized Ni Sepharose (GE Healthcare). The resin was washed using buffer A with 80 mM imidazole. Recombinant PknG was eluted from the column with buffer A containing 400 mM imidazole. The eluted protein was dialyzed against buffer B (10 mM NaH_2PO_4 buffer pH 7.4 and 50 mM NaCl) and concentrated by ultrafiltration. The concentrated fraction was injected on a Superdex 75 10/300 GL (GE Healthcare) size exclusion column previously equilibrated with buffer B. The purity of the sample was assessed by SDS–PAGE.

Results and discussion

Traditional vaccination approaches are based on complete pathogen either live attenuated or inactivated. Among the major problems these vaccines brought are crucial safety concerns, because those pathogens being used for immunization may become activated and cause infection. Moreover due to genetic variation of pathogen strains around the world, vaccines are likely to lose their efficacy in different regions or for a specific population. Novel vaccine approaches like DNA vaccines and epitope based vaccines have the potential to overcome these barriers to create more effective, specific, strong, safe and long lasting immune response without all undesired effects [43]. Next-generation sequencing and proteomic techniques have enabled researchers to mine entire microbial genomes, transcriptomes and proteomes to identify novel candidate immunogens [44]. In silico techniques are the best alternative to find out which regions of a protein out of thousands possible candidates are most likely to evoke immune response [35]. This reverse vaccinology approach has enjoyed considerable success in the past decade, beginning with *Neisseria meningitides*, and continuing with *Streptococcus pneumonia*, pathogenic *E. coli*, and antibiotic resistant *Staphylococcus aureus* [44].

Homology searches

The conservation level between target proteins and proteins of the CMNR group of microorganisms was evaluated by NCBI BLASTP [18] searches in UniprotKB database [19]. This kind of analysis is important for the development of vaccines once they can be used not only for *C. pseudotuberculosis* FRC41 but for other pathogen strains and pathogens of other species. NCBI BLAST Genome searches show the presence of the target protein genes in all 37 *C. pseudotuberculosis* strains currently available in NCBI complete genomes database (data not shown). This indicates that SpaC, SodC, NanH and PknG can potentially be expressed not only in a few strains demonstrating the importance of these proteins for this pathogenic bacterium. Well conserved homologous of the target proteins were also found in microorganisms of the CMNR group (Additional files 1, 2 and 3). These findings are a good indication that a vaccine against *C. pseudotuberculosis* made from the putative virulence factors can be effective not only against numerous strains of the pathogen but also against bacterial pathogens from other species.

The conservation degree among target proteins and mammalian (*Ovis*, *Bos*, *Equus* and *Mus* genera, *Homo sapiens*) proteins was also evaluated by BLASTP searches. The analysis was important to reveal the conservation degree among pathogen proteins and host proteins and so the possibility of undesirable immunological cross-reactions which may induce autoimmunity. The results (Additional files 1, 2 and 3) show that *C. pseudotuberculosis* FRC41 SpaC, SodC, NanH, and PknG sequences share low identity (30 % in average) with mammalian sequences. BLASTP alignments show that most of this weak homology is in conserved domains (data not shown). Thus, regions away from signal peptides and conserved domains are ideal targets for vaccine development.

Primary and secondary structure analysis

The next step was to evaluate the primary and secondary structure features of SpaC, SodC, NanH and PknG as they can predict stability and reveal functional characteristics of the proteins at some extent. Based on ProtParam instability index, SodC was considered the least stable while PknG was the most stable (Table 1). PknG was also the most hydrophilic with the highest GRAVY (−0.211). This same protein also presented the highest aliphatic (92.91) index (Table 1). SOPMA program, used to calculate secondary structure features of the target proteins, reported that SpaC, SodC and NanH were dominated by random coils, consisting in 45.35, 41.26 and 39.05 %, respectively (Table 2). Alpha helix prevailed (44.06 %) in PknG. The differences in secondary structure content and aliphatic character helps to explain the stability indexes estimated for the target proteins. [45].

Subcellular localization and prediction of protective antigens

The candidate molecules from a eukaryotic pathogen expected to induce immunity comprise proteins that are as follows: (i) present on the surface of the pathogen, (ii) excreted/secreted from the pathogen and (iii)

Table 2 Secondary structure content in the target proteins estimated using SOPMA

Secondary structure	SpaC (%)	SodC (%)	NanH (%)	PknG (%)
Alpha helix (Hh)	111 is 13.94	56 is 27.18	228 is 32.85	330 is 44.06
3_{10} helix (Gg)	0.00	0.00	0.00	0.00
Pi helix(Ii)	0.00	0.00	0.00	0.00
Beta bridge (Bb)	0.00	0.00	0.00	0.00
Extended strand (Ee)	244 is 30.65	39 is 18.93	128 is 18.44	114 is 15.22
Beta turn (Tt)	80 is 10.05	26 is 12.62	67 is 9.65	56 is 7.48
Bend region (Ss)	0.00	0.00	0.00	0.00
Random coil (Cc)	361 is 45.35	85 is 41.26	271 is 39.05	249 is 33.24
Ambiguous states (?)	0.00	0.00	0.00	0.00
Other states	0.00	0.00	0.00	0.00

Table 1 Physicochemical properties of the target proteins estimated using ProtParam

Physicochemical property	SpaC	SodC	NanH	PknG
Number of amino acids	796	206	694	749
Molecular weight	85,964.9	21,099.3	74,683.3	83,349.4
Theoretical pI	5.13	5.96	5.05	5.13
Total number of negatively charged residues (Asp + Glu)	96	23	102	101
Total number of positively charged residues (Arg + Lys)	77	17	79	76
Extinction coefficient[a]	93,085	4595[b]	77,600	81,375
Abs 0.1 % (=1 g/l), assuming all pairs of Cys residues form cystines	1.083	0.218[b]	1.039	0.976
Extinction coefficient[a]	92,710	4470[b]	77,350	81,250
Abs 0.1 % (=1 g/l), assuming all Cys residues are reduced	1.078	0.212[b]	1.036	0.975
The estimated half-life				
Mammalian reticulocytes, in vitro	30 h	30 h	30 h	30 h
Yeast, in vivo	>20 h	>20 h	>20 h	>20 h
Escherichia coli, in vivo	>10 h	>10 h	>10 h	>10 h
Instability index (II)	28.21 (stable)	19.62 (stable)	32.92 (stable)	38.18 (stable)
Aliphatic index	80.16	71.65	72.58	92.91
Grand average of hydropathicity (GRAVY)	−0.442	−0.245	−0.485	−0.211

[a] Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm measured in water

[b] This protein does not contain any Trp residues. Experience shows that this could result in more than 10 % error in the computed extinction coefficient

homologous to known proteins involved in pathogenesis and virulence [46]. Signal peptide presence and subcellular localization (Table 3) of SpaC (cell wall), SodC (cytoplasmic membrane) and NanH (extracellular) was as predicted before [8]. They were predicted as protective antigens by VaxiJen. Membrane and secreted proteins are considered potential vaccine targets once they are at the host-pathogen interface. These proteins may interact more directly with host molecules for cell adhesion, invasion, multiplication, immune response evasion, damage generation to the host, and survive to host cell defenses [8, 47, 48].

Table 3 Subcellular localization, signal peptide, and prediction of protective antigen for the target proteins

Parameter (program)	SpaC	SodC	NanH	PknG
Subcellular localization (Psortb)	Cell wall (matched LPXTG; score 9.97)	Cytoplasmic Membrane (matched 61246116: superoxide dismutase Cu–Zn precursor; score 9.68)	Extracellular (matched 585539: sialidase precursor EC 3.2.1.18 NEURAMINIDASE; score 9.70)	Cytoplasmic, (matched 54041713: probable serine/threonine-protein kinase pknG; score 9.89)
Signal peptide (signalp 4.1)[a]	No (D = 0.162 D-cutoff = 0.420)	Yes position: 1–35 (cleavage site between pos. 35 and 36: DSA-DK D = 0.631 D-cutoff = 0.450 networks = signalp-TM)	Yes position: 1–31 (cleavage site between pos. 31 and 32: APA-TL D = 0.562 D-cutoff = 0.450 networks = signalp-TM)	No (D = 0.106 D-cutoff = 0.420)
Prediction of protective antigens (VaxiJen)	Probable ANTIGEN (score 0.6912)	Probable ANTIGEN (score 0.7663)	Probable ANTIGEN (score 0.6967)	Probable NON-ANTIGEN score 0.3686)

[a] For signal peptide prediction, D-cutoff values were set as sensitive (reproduce SignalP 3.0's sensitivity)

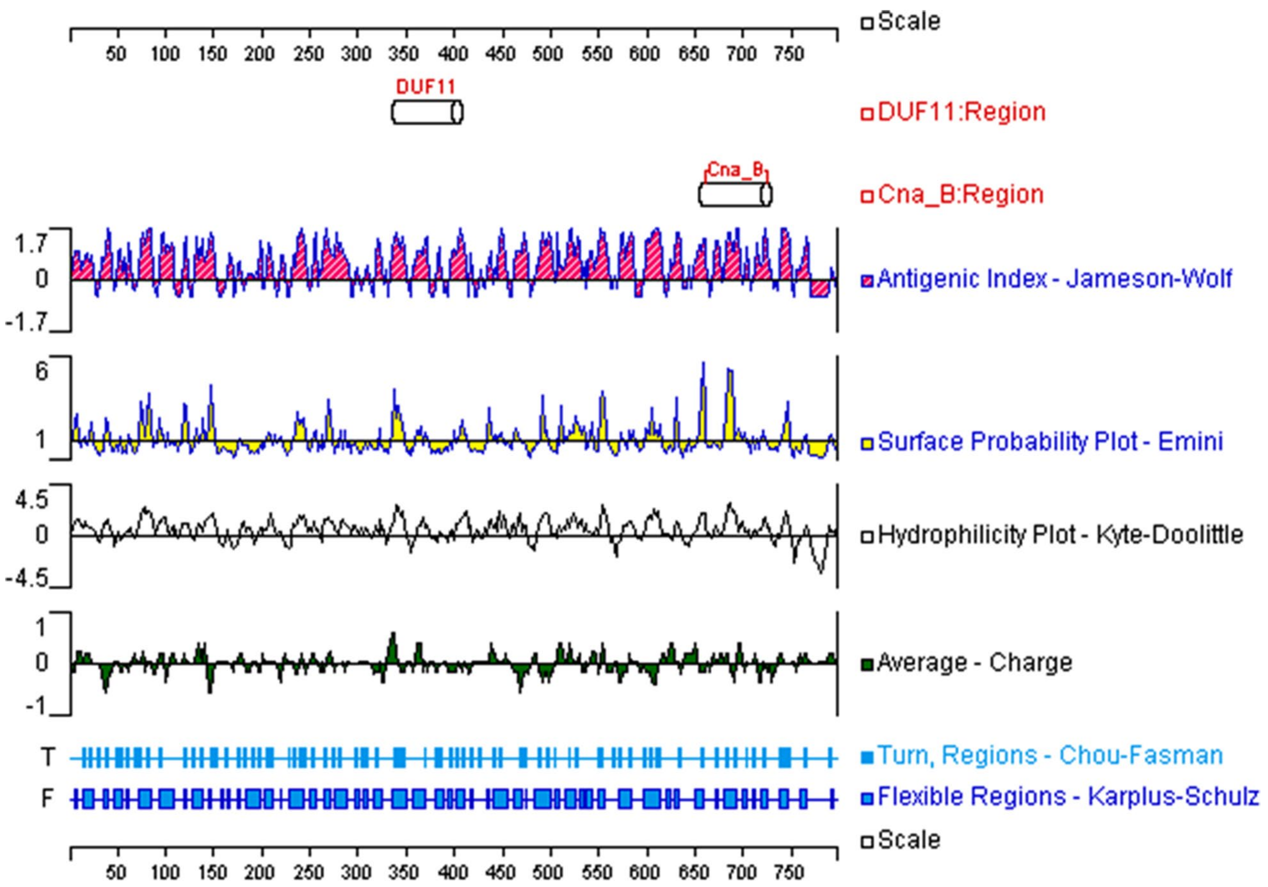


Fig. 1 Graphical outputs of the different methods used to quantitate the physicochemical properties used to predict B-cell epitopes from SpaC. On *top* are the conserved domains of the target protein identified by searching NCBI's Conserved Domain Database (CDD). The *scales* indicate the amino acid positions

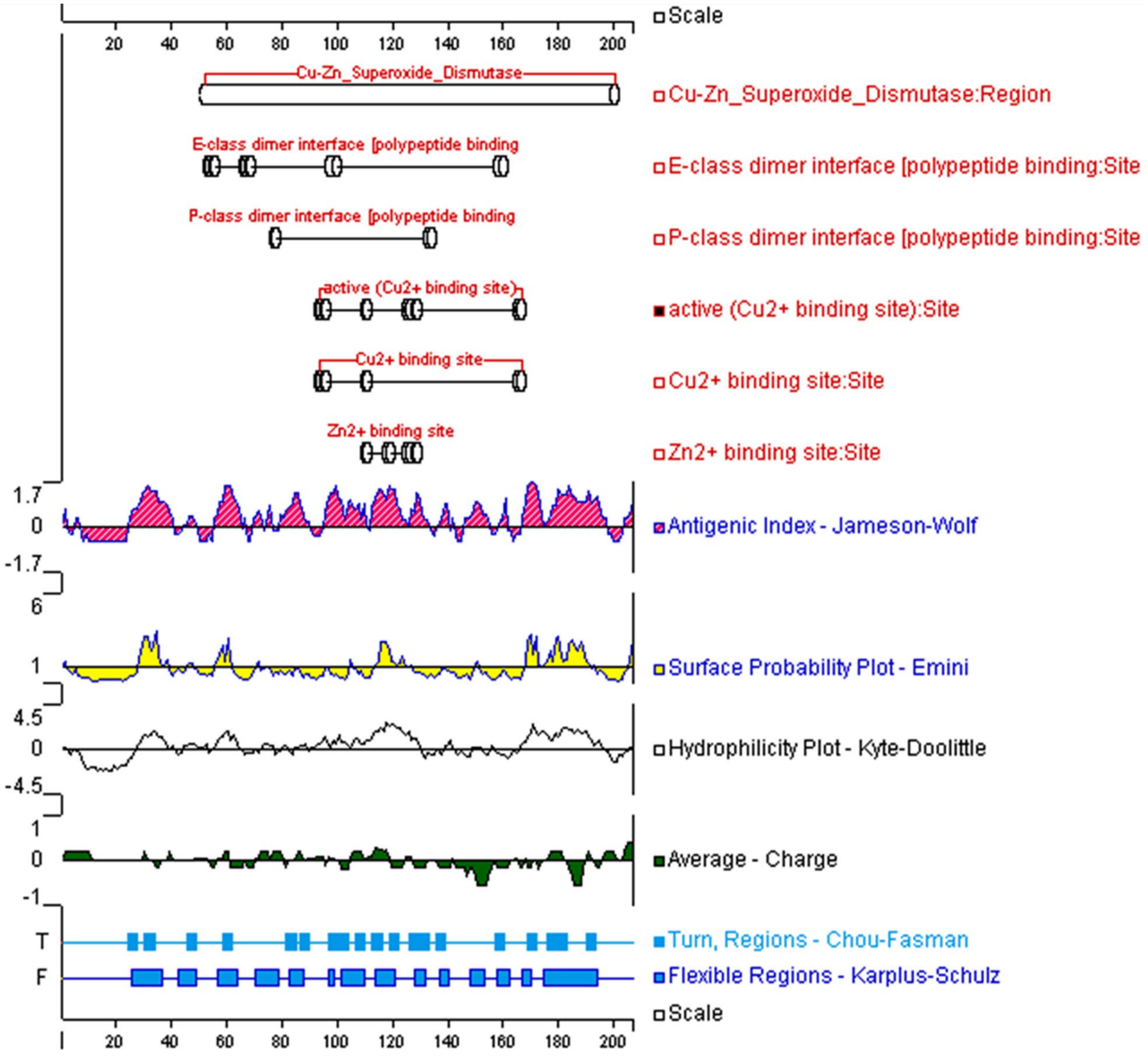


Fig. 2 Graphical outputs of the different methods used to quantitate the physicochemical properties used to predict B-cell epitopes from SodC. On *top* are the conserved domains of the target protein identified by searching NCBI's Conserved Domain Database (CDD). The *scales* indicate the amino acid positions

Like its counterpart in *M. tuberculosis*, which is predominantly found soluble in the cytoplasm [15], PknG was predicted as a cytoplasmic protein (Table 3). However, VaxiJen predicted this *C. pseudotuberculosis* putative serine/threonine protein kinase as non-antigenic. In fact, cytoplasmic proteins have not been widely considered as potential immunogens, since they do not have a close contact to many immune systems' intermediates [49]. Regardless of this, it has been demonstrated that cytoplasmic proteins can be effectively exposed to MHC presentation and may have a key role in the development of a suitable protective immunity. In order to overcome the problem of endogenous antigen access to the MHC II compartment, lysosomal-associated membrane proteins (LAMPs), major lysosomal membrane glycoproteins that contain a cytoplasmic tail targeting sequence that directs the trafficking of the molecule through an endosome/lysossome pathway, including cellular compartments where it is co-localized with MHC II molecules, have been used to induce antigen-trafficking to MHC II compartments and increase the immune response to those antigens [50]. This strategy has shown to elicit enhanced long-term memory response against HIV-1 Gag protein. Besides, a novel mechanism of specific CD8^{+} T cell-mediated

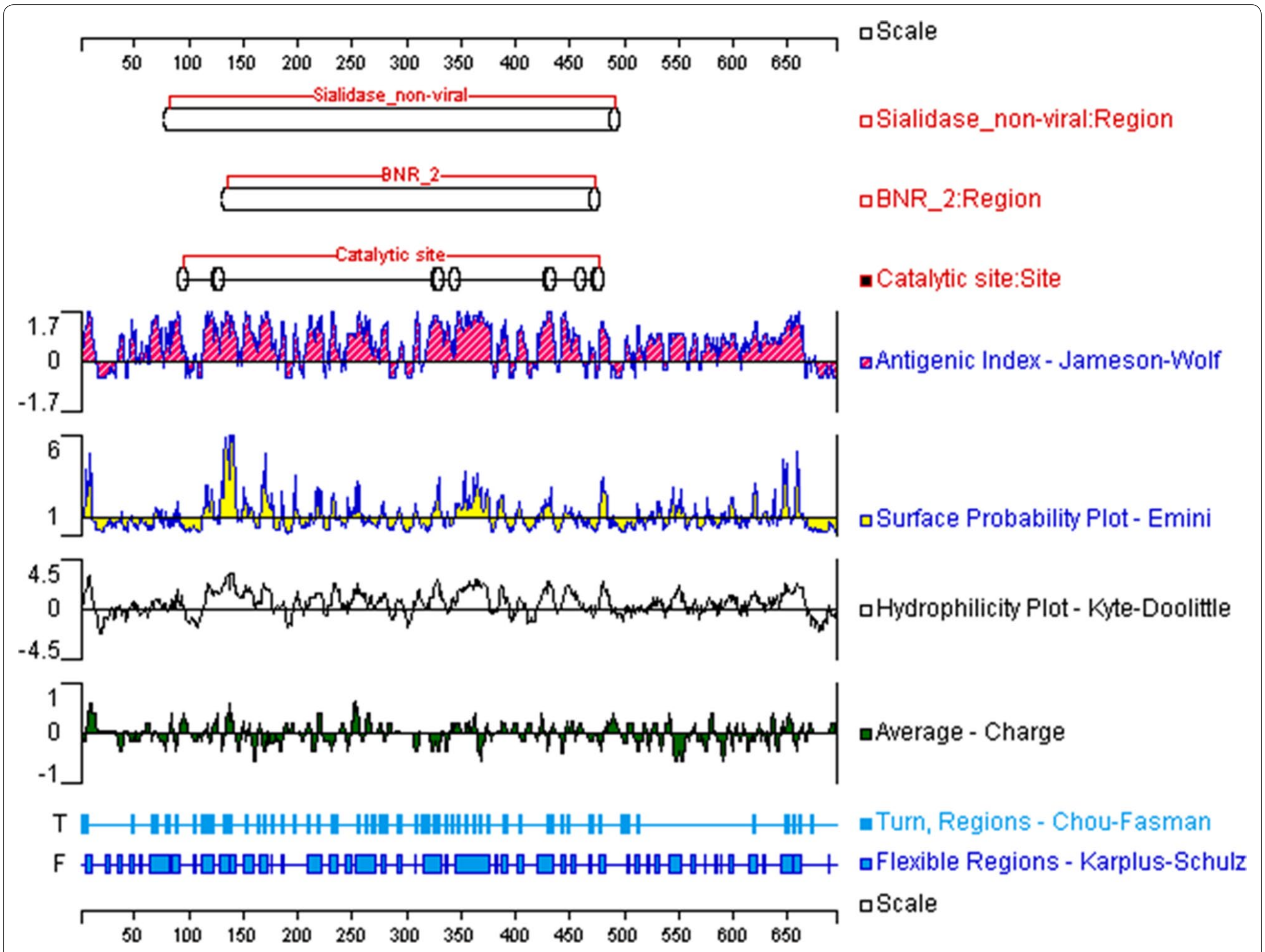


Fig. 3 Graphical outputs of the different methods used to quantitate the physicochemical properties used to predict B-cell epitopes from NanH. On *top* are the conserved domains of the target protein identified by searching NCBI's Conserved Domain Database (CDD). The *scales* indicate the amino acid positions

protective immunity can recognize malaria proteins expressed in the cytoplasm of parasites, form clusters around infected hepatocytes, and protect against parasites [51]. This strongly indicates that cellular and molecular mechanisms underlying the protective immune responses against intracellular parasites need further studies.

Linear B-cell epitope prediction

The general problem in achieving an effective treatment of *C. pseudotuberculosis* infections in animals and humans is probably related to the facultative intracellular lifestyle of this bacterium, as it can survive and multiply in macrophages [52]. The knowledge on the immunity induced by *C. pseudotuberculosis* indicates that the resistance to infection is a complex process involving components of the non-specific and specific host responses, in which humoral and cellular immune responses are both operative [7].

B-cell epitopes can induce both primary and secondary immunity. Although it is believed that the majority of B-cell epitopes are conformational epitopes, experimental determination of epitopes has focused primarily on the identification of linear (non conformational) B-cell epitopes [25]. This is mainly because predictions of conformational epitopes depend on experimentally determined protein structures or homologous protein structures for in silico modeling. So far, there is no protein structure of the target proteins or structures of highly homologous proteins available for modeling.

Most of the existing linear B-cell epitope prediction methods are based on physicochemical properties relating to surface exposure, such as flexibility or hidrophilicity [25, 35], as it is thought that epitopes must lie at the

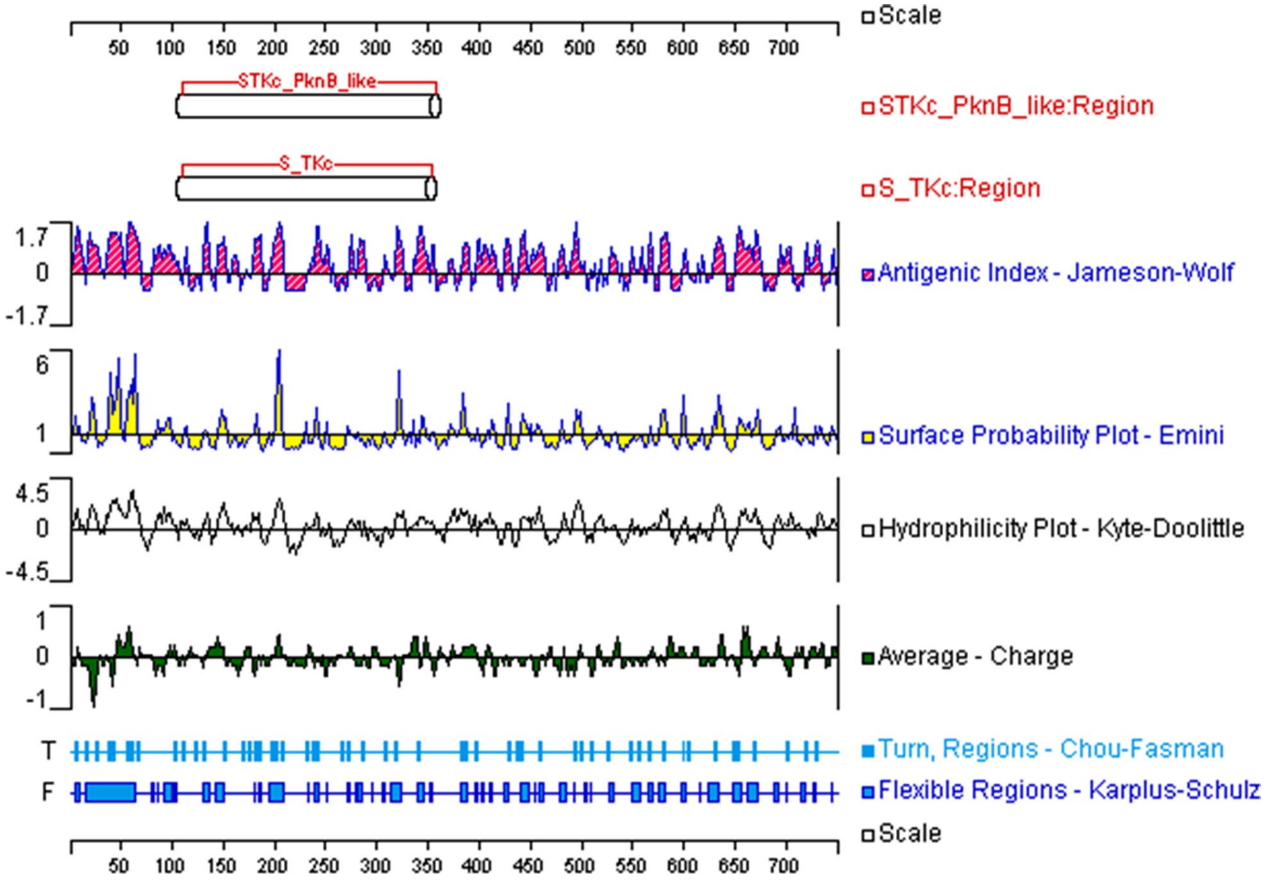


Fig. 4 Graphical outputs of the different methods used to quantitate the physicochemical properties used to predict B-cell epitopes from PknG. On *top* are the conserved domains of the target protein identified by searching NCBI's Conserved Domain Database (CDD). The *scales* indicate the amino acid positions

protein surface for antibody binding to occur. Thus, the target proteins were scanned for B-cell epitopes using several methods designed to quantitate protein physicochemical properties. Graphical outputs of the prediction methods are shown in Figs. 1, 2, 3, and 4. High values gave rise to peaks, whereas valleys correspond to negative properties of the protein. Selected B-cell linear epitopes of target proteins are shown in Table 4. The putative adhesive pili tip protein SpaC, seconded by PknG, presented the highest number of B-cell epitopes. We did pick only one B-cell epitope from SodC since the protein is short (206 aa), has a 35 aa long signal peptide (Table 3) and its highly conserved domain occupies most of the amino acid sequence (Fig. 2).

T-cell epitope prediction

A desirable vaccine preparation should present MHC I and II epitopes for the development of a protective and long lasting immune response to *C. pseudotuberculosis*. MHC I epitopes are presented to CD8+ T cells by cells infected with *C. pseudotuberculosis*, leading to the apoptosis of the host cell and interruption of the bacterial multiplication, and it was already described the injection of anti-CD4 or anti-CD8 monoclonal antibody resulted in significantly increased mortality and a marked suppression of IFN-gamma production in mice [53]. MHC II epitopes are involved in the activation of CD4+ T cells, which will drive the host immune response to a Th1 protective response, as well as to a production of IFN-gamma, that will help macrophages in the fusion of phagosomes and lysosomes, resulting in the destruction of bacteria that underwent phagocytic process [54]. Ultimately, specific high affinity binding should be the main concern since the efficiency of an epitope vaccine greatly relies on the precise interaction between epitope and HLA molecule [55]. Table 5 shows nine mer peptides from target proteins with high affinity (Consensus percentile rank <1 %) for mouse MHC I alleles. Most of them were from SpaC and PknG. SodC peptides were discarded since they were located in conserved regions. The

Table 4 B-cell epitopes predicted from target proteins

Target Protein	Epitope number	B-cell epitopes[a]
SpaC	1	1-MEVPEKTKVEIRFQTGSKISTPSTPSV-27
SpaC	2	70-SQHTNRGETFNDRNSTDLYVQ-90
SpaC	3	116-AYNPKEGYIYAISQGRLKTLQSSKLRIYDEDPNYPAGHLL-155
SpaC	4	234-NDYTSTGKTDSNYVWGI-250
SpaC	5	251-KNSSNPAVLERIDVRDGSRKEFSLDGVKDPLGQNVEKGIYGT-292
SpaC	6	331-IVAKRKGPTSQNNDATSNG-349
SpaC	7	434-KATYKVTANQSISNNEKCLQNTASIYAN-461
SpaC	8	504-GNGLRKVTYKIEVKNPKGFPETKYSLTDTPQFADSV-539
SpaC	9	540-KLERLKVISDYGKKNQEVQAADISV-564
SpaC	10	615-FGLFNSAKLKVGVSEKTSEGCAPIVR-640
SpaC	11	647-QLKKVDAENKETELQATFE-665
SpaC	12	735-PLSKSADQGKDPNLVIL-751
SpaC	13	756-VRVGTLPKTGGHGVAIYLV-774
SodC	1	26-SSSTTTKDSADKAMTS-41
NanH	1	1-MTDSHRRGTRKALVTLTA-18
NanH	2	65-GEGKLPDPVTSEFF-78
NanH	3	520-IEDAKAATAKAEEATAN-536
NanH	4	559-AEAKSAAQDAI-569
NanH	5	595-KAENEAKALAE-605
NanH	6	617-SQDQAKALAEA-627
NanH	7	645-EKEKSGKAGGTDNTENKGFWQE-666
PknG	1	1- MNDPLSRGTEAIPFDPFADDEEDDLSGLLND-31
PknG	1.1	38-DTDTDARSREKSISTFRSRRGTNRDDRTVANG-69
PknG	1.2	79-STAEEMLKDDAYIEQKGLEKPLLHPGD-105
PknG	2	381-SPQRSTFGTKHMVFRTDQLIDGIERNVRITSEEVNA-416
PknG	3	438-YAEPSQTLQTLRDAMAQEEFANSKEIPL-465
PknG	4	479-EARSWLDTLDATLSDDWRHQWYSGVTS-505
PknG	5	576-LTKDPETLRFKALYL-590
PknG	6	627-QVPQNSTHRRMAELTAI-643
PknG	7	651-LSESRIRRAARRLESIPTNEPRFLQIKIA-679
PknG	8	718-DSLRLLARSAPNVHHRYTLV-737

[a] Epitopes in signal peptide and conserved domains were discarded

Table 5 MHC class I epitopes predicted from target proteins

Target Protein	Mouse HLA Allele	Epitope number	Start	End	Peptide (9 mer)	Consensus rank (%)
SpaC	H-2-Db	1	615	623	FGLFNSAKL	0.3
SpaC	H-2-Kk	2	34	42	EEFENTEPI	0.3
SpaC	H-2-Kb	3	90	98	QSFNRNTGL	0.35
SpaC	H-2-Kd	4	124	132	IYAISQGRL	0.4
SpaC	H-2-Kd	5	116	124	AYNPKEGYI	0.5
SpaC	H-2-Kd	6	199	207	RYLVSNSSQ	0.5
SpaC	H-2-Kd	7	771	779	IYLVMGVLL	0.5
SpaC	H-2-Db	8	450	458	KCLQNTASI	0.6
SpaC	H-2-Db	9	208	216	SGTHNLYTL	0.7
SpaC	H-2-Dd	10	48	56	VGPSVDPTV	0.7
SpaC	H-2-Kd	11	458	466	IYANEKDLI	0.8
SpaC	H-2-Kb	12	785	793	SWSLYRNQL	0.85
SpaC	H-2-Kb	13	774	782	VMGVLLVLV	0.95
NanH	H-2-Kk	1	44	52	SEFFDSKVI	0.3
NanH	H-2-Dd	2	39	47	PDPVTSEFF	0.4
NanH	H-2-Dd	3	55	63	VDPAGQRCF	0.4
NanH	H-2-Kk	4	634	642	QELLRIFPG	0.5
NanH	H-2-Dd	5	655	663	GGMQKLLAF	0.6
NanH	H-2-Kb	6	645	653	PIFSFLASI	0.8
PknG	H-2-Kd	1	437	445	SYAEPSQTL	0.2
PknG	H-2-Kk	2	455	463	EEFANSKEI	0.2
PknG	H-2-Db	3	678	686	IAIMNAALT	0.5
PknG	H-2-Ld	4	525	533	LPGEAAPKL	0.5
PknG	H-2-Kb	5	586	594	KALYLYALV	0.55
PknG	H-2-Dd	6	665	673	SIPTNEPRF	0.6
PknG	H-2-Kb	7	685	693	LTWLRQSRL	0.6
PknG	H-2-Db	8	504	512	TSLFLDDYV	0.7
PknG	H-2-Kd	9	379	387	LYSPQRSTF	0.8
PknG	H-2-Kb	10	632	640	STHRRMAEL	0.85
PknG	H-2-Db	11	457	465	FANSKEIPL	0.9
PknG	H-2-Kk	12	21	29	EEDDLSGLL	0.9
PknG	H-2-Kk	13	353	361	LETQLFGIL	0.9

Epitopes in signal peptide and conserved domains were discarded

few strong binding peptides to MHC II were limited to mouse H-2-IAb allele and most of them were from NanH (Table 6). Only two MHC II strong binding peptides were predicted from SodC but both were discarded because they were located in conserved regions of the protein. Additional file 4 shows the MHC class II epitopes predicted from target proteins.

Epitope clustering

All B and T-cell epitopes (MHC I and II) predicted from the target proteins were grouped in clusters of sequence similarity in order to evaluate the redundancy degree among them. A total of 57 clusters were formed from a set of 136 epitopes predicted (Additional file 5). Most of them (34) were single-sequence clusters. Clusters 4 and 5 (PknG) and cluster 12 (SpaC) grouped epitopes for both B and T-cell (MHC I and II). These groups of epitopes can thus potentially stimulate a complete immune response against *C. pseudotuberculosis*. The main goal of vaccination is to induce humoral and cellular immunity by selectively stimulating antigen specific CTLs or B cells together with T_H cells [56]. Several clusters containing B-cell and either MHC I or II epitopes were also formed.

Table 6 Total numbers of MHC class II epitope prediction from target proteins

Target protein	Mouse MHC HLA allele	Number of strong binders[a]	Number of weak binders[a]	Number of peptides[b]
PknG	H-2-IAb	9	35	735
SpaC	H-2-IAb	4	48	782
SodC	H-2-IAb	0	12	192
NanH	H-2-IAb	22	64	680
PknG	H-2-IAd	0	29	735
SpaC	H-2-IAd	0	13	782
SodC	H-2-IAd	2	6	192
NanH	H-2-IAd	0	32	680

See epitope sequences in Additional file 4

[a] Strong binder threshold 50.00. Weak binder threshold 500.00

[b] Peptide length 15 mer

Among them are clusters 9 and 19 formed by epitopes from NanH (Additional file 5). Cluster 14 grouped all SodC weak binding epitopes to H-2-IAb allele.

Protein expression

Large amounts of SpaC, SodC, NanH, and PknG are necessary for future studies on the role of these proteins in *C. pseudotuberculosis* pathogenicity and virulence. *Escherichia coli* remains as one of the most attractive hosts among many systems available for heterologous protein production [57]. Thus, *pknG*, *spaC*, *sodC*, and *nanH* codon-optimized ORFs were cloned into the same expression vector system and individually transformed into BL21(DE3) *E. coli* strains. SDS-PAGE analyses show the successful expression of the target proteins (Fig. 5a). Purification of PknG using affinity and gel chromatography is shown in Fig. 5b.

From the current study we have suggested that several B and T-cell epitopes predicted from SpaC, SodC, NanH and PknG can be used for the development of a multi peptide vaccine to induce a complete immune response against *C. pseudotuberculosis*. The next step will be to evaluate experimentally these epitopes in vitro and in vivo to assess their real protective potential.

Conclusions

The in silico analyses performed show that SpaC, PknG and NanH present good potential as targets for vaccine development. Several epitopes from these proteins can potentially induce both humoral and cellular immune responses against *C. pseudotuberculosis*. The four target proteins were successfully expressed in *E. coli*. The production of these proteins in large amounts represents an important step for future studies on 3-D structure, pathogenicity, virulence, and vaccine development.

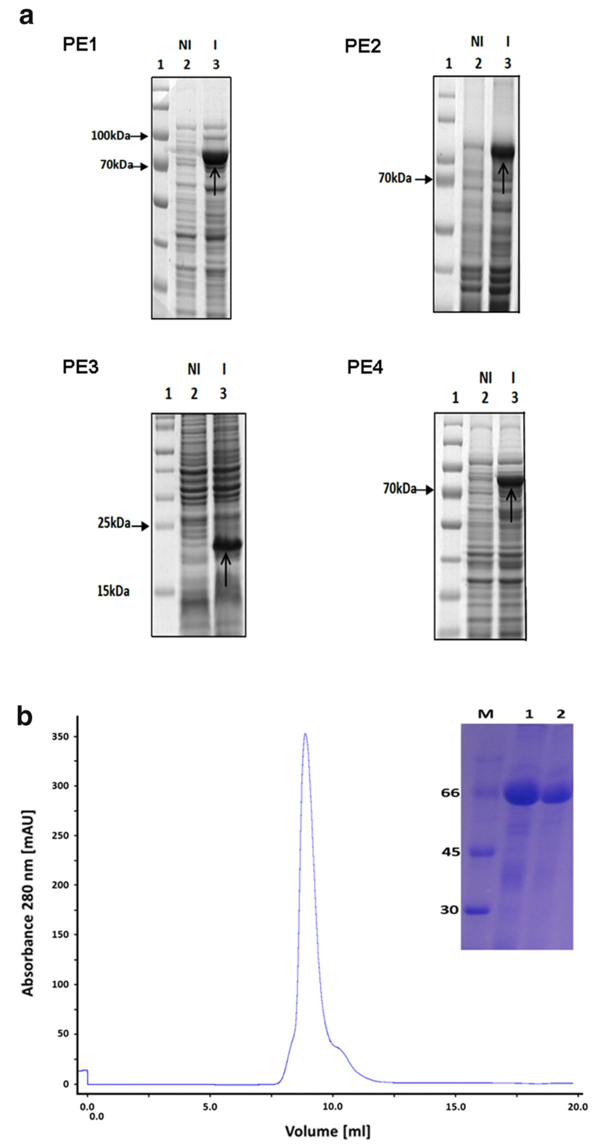


Fig. 5 Heterologous expression of the *C. pseudotuberculosis* FRC41 putative virulence factors in *E. coli* and rPknG purification. **a** Coomassie blue-stained SDS-PAGE analyses of the protein expression experiments: PE1, rPknG expression (83 kDa, 10 % gel) in *E. coli* strain BL21 Star (DE3); PE2, rSpaC expression (86 kDa, 10 % gel) in *E. coli* strain C43 (DE3); PE3, rSodC expression (18 kDa, 15 % gel) in *E. coli* strain BL21 Star (DE3); PE4, rNanH expression (71.5 kDa, 10 % gel) in *E. coli* strain C43 (DE3). 1, pre-stained protein ladder; 2 (NI), non-induced time 0; 3 (I), induced with 1 mM IPTG for 5 h at 37 °C. *Arrows* indicate the recombinant protein position in the gels. **b** Chromatogram of the rPknG purification by gel filtration. SDS–PAGE shows an analysis of the purification steps. *M* molecular-weight markers (kDa); 1, rPknG after affinity chromatography by Ni Sepharose; 2, rPknG purified by gel filtration

Additional file

Additional file 1. Homologous proteins of *C. pseudotuberculosis* FRC41 putative virulence factor PknG in CMNR microorganism and mammalians.

Additional file 2. Homologous proteins of *C. pseudotuberculosis* FRC41 putative virulence factors SpaC and NanH in CMNR microorganism and mammalians.

Additional file 3. Homologous proteins of *C. pseudotuberculosis* FRC41 putative virulence factor SodC in CMNR microorganism and mammalians.

Additional file 4. MHC class II epitopes§ predicted from target proteins.

Additional file 5. Clusters of B and T-cell epitopes predicted from target proteins.

Abbreviations

CLA: caseous lymphadenitis; CMNR: microorganisms from *Corynebacterium*, *Mycobacterium*, *Nocardi,* and *Rhodococcus* genera; PLD: phospholipase D; SOD: superoxide dismutase; ORF: open reading frame; SDS-PAGE: sodium dodecyl sulfate polyacrylamide gel electrophoresis.

Authors' contributions

TMS, KTOSJ and ELA performed the bioinformatic analyses. TMS and RWP made the immunological approaches for the epitopes prediction study. KTOSJ, NRT, RFSS and RBM carried out the protein expression and purification experiments. TMS and KTOSJ drafted the manuscript. VA, TMS, RM and RKA participated in the design and coordination of the study. All authors have read and approved the manuscript.

Author details

[1] Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Avenida Antonio Carlos, 6627, Pampulha, Belo Horizonte 31270-901, Brazil. [2] Uniclon Biotecnologia, Belo Horizonte, MG, Brazil. [3] Multiuser Center for Biomolecular Innovation, Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista "Júlio de Mesquita Filho", São José Do Rio Preto, SP, Brazil. [4] Laboratório de Imunologia e Biologia Molecular, Instituto de Ciências da Saúde, Universidade Federal da Bahia, Salvador, BA, Brazil.

Acknowledgements

This study was supported by CAPES, CNPq, FAPEMIG, and FAPESB.

Competing interests

The authors declare that they have no competing interests.

Received: 14 March 2016 Accepted: 3 May 2016
Published online: 16 May 2016

References

1. Dorella FA, Pacheco LGC, Oliveira SC, Miyoshi A, Azevedo V. *Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. Vet Res. 2006;37:201–18.
2. de Sá Guimarães A, do Carmo FB, Pauletti RB, Seyffert N, Ribeiro D, Lage AP, et al. Caseous lymphadenitis: epidemiology, diagnosis, and control. IIOAB J. 2011;2:33–43.
3. Baird GJ, Fontaine MC. *Corynebacterium pseudotuberculosis* and its role in Ovine Caseous Lymphadenitis. J Comp Pathol. 2007;137:179–210.
4. Fontaine MC, Baird GJ. Caseous lymphadenitis. Small Rumin Res. 2008;76:42–8.
5. Windsor PA. Control of caseous lymphadenitis. Vet Clin North Am-Food Anim Pract. 2011;27:193–202.
6. Oreiby AF. Diagnosis of caseous lymphadenitis in sheep and goat. Small Rumin Res. 2015;123:160–6.
7. Bastos BL, Dias Portela RW, Dorella FA, Ribeiro D, Seyffert N, et al. *Corynebacterium pseudotuberculosis*: immunological responses in animal models and zoonotic potential. J Clin Cell Immunol. 2012;S4:005. doi:10.4172/2155-9899.S4-005
8. Trost E, Ott L, Schneider J, Schröder J, Jaenicke S, Goesmann A, et al. The complete genome sequence of *Corynebacterium pseudotuberculosis* FRC41 isolated from a 12-year-old girl with necrotizing lymphadenitis reveals insights into gene-regulatory networks contributing to virulence. BMC Genom. 2010;11:728.
9. Rogers EA, Das A, Ton-That H. Adhesion by pathogenic corynebacteria. Adv Exp Med Biol. 2011;715:91–103.
10. Kim S, Oh DB, Kang HA, Kwon O. Features and applications of bacterial sialidases. Appl Microbiol Biotechnol. 2011;91:1–15.
11. Kim S, Oh DB, Kwon O, Kang HA. Identification and functional characterization of the NanH extracellular sialidase from *Corynebacterium diphtheriae*. J Biochem. 2010;147:523–33.
12. Dussurget O, Stewart G, Neyrolles O, Pescher P, Young D, Marchal G. Role of *Mycobacterium tuberculosis* copper-zinc superoxide dismutase. Infect Immun. 2001;69:529–33.
13. Piddington DL, Fang FC, Laessig T, Cooper M, Orme IM, Buchmeier NA, et al. Cu, Zn Superoxide Dismutase of *Mycobacterium tuberculosis* contributes to survival in activated macrophages that are generating an oxidative burst cu, zn superoxide dismutase of *Mycobacterium tuberculosis* contributes to survival in activated macrophages. Infect Immun. 2001;69:4980–7.
14. Pereira SFF, Goss L, Dworkin J. Eukaryote-like serine/threonine kinases and phosphatases in bacteria. Microbiol Mol Biol Rev. 2011;75:192–212.
15. Forrellad MA, Klepp LI, Gioffré A, Sabio y García J, Morbidoni HR, de la Paz Santangelo M, et al. Virulence factors of the *Mycobacterium tuberculosis* complex. Virulence. 2013;4:3–66.
16. Walburger A, Koul A, Ferrari G, Nguyen L, Prescianotto-Baschong C, Huygen K, et al. Protein kinase G from pathogenic mycobacteria promotes survival within macrophages. Science. 2004;304:1800–4.
17. Warner DF, Mizrahi V. The survival kit of *Mycobacterium tuberculosis*. Nat Med. 2007;13:282–4.
18. Altschul SF, Gish W, Pennsylvania T, Park U. Basic local alignment search tool 2Department of computer science. J Mol Biol. 1990;215:403–10.
19. Consortium TU. UniProt: a hub for protein information. Nucleic Acids Res. 2014;43:D204–12.
20. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, et al. Protein identification and analysis tools on the ExPASy server. Proteomics protocols handbook. New York City: Humana Press; 2005. p. 571–607.
21. Geourjon C, Deléage G. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Comput Appl Biosci. 1995;11:681–4.
22. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, et al. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics. 2010;26:1608–15.
23. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8:785–6.
24. Doytchinova IA, Flower DR. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. BMC Bioinformatics. 2007;8:4.
25. El-Manzalawy Y, Honavar V. Recent advances in B-cell epitope prediction methods. Immunome Res. London: BioMed Central Ltd; 2010;6:S2.
26. Jameson BA, Wolf H. The antigenic index: a novel algorithm for predicting antigenic determinants. Comput Appl Biosci. 1988;4:181–6.
27. Emini EA, Hughes JV, Perlow DS, Boger J. Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. J Virol. 1985;55:836–9.
28. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982;157:105–32.
29. White A, Handler P, Smith EL, editors. Principles of Biochemistry. 3rd ed. 1964.
30. Rini JM, Schulze-Gahmen U, Wilson IA. Structural evidence for induced fit as a mechanism for antibody-antigen recognition. Science. 1992;255:959–65.
31. Chou BPY, Fasman GD. Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzymol Relat Areas Mol Biol. 1978;47:45–148.

32. Chou PY. Prediction of protein structural classes from amino acid composition. Predict Protein Struct Princ Protein Conform. 1989;549–86.
33. Karplus PA, Schulz GE. Prediction of chain flexibility in proteins synthesis of acetylcholine receptors in xenopus oocytes induced by poly (A)+ -mRNA from locust nervous tissue. Naturwissenschaften. 1985;72:2–3.
34. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, et al. CDD: NCBI's conserved domain database. Nucleic Acids Res. 2015;43:D222–6.
35. Saha S, Saha S, Raghava GPS, Raghava GPS. Prediction methods for B-cell epitopes. Methods Mol Biol. 2007;409:387–94.
36. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, et al. The immune epitope database (IEDB) 3.0. Nucleic Acids Res. 2015;43:D405–12.
37. Moutaftsi M, Peters B, Pasquetto V, Tscharke DC, Sidney J, Bui H-H, et al. A consensus epitope prediction approach identifies the breadth of murine TCD8+ -cell responses to vaccinia virus. Nat Biotechnol. 2006;24:817–9.
38. Nielsen M, Lund O. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. BMC Bioinform. 2009;10:296.
39. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. Mol Syst Biol. 2011;7:539.
40. Sambrook J, Russell DW, editors. Molecular cloning: a laboratory manual. 3rd ed. Cold Spring Harbor: Cold Spring Harbor Laboratory Press; 2001.
41. Welch M, Govindarajan S, Ness JE, Villalobos A, Gurney A, Minshull J, et al. Design parameters to control synthetic gene expression in *Escherichia coli*. PLoS One. 2009;4:e7002.
42. Saïda F, Uzan M, Odaert B, Bontems F. Expression of highly toxic genes in *E. coli*: special strategies and genetic tools. Curr Protein Pept Sci. 2006;7:47–56.
43. Arnon R. A novel approach to vaccine design-pitope-based vaccines. FEBS J. 2006;273:33–4.
44. Grimm SK, Ackerman ME. Vaccine design: emerging concepts and renewed optimism. Curr Opin Biotechnol. 2013;24:1078–88.
45. Trivedi S, Gehlot HS, Rao SR. Protein thermostability in archaea and eubacteria. Genet Mol Res. 2006;5:816–27.
46. Goodswen SJ, Kennedy PJ, Ellis JT. A guide to in silico vaccine discovery for eukaryotic pathogens. Brief Bioinform. 2013;14:753–74.
47. Bhavsar AP, Guttman JA, Finlay BB. Manipulation of host-cell pathways by bacterial pathogens. Nature. 2007;449:827–34.
48. Krachler AM, Orth K. Targeting the bacteria–host interface. Virulence. 2013;4:284–94.
49. Kaufmann SHE, Hess J. Impact of intracellular location of and antigen display by intracellular bacteria: implications for vaccine development. Immunol Lett. 1999;65:81–4.
50. De Arruda LB, Chikhlikar PR, August JT, Marques ETA. DNA vaccine encoding human immunodeficiency virus-1 Gag, targeted to the major histocompatibility complex II compartment by lysosomal-associated membrane protein, elicits enhanced long-term memory response. Immunology. 2004;112:126–35.
51. Kimura K, Kimura D, Matsushima Y, Miyakoda M, Honma K, Yuda M, et al. CD8+ T cells specific for a malaria cytoplasmic antigen form clusters around infected hepatocytes and are protective at the liver stage of infection. Infect Immun. 2013;81:3825–34.
52. McKean S, Davies J, Moore R. Identification of macrophage induced genes of Corynebacterium pseudotuberculosis by differential fluorescence induction. Microbes Infect. 2005;7:1352–63.
53. Lan DT, Taniguchi S, Makino S, Shirahata T, Nakane A. Role of endogenous tumor necrosis factor alpha and gamma interferon in resistance to *Corynebacterium pseudotuberculosis* infection in mice. Microbiol Immunol. 1998;42:863–70.
54. Lan DT, Makino S, Shirahata T, Yamada M, Nakane A. Tumor necrosis factor alpha and gamma interferon are required for the development of protective immunity to secondary *Corynebacterium pseudotuberculosis* infection in mice. J Vet Med Sci. 1999;61:1203–8.
55. Chakraborty S, Chakravorty R, Ahmed M, Rahman A, Waise TZ, Hassan F, et al. A computational approach for identification of epitopes in dengue virus envelope protein: a step towards designing a universal dengue vaccine targeting endemic regions. In Silico Biol. 2010;10:235–46.
56. Dudek NL, Perlmutter P, Aguilar M-I, Croft NP, Purcell AW. Epitope discovery and their use in peptide based vaccines. Curr Pharm Des. 2010;16:3149–57.
57. Sugiki T, Fujiwara T, Kojima C. Latest approaches for efficient protein production in drug discovery. Expert Opin Drug Discov. 2014;1–16.

II.III.5 Proteome scale comparative modeling for conserved drug and vaccine targets identification in *Corynebacterium pseudotuberculosis*.

Hassan SS, Tiwari S, Guimarães LC, Jamal SB, Folador E, Sharma NB, de Castro Soares S, Almeida S, Ali A, Islam A, Póvoa FD, de Abreu VA, Jain N, Bhattacharya A, Juneja L, Miyoshi A, Silva A, Barh D, Turjanski A, **Azevedo V**, Ferreira RS.

BMC Genomics. 2014;15 Suppl 7:S3. doi: 10.1186/1471-2164-15-S7-S3. Epub 2014 Oct 27.

PublisherInfo

PublisherName	:	BioMed Central
PublisherLocation	:	London
PublisherImprintName	:	BioMed Central

Thiago Motta Venancio
João Carlos Setubal
Richard Garratt
Emanuel M Souza
André Fujita
Ricardo Z Vêncio
Hélder I Nakaya
Enrique Medina-Acosta

Proteome scale comparative modeling for conserved drug and vaccine targets identification in *Corynebacterium pseudotuberculosis*

ArticleInfo

ArticleID	:	7174
ArticleDOI	:	10.1186/1471-2164-15-S7-S3
ArticleCitationID	:	S3
ArticleSequenceNumber	:	7
ArticleCategory	:	Research
ArticleFirstPage	:	1
ArticleLastPage	:	19
ArticleHistory	:	RegistrationDate : 2014–10–27 OnlineDate : 2014–10–27
ArticleCopyright	:	Hassan et al.; licensee BioMed Central Ltd.2014 This article is published under license to BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
ArticleGrants	:	
ArticleContext	:	128641515S7S7

Syed Shah Hassan,[Aff1]
Email: hassan_chemist@yahoo.com

Sandeep Tiwari,[Aff1]
Email: sandip_sbtbi@yahoo.com

Luís Carlos Guimarães,[Aff1]
Email: luisguimaraes.bio@gmail.com

Syed Babar Jamal,[Aff1]
Email: syedbabar.jamal@gmail.com

Edson Folador,[Aff1]

Neha Barve Sharma,[Aff4 Aff5]
Email: nehabarve2006@gmail.com

Siomar de Castro Soares,[Aff1]
Email: siomars@gmail.com

Síntia Almeida,[Aff1]
Email: sintiaalmeida@gmail.com

Amjad Ali,[Aff1]
Email: amjad_uni@yahoo.com

Arshad Islam,[Aff6]
Email: arshad.cgl@gmail.com

Fabiana Dias Póvoa,[Aff2]
Email: fabianapovoa@gmail.com

Vinicius Augusto Carvalho de Abreu,[Aff1]
Email: vini.abreu@gmail.com

Neha Jain,[Aff4 Aff5]
Email: mymailtoneha@gmail.com

Antaripa Bhattacharya,[Aff5]
Email: antaripa1210@gmail.com

Lucky Juneja,[Aff4 Aff5]
Email: ljpreet88@gmail.com

Anderson Miyoshi,[Aff1]
Email: miyoshi@icb.ufmg.br

Artur Silva,[Aff3]
Email: asilva@ufpa.br

Debmalya Barh,[Aff5]
Email: dr.barh@gmail.com

Adrian Gustavo Turjanski,[Aff7]
Email: adrian@qi.fcen.uba.ar

Vasco Azevedo,[Aff1]
Email: vasco@icb.ufmg.br

Rafaela Salgado Ferreira,[Aff2]
Corresponding Affiliation: Aff2
Email: rafaelasf@gmail.com

Aff1 Laboratory of Cellular and Molecular Genetics, Department of General Biology, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

Aff2 Departament of Biochemistry and Immunology, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

Aff3 Institute of Biological Sciences, Federal University of Pará, Belém, Para, Brazil

Aff4 School of Biotechnology, Devi Ahilya University, Khandwa Road Campus, Indore, MP, India

Aff5 Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, West Bengal, India

Aff6 Department of Chemistry, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

Aff7 Structural Bioinformatics Group, Institute of Physical Chemistry of Materials, Environment and Energy, University of Buenos Aires, Argentine

Abstract

Corynebacterium pseudotuberculosis (Cp) is a pathogenic bacterium that causes caseous lymphadenitis (CLA), ulcerative lymphangitis, mastitis, and edematous to a broad spectrum of hosts, including ruminants, thereby threatening economic and dairy industries worldwide. Currently there is no effective drug or vaccine available against Cp. To identify new targets, we adopted a novel integrative strategy, which began with the prediction of the modelome (tridimensional protein structures for the proteome of an organism, generated through comparative modeling) for 15 previously sequenced *C. pseudotuberculosis* strains. This pan-modelomics approach identified a set of 331 conserved proteins having 95-100% intra-species sequence similarity. Next, we combined subtractive proteomics and modelomics to reveal a set of 10 Cp proteins, which may be essential for the bacteria. Of these, 4 proteins (tcsR, mtrA, nrdI, and ispH) were essential and non-host homologs (considering man, horse, cow and sheep as hosts) and satisfied all criteria of being putative targets. Additionally, we subjected these 4 proteins to virtual screening of a drug-like compound library. In all cases, molecules predicted to form favorable interactions and which showed high complementarity to the target were found among the top ranking compounds. The remaining 6 essential proteins (adk, gapA, glyA, fumC, gnd, and aspA) have homologs in the host proteomes. Their active site cavities were compared to the respective cavities in host proteins. We propose that some of these proteins can be selectively targeted using structure-based drug design approaches (SBDD). Our results facilitate the selection of *C. pseudotuberculosis* putative proteins for developing broad-spectrum novel drugs and vaccines. A few of the targets identified here have been validated in other microorganisms, suggesting that our modelome strategy is effective and can also be applicable to other pathogens.

ESMHint

Electronic supplementary material

The online version of this article (doi:10.1186/1471-2164-15-S7-S3) contains supplementary material, which is available to authorized users.

Background

Antimicrobial resistance involving a rapid loss of effectiveness in antibiotic treatment and the increasing number of multi-resistant microbial strains pose global challenges and threats. Thereby, efforts to find new drug and/or vaccine targets to control them are becoming indispensible. *Corynebacterium pseudotuberculosis* (Cp) is a pathogen of great veterinary and economic importance, since it affects animal livestock, mainly sheep and goats, worldwide, and its presence is reported in other mammals in several Arabic, Asiatic, East and West African and North and South American countries, as well as in Australia [1]. *C. pseudotuberculosis* is a Gram-positive, facultative intracellular, and pleomorphic organism; it is non-motile, although presenting fimbriae [2]. Based on *rpoB* gene (a β subunit of RNA polymerase), it shows a close phylogenetic relationship with other type strains of CMNR (*Corynebacterium, Mycobacterium, Nocardia* and *Rhodococcus*), a group that comprises genera of great medical, veterinary and biotechnological importance [1, 3]. A recent study showed that phylogenetic analysis for the identification of *Corynebacterium* and other CMNR species based on *rpoB* gene sequences are more accurate than analyses based on 16S rRNA [4]. Its pathogenicity and biological impact have already led to the sequencing of various strains of this pathogen from a wide range of hosts [3]. The pathogen causes several infectious diseases in goat and sheep population (biovar *ovis*), including caseous lymphadenitis (CLA), a chronic contagious disease characterized by abscess formation in superficial lymph nodes and in subcutaneous tissues. In severe cases, biovar *equi* infects the lungs, kidneys, liver and spleen, thereby threatening the herd life of the infected animals [2, 5]. The disease has been rarely reported in humans, as a result of occupational exposure, with symptoms similar to lymphadenitis abscesses [6–8]. The bacteria can survive for several weeks in soil in adverse conditions, what seems to contribute to its resistance and disease transmission [9, 10]. Direct contact to infectious secretions or contaminated materials are the primary sources of pathogen transmission between animals, but most frequently the infection occurs through exposed skin lacerations [5]. Given the medical importance of Cp and a lack of efficient medicines, in this study we applied a computational strategy to search for new molecular targets from this bacterium.

Recently, computational approaches such as reverse vaccinology, differential genome analyses [11], subtractive and comparative microbial genomics have become popular for rapid identification of novel targets in the post genomic era [12], [13]. These approaches were used to identify targets in various human pathogens, like *Mycobacterium tuberculosis* [14], *Helicobacter pylori* [15], *Burkholderia pseudomalleii* [16], *Neisseria gonorrhea* [17], *Pseudomonas aeruginosa* [18] and *Salmonella typhi* [19]. In general, such approaches follow the principle that genes/proteins must be essential to the pathogen and preferably have no homology to the host proteins [20]. Nevertheless, essential targets that are homologous to their corresponding host proteins may also be molecular targets for structure-based selective inhibitors development. In this case, the targets must show significant differences in the active sites or in other druggable pockets, when pathogenic and host proteins are compared [21–23].

Once a molecular target is chosen, the conventional experimental methods for drug discovery consist of testing many synthetic molecules or natural products to identify lead compounds. Such practices are laborious, time consuming and require high investments [24, 25]. On the other hand, computational methods for structure-based rational drug design can expedite the process of ligand identification and molecular understanding of interactions between receptor and ligand [26]. Such approaches are dependent on the availability of the structural information about the target protein. Considering the availability of experimental structures in PDB (Protein Data Bank) only for a low percentage of the known protein sequences, comparative modeling is frequently the method of choice for obtaining 3D coordinates for proteins of interest [27] for the development of specific drugs and docking analyses [28, 29].

In this work, we used a modelomic approach for the predicted proteome of *C. pseudotuberculosis* species. This served to bridge the gap between raw genomic information and the identification of good therapeutic targets based on the three dimensional structures. The novelty of this strategy relies in using the structural information from high-throughput comparative modeling for large-scale proteomics data for inhibitor identification, potentially leading to the discovery of compounds able to prevent bacterial growth. The predicted proteomes of 15 *C. pseudotuberculosis* strains were modeled (pan-modelome) using the MHOLline workflow. Intra-species conserved proteome (core-modelome) with adequate 3D models was further filtered for their essential nature for the bacteria, using the database of essential genes (DEG). This led to the identification of 4 essential bacterial proteins without homologs in the host proteomes, which were employed in virtual screening of compound libraries. Furthermore, we investigated a set of 6 essential host homologs proteins. We observed residues of the predicted bacterial protein cavities that are completely different from the ones found in the homologous domains, and therefore could be specifically targeted. By applying this computational strategy we provide a final list of predicted putative targets in *C. pseudotuberculosis*, in biovar *ovis* and *equi*. They could provide an insight into designing of peptide vaccines, and identification of lead, natural and drug-like compounds that bind to these proteins.

Materials and methods

Genomes selection

Proteomes predicted based on the genomes of fifteen *C. pseudotuberculosis* strains, including both biovar *equi* and biovar *ovis* (Table 1**)** were used in this study. Most of these genomes were sequenced by our group and are available at NCBI. We downloaded the genome sequences in gbk format from the NCBI server (ftp://ftp.ncbi.nih.gov/genomes/Bacteria) and the corresponding protein sequences (curated CDSs) were exported using Artemis Annotation Tool [30] for further analyses.

Table 1 Strains of *C. pseudotuberculosis* employed in the pan-modelome study, and their respective information regarding genomes statistics, disease prevalence and broad-spectrum hosts.

Strains	GPID	NCBI Accession	Genome Size (Mb)	Number of Proteins	G+C%	Hosts/ Location	Nitrate's Reduction/ Biovar	Clinical Manifestation	Sequencing Technology
Cp1/06-A	73235	NC_017308.1	2.28	1,963	52.2	Horse/USA	Positive/*equi*	Abscess	Illumina
Cp31	73223	NC_017730.1	2.3	2,063	52.2	Buffalo/Egypt	Positive/*equi*	Abscess	Ion Torrent, SOLiD v3
Cp258	157069	NC_017945.1	2.31	2,088	52.1	Horse/Belgium	Positive/*equi*	Ulcerative lymphangitis	SOLiD v3
Cp316	71591	NC_016932.1	2.31	2,106	52.1	Horse/USA	Positive/*equi*	Abscess	Ion Torrent
CpCIP52.97	61117	NC_017307.1	2.32	2,057	52.1	Horse/Kenya	Positive/*equi*	Ulcerative Lymphangitis	SOLiD v2
Cp162	89445	NC_018019.1	2.29	2,002	52.2	Camel/UK	Positive/*equi*	Neck Abscess	SOLiD v3
CpP54B96	77871	NC_017031.1	2.34	2,084	52.2	Antelope/S. Africa	Negative/*ovis*	CLA Abscess	Ion Torrent, SOLiD v3
Cp267	73515	NC_017462.1	2.34	2,148	52.2	Lhama/USA	Negative/*ovis*	CLA Abscess	SOLiD v3
Cp1002	40687	NC_017300.1	2.34	2,097	52.2	Goat/Brazil	Negative/*ovis*	CLA Abscess	454, Sanger
Cp42/02-A	73233	NC_017306.1	2.34	2,051	52.2	Sheep/Australia	Negative/*ovis*	CLA Abscess	Illumina
CpC231	40875	NC_017301.1	2.33	2,095	52.2	Sheep/Australia	Negative/*ovis*	CLA Abscess	454, Sanger
CpI19	52845	NC_017303.1	2.34	2,099	52.2	Bovine/Israel	Negative/*ovis*	Bovine Mastitis Abscess	SOLiD v2
Cp3/99-5	73231	NC_016781.1	2.34	2,142	52.2	Sheep/Scotland	Negative/*ovis*	CLA	Illumina
CpPAT10	61115	NC_017305.1	2.34	2,089	52.2	Sheep/Argentina	Negative/*ovis*	Lung Abscess	SOLiD v2
CpFRC41	48979	NC_014329.1	2.34	2,104	52.2	Human/France	Negative/*ovis*	Necrotizing lymphadenitis	SOLiD v3

Pan-modelome construction

A high throughput biological workflow, MHOLline (http://www.mholline.lncc.br), was used to predict the modelome (complete set of protein 3D models for the whole proteome) for each Cp strain. MHOLline uses the program MODELLER [31] for protein 3D structure prediction through comparative modeling. Furthermore, the workflow includes BLASTp (Basic Local Alignment Search Tool for Protein) [32], HMMTOP (Prediction of transmembrane helices and topology of proteins) [33], BATS (Blast Automatic Targeting for Structures), FILTERS, ECNGet (Get Enzyme Commission Number), MODELLER and PROCHECK [34] programs. The protocol used here was modified accordingly from the original work by Capriles et al., 2010 [35]. Briefly, the input files of protein sequences were used in FASTA format for all strains because the MHOLline accepts only .faa format files for the whole process. Firstly, MHOLline selected the template structures available at the Protein data Bank (PDB) via BLASTp (version 2.2.18), using the default parameters (e-value ≤ $10e^{-5}$). Secondly, the program BATS refined the BLASTp search for template sequence identification into different groups namely G0, G1, G2 and G3. Only the protein sequences in the group G2, which are characterized by an e-value ≤ $10e^{-5}$, Identity ≥ 0.25 and LVI ≤ 0.7 (where LVI is a length variation index of the BATS program for sequence coverage, the lower the LVI value, the higher the sequence coverage and vice versa) were selected. Among the MHOLline output files, the group G2 contained the largest number of protein sequences (≥ 50% for each input file). Subsequently, the "Filter" tool classified the group G2 sequences into seven distinct quality models groups, from "Very High" to "Very Low" depending on the quality of the template structure for a given query protein sequence. The program MODELLER then modeled all these groups in an automated manner. The number of sequences in the group G2 varies for each *C. pseudotuberculosis* strain. Only the first four distinct quality model groups of G2 were taken into consideration in this study, these were: 1- Very High quality model sequences (identity ≥ 75%) (LVI ≤ 0.1), 2- High quality model sequences (identity ≥ 50%) and < 75%) (LVI ≤ 0.1), 3- Good quality model sequences (identity ≥ 50%) (LVI > 0.1 and ≤ 0.3) and 4- Medium to Good quality models (identity ≥ 35% and < 50%) (LVI ≤ 0.3) (http://www.mholline.lncc.br). The percentage of identity represents identity between query and template sequences, a LVI ≤ 0.1 is equivalent to coverage of more than 90%, while LVI ≤ 0.3 corresponds to coverage of more than 70%. Therefore, all protein 3D models considered in this study were built from sequences for which there existed a template with identity ≥ 35% and LVI coverage over 70%. Later on, the ECNGet tool assigned an Enzyme Commission (EC) number to each sequence in G2, according to the best PDB template. The MODELLER (v9v5) program performed the automated global alignment and 3D protein model construction. Finally, the program PROCHECK (v3.5.4) evaluated the constructed models based on their stereo-chemical quality. Additionally, transmembrane regions in the input protein sequences were predicted by HMMTOP, for putative vaccine and drug targets identification.

Identification of intra-species conserved genes/proteins

The words genes and proteins are interchangeably used here but they refer to the same protein target of the pathogen. For the identification of highly conserved proteins with 3D models in all Cp strains (≥ 95% sequence identity), the standalone release of NCBI BLASTp+ (v2.2.26) was acquired from the NCBI ftp site (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/), installed on a local machine and a search was performed for all strains using Cp1002 as a reference genome. The highly conserved proteins were selected using a comparative genomics/proteomics approach using an all-against-all BLASTp analysis with cut off values of *E = 0.0001* [12 17 20 36].

Analyses of essential and non-host homologous (ENH) proteins

To select conserved targets that were essential to the bacteria, a subtractive genomics approach was followed [20]. Briefly, the set of core-modelome proteins from *C. pseudotuberculosis* were subjected to the Database of Essential Genes (DEG) for homology analyses. DEG contains experimentally validated essential genes from 20 bacteria [37]. The BLASTp cutoff values used were: *E-value* = 0.0001, *bit score* ≥100, *identity* ≥ 35% [20].

Furthermore, the pool of essential genes was subjected to NCBI-BLASTp (*E-value = 0.0001, bit score ≥100, identity ≥ 35%*) against (human, equine, bovine and ovine proteomes) to identify essential non-host homologs targets [12]. The set of essential non-host homologous proteins were further crosschecked with the NCBI-BLASTp PDB database using default parameters to find any structural similarity with the available host homologs protein structures, keeping cutoff level to ≤ 15% for query coverage. These proteins were checked for their biochemical pathway using KEGG (Kyoto Encyclopedia of Genes and Genomes) [38], virulence using PAIDB (Pathogenicity island database) [39], functionality using UniProt (Universal Protein Resource) [40], and cellular localization using CELLO (subCELlular LOcalization predictor) [41]. The final list of targets was based on 12 criteria as described previously [20].

Analyses of essential and host homologous (EH) proteins

We have extrapolated our analyses and also considered protein targets that were predicted as essential to bacterial survival but showed homology to host proteins. This was based on the possibility to find differences between bacterial and host proteins to rationally design inhibitors. The pool of essential protein targets that showed cut off values equal or higher than those for essential non-host homologs through NCBI-BLASTp was treated as host homologous proteins. These were also analyzed for pathway involvement, virulence, functional annotation and cellular localization like essential non-host homologous proteins. To verify the presence of significant residue differences in druggable protein cavities, a structural comparison was performed for each pathogen and their corresponding host protein through the molecular visualization program PyMOL (v1.5, Schrodinger, LLC) (http://www.pymol.org). The related published data of each template structure for each host homolog was also crosschecked for information about these residues, based on the PDB code of each template structure as input in the PDBelite server [42]. Catalytic Site Atlas (CSA) was also consulted to get robust information of the active site residues for the druggable enzyme targets [43]. CSA is a database documenting enzyme active sites and catalytic residues in enzymes of 3D structure and has 2 types of entry, original hand-annotated entries with literature references and homologous entries, found by PSI-BLAST alignment to an individual original entry, using an *e-value* cut-off of 0.00005. CSA can be accessed via a 4-letter PDB code. The equivalent residue that aligns in the query sequence to the catalytic residue found in the original entry is documented. Though the DoGSiteScorer predicts the druggable protein cavities, the host homologous proteins were further subjected to CASTp (Computed Atlas of Surface Topography of Proteins) [44], Pocket-Finder and Q-SiteFinder [45] to get more reliable and robust results about the druggable cavities of the target proteins.

Prediction of druggable pockets

3D structure information and druggability analyses are important factors for prioritizing and validating putative pathogen targets [46, 47]. As aforementioned, for druggability analyses, the final list of essential non-host and host homologous protein targets in PDB format, were subjected to DoGSiteScorer [48], an automated pocket detection and analysis tool for calculating the druggability of protein cavities. For each cavity detected the program returns the residues present in the pocket and a druggable score ranging from 0 to 1. The closer to 1 the obtained values are, the more druggable the protein cavity is predicted to be, i.e. the cavities are predicted to be more likely to bind ligands with high affinity [48]. The DoGSiteScorer also calculates volume, surface area, lipophilic surface, depth and other related parameters for each predicted cavity.

Virtual screening and docking analyses

The ligand library was obtained from the ZINC database, containing 11,193 drug-like molecules, with Tanimoto cutoff level of 60% [49]. Proteins were inspected for structural errors such as missing atoms or erroneous bonds and protonation states in MVD (Molegro Virtual Docker) [50]. The cavities predicted with DogSiteScorer (druggability ≥ 0.80) for all protein targets, were compared with the cavities detected by MVD. The most druggable cavity, according to DogSiteScorer, was subjected to virtual screening. MVD includes three search algorithms for molecular docking namely MolDock Optimizer [50], MolDock Simplex Evolution (SE), and Iterated Simplex (IS). In this work the MolDock Optimizer search algorithm, which is based on a differential evolutionary algorithm, was employed. The default parameters used for the guided differential evolution algorithm are a) population size = 50, b) crossover rate = 0.9, and c) scaling factor = 0.5. The top ranked 200 compounds for each protein were analyzed in Chimera for shape complementarity and hydrogen bond interactions, leading to the selection of a final set of 10 compounds for each target protein.

Results and discussion

Modelome and common targets in *C. pseudotuberculosis*species

Here we report the identification of common putative targets among 15 strains of *C. pseudotuberculosis* species based on the construction of genome scale protein three-dimensional structural models. Structural information of target proteins can aid in drug and/or vaccine design and in the discovery of new lead compounds [51]. The approach employed here generated high-confidence structural models through the MHOLline workflow **(**Figure 1**)** from orthologous protein. To identify the common conserved proteins with a sequence similarity of 95-100%, a comparative genomics approach was performed where all the BATS classified G2 sequences from "Very High" to "Medium to Good" quality, from 14 Cp strains, were aligned to the G2 sequences of Cp1002, assumed as a reference genome for this study. In total, a set of 331 protein sequences was selected, being conserved in all strains. An overview of the different steps involved in this computational approach for genome scale modelome and prioritization of putative drug and vaccine targets is given in Figure 2a-b.

Figure 1 High-throughputness (efficiency) of the MHOLline biological workflow for genome-scale modelome (3D models) prediction. Predicted proteomes from the genomes of 15 *C. pseudotuberculosis* strains were fed to the MHOLline workflow in FASTA format. The blue line represents the number of input data, according to the left-hand side y-axis. The bars show the number in the form of MHOLline output data (according to the right-hand side y-axis) of: not aligned sequences (G0, green bars); sequences for which there is a template structure available at RCSB PDB (yellow bars); sequences with acceptable template structures that where modeled in the MHOLline workflow (G2, red bars); sequences with predicted transmembrane regions (HMMTOP, purple bars) and the number of sequences that were predicted as enzymes in each genome and were assigned an EC number (ECNGet, gray bars). The x-axis represents the *C. pseudotuberculosis* genomes used in this study.

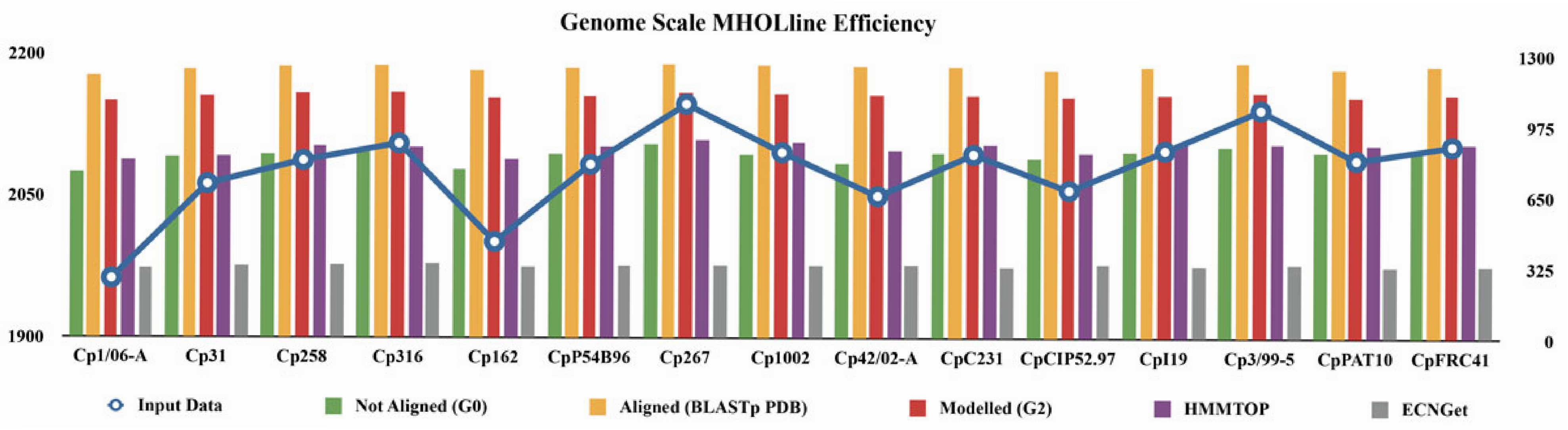


Figure 2 Overview of different computational steps employed in the identification of putative essential targets (non-host homologous and host homologous) for drugs and vaccines from the core-proteome of 15 *C* . pseudotuberculosis strains. **Figure 2b**. Intra-species subtractive modelomics workflow for conserved targets identification in *C*. pseudo tuberculosis species. The table (from left to right) represents the total number of protein sequences as an input data in fasta format fed to the MHOLline workflow (upper forward arrow). The remaining columns show the output data of group G2 (upper backward arrow), first by BATS and then by Filter tools of the MHOLline workflow respectively. Columns 4th-7th constitute the number of protein sequences of different qualities of all 15 Cp strains, where the sequences of 14 Cp strains were compared using BLASTp, to the sequences of Cp1002 strain as reference, for the identification of conserved protein targets (core-modelome). The funnel shows how this workflow processes and filters a large quantity of genomic data for putative drug and vaccine targets identification of a pathogen.

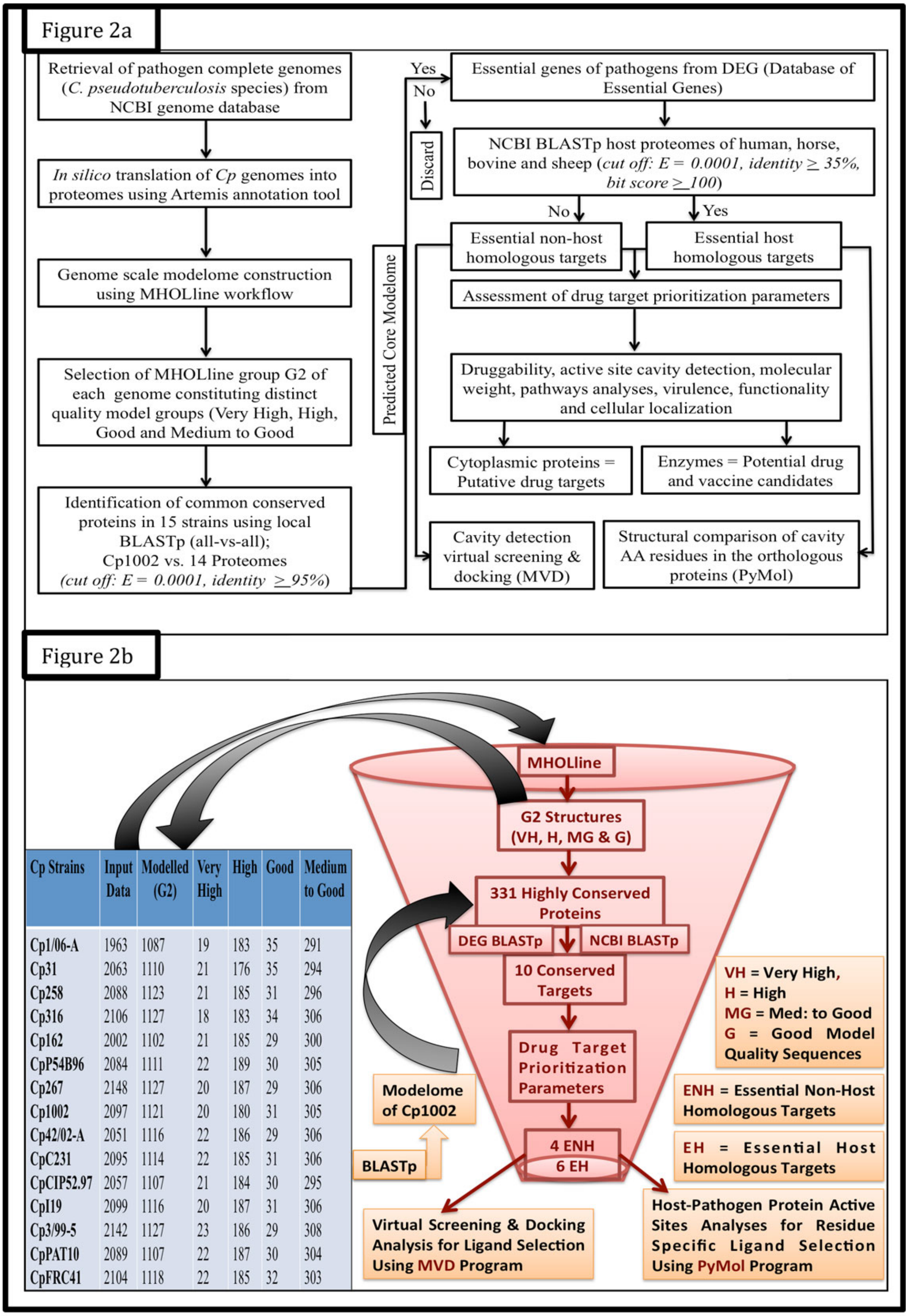


Figure 2a
Retrieval of pathogen complete genomes (*C. pseudotuberculosis* species) from NCBI genome database
In silico translation of *Cp* genomes into proteomes using Artemis annotation tool
Genome scale modelome construction using MHOLline workflow
Selection of MHOLline group G2 of each genome constituting distinct quality model groups (Very High, High, Good and Medium to Good
Identification of common conserved proteins in 15 strains using local BLASTp (all-vs-all); Cp1002 vs. 14 Proteomes (*cut off: E = 0.0001, identity ≥95%*)
Predicted Core Modelome
Yes
No
Discard
Essential genes of pathogens from DEG (Database of Essential Genes)
NCBI BLASTp host proteomes of human, horse, bovine and sheep (*cut off: E = 0.0001, identity ≥ 35%, bit score ≥100*)
No
Yes
Essential non-host homologous targets
Essential host homologous targets
Assessment of drug target prioritization parameters
Druggability, active site cavity detection, molecular weight, pathways analyses, virulence, functionality and cellular localization
Cytoplasmic proteins = Putative drug targets
Enzymes = Potential drug and vaccine candidates
Cavity detection virtual screening & docking (MVD)
Structural comparison of cavity AA residues in the orthologous proteins (PyMol)
Figure 2b
MHOLline
G2 Structures (VH, H, MG & G)
331 Highly Conserved Proteins
DEG BLASTp
NCBI BLASTp
10 Conserved Targets
Drug Target Prioritization Parameters
4 ENH
6 EH
Modelome of Cp1002
BLASTp
VH = Very High,
H = High
MG = Med: to Good
G = Good Model Quality Sequences
ENH = Essential Non-Host Homologous Targets
EH = Essential Host Homologous Targets
Virtual Screening & Docking Analysis for Ligand Selection Using MVD Program
Host-Pathogen Protein Active Sites Analyses for Residue Specific Ligand Selection Using PyMol Program

Cp Strains	Input Data	Modelled (G2)	Very High	High	Good	Medium to Good
Cp1/06-A	1963	1087	19	183	35	291
Cp31	2063	1110	21	176	35	294
Cp258	2088	1123	21	185	31	296
Cp316	2106	1127	18	183	34	306
Cp162	2002	1102	21	185	29	300
CpP54B96	2084	1111	22	189	30	305
Cp267	2148	1127	20	187	29	306
Cp1002	2097	1121	20	180	31	305
Cp42/02-A	2051	1116	22	186	29	306
CpC231	2095	1114	22	185	31	306
CpCIP52.97	2057	1107	21	184	30	295
CpI19	2099	1116	20	187	31	306
Cp3/99-5	2142	1127	23	186	29	308
CpPAT10	2089	1107	22	187	30	304
CpFRC41	2104	1118	22	185	32	303

Identification of ENH and EH proteins as putative drug and/or vaccine targets

To identify essential proteins as putative therapeutic targets in *C. pseudotuberculosis*, from the set of core-modelome, these were compared to the Database of Essential Genes (DEG). Based on this filter, the number of selected targets was reduced drastically to a final set of only 10 targets. These were compared to the aforementioned corresponding host proteomes, leading to the identification of 4 essential non-host homologous proteins (ENH, Table 2) and 6 essential host homologous proteins (EH, Table 3).

Table 2 Drug and/or vaccine targets prioritization parameters and functional annotation of the four essential non-host homologous putative targets.

Gene and protein codes	Official full name	Number of cavities with Drug Score[a] > 0.80	Number of cavities with Drug Score[a] > 0.60 and < 0.80	Mol. Wt (KDa)[b]	Functions[c]	Cellular component[d]	Pathways[e]	Virulence[f]
Cp1002_0515 **MtrA**	DNA-binding response regulator mtrA	1	2	25.97	**MF:** DNA binding, two-component response regulator activity. **BP:** Intracellular signal transduction, regulation of transcription, DNA-dependent	Intracellular/ Cytoplasm	Two-component signaling systems	Yes
Cp1002_0742 **IspH**	4-hydroxy-3-methylbut-2-enyl diphosphatereductase	1	4	36.59	**MF:** Metal ion binding, 4-hydroxy-3-methylbut-2-en-1-yl diphosphate reductase activity, 3 iron, 4 sulfur cluster binding **EC: 1.17.1.2** **BP:** Dimethylallyldiphosphate biosynthetic process, isopentenyldiphosphate biosynthetic process, mevalonate-independent pathway	Cytoplasm	Inositol phosphate metabolism/ Pentose phosphate pathway/Terpene metabolism	Yes
Cp1002_1648 **TcsR**	Two-component system transcriptional regulatory protein	3	2	21.93	**MF:** Sequence-specific DNA binding, two-component response regulator activity, sequence-specific DNA binding transcription factor activity **BP:** Intracellular signal transduction, transcription, DNA-dependent	Intracellular/ Cytoplasm	Two-component system	Yes
Cp1002_1676 **Nrdl**	Ribonucleoside-diphosphatereductase alpha chain	1	1	88.02	**MF:** ATP binding, ribonucleoside-diphosphate reductase activity, thioredoxin disulfide as acceptor **BP:** DNA replication	Cytoplasm	Pyrimidine metabolism/ Purine metabolism	Yes

[a] Druggability predicted with DoGSiteScorer software. A druggability score above 0.60 is considered to be good, but a score above 0.80 is favored [48].

[b] Molecular weight was determined using ProtParam tool (http://web.expasy.org/protparam/).

[c] Molecular function (MF) and biological process (BP) for each target protein was determined using UniProt.

[d] Cellular localization of pathogen targets was performed using CELLO.

[e] KEGG was used to find the role of these targets in different cellular pathways.

[f] PAIDB was used to check if the putative targets are involved in pathogen's virulence.

Table 3 Drug and/or vaccine targets prioritization parameters and functional annotation of the six essential host homologous putative targets.

Gene and protein codes	Official full name	Number of cavities with Drug Score[a] > 0.80	Number of cavities with Drug Score[a] > 0.60 and < 0.80	Mol. Wt (KDa)[b]	Functions[c]	Cellular component[d]	Pathways[e]	Virulence[f]
Cp1002_0385 **Adk**	Adenylate kinase	**1**	0	24.120	**MF:** Kinase, Transferase, ATP binding **BP:** Nucleotide biosynthesis **EC 2.7.4.3**	Cytoplasm	Purine metabolism; AMP biosynthesis via salvage pathway	**Yes**
Cp1002_0692 **GapA**	Glyceraldehyde-3-phosphate dehydrogenase A	**2**	1	51.918	**MF:** Oxidoreductase, NAD binding, NADP binding, **BP:** glucose metabolic process **EC 1.2.1.13**	Cytoplasm	Glycolysis/Gluconeogenesis	**Yes**
Cp1002_0728 **GlyA**	Serine hydroxymethyltransferase	**2**	1	46.187	**MF:** Methyltransferase, Transferase **BP:** Amino-acid biosynthesis One-carbon metabolism **EC 2.1.2.1**	Cytoplasm	Amino-acid biosynthesis; glycine biosynthesis; One-carbon metabolism; tetrahydrofolate interconversion.	**Yes**
Cp1002_0738 **FumC**	Fumaratehydratase class II	**2**	0	49.767	**MF:** Lyase **BP:** Tricarboxylic acid cycle **EC 4.2.1.2**	Cytoplasm	Carbohydrate metabolism; tricarboxylic acid cycle; (S)-malate from fumarate	**Yes**
Cp1002_1005 **Gnd**	6-phosphogluconate dehydrogenase	**3**	5	53.669	**MF:** Oxidoreductase **BP:** Pentose shunt **EC 1.1.1.44**	Cytoplasm	Carbohydrate degradation; pentose phosphate pathway;	**No**
Cp1002_1042 **AspA**	Aspartate ammonia-lyase	**2**	4	52.277	**MF: Lyase** **EC 4.3.1.1**	Cytoplasm	Alanine, aspartate and glutamate metabolism, Nitrogen metabolism	**Yes**

[a] Druggability predicted with DoGSiteScorer software. A druggability score above 0.60 is usually considered, but a score above 0.80 is favored [48].

[b] Molecular weight was determined using ProtParam tool (http://web.expasy.org/protparam/).

[c] Molecular function (MF) and biological process (BP) for each target protein was determined using UniProt.

[d] Cellular localization of pathogen targets was performed using CELLO.

[e] KEGG was used to find the role of these targets in different cellular pathways.

[f] PAIDB was used to check if the putative targets are involved in pathogen's virulence.

Among the ENH proteins, two targets were selected from a bacterial unique pathway, the two component signaling system. These targets are tcsR (two-component response regulator) and mtrA (two component sensory transduction transcriptional regulatory protein). While the tcsR is a novel protein target, as it is has not been described so far as a target in any organism, mtrA has been already reported as a target in *Mycobacterium* [52] and provides multidrug resistance to *Mycobacterium avium* [53]. Therefore, targeting mtrA in *C. pseudotuberculosis* may also be effective in controlling the infection of CLA. The remaining ENH protein targets, nrdI and ispH, also participate in biochemical pathways. NrdI (ribonucleoside-diphosphate reductase alpha chain) is a flavodoxin which contains a diferric-tyrosyl radical cofactor and it is involved in nucleotide metabolism in *E. coli* [54]. It has been reported as a putative target in several pathogens including *C. pseudotuberculosis, Corynebacterium diphtheriae* and *Mycobacterium tuberculosis* [20]. The target ispH (4-hydroxy-3-methylbut-2-enyl diphosphate reductase; EC 1.17.1.2) is an essential cytoplasmic enzyme in *Escherichia coli* [55]. This iron-sulfur protein plays a crucial role in terpene metabolism of various pathogenic bacteria [56, 57] and it is a predicted target in *Salmonella tyhpimurium* [58] and *Plasmodium falciparum* [59]. It should be noted that according to the cut off threshold for NCBI-BLASTp that we have followed, ispH shows homology only to the human host. So, if human is not considered as a possible host, ispH can also be considered as a common putative target. The roles of these proteins in different metabolic pathways was confirmed from KEGG [38] and METACYC [60] databases.

Prioritization parameters of drug and/or vaccine targets

Previous studies have shown several factors that can aid in determining the suitability of therapeutic targets [46]. The availability of 3D structural information, the main approach of our study, is very helpful in drug development. Other important factors for drug targets include preferred low MW and high druggability. On the other hand, for vaccine targets the information about subcellular localization is important and proteins that contain transmembrane motifs are preferred [36, 46, 61, 62]. We have determined most of these prioritizing properties for the 10 essential proteins (Table 2 &3). Interestingly, according to the target-prioritizing criterion, all targets have a low MW, and are predicted to be localized in the cytoplasmic compartment of the Cp. Druggability evaluation with DoGSiteScorer [48] for all conserved targets allowed the prediction of numerous druggable cavities with at least one druggable cavity for each Cp target. For the 4 ENH proteins tcsR, mtrA, nrdI, and ispH, 3, 5, 5 and 2 cavities with score ≥ 0.80 were observed respectively. For each protein, the cavity that exhibited the highest druggability score was selected for docking analyses. For 6 EH targets, adk, gapA, glyA, fumC, gnd, and aspA, 1, 3, 3, 2, 8 and 6 cavities were observed respectively according to the aforementioned druggability score criteria (Table 2 &3). Here, in each case, the most druggable predicted cavity was structurally compared with the cavities in respective host proteins.

Virtual screening and molecular docking analyses of ENH targets

For each ENH target protein (mtrA, ispH, tcsR and nrdl), the top 200 drug-like molecules from virtual screening were visually inspected to select 10 molecules that showed favorable interactions with the target. The biological importance of each target and an analysis of the predicted protein-ligand interaction are described below. ZINC codes and MolDock scores of selected ligands, the number of hydrogen bonds as well as protein residues involved in these interactions, are shown in a table for each target protein (Tables 4, 5, 6, 7. Figures showing the predicted binding mode for one of the 10 selected ligands are also shown for each target (Additional files 1, 2, 3, 4, 5).

Table 4 ZINC codes, MolDock scores and predicted hydrogen bonds for the ten compounds selected among the top ranking 200 molecules against **Cp1002_0515** (**MtrA**, DNA-binding response regulator).

ZINC IDs	MolDock score	Number of H-bonds/ residues interacting with the compound
75109074	-130.402	3 Thr73, Asp48, Arg116
12117405	-115.838	3 Arg119, Arg118, Ala115
02546720	-113.761	3 Thr73, Arg119
40266587	-116.119	2 Asp48, Leu117
71405274	-113.264	2 Arg116, Asp97
05687366	-111.376	2 Arg119, Asp48
04730243	-109.609	2 Arg119, Asp157
19720976	-109.061	2 Arg119
72342680	-108.299	2 Arg119, Asp157

Table 5 ZINC codes, MolDock scores and predicted hydrogen bonds for the ten compounds selected among the top ranking 200 molecules against **Cp1002_0742** (**IspH**, 4-hydroxy-3-methyl but-2-enyl diphosphate reductase).

ZINC IDs	MolDock score	Number of H-bonds/ residues interacting with the compound
00510419	-151.376	7 Cys39, His68, Thr225, Ser250, Asn252
00529019	-129.348	5 His68, Ser250, Asn252, Thr193, Thr193
04344036	-135.156	8 Thr193, His151, His68, Ser251, Asn252, Ser250, Asn252
04632419	-136.984	6 Cys39, Gly41, Ala100, Cys222, Thr193, Asn252
04730243	-129.414	10 Cys222, Thr193, Asn252, His151, Ser250, His68, Ser250
05479451	-129.963	9 Asn252, Ser250, Ser251, His68, Cys123, Cys39, Gly41
05775454	-161.806	3 Asn252, Thr193, His68
16941408	-126.163	6 Thr193, Asn252, Asn252, Ser250
04622741	-127.816	12 Cys39, Cys123, His68, Cys222, Ser250, Ser251, His151, Thr193
14017317	-129.664	8 Cys39, Glu153, His68, Asn252, Ser251, Asn252, His151, Thr193

Table 6 ZINC codes, MolDock scores and predicted hydrogen bonds for the ten compounds selected among the top ranking 200 molecules against **Cp1002_1648** (**TcsR,**Two component transcriptional regulator).

ZINC IDs	MolDock score	Number of H-bonds/ residues interacting with the compound
00510419	-167.633	3 Val76, Gln185, Asn193
01617096	-146.178	3 Ala74, Gln185, Arg191
32911447	-148.424	3 Gln185, Ala70, Arg193
00091802	-143.287	3 Val76, Ala51
67847806	-156.655	4 Thr75, Thr75, Pro48, Val76
19399766	-160.743	3 Val76, Val76, Ala51
16980834	-147.631	4 Ala66, Val76
06269029	-145.277	4 Thr75, Pro48, Val76
05934077	-145.785	3 Arg191, Gln185
01647971	-167.152	3 Thr75, Val76, Ala51

Table 7 ZINC codes, MolDock scores and predicted hydrogen bonds for the ten compounds selected among the top ranking 200 molecules against **Cp1002_1676** (**NrdI**).

ZINC IDs	MolDock score	Number of H-bonds/ residues interacting with the compound
01585114	-151.406	6 Ser8, Ser7, Thr13, Asn12
04721321	-144.134	7 Ser8, Ser7, Thr13, Leu116
17023683	-140.718	6 Ser7, Ser8, Thr13, Thr13, Thr13
00510419	-154.064	4 Thr10, Thr13, Ser8
01417445	-138.997	4 Thr13, Tyr49, Ser8, Ser7

00042420	-135.363	6 Tyr49, Ser7, Ser8, Thr13, Thr13, Thr13
00408361	-133.535	6 Thr13, Ser7, Ser8, Tyr49, Thr48
15830653	-153.83	4 Ser7, Thr13, Tyr49
00032839	-139.327	6 Ser8, Ser7, Thr13, Thr13, Thr13
48212336	-137.675	6 Ser7, Ser8, Ser8, Thr13, Ser54

Cp1002_0515 (**MtrA**, DNA-binding response regulator) is part of the two-component signal transduction system consisting of the sensor kinase (Histidine protein kinases, HKs) and the response regulator, MtrB and MtrA respectively. This system is highly conserved in *Corynebacteria* and *Mycobacteria* and it is essential for their survival to adapt to environmental changes. Homologs of MtrA and MtrB are present in many species of the genera *Corynebacterium, Mycobacterium, Nocardia, Rhodococcus* (CMNR), and others like *Thermomonospora, Leifsonia, Streptomyces, Propionibacterium*, and *Bifidobacterium* [63]. MtrA represents the fourth family member of the OmpR/PhoB family of response regulators. Like other family members, MtrA has been reported to be essential in *M. tuberculosis* [64]. It possesses an N-terminal regulatory domain and a C-terminal helix-turn-helix DNA-binding domain, already indicating that this response regulator functions as a transcriptional regulator, with phosphorylation of the regulatory domain modulating the activity of the protein [65]. Based on a comparison with a crystallographic structure of the MtrA template (2GWR, MtrA from *M. tuberculosis*), the active site residues involved in H-bond interactions with the crystallographic ligand are Val145, Gln151, Ile152 and Leu154. Although none of these residues is predicted to form hydrogen bonds with the ten selected docked ligands, these molecules were predicted to interact with other residues in the pocket. Table 4 shows the 10 selected ligands according to their minimum energy values and number of hydrogen bond interactions. **ZINC75109074** (N-benzyl-N-[[2-(2-thienyl)-1H-imidazol-4-yl] methyl] prop-2-en-1-amine) is shown here as the top scoring ligand (Additional file 1**)**.

Cp1002_0742 (**IspH**, 4-hydroxy-3-methylbut-2-enyl diphosphate reductase) is an iron-sulfur oxidoreductase enzyme that plays a key role in the metabolism of terpenes in several pathogens. Terpenes constitute a large class of natural compounds. Their biosynthesis initiates with the building blocks isopentenyl-diphosphate (IPP) and dimethylallyldiphosphate (DMAPP), and differs in bacteria and mammals [57]. In bacteria and other pathogenic microorganisms the enzyme IspH catalyzes the last step in the production of IPP and DMAPP. The three structural units of the enzyme harbor a cubic iron-sulfur cluster at their center, enabling the enzyme to accomplish a challenging reaction by converting an allyl alcohol to two isoprene components. The iron-sulfur proteins normally participate in electron transfers. The IspH enzyme, thereby, in a similar fashion, binds the substrate directly to the iron-sulfur cluster [57]. In the template crystal structure of IspH (PDB 3KE8), it has been shown that His41, His74, His124, Thr167, Ser225, Ser226, Asn227 and Ser269 are the active site residues that are involved in hydrogen bond interactions with the ligand 4-hydroxy-3-methylbutyldiphosphate (EIP). Also, Cys12, Cys96, Cys197 and EIP have been shown to make metal interaction with the Fe_4S_4 (Iron/Sulfur Cluster). Although the ten selected drug-like compounds (Table 5**)** did not show any interaction with the aforementioned IspH residues, they are predicted to make very good hydrogen bond interactions with other surrounding residues of the predicted cavity. The predicted binding mode of the best scoring compound, **ZINC00510419** is shown in Additional file 2. Good shape complementarity and 6 hydrogen bond interactions are observed in this complex.

Cp1002_1648 (**TcsR,** Two component transcriptional regulator) is a novel target without host homologs proteins. Differently from MtrA and IspH, in this case the template structure from *Escherichia coli* for TcsR did not contain any ligand (PDB 1A04), and no reported information was found about the ligand-residues interactions in their cavities. Therefore, among the cavities identified by MVD, the best cavity for virtual screening analysis was simply chosen based on the highest druggability score by the DogSiteScorer. Compound **ZINC00510419** (Additional file 3**)** was the top-ranking compound, forming a network of 3 hydrogen bonds with Val76, Gln185 and Asn193. Table 6 lists the 10 compounds selected for this target.

Cp1002_1676 (**NrdI,** protein) belongs to the nrdI protein family, a unique group of metalloenzymes that are essential for cell-proliferation [66]. It is classified as a ribonucleotide reductase (RNR), an iron-dependent enzyme that belongs to class Oxidoreductases (EC 1.17.4.1) acting on CH or CH_2 groups with a disulfide as acceptor [67]. The class Ia enzyme supplies deoxynucleotides during normal aerobic growth. The class Ib RNR plays a similar role although its function in *E. coli* is not clear, but it is reported to be expressed under oxidative stress and iron-limited conditions [68]. Class I RNR enzymes have two homodimeric subunits, α2 (NrdE), where nucleotide reduction takes place, and β2 (NrdF) containing an unidentified metallocofactor for initiating nucleotide reduction in α2. Although the exact function of NrdI within RNR has not yet been fully characterized, it is found in the same operon as NrdE and NrdF, and encodes an unusual flavodoxin, a bacterial electron-transfer protein that includes a flavin mononucleotide that has been proposed to be involved in metallocofactor biosynthesis and/or maintenance. It has also been proposed that NrdI plays an important role in *E. coli* class Ib RNR cluster assembly. Recent *in vitro* studies have shown that a stable diferric-tyrosyl radical (FeIII2-Y·) and dimanganese (III)-Y· (MnIII2-Y·) cofactors are active in nucleotide reduction [69]. The first one can be formed by self-assembly from FeII and O_2 while the later cofactor can be generated from MnII-2-NrdF, but only in the presence of O_2 and NrdI protein [54, 69]. RNR is responsible for the *de novo* conversion of ribonucleoside diphosphates into deoxyribonucleoside diphosphates and it is essential for DNA synthesis and repair [70]. The active site residues of RNR, in the template structure of NrdI protein (PDB 3N3A), include Ser8, Ser9, Ser11, Ser48, Asn13, Asn83, Thr14, Tyr49, Ala89 and Gly91, all of which are involved in a hydrogen bond network with the cofactor flavin mononucleotide isoalloxazine ring (FMN, PDB 3N3A) [71]. Interestingly, two of these residues, Ser8 and Tyr49, were predicted to make hydrogen bonds with all 10 selected ligands (Table 7**)**. The interaction between the top scoring compound **ZINC01585114** (5-nitro-3, 4-diphenyl-2-furamide) and the residues from the predicted target cavities are shown in Additional file 4.

Furthermore, the drug-like molecule **ZINC00510419** (3,4-bis (5-methylisoxazole-3-carbonyl)-1,2,5-oxadiazole 2-oxide was among the top ten selected molecules for three of the pathogen target proteins, showing good H-bond interactions. It ranked first against the targets Cp1002_0742 (MolDock score = -151.376, no. of H-bonds = 7) and Cp1002_1648 (MolDock score = -167.633, no. of H-bonds = 3) and ranked fourth against the target Cp1002_1676 (MolDock score = -154.064, no. of H-bonds = 4).

Essential host homologous as putative targets

To compare the predicted EH protein targets to their host homologs, two approaches were taken. First, ClustalX (v2.1, http://www.clustal.org), a multiple sequence alignment program, was used to find different residues between bacterial and host proteins. As expected, a high percentage of residues was found to be conserved, but significant differences were also observed. Most percentage identities are between 35 and 50 (Table 8), except for fumarate hydratase, which shows 54% sequence identity to human and equine homologous proteins, but no hits in bovine and ovine proteomes.

Table 8 Percentage of sequence identity between *C.* pseudotuberculosis and host homologous proteins.

Protein Locus tag	Official full name	Percentage of Sequence Identity[#]			
		HS*	EC*	BT*	OA*
Cp1002_0385 **Adk**	Adenylate kinase	38	36	35	35
Cp1002_0692 **GapA**	Glyceraldehyde-3-phosphate dehydrogenase A	39	40	41	41
Cp1002_0728 **GlyA**	Serine hydroxymethyltransferase	43	45	45	45
Cp1002_0738 **FumC**	Fumaratehydratase class II	54	54	No Hits	No Hits
Cp1002_1005 **Gnd**	6-phosphogluconate dehydrogenase	48	48	48	48
Cp1002_1042 **AspA**	Aspartate ammonia-lyase	39	39	39	39

Next, to determine if the observed differences could be exploited in rational design of ligands selective to bacterial proteins, we focused on the predicted druggable cavities. A structural alignment to the host homologous proteins was performed and the cavities were compared in PyMol. In most cases, the DogSiteScorer predicted more than one cavity for each input Cp protein structure. The number of residues in the bacterial predicted cavity that differ from the residues in the cavity of the host protein, for all druggable pockets, varied from zero to seven (Table 9).

Table 9 Comparison of the residues from druggable cavities in *C. pseudotuberculosis* proteins and the corresponding residues in structurally aligned host protein cavities.

Protein Loci	Bacterial Residues for the Most Druggable Cavity Predicted by DGSS Server[#]	HS*	EC*	BT*	OA*
Cp1002_0692 (Glyceralderayde 3-phosphate dehydrogenase)	Lys157	Asp35	Asp33	Asp33	Asp33
	Val174	Thr52	Thr50	Thr50	Thr50
	Arg229	Thr103	Thr101	Thr101	Thr101
	Asn311	Ala183	Ala181	Ala181	Ala181
Cp1002_0385 (Adenylate kinase)	Phe35	Leu50	Leu52	Leu52	Leu43
	Ile53	Met68	Met70	Met70	Met61
	Thr64	Val79	Val81	Val81	Val72
Cp1002_0728 (Serine hydroxymethyltransferase)	Cys70	Ala88	Thr86	Thr86	Thr86
	Ala99	Ser121	Ser119	Ser119	Ser119
	Ala101	Ser123	Ser121	Ser121	Ser121
	Trp177	Thr204	Thr202	Thr202	Thr202
	Pro361	Ala397	Ala395	Ala395	Ala395
Cp1002_1005 (6-phosphogluconate dehydrogenase)	Ser55	Thr35	Thr161	Thr35	Thr35
	Met94	Leu74	Leu200	Leu74	Leu74
	Gln96	Lys76	Lys202	Lys76	Lys76
	Val104	Phe84	Phe210	Phe84	Phe84
	Ile148	Val128	Val254	Val128	Val128
	Gln268	Lys248	Lys374	Lys248	Lys248
	Pro269	His249	Tyr375	His249	His249
Cp1002_1042 **(Aspartate ammonia-lyase)**	Gln193	His235	His257	His235	His235
	Ile428	Lys470	Lys492	Lys470	Lys470
	His447	Leu489	Leu511	Leu489	Leu489

[#] Drug score ≥ 0.80
*HS = Homo sapiens, EC = Equus caballus, BT = Bos taurus, OA = Ovis aries

For conserved host-homologous targets Cp1002_0385 (adk, Adenylate kinase), Cp1002_0692 (gapA, Glyceraldehyde 3-phosphate dehydrogenase), Cp1002_0728 (glyA, Serine hydroxymethyltransferase), Cp1002_0738 (fumC, Fumarate hydratase class II/fumarase), Cp1002_1005 (gnd, 6-Phosphogluconate dehydrogenase) and Cp1002_1042 (aspA, Aspartate ammonia-lyase/aspartase), three, four, five, zero, seven and three different residues were observed, respectively. Then, a more detailed analysis was performed for the predicted highest druggable cavity for each protein. The results are described below, together with information about the biological importance of each target protein.

Cp1002_0692 (**GapA,** Glyceraldehyde 3-phosphate dehydrogenase, GAPDH/G3PDH, EC 1.2.1.12) catalyzes the sixth step of glycolysis. In addition, GAPDH has recently been shown to be involved in several non-metabolic processes, including transcription activation, initiation of apoptosis [72] fast axonal or axoplasmic transport and endoplasmic reticulum to Golgi vesicle shuttling [73, 74]. This enzyme has been reported as an anti-trypanosomatid and anti-leishmania drug target in structure-based drug design efforts [21–23]. Furthermore, it has been shown as an interesting putative drug and vaccine target in malaria pathogenesis [75]. Comparison of protein cavities reveals significant differences between bacterial and host proteins, with replacement of bacterial Lys157, Arg229 and Asn311 by Asp, Thr and Ala, respectively. Such differences result in a more basic cavity in bacteria, making it possible to rationally design selective ligands, especially negatively charged molecules, which interact with Lys157 and Arg229, or compounds able to form hydrogen bond to Asn311 (Additional file 5a).

Nucleoside monophosphate kinases vitally participate in sustaining the intracellular nucleotide pools in all living organisms. **Cp1002_0385** (**Adk,** Adenylate kinase, EC 2.7.4.3) is a ubiquitous enzyme, which catalyzes the reversible $Mg2^{+}$-dependent transfer of the terminal phosphate group from ATP to AMP, releasing two molecules of ADP [76]. Only one highly druggable cavity was predicted for adenylate kinase, with a druggability score = 0.81. Three residues in the bacteria cavity were different from the hosts: Leu, Met and Val in the hosts replaced Phe35, Ile53 and Thr64, respectively (Additional file 5b). These differences impact the cavity volume, since aromatic and bulky Phe is replaced by Leu, and the ability to make hydrogen bonds, through the replacement of a Thr by a Val. Therefore; the bacterial cavity is smaller and more hydrophilic, making it possible to envision rational design of selective ligands that interact with Thr64.

Cp1002_0728 (GlyA, Serine hydroxymethyltransferase EC 2.1.2.1) is an enzyme that plays an important role in cellular one-carbon pathways by catalyzing the reversible, simultaneous conversions of L-serine to glycine (retro-aldol cleavage) and tetrahydrofolate to 5,10-methylenetetrahydrofolate [77]. In Plasmodium, serine hydroxymethyltransferase (SHMT) has been reported as an attractive drug target [78]. For this protein 3 residues were observed different between bacteria and host: Ala99 and Ala101 replaced two Ser residues while Trp177 replaced Thr (Additional file 5c). At first glance these changes could have a big impact in the active site, generating a considerably more hydrophilic pocket in the hosts. However, careful inspection of the pocket reveals that the side chains of these residues are not turned towards the pocket, in such a way that these differences probably would not allow rational design of selective ligands.

Cp1002_0738 (**FumC,** Fumaratehydratase class II/fumarase EC 4.2.1.2) catalyzes the reversible hydration/dehydration of fumarate to S-malate during the ubiquitous Krebs cycle, through the aci-carboxylate intermediate subsequent to olefin production [79]. There are two classes of fumarases; Class I fumarases, composed of heat-labile, iron-sulfur (4Fe-4S) homodimeric enzymes, only found in prokaryotes; and Class II fumarases, made of thermostable homotetrameric enzymes [80] found in both prokaryotic and eukaryotic mitochondria. Class II belongs to a superfamily that also includes aspartate-ammonia lyases, arginino-succinatases, d-crystallins and 3-carboxy-cis, cis-muconate lactonizing enzymes. All these enzymes release fumarate from different substrates, ranging from adenylosuccinate to malate [81–84]. FumC of *Escherichia coli* is the first member of class II fumarases family whose structure has been solved and provided most of the structural information [85]. Inhibition of fumarase in the tricarboxylic acid cycle (TCA) has been reported as a potential molecular target of bismuth drugs in *Helicobacter pylori* [86]. Comparison of the active site cavity of this protein, which is formed in the interface of three monomers, revealed no differences between bacteria and hosts (additional file 5d).

Cp1002_1005 (**Gnd,** 6-Phosphogluconate dehydrogenase EC 1.1.1.44) is an enzyme from the pentose phosphate pathway. It forms ribulose 5-phosphate from 6-phosphogluconate. The enzyme 6-phosphogluconate dehydrogenase is a potential drug target for the parasitic protozoan *Trypanosoma brucei*, the causative organism of human African trypanosomiasis [87]. Three druggable sites with score > 0.80 were detected in this protein. As opposed to the observation for other proteins, the most druggable predicted cavity (score = 0.88) was not the active site. Leu, Lys and Val residues in the hosts replace residues Met94, Gln96 and Ile148 in the bacterial cavity, respectively (Additional file 5e). The most significant of these differences is the replacement of Gln by Lys, which could make binding of negative molecules more favorable to the host proteins.

Cp1002_1042 (**AspA,** Aspartate ammonia-lyase/aspartase EC 4.3.1.1) catalyzes the deamination of aspartic acid to form fumarate and ammonia [88]. Recent progresses to prepare enantiopure l-aspartic acid derivatives, highly valuable tools for biological research and chiral building blocks for pharmaceuticals and food additives, make it a target of interest for industrial applications. On the other hand, the important role that it plays in microbial nitrogen metabolism makes it a putative drug target in overcoming bacterial pathogenesis [89]. Based on the sequence alignment for this protein, two significant differences in residues are observed in the most druggable pocket: bacterial His447 and Ile428 are replaced by Leu and Lys in host proteins. Such differences should allow rational ligand design. It is interesting to note that additional differences in the position of helices that contain these residues increase the difference between the active sites (Additional file 5f).

Based on the above-mentioned analyses, we conclude that it would be difficult to rationally design selective ligands for **Cp1002_0738** (**FumC,** Fumaratehydratase class II), since no residue differences were observed in the most druggable cavity, and for **Cp1002_0728 (GlyA**, Serine hydroxymethyltransferase), where the side chains of differing residues are not turned toward the druggable pocket. On the other hand, for putative essential and homologous targets that include **Cp1002_0692** (**GapA,** Glyceraldehyde 3-phosphate dehydrogenase), **Cp1002_0385** (**Adk,** Adenylate kinase), **Cp1002_1005** (**Gnd,** 6-Phosphogluconate dehydrogenase) and **Cp1002_1042** (**AspA,** Aspartate ammonia-lyase), significant differences were observed in druggable pockets, suggesting that despite the existence of a host homologous protein they could be good targets for the design of ligands, selective only to the bacterial proteins.

Conclusion

Here, for the first time, the genomic information was used to determine the conserved predicted proteome of 15 strains of *C. pseudotuberculosis*, along with their three-dimensional structural information. Even though the structural information discussed is fully computationally predicted, and could therefore deviate from eventually solved experimental structures, we have been careful to concentrate on the analysis of protein models for which there were good templates which provided high quality models, minimizing this concern. The data presented here can effectively contribute in guiding further research for antibiotics and vaccines development. The final dataset can provide valuable information in designing molecular biology and immunization experiments in animal models for validating the targets of a pathogen, as well as in experimental structure determination protocols.

The criterion for target selection in *C. pseudotuberculosis* was stringent, resulting in a small set of prioritized putative drug and vaccine targets, of which four are essential and non-homologous and six are essential and host homologous proteins. For the latter, a detailed structural comparison between the residues of the predicted cavities of host and pathogen proteins has been performed, showing in most cases the potential for the development of selective ligands. Therefore, we suggest that the whole set can be considered for antimicrobial chemotherapy, especially the four essential non-host homologous targets.

The *in silico* approaches followed in this study might aid in the development of novel therapeutic drugs and vaccines in a broad-spectrum of hosts at intraspecies level against *C. pseudotuberculosis*. Furthermore, the strategy described here could also be applied to other pathogenic microorganisms.

Conflict of interest

The authors declare that they have no competing interests.

Misc

Authors' contributions

Coordinated entire work: SSH RSF VA DB. Performed all *in silico* analyses: SSH RSF ST SBJ NBS FDP LCG. Cross-analyzed genome contents, pan-modelome construction, conserved pan-modelome, subtractive modelome approach, virtual screening & docking analyses and residue level structural comparison: SSH RSF ST FDP AI SCS SA DB AGT. Provided timely consultation and reviewed the manuscript: VA AI SCS SA DB NBS LCG AA AM AS VACA AGT. Read and approved the final manuscript: RSF SSH ST AI SCS SBJ SA DB NBS LCG AGTAA AM AS VA. Conceived and designed the work: SSH RSF VA DB. Analyzed the data: SSH RSF ST AI SCS SBJ SA DB NBS LCG AA AB LJ AGTAM AS VA. Wrote the paper: SSH RSF ST.

Acknowledgements

We acknowledge financial support from the funding agencies CNPq, CAPES and FAPEMIG. Hassan S.S acknowledges the receipt of fellowship under "TWAS-CNPq Postgraduate Fellowship Program" for doctoral studies.

This article has been published as part of *BMC Genomics* Volume 15 Supplement 7, 2014: Proceedings of the 9th International Conference of the Brazilian Association for Bioinformatics and Computational Biology (X-Meeting 2013). The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/15/S7.

Electronic supplementary material

MediaObjects/12864_2014_7174_MOESM1_ESM.pdf
Additional file 1: Docking representation of the best drug-like compound **ZINC75109074** in the most druggable protein cavity of **Cp1002_0515** (**MtrA**, DNA-binding response regulator). Three hydrogen bonds were observed with Thr73, Asp48 and Arg116. (PDF 1 MB)

MediaObjects/12864_2014_7174_MOESM2_ESM.pdf
Additional file 2: Docking representation of compound **ZINC00510419** in the most druggable protein cavity of **Cp1002_0742** (**IspH**, 4-hydroxy-3-methyl but-2-enyl diphosphate reductase). Residues Cys39, Thr225, Ser250, His68 and Asn252 are predicted to make seven hydrogen bonds to this ligand. (PDF 932 KB)

MediaObjects/12864_2014_7174_MOESM3_ESM.pdf
Additional file 3: Docking representation of the best drug-like compound **ZINC00510419** in the most druggable protein cavity of **Cp1002_1648** (**TcsR,** Two component transcriptional regulator). Hydrogen bonds were observed with residues Val76, Gln185 and Asn193. (PDF 1 MB)

MediaObjects/12864_2014_7174_MOESM4_ESM.pdf
Additional file 4: Docking representation of the best drug-like compound **ZINC04721321** in the most druggable protein cavity of **Cp1002_1676** (**NrdI** protein). Hydrogen bonds were observed with residues Ser8, Thr13 and Leu116. (PDF 1 MB)

MediaObjects/12864_2014_7174_MOESM5_ESM.pdf
Additional file 5 (a-f): Comparison among the most druggable cavities from essential bacterial and the respective host homologue proteins. Protein structures are shown as cartoon (green for the bacterial protein and gray for *Ovis aries* host protein). Other host proteins are not shown for simplicity, but the same substitutions were present in all host proteins analyzed. Residues that differ in the bacterial and host cavity are highlighted in sticks and labeled (bacterial labels in green and host labels in black). a) **Cp1002_0692** (Glyceralderayde 3-phosphate dehydrogenase); b) **Cp1002_0385** (adenylate kinase); c) **Cp1002_0728** (serine hydroxymethyltransferase); d) **Cp1002_0738** (fumarate hydratase class II) the site shown is formed by three monomers, which are represented in green, blue and orange. No residues are highlighted, since the active sites are identical between bacteria and host; e) **Cp1002_1005** (6-phosphogluconate dehydrogenase); f) **Cp1002_1042** (aspartate ammonia-lyase). Figures were prepared with the PyMol. (PDF 4 MB)

References

1. Hassan SS, Schneider MP, Ramos RT, Carneiro AR, Ranieri A, Guimaraes LC, Ali A, Bakhtiar SM, Pereira Ude P, dos Santos AR, et al: Whole-genome sequence of *Corynebacterium pseudotuberculosis* strain Cp162, isolated from camel. Journal of bacteriology. 2012, 194 (20): 5718-5719. 10.1128/JB.01373-12.

2. Dorella FA, Pacheco LG, Oliveira SC, Miyoshi A, Azevedo V: *Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. Veterinary research. 2006, 37 (2): 201-218. 10.1051/vetres:2005056.

3. Soares SC, Trost E, Ramos RT, Carneiro AR, Santos AR, Pinto AC, Barbosa E, Aburjaile F, Ali A, Diniz CA, et al: Genome sequence of Corynebacterium pseudotuberculosis biovar equi strain 258 and prediction of antigenic targets to improve biotechnological vaccine production. Journal of biotechnology. 2012

4. Khamis A, Raoult D, La Scola B: Comparison between rpoB and 16S rRNA gene sequencing for molecular identification of 168 clinical isolates of Corynebacterium. Journal of clinical microbiology. 2005, 43 (4): 1934-1936. 10.1128/JCM.43.4.1934-1936.2005.

5. Williamson LH: Caseous lymphadenitis in small ruminants. Vet Clin North Am Food Anim Pract. 2001, 17 (2): 359-371. vii

6. Peel MM, Palmer GG, Stacpoole AM, Kerr TG: Human lymphadenitis due to Corynebacterium pseudotuberculosis: report of ten cases from Australia and review. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 1997, 24 (2): 185-191. 10.1093/clinids/24.2.185.

7. Luis MA, Lunetta AC: [Alcohol and drugs: preliminary survey of Brazilian nursing research]. Revista latino-americana de enfermagem. 2005, 13: Spec No:1219-1230

8. Mills AE, Mitchell RD, Lim EK: Corynebacterium pseudotuberculosis is a cause of human necrotising granulomatous lymphadenitis. Pathology. 1997, 29 (2): 231-233. 10.1080/00313029700169944.

9. Augustine JL, Renshaw HW: Survival of Corynebacterium pseudotuberculosis in axenic purulent exudate on common barnyard fomites. American journal of veterinary research. 1986, 47 (4): 713-715.

10. Yeruham I, Friedman S, Perl S, Elad D, Berkovich Y, Kalgard Y: A herd level analysis of a Corynebacterium pseudotuberculosis outbreak in a dairy cattle herd. Veterinary dermatology. 2004, 15 (5): 315-320. 10.1111/j.1365-3164.2004.00388.x.

11. Perumal D, Lim CS, Sakharkar KR, Sakharkar MK: Differential genome analyses of metabolic enzymes in Pseudomonas aeruginosa for drug target identification. In silico biology. 2007, 7 (4-5): 453-465.

12. Barh D, Gupta K, Jain N, Khatri G, Leon-Sicairos N, Canizalez-Roman A, Tiwari S, Verma A, Rahangdale S, Shah Hassan S, et al: Conserved host-pathogen PPIs. Integrative biology : quantitative biosciences from nano to macro. 2013

13. Pizza M, Scarlato V, Masignani V, Giuliani MM, Arico B, Comanducci M, Jennings GT, Baldi L, Bartolini E, Capecchi B, et al: Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. Science. 2000, 287 (5459): 1816-1820. 10.1126/science.287.5459.1816.

14. Asif SM, Asad A, Faizan A, Anjali MS, Arvind A, Neelesh K, Hirdesh K, Sanjay K: Dataset of potential targets for Mycobacterium tuberculosis H37Rv through comparative genome analysis. Bioinformation. 2009, 4 (6): 245-248. 10.6026/97320630004245.

15. Dutta A, Singh SK, Ghosh P, Mukherjee R, Mitter S, Bandyopadhyay D: In silico identification of potential therapeutic targets in the human pathogen Helicobacter pylori. In silico biology. 2006, 6 (1-2): 43-47.

16. Chong CE, Lim BS, Nathan S, Mohamed R: In silico analysis of Burkholderia pseudomallei genome sequence for potential drug targets. In silico biology. 2006, 6 (4): 341-346.

17. Barh D, Kumar A: In silico identification of candidate drug and vaccine targets from various pathways in Neisseria gonorrhoeae. In silico biology. 2009, 9 (4): 225-231.

18. Sakharkar KR, Sakharkar MK, Chow VT: A novel genomics approach for the identification of drug targets in pathogens, with special reference to Pseudomonas aeruginosa. In silico biology. 2004, 4 (3): 355-360.

19. Rathi B, Sarangi AN, Trivedi N: Genome subtraction for novel target definition in Salmonella typhi. Bioinformation. 2009, 4 (4): 143-150. 10.6026/97320630004143.

20. Barh D, Jain N, Tiwari S, Parida BP, D'Afonseca V, Li L, Ali A, Santos AR, Guimaraes LC, de Castro Soares S, et al: A novel comparative genomics analysis for common drug and vaccine targets in Corynebacterium pseudotuberculosis and other CMN group of human pathogens. Chemical biology & drug design. 2011, 78 (1): 73-84. 10.1111/j.1747-0285.2011.01118.x.

21. Aronov AM, Verlinde CL, Hol WG, Gelb MH: Selective tight binding inhibitors of trypanosomal glyceraldehyde-3-phosphate dehydrogenase via structure-based drug design. Journal of medicinal chemistry. 1998, 41 (24): 4790-4799. 10.1021/jm9802620.

22. Singh S, Malik BK, Sharma DK: Molecular modeling and docking analysis of Entamoeba histolytica glyceraldehyde-3 phosphate dehydrogenase, a potential target enzyme for anti-protozoal drug development. Chemical biology & drug design. 2008, 71 (6): 554-562. 10.1111/j.1747-0285.2008.00666.x.

23. Suresh S, Bressi JC, Kennedy KJ, Verlinde CL, Gelb MH, Hol WG: Conformational changes in Leishmania mexicana glyceraldehyde-3-phosphate dehydrogenase induced by designed inhibitors. Journal of molecular biology. 2001, 309 (2): 423-435. 10.1006/jmbi.2001.4588.

24. Adams CP, Brantner VV: Estimating the cost of new drug development: is it really 802 million dollars?. Health affairs. 2006, 25 (2): 420-428. 10.1377/hlthaff.25.2.420.

25. Kola I, Landis J: Can the pharmaceutical industry reduce attrition rates?. Nature reviews Drug discovery. 2004, 3 (8): 711-715. 10.1038/nrd1470.

26. Congreve M, Murray CW, Blundell TL: Structural biology and drug discovery. Drug discovery today. 2005, 10 (13): 895-907. 10.1016/S1359-6446(05)03484-7.

27. Baker D, Sali A: Protein structure prediction and structural genomics. Science. 2001, 294 (5540): 93-96. 10.1126/science.1065659.

28. Cavasotto CN, Phatak SS: Homology modeling in drug discovery: current trends and applications. Drug discovery today. 2009, 14 (13-14): 676-683. 10.1016/j.drudis.2009.04.006.

29. Behera DK, Behera PM, Acharya L, Dixit A, Padhi P: In silico biology of H1N1: molecular modelling of novel receptors and docking studies of inhibitors to reveal new insight in flu treatment. Journal of biomedicine & biotechnology. 2012, 2012: 714623-

30. Mural RJ: ARTEMIS: a tool for displaying and annotating DNA sequence. Briefings in bioinformatics. 2000, 1 (2): 199-200. 10.1093/bib/1.2.199.

31. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A: Comparative protein structure modeling using MODELLER. Current protocols in protein science / editorial board, John E Coligan [et al]. 2007, Chapter 2:Unit 2 9

32. Mount DW: Using the Basic Local Alignment Search Tool (BLAST). CSH protocols. 2007, 2007:pdb top17

33. Tusnady GE, Simon I: The HMMTOP transmembrane topology prediction server. Bioinformatics. 2001, 17 (9): 849-850. 10.1093/bioinformatics/17.9.849.

34. Laskowski RA, Macarthur MW, Moss DS, Thornton JM: Procheck - a Program to Check the Stereochemical Quality of Protein Structures. J Appl Crystallogr. 1993, 26: 283-291. 10.1107/S0021889892009944.

35. Capriles PV, Guimaraes AC, Otto TD, Miranda AB, Dardenne LE, Degrave WM: Structural modelling and comparative analysis of homologous, analogous and specific proteins from Trypanosoma cruzi versus Homo sapiens: putative drug targets for chagas' disease treatment. BMC genomics. 2010, 11: 610-10.1186/1471-2164-11-610.

36. Abadio AK, Kioshima ES, Teixeira MM, Martins NF, Maigret B, Felipe MS: Comparative genomics allowed the identification of drug targets against human fungal pathogens. BMC genomics. 2011, 12: 75-10.1186/1471-2164-12-75.

37. Zhang R, Ou HY, Zhang CT: DEG: a database of essential genes. Nucleic acids research. 2004, 32 (Database): D271-272.

38. Kanehisa M, Goto S: KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000, 28 (1): 27-30. 10.1093/nar/28.1.27.

39. Yoon SH, Park YK, Lee S, Choi D, Oh TK, Hur CG, Kim JF: Towards pathogenomics: a web-based resource for pathogenicity islands. Nucleic acids research. 2007, 35 (Database): D395-400. 10.1093/nar/gkl790.

40. Magrane M, Consortium U: UniProt Knowledgebase: a hub of integrated protein data. Database : the journal of biological databases and curation. 2011, 2011: bar009-

41. Yu CS, Lin CJ, Hwang JK: Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein science : a publication of the Protein Society. 2004, 13 (5): 1402-1406. 10.1110/ps.03479604.

42. Velankar S, Alhroub Y, Best C, Caboche S, Conroy MJ, Dana JM, Fernandez Montecelo MA, van Ginkel G, Golovin A, Gore SP, et al: PDBe: Protein Data Bank in Europe. Nucleic acids research. 2012, 40 (Database): D445-452.

43. Porter CT, Bartlett GJ, Thornton JM: The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic acids research. 2004, 32 (Database): D129-133.

44. Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J: CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic acids research. 2006, 34 (Web Server): W116-118. 10.1093/nar/gkl282.

45. Laurie AT, Jackson RM: Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics. 2005, 21 (9): 1908-1916. 10.1093/bioinformatics/bti315.

46. Aguero F, Al-Lazikani B, Aslett M, Berriman M, Buckner FS, Campbell RK, Carmona S, Carruthers IM, Chan AW, Chen F, et al: Genomic-scale prioritization of drug targets: the TDR Targets database. Nature reviews Drug discovery. 2008, 7 (11): 900-907. 10.1038/nrd2684.

47. Butt AM, Nasrullah I, Tahir S, Tong Y: Comparative genomics analysis of Mycobacterium ulcerans for the identification of putative essential genes and therapeutic candidates. PloS one. 2012, 7 (8): e43080-10.1371/journal.pone.0043080.

48. Volkamer A, Kuhn D, Rippmann F, Rarey M: DoGSiteScorer: a web server for automatic binding site prediction, analysis and druggability assessment. Bioinformatics. 2012, 28 (15): 2074-2075. 10.1093/bioinformatics/bts310.

49. Voigt JH, Bienfait B, Wang S, Nicklaus MC: Comparison of the NCI open database with seven large chemical structural databases. Journal of chemical information and computer sciences. 2001, 41 (3): 702-712.

50. Thomsen R, Christensen MH: MolDock: a new technique for high-accuracy molecular docking. Journal of medicinal chemistry. 2006, 49 (11): 3315-3321. 10.1021/jm051197e.

51. Hopkins AL, Groom CR: The druggable genome. Nature reviews Drug discovery. 2002, 1 (9): 727-730. 10.1038/nrd892.

52. Li Y, Zeng J, He ZG: Characterization of a functional C-terminus of the Mycobacterium tuberculosis MtrA responsible for both DNA binding and interaction with its two-component partner protein, MtrB. Journal of biochemistry. 2010, 148 (5): 549-556. 10.1093/jb/mvq082.

53. Cangelosi GA, Do JS, Freeman R, Bennett JG, Semret M, Behr MA: The two-component regulatory system mtrAB is required for morphotypic multidrug resistance in Mycobacterium avium. Antimicrobial agents and chemotherapy. 2006, 50 (2): 461-468. 10.1128/AAC.50.2.461-468.2006.

54. Cotruvo JA, Stubbe J: NrdI, a flavodoxin involved in maintenance of the diferric-tyrosyl radical cofactor in Escherichia coli class Ib ribonucleotide reductase. Proceedings of the National Academy of Sciences of the United States of America. 2008, 105 (38): 14383-14388. 10.1073/pnas.0807348105.

55. McAteer S, Coulson A, McLennan N, Masters M: The lytB gene of Escherichia coli is essential and specifies a product needed for isoprenoid biosynthesis. Journal of bacteriology. 2001, 183 (24): 7403-7407. 10.1128/JB.183.24.7403-7407.2001.

56. Eberl M, Hintz M, Reichenberg A, Kollas AK, Wiesner J, Jomaa H: Microbial isoprenoid biosynthesis and human gammadelta T cell activation. FEBS letters. 2003, 544 (1-3): 4-10. 10.1016/S0014-5793(03)00483-6.

57. Span I, Wang K, Wang W, Zhang Y, Bacher A, Eisenreich W, Li K, Schulz C, Oldfield E, Groll M: Discovery of acetylene hydratase activity of the iron-sulphur protein IspH. Nature communications. 2012, 3: 1042-

58. Plaimas K, Eils R, Konig R: Identifying essential genes in bacterial metabolic networks with machine learning methods. BMC systems biology. 2010, 4: 56-10.1186/1752-0509-4-56.

59. Vinayak S, Sharma YD: Inhibition of Plasmodium falciparum ispH (lytB) gene expression by hammerhead ribozyme. Oligonucleotides. 2007, 17 (2): 189-200. 10.1089/oli.2007.0075.

60. Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, Kaipa P, Karthikeyan AS, Kothari A, Krummenacker M, et al: The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic acids research. 2010, 38 (Database): D473-479. 10.1093/nar/gkp875.

61. Caffrey CR, Rohwer A, Oellien F, Marhofer RJ, Braschi S, Oliveira G, McKerrow JH, Selzer PM: A comparative chemogenomics strategy to predict potential drug targets in the metazoan pathogen, Schistosoma mansoni. PloS one. 2009, 4 (2): e4413-10.1371/journal.pone.0004413.

62. Crowther GJ, Shanmugam D, Carmona SJ, Doyle MA, Hertz-Fowler C, Berriman M, Nwaka S, Ralph SA, Roos DS, Van Voorhis WC, et al: Identification of attractive drug targets in neglected-disease pathogens using an in silico approach. PLoS neglected tropical diseases. 2010, 4 (8): e804-10.1371/journal.pntd.0000804.

63. Brocker M, Mack C, Bott M: Target genes, consensus binding site, and role of phosphorylation for the response regulator MtrA of Corynebacterium glutamicum. Journal of bacteriology. 2011, 193 (5): 1237-1249. 10.1128/JB.01032-10.

64. Zahrt TC, Deretic V: An essential two-component signal transduction system in Mycobacterium tuberculosis. Journal of bacteriology. 2000, 182 (13): 3832-3838. 10.1128/JB.182.13.3832-3838.2000.

65. Friedland N, Mack TR, Yu M, Hung LW, Terwilliger TC, Waldo GS, Stock AM: Domain orientation in the inactive response regulator Mycobacterium tuberculosis MtrA provides a barrier to activation. Biochemistry. 2007, 46 (23): 6733-6743. 10.1021/bi602546q.

66. Lammers M, Follmann H: The Ribonucleotide Reductases - a Unique Group of Metalloenzymes Essential for Cell-Proliferation. Struct Bond. 1983, 54: 27-91. 10.1007/BFb0111318.

67. Nordlund P, Reichard P: Ribonucleotide reductases. Annual review of biochemistry. 2006, 75: 681-706. 10.1146/annurev.biochem.75.103004.142443.

68. Monje-Casas F, Jurado J, Prieto-Alamo MJ, Holmgren A, Pueyo C: Expression analysis of the nrdHIEF operon from Escherichia coli. Conditions that trigger the transcript level in vivo. The Journal of biological chemistry. 2001, 276 (21): 18031-18037. 10.1074/jbc.M011728200.

69. Cotruvo JA, Stubbe J: An active dimanganese(III)-tyrosyl radical cofactor in Escherichia coli class Ib ribonucleotide reductase. Biochemistry. 2010, 49 (6): 1297-1309. 10.1021/bi902106n.

70. Elledge SJ, Zhou Z, Allen JB: Ribonucleotide reductase: regulation, regulation, regulation. Trends in biochemical sciences. 1992, 17 (3): 119-123. 10.1016/0968-0004(92)90249-9.

71. Boal AK, Cotruvo JA, Stubbe J, Rosenzweig AC: Structural basis for activation of class Ib ribonucleotide reductase. Science. 2010, 329 (5998): 1526-1530. 10.1126/science.1190187.

72. Tarze A, Deniaud A, Le Bras M, Maillier E, Molle D, Larochette N, Zamzami N, Jan G, Kroemer G, Brenner C: GAPDH, a novel regulator of the pro-apoptotic mitochondrial membrane permeabilization. Oncogene. 2007, 26 (18): 2606-2620. 10.1038/sj.onc.1210074.

73. Zala D, Hinckelmann MV, Yu H, Lyra da Cunha MM, Liot G, Cordelieres FP, Marco S, Saudou F: Vesicular glycolysis provides on-board energy for fast axonal transport. Cell. 2013, 152 (3): 479-491. 10.1016/j.cell.2012.12.029.

74. Bressi JC, Verlinde CL, Aronov AM, Shaw ML, Shin SS, Nguyen LN, Suresh S, Buckner FS, Van Voorhis WC, Kuntz ID, et al: Adenosine analogues as selective inhibitors of glyceraldehyde-3-phosphate dehydrogenase of Trypanosomatidae via structure-based drug design. Journal of medicinal chemistry. 2001, 44 (13): 2080-2093. 10.1021/jm000472o.

75. Pal-Bhowmick I, Andersen J, Srinivasan P, Narum DL, Bosch J, Miller LH: Binding of aldolase and glyceraldehyde-3-phosphate dehydrogenase to the cytoplasmic tails of Plasmodium falciparum merozoite duffy binding-like and reticulocyte homology ligands. mBio. 2012, 3 (5):

76. Bellinzoni M, Haouz A, Grana M, Munier-Lehmann H, Shepard W, Alzari PM: The crystal structure of Mycobacterium tuberculosis adenylate kinase in complex with two molecules of ADP and Mg2+ supports an associative mechanism for phosphoryl transfer. Protein science : a publication of the Protein Society. 2006, 15 (6): 1489-1493. 10.1110/ps.062163406.

77. Appaji Rao N, Ambili M, Jala VR, Subramanya HS, Savithri HS: Structure-function relationship in serine hydroxymethyltransferase. Biochimica et biophysica acta. 2003, 1647 (1-2): 24-29. 10.1016/S1570-9639(03)00043-8.

78. Sopitthummakhun K, Thongpanchang C, Vilaivan T, Yuthavong Y, Chaiyen P, Leartsakulpanich U: Plasmodium serine hydroxymethyltransferase as a potential anti-malarial target: inhibition studies using improved methods for enzyme production and assay. Malaria journal. 2012, 11: 194-10.1186/1475-2875-11-194.

79. Mechaly AE, Haouz A, Miras I, Barilone N, Weber P, Shepard W, Alzari PM, Bellinzoni M: Conformational changes upon ligand binding in the essential class II fumarase Rv1098c from Mycobacterium tuberculosis. FEBS letters. 2012, 586 (11): 1606-1611. 10.1016/j.febslet.2012.04.034.

80. Woods SA, Schwartzbach SD, Guest JR: Two biochemically distinct classes of fumarase in Escherichia coli. Biochimica et biophysica acta. 1988, 954 (1): 14-26.

81. Sampaleanu LM, Vallee F, Slingsby C, Howell PL: Structural studies of duck delta 1 and delta 2 crystallin suggest conformational changes occur during catalysis. Biochemistry. 2001, 40 (9): 2732-2742. 10.1021/bi002272k.

82. Yang J, Wang Y, Woolridge EM, Arora V, Petsko GA, Kozarich JW, Ringe D: Crystal structure of 3-carboxy-cis,cis-muconate lactonizing enzyme from Pseudomonas putida, a fumarase class II type cycloisomerase: enzyme evolution in parallel pathways. Biochemistry. 2004, 43 (32): 10424-10434. 10.1021/bi036205c.

83. Toth EA, Yeates TO: The structure of adenylosuccinate lyase, an enzyme with dual activity in the de novo purine biosynthetic pathway. Structure. 2000, 8 (2): 163-174. 10.1016/S0969-2126(00)00092-7.

84. Tsai M, Koo J, Yip P, Colman RF, Segall ML, Howell PL: Substrate and product complexes of Escherichia coli adenylosuccinate lyase provide new insights into the enzymatic mechanism. Journal of molecular biology. 2007, 370 (3): 541-554. 10.1016/j.jmb.2007.04.052.

85. Weaver TM, Levitt DG, Donnelly MI, Stevens PP, Banaszak LJ: The multisubunit active site of fumarase C from Escherichia coli. Nature structural biology. 1995, 2 (8): 654-662. 10.1038/nsb0895-654.

86. Chen Z, Zhou Q, Ge R: Inhibition of fumarase by bismuth(III): implications for the tricarboxylic acid cycle as a potential target of bismuth drugs in Helicobacter pylori. Biometals : an international journal on the role of metal ions in biology, biochemistry, and medicine. 2012, 25 (1): 95-102. 10.1007/s10534-011-9485-7.

87. Ruda GF, Campbell G, Alibu VP, Barrett MP, Brenk R, Gilbert IH: Virtual fragment screening for novel inhibitors of 6-phosphogluconate dehydrogenase. Bioorganic & medicinal chemistry. 2010, 18 (14): 5056-5062. 10.1016/j.bmc.2010.05.077.

88. Shi W, Dunbar J, Jayasekera MM, Viola RE, Farber GK: The structure of L-aspartate ammonia-lyase from Escherichia coli. Biochemistry. 1997, 36 (30): 9136-9144. 10.1021/bi9704515.

89. de Villiers M, Puthan Veetil V, Raj H, de Villiers J, Poelarends GJ: Catalytic mechanisms and biocatalytic applications of aspartate and methylaspartate ammonia lyases. ACS chemical biology. 2012, 7 (10): 1618-1628. 10.1021/cb3002792.

II.III.6 Label-free proteomic analysis to confirm the predicted proteome of *Corynebacterium pseudotuberculosis* under nitrosative stress mediated by nitric oxide.

Silva WM, Carvalho RD, Soares SC, Bastos IF, Folador EL, Souza GH, Le Loir Y, Miyoshi A, Silva A, **Azevedo V**.
BMC Genomics. 2014 Dec 4;15:1065. doi: 10.1186/1471-2164-15-1065.

Durante o processo de infecção *C. pseudotuberculosis* se depara com diferentes condições de estresse, incluindo o estresse nitrosativo que é causado pelo óxido nítrico. Analises *in silico* do genoma de *C. pseudotuberculosis* biovar Ovis linhagem 1002 demonstrou que esta linhagem possui genes que podem estar relacionados ao processo de resistência deste patógeno a este tipo de estresse. Com o objetivo de identificar proteínas que podem contribuir para este processo de resistência ao estresse nitrosativo, nós utilizamos *high-throughput proteomics* com a abordagem *Label-free proteomics* para caracterizar o genoma funcional da linhagem 1002_ovis, após exposição ao agente gerador de estresse *NO-donor Diethylenetriamine /nitricoxide adduct* (DETA/NO). A partir de nossa análise proteômica foram caracterizadas 835 proteínas, representando aproximadamente 41% do genoma predito da linhagem 1002_ovis. Quando comparado o proteoma das duas condições: (i) estresse e (ii) controle foram identificadas 102 e 34 proteínas exclusivas para a condição, respectivamente. Além disso, 58 proteínas foram diferencialmente expressas entre as duas condições. Para complementar nosso entendimento a cerca das proteínas relacionados a este processo de resistência ao estresse nitrosativo uma analise *protein-protein interaction* (PPI) também foi realizada. Finalmente, as proteínas diferencialmente expressas foram analisadas pelo software Blast2Go e foi observado à indução de proteínas relacionadas principalmente aos processos de detoxificação, regulação transcricional, síntese e reparo de DNA. Este estudo proteômico promoveu uma análise global do genoma da linhagem 1002_Ovis na presença do estresse nitrosativo, o que permitiu a identificação de várias proteínas que podem contribuir para resistência e sobrevivência deste patógeno, durante a exposição ao óxido nítrico

BMC Genomics

RESEARCH ARTICLE Open Access

Label-free proteomic analysis to confirm the predicted proteome of *Corynebacterium pseudotuberculosis* under nitrosative stress mediated by nitric oxide

Wanderson M Silva[1,4,5], Rodrigo D Carvalho[1], Siomar C Soares[1], Isabela FS Bastos[1], Edson L Folador[1], Gustavo HMF Souza[3], Yves Le Loir[4,5], Anderson Miyoshi[1], Artur Silva[2] and Vasco Azevedo[1*]

Abstract

Background: *Corynebacterium pseudotuberculosis* biovar *ovis* is a facultative intracellular pathogen, and the etiological agent of caseous lymphadenitis in small ruminants. During the infection process, the bacterium is subjected to several stress conditions, including nitrosative stress, which is caused by nitric oxide (NO). *In silico* analysis of the genome of *C. pseudotuberculosis ovis* 1002 predicted several genes that could influence the resistance of this pathogen to nitrosative stress. Here, we applied high-throughput proteomics using high definition mass spectrometry to characterize the functional genome of *C. pseudotuberculosis ovis* 1002 in the presence of NO-donor Diethylenetriamine/nitric oxide adduct (DETA/NO), with the aim of identifying proteins involved in nitrosative stress resistance.

Results: We characterized 835 proteins, representing approximately 41% of the predicted proteome of *C. pseudotuberculosis ovis* 1002, following exposure to nitrosative stress. In total, 102 proteins were exclusive to the proteome of DETA/NO-induced cells, and a further 58 proteins were differentially regulated between the DETA/NO and control conditions. An interactomic analysis of the differential proteome of *C. pseudotuberculosis* in response to nitrosative stress was also performed. Our proteomic data set suggested the activation of both a general stress response and a specific nitrosative stress response, as well as changes in proteins involved in cellular metabolism, detoxification, transcriptional regulation, and DNA synthesis and repair.

Conclusions: Our proteomic analysis validated previously-determined *in silico* data for *C. pseudotuberculosis ovis* 1002. In addition, proteomic screening performed in the presence of NO enabled the identification of a set of factors that can influence the resistance and survival of *C. pseudotuberculosis* during exposure to nitrosative stress.

Keywords: *Corynebacterium pseudotuberculosis*, Caseous lymphadenitis, Proteomics, Label-free proteomics, Nitrosative stress, Nitric oxide

Background

Corynebacterium pseudotuberculosis is a Gram-positive, facultative, intracellular pathogen belonging to the *Corynebacterium*, *Mycobacterium*, *Nocardia*, or CMN, group. This group belongs to the phylum Actinobacteria. The defining characteristics of the CMN group are a specific cell wall organization, consisting of peptidoglycan, arabinogalactan, and mycolic acids, and a high chromosomal G + C content [1]. *C. pseudotuberculosis ovis* is the etiological agent of the chronic infectious disease caseous lymphadenitis, which affects small ruminants worldwide. As a result, *C. pseudotuberculosis ovis* is responsible for significant economic losses in the goat and sheep industries, mainly stemming from decreased meat, wool, and milk production, reproductive disorders, and carcass contamination [1,2]. Bacterial factors that contribute to the virulence of *C. pseudotuberculosis* include phospholipase D [3], toxic cell wall lipids [4], and the iron transporter *fagABC* complex [5].

* Correspondence: vasco@icb.ufmg.br
[1] Depto de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
Full list of author information is available at the end of the article

In silico analysis of the genome of *C. pseudotuberculosis ovis* 1002 [6], as well as the pan-genome analysis of 15 other strains of *C. pseudotuberculosis* [7], identified genes involved in the response of this pathogen to different types of stress. Recently, the functional genome of *C. pseudotuberculosis ovis* 1002 was evaluated at the transcriptional level following exposure to different types of abiotic stress, including heat, osmotic, and acid stresses [8]. This allowed the characterization of several genes involved in distinct biological processes that favor the survival of the pathogen under the given stress condition.

However, during the infection process, *C. pseudotuberculosis* encounters nitrosative stress, caused by nitric oxide (NO), in the macrophage intracellular environment. A reactive nitrogen species (RNS) found in mammalian systems, NO is produced from L-arginine by NO synthases (NOS), and is present in three isoforms: endothelial NOS, neuronal NOS, involved in blood pressure control and neural signaling, and inducible NOS, associated with host defenses [9,10]. The NO produced during bacterial infection has antimicrobial properties, killing pathogens by causing damage to DNA, RNA, and proteins [11]. However, several pathogens contain pathways that allow bacterial survival under nitrosative stress conditions, including NO-sensitive transcriptional regulators [12], DNA and protein repair systems [13], and antioxidant systems [14].

Currently, little is known about the factors involved in the resistance of *C. pseudotuberculosis* to nitrosative stress. Pacheco et al. [15] showed that the alternative sigma (σ) factor, σ^E, plays a role in the survival of *C. pseudotuberculosis* in the presence of RNS. A σ^E null strain showed increased susceptibility to nitric oxide compared with the wild-type, and, in an *in vivo* assay, was unable to persist in mice. However, in iNOS-deficient mice, the mutant strain maintained its virulence [15]. In the same study, the extracellular proteome of *C. pseudotuberculosis* was analyzed in response to nitrosative stress, allowing the characterization of proteins that contribute to the adaptive processes of the pathogen in this environment [15].

To complement the results obtained in previous studies, and to identify factors involved in the survival of *C. pseudotuberculosis* under nitrosative stress conditions, we applied high-throughput proteomics using an liquid chromatograph high definition mass spectrometry (LC-$HDMS^E$) (data-independent acquisition, in ion mobility mode) approach to evaluate the global expression of the functional genome of *C. pseudotuberculosis ovis* 1002 at the protein level under nitrosative stress conditions.

Methods

Bacterial strain and growth conditions

C. pseudotuberculosis biovar *ovis* strain 1002, isolated from a goat, was maintained in brain heart infusion broth (BHI; HiMedia Laboratories Pvt. Ltd., Mumbai, India) at 37°C. For stress-resistance assays, strain 1002 was cultivated in a chemically-defined medium (CDM), containing $Na_2HPO_4.7H_2O$ (12.93 g/l), KH_2PO_4 (2.55 g/l), NH_4Cl (1 g/l), $MgSO_4.7H_2O$ (0.20 g/l), $CaCl_2$ (0.02 g/l), 0.05% (v/v) Tween 80, 4% (v/v) MEM vitamin solution (Invitrogen, Gaithersburg, MD, USA), 1% (v/v) MEM amino acid solution (Invitrogen), 1% (v/v) MEM non-essential amino acid solution (Invitrogen), and 1.2% (w/v) glucose, at 37°C [16].

Nitric oxide assay and preparation of whole bacterial lysates

Diethylenetriamine/nitric oxide adduct (DETA/NO) resistance of *C. pseudotuberculosis* was characterized as previously described [15]. When strain 1002 reached exponential growth phase ($OD_{600} = 0.6$) in the chemically-defined medium, the culture was divided into two aliquots (control condition, strain 1002_Ct; NO exposure, strain 1002_*DETA/NO*), and DETA/NO was added to the appropriate aliquot to a final concentration of 0.5 mM. The growth of strain 1002 in the presence of DETA/NO was then evaluated for 10 h. For proteomic analysis, protein was extracted after 1 h of exposure to DETA/NO. Both the control and DETA/NO cultures were centrifuged at 4,000 × *g* for 10 min at 4°C. The cell pellets were washed in phosphate buffered saline and then resuspended in 1 ml of lysis buffer (7 M urea, 2 M thiourea, 4% (w/v) CHAPS, and 1 M dithiothreitol (DTT)). The cells were then sonicated using five 1-min cycles on ice. The resulting lysates were centrifuged at 14,000 × *g* for 30 min at 4°C. The extracted proteins were then submitted to centrifugation at 13,000 × *g* for 10 min using a spin column with a threshold of 10 kDa (Millipore, Billerica, USA). Proteins were denatured with (0.1% (w/v) *Rapi*GEST SF surfactant at 60°C for 15 min (Waters, Milford, CA, USA), reduced using 10 mM DTT for 30 min at 60°C, and alkylated with 10 mM iodoacetamide in a dark chamber at 25°C for 30 min. Next, the proteins were enzymatically digested with 1:50 (w/w) trypsin at 37°C for 16 hours (sequencing grade modified trypsin; Promega, Madison, WI, USA). The digestion process was stopped by adding 10 μl of 5% (v/v) Trifluoroacetic acid (TFA) (Fluka, Buchs, Germany). Glycogen phosphorylase was added to the digests to a final concentration of 20 fmol/μl as an internal standard for normalization prior to each replicate injection. Analysis was carried out using a two-dimensional reversed phase (2D RP-RP) nanoUPLC-MS (Nano Ultra Performance Liquid Chromatography) approach, using multiplexed $HDMS^E$ label-free quantitation as described previously [17].

LC-$HDMS^E$ analysis and data processing

Qualitative and quantitative by 2D nanoUPLC tandem nanoESI-$HDMS^E$ (Nano Electrospray High Definition Mass Spectrometry) experiments were conducted using a 1-h reversed phase (RP) acetonitrile (0.1% v/v formic acid) gradient (7–40% (v/v)) at 500 nl/min on a nanoACQUITY UPLC 2D RP × RP Technology system [18]. A nanoACQUITY UPLC High Strength Silica (HSS) T3 1.8 μm 75 μm × 15 cm column (pH 3) was used in conjunction with a RP XBridge BEH130 C18 5 μm 300 μm × 50 mm nanoflow column (pH 10). Typical on-column sample loads were 250 ng of the total protein digests for each of the five fractions (250 ng/ fraction/load). For all measurements, the mass spectrometer was operated in resolution mode, with a typical effective *m/z* conjoined ion-mobility resolving power of at least 1.5 M FWHM, an ion mobility cell filled with nitrogen gas, and a cross-section resolving power at least 40 Ω/ΔΩ. All analyses were performed using nanoelectrospray ionization in the positive ion mode nanoESI (+), and a NanoLockSpray (Waters) ionization source. The lock mass channel was sampled every 30 s. The mass spectrometer was calibrated with a MS/MS spectrum of [Glu^1]-fibrinopeptide B (Glu-Fib) human solution (100 fmol/μl) delivered though the reference sprayer of the NanoLockSpray source. The double-charged ion ($[M + 2H]^{2+} = 785.8426$) was used for initial single-point calibration, and MS/MS fragment ions of Glu-Fib were used to obtain the final instrument calibration. Multiplexed data-independent scanning with added specificity and selectivity of a non-linear "T-wave" ion mobility (HDMS^E) experiments were performed using a Synapt G2-S HDMS mass spectrometer (Waters). The mass spectrometer was set to switch automatically between standard MS (3 eV) and elevated collision energies HDMS^E (19–45 eV) applied to the transfer "T-wave" collision-induced dissociation cell with argon gas. The trap collision cell was adjusted for 1 eV using a millisecond scan time adjusted based on the linear velocity of the chromatography peak delivered though nanoACQUITY UPLC, to obtain a minimum of 20 scan points for each single peak at both low-energy and high-energy transmission, followed by an orthogonal acceleration time-of-flight from 50–2000 *m/z*. The radio frequency (RF) offset (MS profile) was adjusted so that the nanoUPLC-HDMS^E data were effectively acquired from an *m/z* range of 400–2000, which ensured that any masses observed in the high energy spectra of less than 400 *m/z* arose from dissociations in the collision cell.

Data processing

Protein identification and quantitative data packaging were generated using dedicated algorithms [19,20], and by searching against a *C. pseudotuberculosis* database with default parameters for ion accounting [21]. The databases were reversed "on-the fly" during the database query searches, and appended to the original database to assess the false positive rate of identification. For proper processing of spectra and database searching conditions, ProteinLynxGlobalServer v.2.5.2 (PLGS) with Identity^E and Expression^E informatics v.2.5.2 (Waters) were used. UniProtKB (release 2013_01) with manually-reviewed annotations was also used, and the search conditions were based on taxonomy (*C. pseudotuberculosis*), maximum missed cleavages by trypsin allowed up to one, and variable carbamidomethyl, acetyl N-terminal, phosphoryl, and oxidation (M) modifications [21,22]. The Identity^E algorithm with Hi3 methodology was used for protein quantitation. The search threshold for accepting each individual spectrum was set to the default value, with a false-positive value of 4%. Biological variability was addressed by analyzing each culture three times. Normalization was performed using the Expression^E tool with a housekeeping protein that showed no significant difference in abundance across all injections. The proteins obtained were organized by the PLGS Expression^E tool algorithm into a statistically significant list corresponding to increased and decreased regulation ratios among the different groups. The quantitation values were averaged over all of the samples, and the quoted standard deviations at $p \leq 0.05$ in the Expression^E software refer to the differences between biological replicates. Only proteins with a differential expression $\log_2$ ratio between the two conditions greater than or equal to 1.2 were considered [23].

Bioinformatics analysis

The identified proteins were analyzed using the prediction tools SurfG+ v1.0 [24], to predict sub-cellular localization, and Blast2GO, to predict gene ontology functional annotations [25]. The PIPS software predicted proteins present in pathogenicity islands [26]. The protein-protein interaction network was constructed using interolog mapping methodology and metrics according to Rezende et al. [27]. A preview of the interaction network was generated using Cytoscape version 2.8.3 [28], with a spring-embedded layout. CMRegNet was used to predict gene regulatory networks [29].

Results

Effects of nitric oxide on the growth of *C. pseudotuberculosis*

In this study, we examined the exponential growth of *C. pseudotuberculosis* strain 1002 under nitrosative stress. The growth and cell viability of strain 1002 was monitored for 10 h with and without DETA/NO supplementation (Figure 1). The control culture reached stationary phase by 5 h post-inoculation, while the culture containing DETA/NO did not reach stationary phase until approximately 10 h post-inoculation. However, these results showed that although DETA/NO (0.5 mM) affected the growth rate, *C. pseudotuberculosis* likely contains factors that promote survival in the presence of RNS.

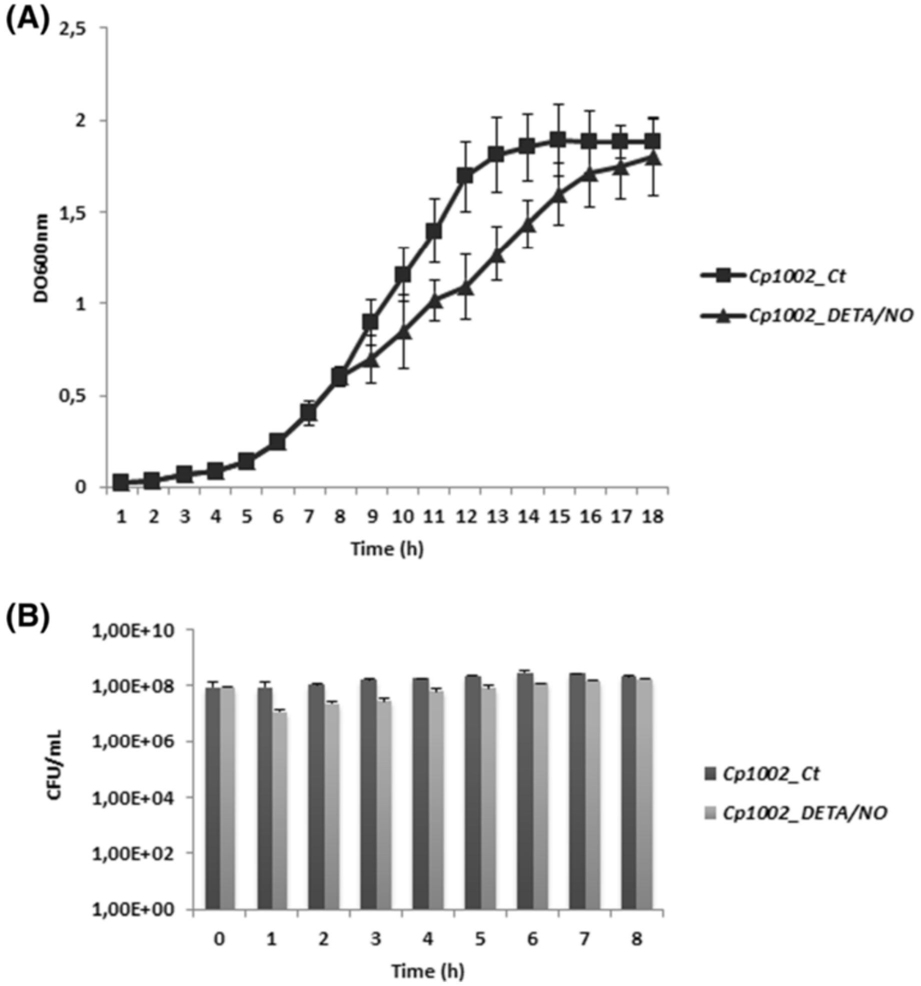


Figure 1 Growth and survival profile of *C. pseudotuberculosis* during NO exposure. (A) Growth of *C. pseudotuberculosis* after 10 h exposure to 0.5 mM DETA/NO. **(B)** Survival of *C. pseudotuberculosis* evaluated by colony forming units. The results shown in A and B represent an average of three independent experiments.

Label-free proteomic analysis of *C. pseudotuberculosis* grown under nitrosative stress conditions

Total proteome digests from three biological replicates of each individual condition were subjected to LC/MS^E. In total, we identified more than 31,000 peptides, with a normal distribution of 10 ppm error of the total identified peptides. Peptides as source fragments, peptides with a charge state of at least $[M + 2H]^{2+}$, and the absence of decoys were factors considered to increase data quality. A combined total of 2,063 proteins were present in at least two of the three biological replicates for the two conditions tested, with an average of 15 peptides per protein, and a false discovery rate (FDR) of 0% when decoy detection was set at agreement of two out of three replicates. The proteins referred to as exclusive to one condition or another was only identified in one condition within the detection limits of the experiment (LOD). The dynamic range of the quantified proteins is about 3 logs, and proteins unique to one condition or another were only observed above the LOD of the experiment, which was determined by the sample normalization prior to injection. Therefore, in our study, all samples were normalized using "scouting runs" taking into account the stoichiometry between the intensity and molarity proportion prior to the replicate runs per condition. The dynamic range was similar for each sample, and the total amount of sample used in fmol was nearly the same. We generate a graph of protein amounts of the identified proteins from all samples against protein ranks (Figure 2A).

After, analysis by PLGS v2.5.2 software, the 2,063 proteins originally identified in two out of three replicates were narrowed down to 699 proteins with $p \leq 0.05$. Among these proteins, 44 were up-regulated in the presence of DETA/NO, while 14 proteins were down-regulated (Table 1, Figure 2B and C). The remaining 641 proteins with $p \leq 0.05$ and $\log_2 < 1.2$ that were common to

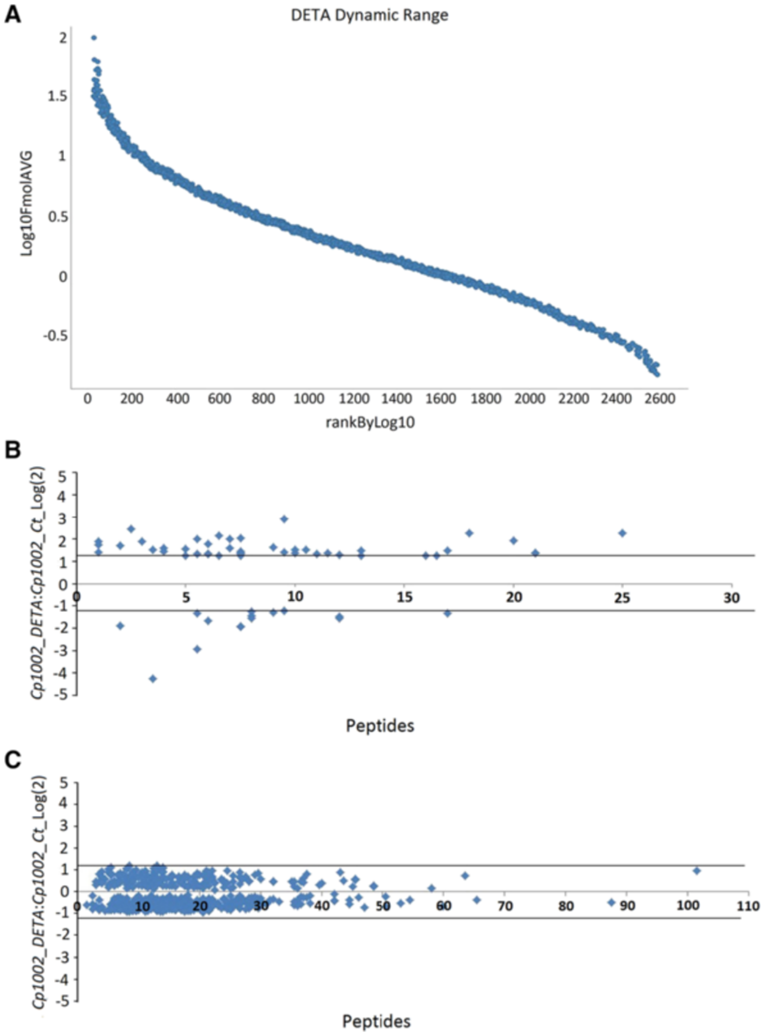


Figure 2 2D nanoUPLC HDMSE analysis showing: (A) Dynamic range of the method based on protein abundance estimates, data points derived from LC-HDMS^E analysis. (B and C) Proteins that were significantly differentially-regulated during NO exposure. The distribution of identified proteins with $p < 0.05$, and differentially-regulated proteins with an I:C $\log_2$ ratio < 1.2 in relation to the number of peptides identified for each protein. **(B)** Proteins with $p < 0.05$ and an I:C $\log_2$ ratio < 1.2. **(C)** Proteins with $p < 0.05$ and an I:C $\log_2$ ratio > 1.2.

the two treatments are summarized in Additional file 1. In addition to the 699 identified proteins that were present under both control and stress conditions, 34 proteins were exclusively expressed under the control conditions, and 102 proteins were exclusively expressed in response to DETA/NO stress (Additional files 2 and 3). Thus, our final list of proteins is composed of 835 proteins from *C. pseudotuberculosis*.

In silico analysis of LC-HDMS^E data

The 835 proteins were then analyzed using the SurfG+ tool to predict sub-cellular localization. According with SurfG+, our data set included approximately 41% of the predicted proteome of strain 1002 (Figure 3A). In addition, we characterized proteins belonging to the following cell fractions: cytoplasmic (CYT) (668 proteins), membrane (MEM) (59 proteins), potentially surface-exposed (PSE) (69 proteins), and secreted (SEC) (39 proteins) (Figure 3B).

To evaluate whether the proteins identified in our proteomic analysis could represent a protein set expressed by *C. pseudotuberculosis* during exposure to nitrosative stress, we correlated our proteomic data with the predicted core-genomes of 15 *C. pseudotuberculosis* strains [7]. Of the open reading frames (ORFs) coding for the differentially-regulated proteins and exclusive proteome of DETA/NO-exposed cells, 86% (50/58 proteins) and 82% (84/102 proteins) were identified, respectively, in the core-genome of *C. pseudotuberculosis* (Figure 3C and D). In addition, of the 835 total proteins identified from the proteome of strain 1002 following exposure to nitrosative stress, 83% (696 proteins) of the ORFs coding for these proteins were present in the core-genome of *C. pseudotuberculosis*, this result correspond approximately 46% of the predicted core-genome of *C. pseudotuberculosis* (Figure 3E).

Functional classification of the proteome of *C. pseudotuberculosis* expressed under exposure to nitrosative stress

The strain 1002 proteome was functionally classified using the Blast2Go tool [24]. A large proportion of the differentially-regulated proteins and those exclusive to one condition were identified as hypothetical proteins. According to the biological function prediction, 18 biological processes were classified as differentially regulated (Figure 4A). In addition, the analysis of the exclusive proteome of each condition revealed 12 common processes between the control and stress conditions (Figure 4B). However, seven biological processes were identified only in stress-exposed cells. These processes were antibiotic metabolism (six proteins), nucleotide metabolism (five proteins), oxidative phosphorylation (three proteins), translation (three proteins), glycolysis pathways (one protein), iron-sulfur clusters (one protein), and starch and sucrose metabolism (one protein). Among all processes identified, DNA synthesis and repair proteins (14 proteins) were most common. An overview of the *C. pseudotuberculosis* response to nitrosative stress according with the proteins identified is shown in Figure 5.

The proteins that were grouped into of transcriptional process were evaluated by CMRegNet and among regulators identified; we identified the GntR- family regulatory protein (D9Q5B7_CORP1), genes regulated by GntR-type regulators are usually involved in carbohydrate metabolism. The CMRegNet analysis showed that of the four genes under the control of this regulator, the N-acetylglucosamine kinase (D9Q5B6_CORP1) protein was highly expressed by *C. pseudotuberculosis* in response to DETA/NO. We identified other regulator the LexA repressor (D9Q8W2_CORP1) that was down regulated in the DETA/NO condition. According with CMRegNet, two proteins regulated by this repressor were detected in the DETA/NO proteome specific, pyridoxal biosynthesis lyase (PdxS; D9Q5T9_CORP1) and DNA translocase (D9Q8Z6_CORP1). Others proteins under the control of this repressor was detected, however not presented significant differential regulation like RecA protein

Protein-protein interaction network

To investigate the interactions among the proteins identified as exclusive and differentially regulated in cells exposed to DETA/NO, we generated a protein interaction network using Cytoscape. The interactome analysis revealed 67 protein-protein interactions (Figure 6). DnaB/DNA helicase (D9Q578_CORP1), identified in the exclusive proteome for strain 1002_*DETA/NO*, and PyrE/orotate phosphoribosyltransferase (D9Q4S2_CORP1), which was down-regulated in strain 1002_*DETA/NO*, showed the greatest number of interactions with other proteins (eight interactions each). Moreover, both of these proteins interact with proteins that are involved in metabolic processes, DNA processes, antibiotic metabolism, cell cycling, and translation.

Discussion

C. pseudotuberculosis is exposed to different forms of oxidative and nitrosative stress during the infection process. A previous study showed that *C. pseudotuberculosis* resists nitrosative stress generated by the NO-donor DETA/NO, and that a low concentration of DETA/NO (100 μM) induces a change in the extracellular proteome this pathogen [15]. To better understand the physiology of *C. pseudotuberculosis* in response to nitrosative stress, we analyzed the proteome of whole bacterial lysates of *C. pseudotuberculosis* in response to exposure to DETA/NO (0.5 mM).

The strain 1002 proteome under nitrosative stress reveals proteins involved in bacterial defense against DNA damage

Proteomic analysis identified proteins involved in DNA repair systems in both the exclusive proteome of DETA/NO-exposed cells and in the differentially-regulated proteome. We detected the proteins formamidopyrimidine-DNA glycosylase (Fpg) (D9Q598_CORP1), RecB (D9Q8C9_CORP1), and methylated-DNA-protein-cysteine methyltransferase (Ada) (D9Q923_CORP1), the genes for which were previously identified in a transcriptome analysis of strain 1002 in response to different abiotic stresses [8]. Activation of these proteins in response to nitrosative stress confirms that they belong a group of general stress-response proteins in *C. pseudotuberculosis*.

The expression of Fpg was up-regulated in response to acid stress [8]. We also identified endonuclease III (Endo III) (D9Q615_CORP1), which, in addition to Fpg, is involved in the base excision repair (BER) system of various bacteria. This system cleaves N-glycosidic bonds from damaged bases, allowing their excision and replacement. In *Salmonella enterica* serovar Typhimurium, the BER system repairs DNA damaged by exposure to NO. In addition, an *S.* Typhimurium strain defective in Fpg demonstrated reduced virulence in a murine model [30]. Our interactome analysis showed that Endo III had one of the highest numbers of interactions with other proteins, including interactions with proteins involved in DNA replication such zinc metalloprotease (D9Q378_CORP1) and DNA translocase (D9Q8Z6_CORP1), suggesting that this protein could play an important role in the defense pathway against RNS.

The Ada and RecB protein were up-regulated in response to osmotic stress [8]. Ada is involved in the repair of DNA-methylation damage, this protein have plays important in the pathway DNA damage [31]. RecB is a component of the RecBC system, which is part of

Table 1 Proteins identified as differentially-expressed following exposure to nitrosative stress

Uniprot access	Proteins	Score	Peptides	$\log_2$ DETA: CT[a]	*p*-value[a]	Subcellular localization[c]	Gene name	Genome[b]
Transport								
F9Y2Z3_CORP1	Cell wall channel	5321.88	4	1.42	1	CYT	*porH*	Shared
Cell division								
D9Q7G2_CORP1	Hypothetical protein	2417.8	21	1.34	1	CYT	*Cp1002_0716*	Core
DNA synthesis and repair								
D9Q5V6_CORP1	Nucleoid-associated protein	2327.08	5	1.52	1	CYT	*ybaB*	Core
D9Q923_CORP1	Methylated-DNA-protein-cysteine methyltransferase	6332.83	8	1.22	1	CYT	*ada*	Core
D9Q4P0_CORP1	7,8-dihydro-8-oxoguanine-triphosphatase	1640.23	8	−1.97	0	CYT	*mutT*	Core
Transcription								
D9Q8W2_CORP1	LexA repressor	800.31	6	−1.37	0.04	CYT	*lexA*	Shared
D9Q5L4_CORP1	ECF family sigma factor k	364.82	8	−1.58	0	CYT	*sigK*	Core
Translation								
D9Q753_CORP1	Fkbp-type peptidyl-prolyl cis-trans isomerase	7113.34	3	2.43	1	CYT	*fkbP*	Core
D9Q830_CORP1	50S ribosomal protein L35	2271.66	1	1.36	1	CYT	*rpmI*	Core
D9Q7W1_CORP1	Aspartyl glutamyl-tRNA amidotransferase subunit C	3100.8	7	1.24	0.99	CYT	*gatC*	Core
D9Q582_CORP1	50S ribosomal protein L9	41082.46	10	−1.25	0	CYT	*rplI*	
D9Q6H6_CORP1	30S ribosomal protein S8	45333.23	9	−1.34	0	CYT	*rpsH*	Core
Cell communication								
D9Q559_CORP1	Hypothetical protein	1402.27	6	1.99	1	PSE	*Cp1002_2005*	Core
D9Q5U9_CORP1	Thermosensitive gluconokinase	2068.35	7	1.96	0.99	CYT	*gntK*	Core
D9Q668_CORP1	Sensory transduction protein RegX3	2540.92	13	1.45	1	CYT	*regX3*	Core
Detoxification								
D9Q7U6_CORP1	Thioredoxin	1835.7	11	1.50	1	CYT	*trxA*	Core
D9Q4E5_CORP1	Glutathione peroxidase	1426.27	10	1.47	1	CYT	*Cp1002_1731*	Core
D9Q5T5_CORP1	Glyoxalase bleomycin resistance protein dihydroxybiphenyl dioxygenase	2417.77	11	1.28	1	CYT	*Cp1002_0124*	Shared
D9Q5N2_CORP1	NADH dehydrogenase	7030.94	12	1.25	1	CYT	*noxC*	Shared
D9Q680_CORP1	Glutaredoxin-like domain protein	292.69	2	−1.91	0	CYT	*Cp1002_0272*	Core
Glycolysis pathways								
D9Q5B6_CORP1	N-Acetylglucosamine kinase	228.69	6	1.74	0.98	CYT	*nanK*	Core
D9Q4U9_CORP1	Alcohol dehydrogenase	236.02	17	1.22	1	CYT	*adhA*	Shared
Iron-sulfur clusters								
D9Q7L6_CORP1	Ferredoxin	36927.57	7	2.10	1	CYT	*fdxA*	Core
Antibiotic resistance								
D9Q827_CORP1	Metallo-beta-lactamase superfamily protein	657.33	6	−2.95	0	CYT	*Cp1002_0937*	Core

Table 1 Proteins identified as differentially-expressed following exposure to nitrosative stress *(Continued)*

Amino acid metabolism								
D9Q622_CORP1	Phosphoserine phosphatase	949.15	9	1.58	0.99	PSE	*serB*	Core
D9Q4N1_CORP1	Carboxylate-amine ligase	205.54	16	1.24	1	CYT	*Cp1002_1819*	Core
D9Q6H4_CORP1	L-serine dehydratase I	284.11	17	−1.37	0	MEM	*sdaA*	Core
Lipid metabolism								
D9Q520_CORP1	Glycerophosphoryl diester phosphodiesterase	2417.8	21	1.34	1	PSE	*glpQ*	Core
Oxidative phosphorylation								
D9Q8I5_CORP1	Cytochrome aa3 controlling protein	676.2	6	1.28	1	MEM	*Cp1002_1095*	Core
Specific metabolic pathways								
D9Q5M9_CORP1	Inositol-3-phosphate synthase	7473.38	18	2.25	1	CYT	*ino1*	Core
D9Q721_CORP1	Hypothetical protein	4602.9	17	1.44	1	SEC	*Cp1002_0573*	Core
D9Q689_CORP1	3-Hydroxyisobutyrate dehydrogenase	2137.24	12	1.34	1	CYT	*mmsB*	Core
D9Q4X1_CORP1	Urease accessory protein UreG	1532.39	12	−1.6	0	CYT	*ureG*	Core
Nucleotide metabolism								
D9Q4S2_CORP1	Orotate phosphoribosyltransferase	2618.52	8	−1.26	0	CYT	*pyrE*	Core
Unknown function								
D9Q6Y9_CORP1	Hypothetical protein	491.89	10	2.87	1	CYT	*Cp1002_0540*	Core
D9Q6C7_CORP1	Hypothetical protein	689.6	25	2.25	1	PSE	*Cp1002_0320*	Core
D9Q3P3_CORP1	Hypothetical protein	5703.38	3	1.87	1	CYT	*Cp1002_1474*	Core
D9Q5V4_CORP1	Hypothetical protein	994.52	1	1.7	1	CYT	*Cp1002_0143*	Core
D9Q610_CORP1	Hypothetical protein	27217.36	2	1.67	1	CYT	*Cp1002_0202*	Core
D9Q8D8_CORP1	Hypothetical protein	2324.12	7	1.57	0.98	CYT	*Cp1002_1048*	Shared
D9Q6W1_CORP1	Hypothetical protein	9303.91	4	1.54	1	CYT	*Cp1002_0512*	Core
D9Q6V5_CORP1	Hypothetical protein	1346.2	4	1.5	0.99	CYT	*Cp1002_0506*	Core
D9Q5R7_CORP1	Hypothetical protein	2090.7	8	1.42	1	CYT	*Cp1002_0105*	Core
D9Q917_CORP1	Hypothetical protein	555.89	10	1.37	1	PSE	*Cp1002_1281*	Core
D9Q3P5_CORP1	Hypothetical protein	1121.7	6	1.29	1	SEC	*Cp1002_1476*	Core
D9Q7U5_CORP1	Hypothetical protein	517.06	8	1.28	1	CYT	*Cp1002_0852*	Core
D9Q7L1_CORP1	Hypothetical protein	15693.97	6	1.28	1	SEC	*Cp1002_0766*	Core
D9Q3P6_CORP1	Hypothetical protein	1729.59	5	1.22	1	CYT	*Cp1002_1477*	Core
D9Q6Z7_CORP1	Hypothetical protein	1835.7	13	1.22	1	CYT	*Cp1002_0548*	Core
D9Q8V8_CORP1	Hypothetical protein	293.23	8	−1.48	0		*Cp1002_1221*	Core
D9Q6C8_CORP1	Hypothetical protein	413.31	12	−1.52	0	PSE	*Cp1002_0321*	Core
D9Q5H0_CORP1	Hypothetical protein	12376.2	6	−1.71	0	CYT	*Cp1002_0007*	Core
D9Q4D5_CORP1	Hypothetical protein	10161.64	4	−4.29	0	CYT	*Cp1002_1721*	Shared
Others								
D9Q5N5_CORP1	Iron-regulated MEM protein	992.54	8	2.01	0	PSE	*piuB*	Core
D9Q922_CORP1	CobW/HypB/UreG, nucleotide-binding	1771.22	20	1.88	1	CYT	*Cp1002_1286*	Core

Table 1 Proteins identified as differentially-expressed following exposure to nitrosative stress *(Continued)*

D9Q8C4_CORP1	Prokaryotic ubiquitin-like protein Pup	2194.86	1	1.84	1	CYT	*pup*	Core
D9Q7B8_CORP1	Ribosomal-protein-alanine n-acetyltransferase	2791.1	10	1.34	1	CYT	*rimJ*	Shared
D9Q7K9_CORP1	Arsenate reductase	5147.54	8	1.32	1	CYT	*arsC*	Core

(a) Ratio values to: strain 1002_*DETA/NO*:strain 1002_*Ct*, Log(2) Ratio > 1.5, *p* > 0.95 = up-regulation, *p* < 0.05 = down-regulation.
(b) Core-genome analysis of 15 strains of *C. pseudotuberculosis*: shared = present in two or more strains; core = present in 15 strains of *C. pseudotuberculosis*.
(c) CYT =cytoplasmic, MEM = membrane, PSE = potentially surface-exposed, SEC = secreted.

the SOS response the more regulatory network encoded by prokaryotic involved in DNA repair [32]. The RecBC system acts in the recombination or degradative repair of arrested DNA replication forks. Studies in *S.* Typhimurium showed that *recBC* mutant strains are more attenuated than *recA* mutants in a murine model of infection [33]. In addition, unlike *recA* mutants, *recBC* mutants were susceptible to RNS [34], indicating that RecBC is highly important in the bacterial response to nitrosative stress. The LexA repressor (D9Q8W2_CORP1), which forms part of the general SOS system along with RecA [35], was down-regulated in *C. pseudotuberculosis* cells exposed to DETA/NO. We also detected the RecA protein (D9Q8Y3_CORP1); however, despite having a *p*-value <0.05, the fold-change of −0.50 showed that this protein was not activated under the experimental conditions. Studies performed in *Mycobacterium tuberculosis* showed that *recA* was not induced until cells had been exposed to DETA/NO (0.5 mM) for 4 h, but that hydrogen peroxide induced the immediate expression of *recA* [36], suggesting

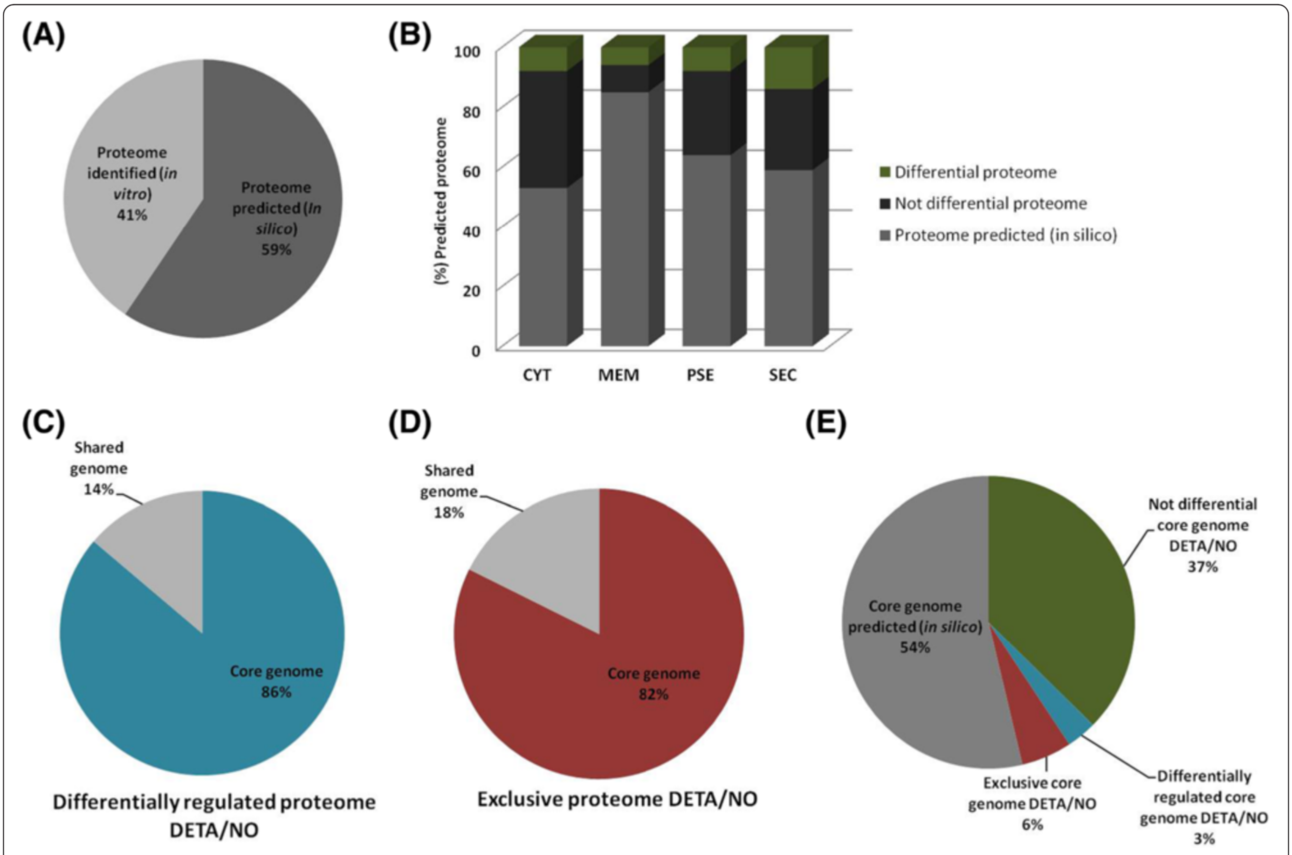


Figure 3 Correlation of *in silico* predicted data with proteome results. (A) Percentage of coverage of the *C. pseudotuberculosis* 1002 *in silico* proteome. **(B)** Prediction of the subcellular localization of the proteins identified by LC/MS. **(C)** Analysis of the differentially-regulated proteins of cells exposed to DETA/NO in relation to the core genome of *C. pseudotuberculosis* (shared genome: present in only two strains; core genome: present in 15 strains of *C. pseudotuberculosis*). **(D)** Analysis of the exclusive proteome of cells exposed to DEA/NO in relation to the core-genome of *C. pseudotuberculosis* (shared genome: present in only two strains; core genome: present in 15 strains of *C. pseudotuberculosis*). **(E)** Percent coverage of the core-genome of *C. pseudotuberculosis* in relation to the characterized proteome *in vitro*.

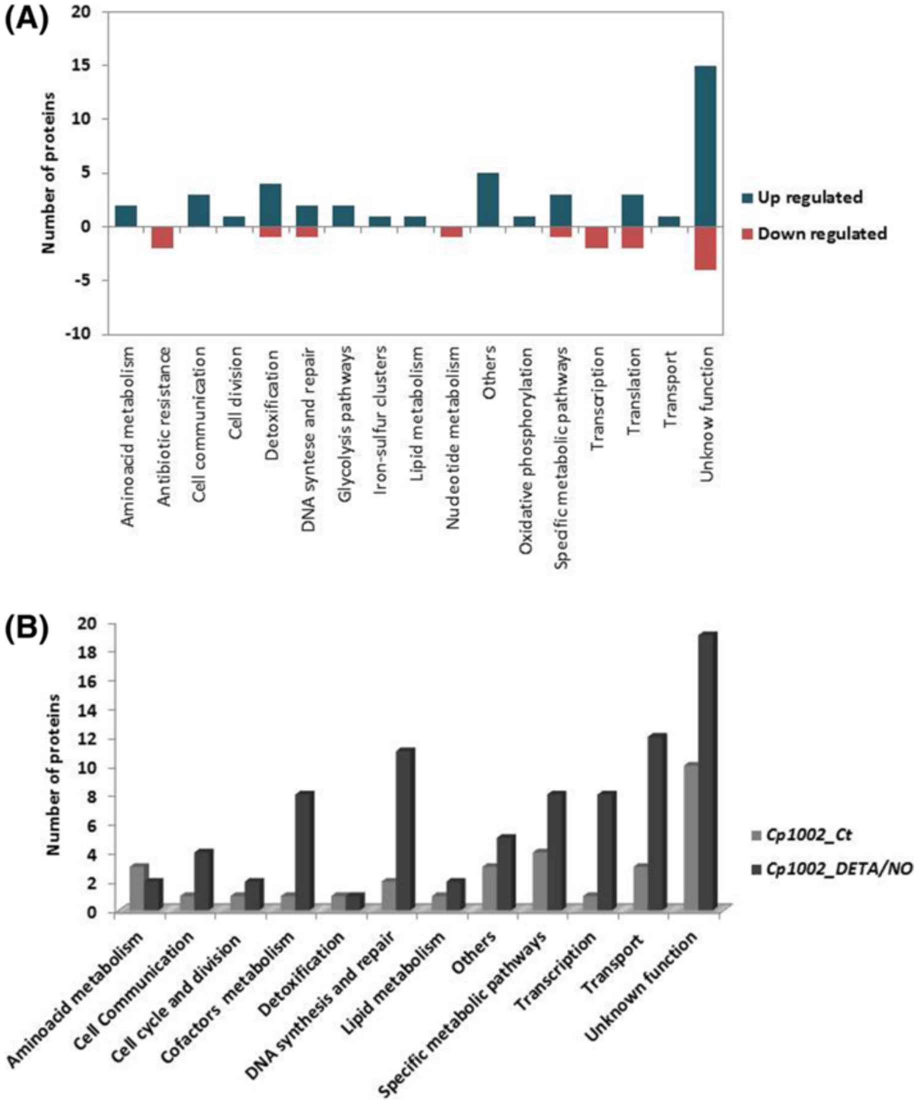


Figure 4 Comparison of biological processes between control and DETA/NO conditions. A representation of the biological processes in relation to a set list of proteins identified as **(A)** differentially-regulated in DETA/NO-stressed cells and **(B)** comparison of exclusive biological process between the two test conditions.

that RecA is involved in the later stages of the nitrosative stress response. Nevertheless, CMRegNet analysis identified other proteins that are regulated by LexA in the DETA/NO-specific proteome, including pyridoxal biosynthesis lyase (PdxS; D9Q5T9_CORP1) and DNA translocase (D9Q8Z6_CORP1).

NO-sensitive transcriptional regulators are activated in the presence of NO

To activate these DNA repair systems, it is essential that bacteria can detect ROS and RNS, and concomitantly activate the transcriptional regulators needed for the expression of genes involved in protection against these compounds. In the DETA/NO-specific proteome, we detected the transcription factor WhiB (D9Q6Y2_CORP1). The WhiB transcriptional family is composed of iron-sulfur (Fe-S) cluster proteins. These proteins are O_2- and NO-sensitive, and allow the sensing of both external environmental signals and the redox state for intracellular bacteria [37,38]. In *M. tuberculosis,* the reaction of the iron-sulfur cluster of WhiB3 with NO generates a dinitrosyl iron complex (DNIC), which activates a sensing mechanism in response to the NO, consequently activating a system of defense against nitrosative stress [12]. In addition, other *in vivo* and *in vitro* studies have also demonstrated that WhiB regulators play a role in the adaptation and survival of *M. tuberculosis* during exposure to redox environments [12,39-41].

We identified other regulators that are activated in response to environmental stimuli, such as a MerR-family transcriptional regulator (D9Q889_CORP1) and a LysR-

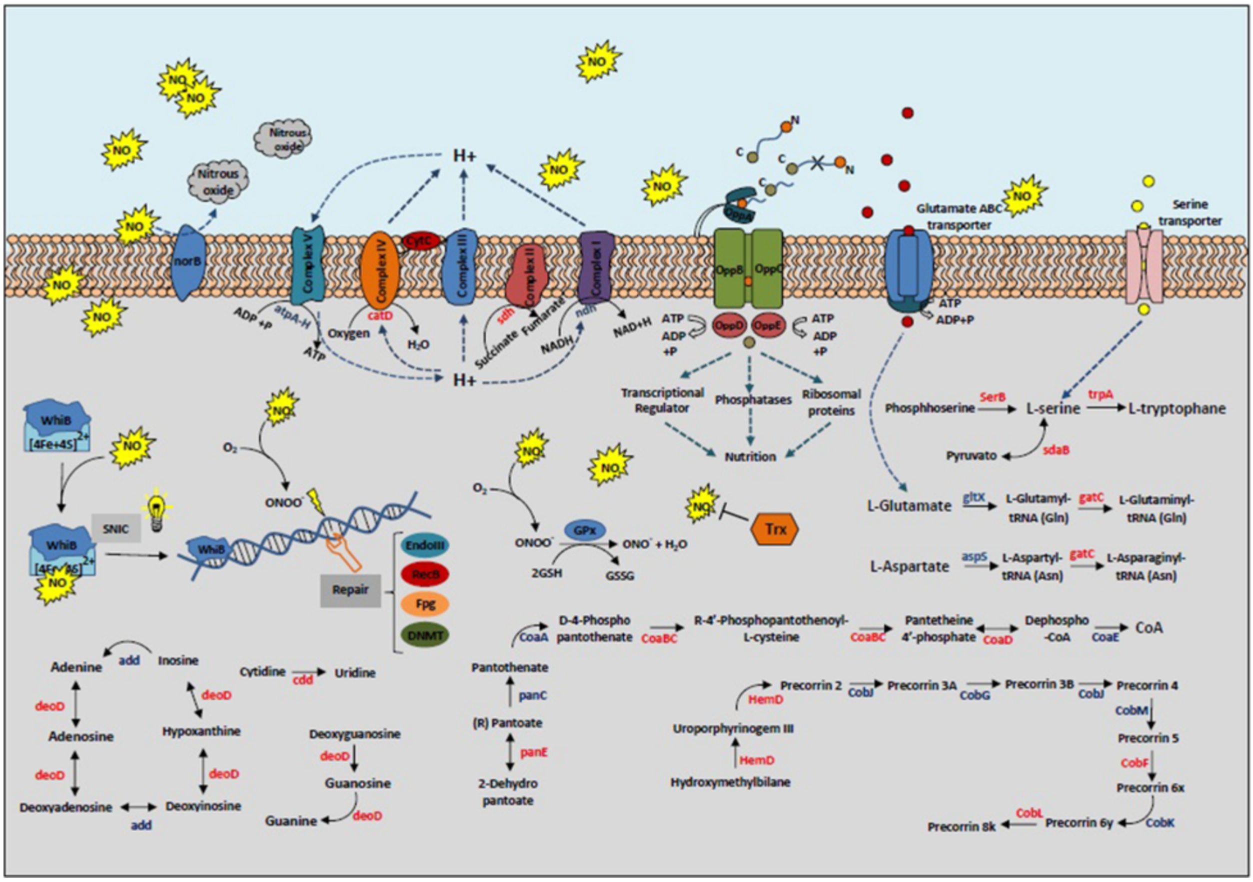


Figure 5 Overview of *C. pseudotuberculosis* response to nitrosative stress. All proteins detected by proteomic analysis are marked in red (differentially-regulated proteins or exclusive to the proteome of DETA/NO-stressed cells).

type transcriptional regulator (LTTR) (D9Q7H8_CORP1). This regulator was also highly expressed in the transcriptional response of *C. pseudotuberculosis* 1002 to acid stress [8]. MerR-type regulators have been described in the detoxification of toxic metal in several pathogenic and non-pathogenic bacteria [42]. Other studies have shown that this class of regulator plays a role in bacterial resistance to oxidative and nitrosative stress [43,44]. LTTRs are associated with the regulation of several biological processes, as well as in the adaptive response of bacteria to different types of stress [45]. In *Vibrio cholerae*, LTTRs are associated with efflux pump regulation, which contribute to antimicrobial resistance, and are involved in colonization of the human host [46]. In pathogens like *E. coli* [47], *Enterococcus faecalis* [48], *S. enterica* [49], and *Pseudomonas aeruginosa* [50], LTTRs are involved in resistance to oxidative stress.

The detoxification pathways of *C. pseudotuberculosis* following NO exposure

Our proteomic analysis identified proteins specifically expressed by cells exposed to DETA/NO that are involved in the detoxification process. Two of these proteins were thioredoxin (*trxA*) (D9Q7U6_CORP1) and glutathione peroxidase (D9Q4E5_CORP1). The thioredoxin and glutathione systems play major roles in thiol and disulfide balance, respectively [14]. In pathogens such as *Helicobacter pylori*, *Streptococcus pyogenes*, and *M. tuberculosis*, this system is of great importance in combating the presence of ROS/RNS [36,51,52]. A glyoxalase/dioxygenase (D9Q5T5_CORP1) was identified in the differential proteome of cells exposed to DETA/NO. This protein was previously detected in the proteome of *C. pseudotuberculosis* strain 1002 in response to 0.1 mM DETA/NO [15]. The presence of this protein suggests that glyoxalase/dioxygenase plays a role in the resistance of this pathogen to nitrosative stress.

Nevertheless, unlike *P. aeruginosa*, which contains a complete denitrification pathway [53], the predicted genome of *C. pseudotuberculosis ovis* 1002 revealed a truncated denitrification pathway. However, we detected the nitric-oxide reductase cytochrome b (NorB) (D9Q5T6_CORP1) in the exclusive proteome of DETA/NO-stressed cells. *norB*, which codes for this nitric-oxide reductase, is organized into the *norCBQDEF* operon in *Paracoccus*

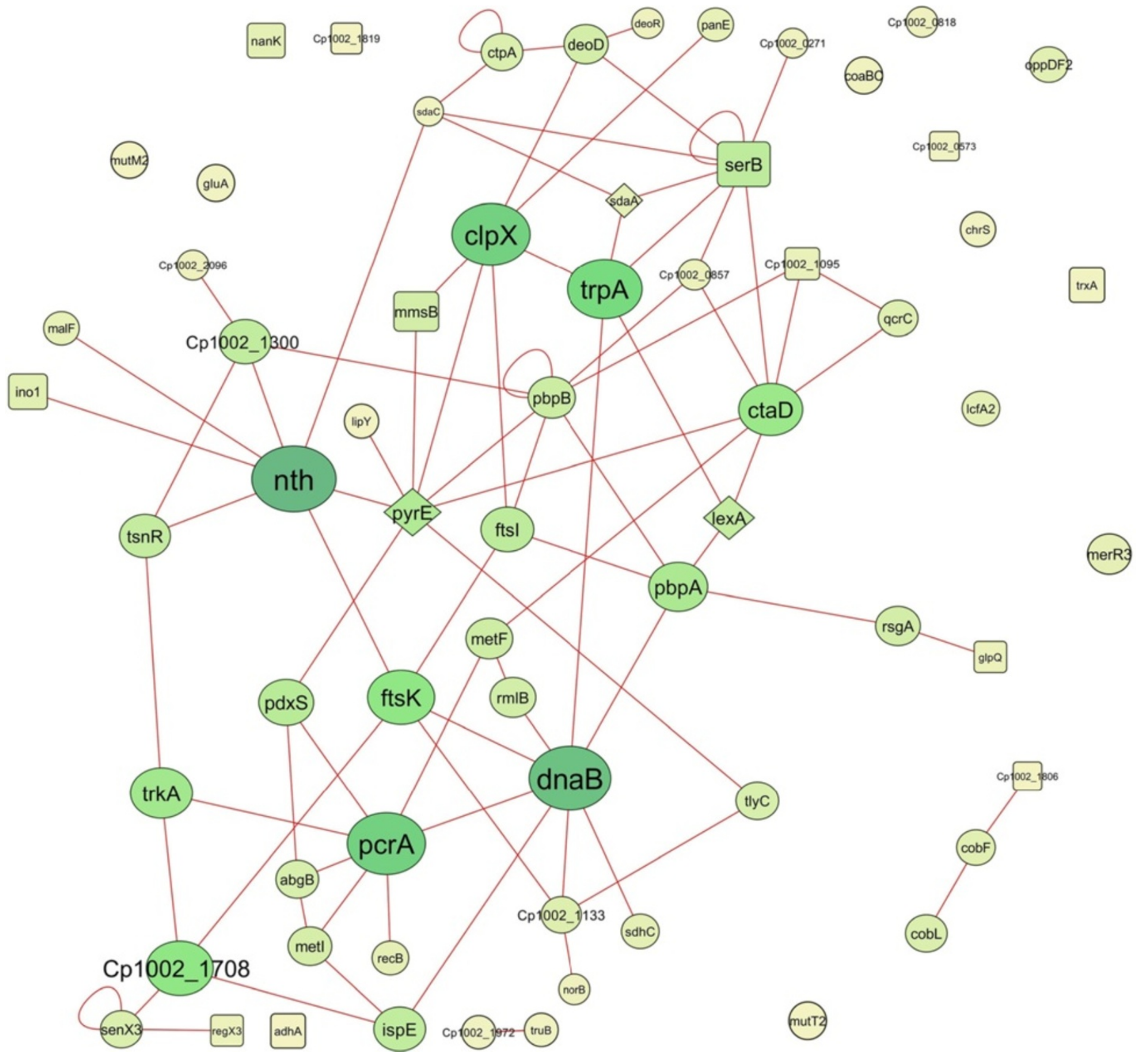


Figure 6 Protein-protein interactions. Protein-protein interactions of the proteins identified in DETA/NO-exposed cells. Exclusive proteome, circle; up-regulated, square; and down-regulated, rhombus. The sizes of the nodes represent the degree of interaction for each gene/protein; the major nodes demonstrate greater interactions. The colors of nodes and lines are in an increasing gradient scale from yellow to green to blue. The networks were visualized using Cytoscape.

denitrificans [54], and into the *norCBD* operon in *P. aeruginosa* [55]. The *C. pseudotuberculosis* genome was predicted to only contain *norB*. Moreover, *norB* is located in the *Cp1002PiCp12* pathogenicity island, suggesting horizontal acquisition of the gene by this pathogen. Nitric-oxide reductase is an important protein in the denitrification process of some bacteria [56]. In *P. aeruginosa*, NorB plays a role in both the growth of the pathogen in the presence of NO, and in its survival in macrophages [55]. The flavohemoglobin Hmp is involved in the NO detoxification pathway in *S.* Typhimurium, and levels of Hmp are increased approximately two-fold in macrophages [57]. Interestingly, in *N. meningitidis,* NorB levels are increased ten-fold in macrophages [58], demonstrating the great power of this protein in the detoxification process.

Metabolic profile of *C. pseudotuberculosis* in response to nitrosative stress

In addition to the presence of proteins involved in bacterial defense and detoxification pathways, strain 1002 needs to undergo metabolic adaptation to favor bacterial survival. We observed a metabolic readjustment in this pathogen in the proteomic analysis. Of the proteins involved in central carbohydrate metabolism, we detected only phosphoglycerate mutase (D9Q533_CORP1) and N-acetylglucosamine kinase (D9Q5B6_CORP1) in the proteome of DETA/NO-exposed cells. Other essential proteins involved in glycolysis (the Embdem-Meyerhof pathway), the pentose phosphate pathway, and the citric acid cycle were not detected. Similar results were found in a metabolomic study of *V. cholerae* in response to nitrosative stress [59].

However, we hypothesized that *C. pseudotuberculosis* uses oxidative phosphorylation to obtain energy. This is supported by the presence of cytochrome C oxidase polypeptide I (D9Q486_CORP1), succinate dehydrogenase cytochrome b556 subunit (D9Q650_CORP1), and ubiquinol-cytochrome C reductase cytochrome C subunit (D9Q3J7_CORP1) in the exclusive proteome of DETA/NO-stressed cells, and by the up-regulation of the cytochrome oxidase assembly protein (D9Q8I5_CORP1) under the same conditions. However, this oxidative phosphorylation may be associated with the bacterial culture conditions used in this work, in which *C. pseudotuberculosis* was cultivated in the presence of DETA/NO under aerobic conditions. Studies have shown that growing *M. tuberculosis* in a low concentration of NO with low levels of O_2 can induce anaerobic respiration as a result of the inhibition of the respiratory proteins cytochrome *c* oxidase and NADH reductase by irreversible ligation of NO. The ligation of NO to the respiratory proteins is an effect that may be both short-term reversible and long-term irreversible [60]. Thus, we suggest that activation of the oxidative phosphorylation system may be a more effective pathway for this pathogen to obtain energy [61].

Another metabolic adjustment was observed in relation to amino acid biosynthesis. Transporters and enzymes involved in the synthesis of methionine, tryptophan, and serine were identified. However, the presence of these proteins can be associated with the bioavailability of these amino acids during exposure to NO. In addition, we detected two oligopeptide transport ATP-binding proteins (OppD) (D9Q6G5_CORP1/D9Q3X0_CORP1) that compose the oligopeptide permease system (Opp). This complex is associated with the internalization of peptides from the extracellular environment to be used as a source of carbon and nitrogen in bacterial nutrition [62]. We also identified proteins that are cofactors of metabolism, such as CoaBC (D9Q8L2_CORP1), phosphopantetheine adenylyltransferase (D9Q809_CORP1), and 2-dehydropantoate 2-reductase (D9Q7J9_CORP1). The presence of these proteins demonstrates activity in pantothenic acid metabolism and the biosynthesis of coenzyme A (CoA). Studies performed in species such as *Corynebacterium diphtheriae* [63], *Streptococcus haemolyticus* [64], and *M. tuberculosis* [65] showed that pantothenic acid and CoA could have an important role in the growth and viability of these pathogens.

Conclusions

In this work, we applied high-throughput proteomics to characterize the proteome of *C. pseudotuberculosis ovis* 1002 following exposure to NO. Our proteomic analysis generated two profiles, which together validated findings from previous *in silico* analyses of *C. pseudotuberculosis ovis* 1002. The proteomic profile generated after the addition of the NO-donor, DETA/NO (0.5 mM), revealed a set of proteins that are involved in distinct biological process. We detected proteins related to both the general stress response and to a more specific nitrosative stress response, which together form a network of factors that promote the survival of this pathogen under stress conditions. However, more detailed studies are needed to assess the true role of these proteins in response to nitrosative stress in *C. pseudotuberculosis*. In conclusion, this functional analysis of the genome of *C. pseudotuberculosis* shows the versatility of this pathogen in the presence of NO. Moreover, the results presented in this study provide insights into the processes of resistance of *C. pseudotuberculosis* during exposure to nitrosative stress.

Additional files

Additional file 1: Table S1. Complete list of proteins identified as significantly altered ($p < 0.05$).

Additional file 2: Table S2. Unique proteins identified in strain 1002_*DETA/NO*.

Additional file 3: Table S3. Unique proteins identified in strain 1002 control condition.

Abbreviations

G: Guanine; C: Citosine; NO: Nitric oxide; RNS: Reactive nitrogen species; NOS: Nitric oxide synthases; LC-HDMS^E: Liquid chromatograph high definition mass spectrometry; LC/MS: Liquid chromatograph mass spectrometry; CDM: Chemically-defined médium; DETA/NO: Diethylenetriamine/nitric oxide adduct; DTT: Dithiothreitol; 2D-RP: Two-dimensional reversed phase; nanoUPLC: Nano Ultra performance liquid chromatography; nanoESI-HDMS: Nano electrospray high definition mass spectrometry; HSS: High strength silica; PLGS: Protein lynx global server; FDR: False discovery rate.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

WMS, RDC, and IFSB performed microbiological analyses and sample preparation for proteomic analysis. GHMFS and WMS conducted the proteomic analysis. SCS and ELF performed bioinformatics analysis of the data. YLL and AM contributed substantially to data interpretation and revisions. VA and AS participated in all steps of the project as coordinators, and critically reviewed the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the Genomics and Proteomics Network of the State of Pará of the Federal University of Pará, the Amazon Research Foundation (FAPESPA), the National Council for Scientific and Technological

Development (CNPq), the Brazilian Federal Agency for the Support and Evaluation of Graduate Education (CAPES), the Minas Gerais Research Foundation (FAPEMIG), and Waters Corporation, Brazil. Yves Le Loir is the recipient of a PVE grant (71/2013) from Programa Ciências sem Fronteiras.

Author details

[1]Depto de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [2]Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém, Pará, Brazil. [3]Waters Corporation, MS Applications and Development Laboratory, São Paulo, Brazil. [4]Institut National de la Recherche Agronomique - INRA, UMR1253 STLO, Rennes 35042, France. [5]Agrocampus Ouest, UMR1253 STLO, Rennes 35042, France.

Received: 4 September 2014 Accepted: 24 November 2014
Published: 4 December 2014

References

1. Dorella FA, Pacheco LG, Oliveira SC, Miyoshi A, Azevedo V: ***Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence.** *Vet Res* 2006, **37:**201–218.
2. Baird GJ, Fontaine MC: ***Corynebacterium pseudotuberculosis* and its role in ovine caseous lymphadenitis.** *J Comp Pathol* 2007, **137:**179–210.
3. Hodgson AL, Bird P, Nisbet IT: **Cloning, nucleotide sequence, and expression in *Escherichia coli* of the phospholipase D gene from *Corynebacterium pseudotuberculosis*.** *J Bacteriol* 1990, **172:**1256–1261.
4. Hard GC: **Comparative toxic effect of the surface lipid of *Corynebacterium ovis* on peritoneal macrophages.** *Infect Immun* 1975, **12:**1439–1449.
5. Billington SJ, Esmay PA, Songer JG, Jost BH: **Identification and role in virulence of putative iron acquisition genes from *Corynebacterium pseudotuberculosis*.** *J Bacteriol* 2002, **180:**3233–3236.
6. Ruiz JC, D'Afonseca V, Silva A, Ali A, Pinto AC, Santos AR, Rocha AA, Lopes DO, Dorella FA, Pacheco LG, Costa MP, Turk MZ, Seyffert N, Moraes PM, Soares SC, Almeida SS, Castro TL, Abreu VA, Trost E, Baumbach J, Tauch A, Schneider MP, McCulloch J, Cerdeira LT, Ramos RT, Zerlotini A, Dominitini A, Resende DM, Coser EM, Oliveira LM, *et al*: **Evidence for reductive genome evolution and lateral acquisition of virulence functions in two *Corynebacterium pseudotuberculosis* strains.** *PLoS One* 2011, **18:**e18551.
7. Soares SC, Silva A, Trost E, Blom J, Ramos R, Carneiro A, Ali A, Santos AR, Pinto AC, Diniz C, Barbosa EG, Dorella FA, Aburjaile F, Rocha FS, Nascimento KK, Guimarães LC, Almeida S, Hassan SS, Bakhtiar SM, Pereira UP, Abreu VA, Schneider MP, Miyoshi A, Tauch A, Azevedo V: **The pan-genome of the animal pathogen *Corynebacterium pseudotuberculosis* reveals differences in genome plasticity between the biovar *ovis* and *equi* strains.** *PLoS One* 2013, **8:**e53818.
8. Pinto AC, de Sá PH, Ramos RT, Barbosa S, Barbosa HP, Ribeiro AC, Silva WM, Rocha FS, Santana MP, de Paula Castro TL, Miyoshi A, Schneider MP, Silva A, Azevedo V: **Differential transcriptional profile of *Corynebacterium pseudotuberculosis* in response to abiotic stresses.** *BMC Genomics* 2014, **9:**14.
9. Marletta MA: **Nitric oxide synthase: aspects concerning structure and catalysis.** *Cell* 1994, **23:**927–930.
10. Griffith OW, Stueh DJ: **Nitric oxide synthases: properties and catalytic mechanism.** *Annu Rev Physiol* 1995, **57:**707–736.
11. Nathan C, Shiloh MU: **Reactive oxygen and nitrogen intermediates in the relationship between mammalian hosts and microbial pathogens.** *Proc Natl Acad Sci USA* 2000, **1:**8841–8848.
12. Singh A, Guidry L, Narasimhulu KV, Mai D, Trombley J, Redding KE, Giles GI, Lancaster JR Jr, Steyn AJ: ***Mycobacterium tuberculosis* WhiB3 responds to O$_2$ and nitric oxide via its [4Fe-4S] cluster and is essential for nutrient starvation survival.** *Proc Natl Acad Sci USA* 2007, **10:**11562–11567.
13. Ehrt S, Schnappinger D: **Mycobacterial survival strategies in the phagosome: defence against host stresses.** *Cell Microbiol* 2009, **11:**1170–1178.
14. Lu J, Holmgren A: **The thioredoxin antioxidant system.** *Free Radic Biol Med* 2013, **8:**75–87.
15. Pacheco LG, Castro TL, Carvalho RD, Moraes PM, Dorella FA, Carvalho NB, Slade SE, Scrivens JH, Feelisch M, Meyer R, Miyoshi A, Oliveira SC, Dowson CG, Azevedo V: **A role for sigma factor σ(^{E}) in *Corynebacterium pseudotuberculosis* resistance to nitric oxide/peroxide stress.** *Front Microbiol* 2012, **3:**126.
16. Moura-Costa LF, Paule BJA, Freire SM, Nascimento I, Schaer R, Regis LF, Vale VLC, Matos DP, Bahia RC, Carminati R, Meyer R: **Chemically defined synthetic medium for *Corynebacterium pseudotuberculosis* culture.** *Rev Bras Saúde Prod An* 2002, **3:**1–9.
17. Silva JC, Gorenstein MV, Li GZ, Vissers JP, Geromanos SJ: **Absolute quantification of proteins by LC/MS^E: a virtue of parallel MS acquisition.** *Mol Cell Proteomics* 2006, **5:**144–156.
18. Gilar M, Olivova P, Daly AE, Gebler JC: **Two-dimensional separation of peptides using RP-RP-HPLC system with different pH in first and second separation dimensions.** *J Sep Sci* 2005, **28:**1694–1703.
19. Silva JC, Denny R, Dorschel CA, Gorenstein M, Kass IJ, Li GZ, McKenna T, Nold MJ, Richardson K, Young P, Geromanos S: **Quantitative proteomic analysis by accurate mass retention time pairs.** *Anal Chem* 2005, **77:**2187–2000.
20. Geromanos SJ, Vissers JP, Silva JC, Dorschel CA, Li GZ, Gorenstein MV, Bateman RH, Langridge JI: **The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS.** *Proteomics* 2009, **9:**1683–1695.
21. Li GZ, Vissers JP, Silva JC, Golick D, Gorenstein MV, Geromanos SJ: **Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures.** *Proteomics* 2009, **9:**1696–1719.
22. Curty N, Kubitschek-Barreira PH, Neves GW, Gomes D, Pizzatti L, Abdelhay E, Souza GH, Lopes-Bezerra LM: **Discovering the infectome of human endothelial cells challenged with *Aspergillus fumigatus* applying a mass spectrometry label-free approach.** *J Proteomics* 2014, **31:**126–140.
23. Levin Y, Hadetzky E, Bahn S: **Quantification of proteins using data-independent analysis (MSE) in simple and complex samples: a systematic evaluation.** *Proteomics* 2011, **11:**3273–3287.
24. Barinov A, Loux V, Hammani A, Nicolas P, Langella P, Ehrlich D, Maguin E, van de Guchte M: **Prediction of surface exposed proteins in *Streptococcus pyogenes*, with a potential application to other Gram-positive bacteria.** *Proteomics* 2009, **9:**61–73.
25. Conesa A, Gotz S, García-Gómez JM, Terol J, Talón M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics* 2005, **15:**3674–3676.
26. Soares SC, Abreu VA, Ramos RT, Cerdeira L, Silva A, Baumbach J, Trost E, Tauch A, Hirata R Jr, Mattos-Guaraldi AL, Miyoshi A, Azevedo V: **PIPs: pathogenicity island prediction software.** *PLoS One* 2012, **7:**e30848.
27. Rezende AM, Folador EL, Resende Dde M, Ruiz JC: **Computational prediction of protein-protein interactions in *Leishmania* predicted proteomes.** *PLoS One* 2012, **7:**e51304.
28. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13:**2498–2504.
29. Pauling J, Röttger R, Tauch A, Azevedo V, Baumbach J: **CoryneRegNet 6.0 -Updated database content, new analysis methods and novel features focusing on community demands.** *Nucleic Acids Res* 2012, **40:**D610–614.
30. Richardson AR, Soliven KC, Castor ME, Barnes PD, Libby SJ, Fang FC: **The base excision repair system of *Salmonella enterica* serovar *typhimurium* counteracts DNA damage by host nitric oxide.** *PLoS Pathog* 2009, **5:**e1000451.
31. Sedgwick B: **Repairing DNA-methylation damage.** *Nat Rev Mol Cell Biol* 2004, **5:**148–157.
32. Baharoglu Z, Mazel D: **SOS, the formidable strategy of bacteria against aggressions.** *FEMS Microbiol Rev* 2014, **38:**1126–1145.
33. Cano DA, Pucciarelli MG, García-del Portillo F, Casadesús J: **Role of the *recBCD* recombination pathway in *Salmonella* virulence.** *J Bacteriol* 2002, **184:**592–595.
34. Koskiniemi S, Andersson DI: **Translesion DNA polymerases are required for spontaneous deletion formation in *Salmonella typhimurium*.** *Proc Natl Acad Sci U S A.* 2009, **23:**10248–10253.
35. Butala M, Zgur-Bertok D, Busby SJ: **The bacterial LexA transcriptional repressor.** *Cell Mol Life Sci* 2009, **66:**82–93.
36. Voskuil MI, Bartek IL, Visconti K, Schoolnik GK: **The response of *Mycobacterium tuberculosis* to reactive oxygen and nitrogen species.** *Front Microbiol* 2011, **13:**105.
37. Green J, Paget MS: **Bacterial redox sensors.** *Nat Rev Microbiol* 2004, **2:**954–966.
38. Green J, Rolfe MD, Smith LJ: **Transcriptional regulation of bacterial virulence gene expression by molecular oxygen and nitric oxide.** *Virulence* 2014, **4:**5(4).
39. Singh A, Crossman DK, Mai D, Guidry L, Voskuil MI, Renfrow MB, Steyn AJ: ***Mycobacterium tuberculosis* WhiB3 maintains redox homeostasis by regulating virulence lipid anabolism to modulate macrophage response.** *PLoS Pathog* 2009, **5:**e1000545.
40. Chawla M, Parikh P, Saxena A, Munshi M, Mehta M, Mai D, Srivastava AK, Narasimhulu KV, Redding KE, Vashi N, Kumar D, Steyn AJ, Singh A:

Mycobacterium tuberculosis WhiB4 regulates oxidative stress response to modulate survival and dissemination *in vivo*. *Mol Microbiol* 2012, **85**:1148–1165.

41. Larsson C, Luna B, Ammerman NC, Maiga M, Agarwal N, Bishai WR: Gene expression of *Mycobacterium tuberculosis* putative transcription factors WhiB1-7 in redox environments. *PLoS One* 2012, **7**:e37516.
42. Hobman JL: MerR family transcription activators: similar designs, different specificities. *Mol Microbiol* 2007, **63**:1275–1278.
43. McEwan AG, Djoko KY, Chen NH, Couñago RL, Kidd SP, Potter AJ, Jennings MP: Novel bacterial MerR-like regulators their role in the response to carbonyl and nitrosative stress. *Adv Microb Physiol* 2011, **58**:1–22.
44. Brown NL, Stoyanov JV, Kidd SP, Hobman JL: The MerR family of transcriptional regulators. *FEMS Microbiol Rev* 2003, **27**:145–163.
45. Maddocks SE, Oyston PC: Structure and function of the LysR-type transcriptional regulator (LTTR) family proteins. *Microbiology* 2008, **154**:3609–3623.
46. Chen S, Wang H, Katzianer DS, Zhong Z, Zhu J: LysR family activator-regulated major facilitator superfamily transporters are involved in *Vibrio cholerae* antimicrobial compound resistance and intestinal colonisation. *Int J Antimicrob Agents* 2013, **41**:188–192.
47. Gonzalez-Flecha B, Demple B: Role for the *oxyS* gene in regulation of intracellular hydrogen peroxide in *Escherichia coli*. *J Bacteriol* 1999, **181**:3833–3836.
48. Verneuil N, Rincé A, Sanguinetti M, Posteraro B, Fadda G, Auffray Y, Hartke A, Giard JC: Contribution of a PerR-like regulator to the oxidative-stress response and virulence of *Enterococcus faecalis*. *Microbiology* 2005, **151**:3997–4004.
49. Lahiri A, Das P, Chakravortty D: The LysR-type transcriptional regulator Hrg counteracts phagocyte oxidative burst and imparts survival advantage to *Salmonella enterica* serovar Typhimurium. *Microbiology* 2008, **154**:2837–2846.
50. Reen FJ, Haynes JM, Mooij MJ, O'Gara F: A non-classical LysR-type transcriptional regulator PA2206 is required for an effective oxidative stress response in *Pseudomonas aeruginosa*. *PLoS One* 2013, **8**:e54479.
51. Comtois SL, Gidley MD, Kelly DJ: Role of the thioredoxin system and the thiol-peroxidases Tpx and Bcp in mediating resistance to oxidative and nitrosative stress in *Helicobacter pylori*. *Microbiology* 2003, **149**:121–129.
52. Brenot A, King KY, Janowiak B, Griffith O, Caparon MG: Contribution of glutathione peroxidase to the virulence of *Streptococcus pyogenes*. *Infect Immun* 2004, **72**:408–413.
53. Kalkowski I, Conrad R: Metabolism of nitric oxide in denitrifying *Pseudomonas aeruginosa* and nitrate-respiring *Bacillus cereus*. *FEMS Microbiol Lett.* 1991, **15**:107–111.
54. de Boer AP, van der Oost J, Reijnders WN, Westerhoff HV, Stouthamer AH, van Spanning RJ: Mutational analysis of the *nor* gene cluster which encodes nitric-oxide reductase from *Paracoccus denitrificans*. *Eur J Biochem* 1996, **15**:592–600.
55. Kakishima K, Shiratsuchi A, Taoka A, Nakanishi Y, Fukumori Y: Participation of nitric oxide reductase in survival of *Pseudomonas aeruginosa* in LPS-activated macrophages. *Biochem Biophys Res Commun* 2007, **6**:587–591.
56. Hendriks J, Oubrie A, Castresana J, Urbani A, Gemeinhardt S, Saraste M: Nitric oxide reductases in bacteria. *Biochim Biophys Acta* 2000, **15**:266–273.
57. Stevanin TM, Poole RK, Demoncheaux EA, Read RC: Flavohemoglobin Hmp protects *Salmonella enterica* serovar Typhimurium from nitric oxide-related killing by human macrophages. *Infect Immun* 2002, **70**:4399–4405.
58. Stevanin TM, Moir JW, Read RC: Nitric oxide detoxification systems enhance survival of *Neisseria meningitidis* in human macrophages and in nasopharyngeal mucosa. *Infect Immun* 2005, **73**:3322–3329.
59. Stern AM, Liu B, Bakken LR, Shapleigh JP, Zhu J: A novel protein protects bacterial iron-dependent metabolism from nitric oxide. *J Bacteriol* 2013, **195**:4702–4708.
60. Brown GC: Regulation of mitochondrial respiration by nitric oxide inhibition of cytochrome C oxidase. *Biochim Biophys Acta* 2001, **1**:46–57.
61. Kadenbach B: Intrinsic and extrinsic uncoupling of oxidative phosphorylation. *Biochim Biophys Acta* 2003, **5**:77–94.
62. Payne JW, Smith MW: Peptide transport by micro-organisms. *Adv Microb Physiol* 1994, **36**:1–80.
63. Mueller JH, Klotz AW: Pantothenic acid as a growth factor for the diphtheria bacillus. *J Am Chem Soc* 1938, **60**:3086–3087.
64. McIlwain H: Pantothenic acid and the growth of *Streptococcus haemolyticus*. *Br J Exp Pathol* 1939, **20**:330–333.
65. Sassetti CM, Boyd DH, Rubin EJ: Genes required for mycobacterial growth defined by high density mutagenesis. *Mol Microbiol* 2003, **48**:77–84.

doi:10.1186/1471-2164-15-1065
Cite this article as: Silva *et al.*: Label-free proteomic analysis to confirm the predicted proteome of *Corynebacterium pseudotuberculosis* under nitrosative stress mediated by nitric oxide. *BMC Genomics* 2014 **15**:1065.

II.III.7 Quantitative Proteomic Analysis Reveals Changes in the Benchmark *Corynebacterium pseudotuberculosis* Biovar Equi Exoproteome after Passage in a Murine Host.

Silva WM, Carvalho RDO, Dorella FA, Folador EL, Souza GHMF, Pimenta AMC, Figueiredo HCP, Le Loir Y, Silva A and **Azevedo V**.

Front. Cell. Infect. Microbiol. 7:325. doi: 10.3389/fcimb.2017.00325

Neste trabalho utilizamos a proteômica para caracterizar pela primeira vez o exoproteoma de uma linhagem pertencente ao biovar *equi* de *C. pseudotuberculosis*. Para identificar proteínas que poderiam estar envolvidas na virulência deste patógeno, nós combinamos um processo de passagem experimental da linhagem 258_*equi* em um modelo murino e *Label-free proteomics* para identificar e quantificar o exoproteoma desta linhagem. Interessantemente, a recuperação desta linhagem do baço de camundongo infectado induziu uma alteração no seu potencial de virulência, tornando-se mais virulenta, em um segundo ensaio de infecção. O *screening* proteômico realizado a partir do sobrenadante da condição controle e recuperado revelou 104 proteínas com valores estatísticos significativos, sendo que 39 proteínas foram induzidas e 16 *down-regulated* na condição recuperada. Analises de enriquecimento a partir do *KEGG database* demonstrou que transportadores ABC, sistemas bacterianos de secreção e mecanismos de exportação foram significativamente alterados na condição recuperada. Estes achados demonstram que a exportação de proteínas e as proteínas exportadas são elementos importantes na virulência e patogêneses de *C. pseudotuberculosis*. Coletivamente, proteínas relacionadas à patogênese bacteriana como: proteínas de adesão, crescimento celular e evasão do sistema imune foram identificados. Assim, este estudo promoveu informações importantes a respeito dos fatores que podem influenciar na patogênese de *C. pseudotuberculosis*.

frontiers
in Cellular and Infection Microbiology

ORIGINAL RESEARCH
published: 25 July 2017
doi: 10.3389/fcimb.2017.00325

Quantitative Proteomic Analysis Reveals Changes in the Benchmark *Corynebacterium pseudotuberculosis* Biovar *Equi* Exoproteome after Passage in a Murine Host

Wanderson M. Silva[1, 2, 3], Rodrigo D. De Oliveira Carvalho[1], Fernanda A. Dorella[4], Edson L. Folador[5], Gustavo H. M. F. Souza[6], Adriano M. C. Pimenta[7], Henrique C. P. Figueiredo[4], Yves Le Loir[2, 3], Artur Silva[8] and Vasco Azevedo[1]*

[1] Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, [2] Institut National de la Recherche Agronomique (INRA), UMR1253 Science & Technologie du Lait & de l'Oeuf (STLO), Rennes, France, [3] Agrocampus Ouest, UMR1253 Science & Technologie du Lait & de l'Oeuf (STLO), Rennes, France, [4] Escola de Veterinária, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, [5] Centro de Biotecnologia, Universidade Federal da Paraíba, João Pessoa, Brazil, [6] Waters Corporation, Waters Technologies Brazil, MS Applications Laboratory, São Paulo, Brazil, [7] Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, [8] Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém, Brazil

OPEN ACCESS

Edited by:
Nora Lía Padola,
National University of Central Buenos Aires, Argentina

Reviewed by:
Javier Bezos,
Complutense University of Madrid, Spain
Travis Bourret,
Creighton University, United States

***Correspondence:**
Vasco Azevedo
vasco@icb.ufmg.br

Received: 07 April 2017
Accepted: 03 July 2017
Published: 25 July 2017

Citation:
Silva WM, Carvalho RDO, Dorella FA, Folador EL, Souza GHMF, Pimenta AMC, Figueiredo HCP, Le Loir Y, Silva A and Azevedo V (2017) Quantitative Proteomic Analysis Reveals Changes in the Benchmark Corynebacterium pseudotuberculosis Biovar Equi Exoproteome after Passage in a Murine Host. Front. Cell. Infect. Microbiol. 7:325. doi: 10.3389/fcimb.2017.00325

Corynebacterium pseudotuberculosis biovar *equi* is the etiologic agent of ulcerative lymphangitis. To investigate proteins that could be related to the virulence of this pathogen, we combined an experimental passage process using a murine model and high-throughput proteomics with a mass spectrometry, data-independent acquisition (LC-MS^E) approach to identify and quantify the proteins released into the supernatants of strain 258_*equi*. To our knowledge, this approach allowed characterization of the exoproteome of a *C. pseudotuberculosis equi* strain for the first time. Interestingly, the recovery of this strain from infected mouse spleens induced a change in its virulence potential, and it became more virulent in a second infection challenge. Proteomic screening performed from culture supernatant of the control and recovered conditions revealed 104 proteins that were differentially expressed between the two conditions. In this context, proteomic analysis of the recovered condition detected the induction of proteins involved in bacterial pathogenesis, mainly related to iron uptake. In addition, KEGG enrichment analysis showed that ABC transporters, bacterial secretion systems and protein export pathways were significantly altered in the recovered condition. These findings show that secretion and secreted proteins are key elements in the virulence and adaptation of *C. pseudotuberculosis*. Collectively, bacterial pathogenesis-related proteins were identified that contribute to the processes of adherence, intracellular growth and evasion of the immune system. Moreover, this study enhances our understanding of the factors that may influence the pathogenesis of *C. pseudotuberculosis*.

Keywords: *Corynebacterium pseudotuberculosis*, label-free proteome, ulcerative lymphangitis, bacterial proteome, bacterial virulence, serial passage

INTRODUCTION

Corynebacterium pseudotuberculosis is a gram-positive, facultative intracellular pathogen that is globally distributed and can infect horses, cattle, sheep, goats, buffalos, and occasionally humans. *C. pseudotuberculosis* biovar *ovis* is the etiologic agent of caseous lymphadenitis in small ruminants (Dorella et al., 2006). *Corynebacterium pseudotuberculosis* strains belonging to biovar *equi*, however, cause edematous skin illness in buffalos. In horses, the infection can manifest through one of two forms: (i) externally, which usually presents with chronic ventral and pectoral lymph node abscesses and, in a more advanced stage, generates an illness denominated ulcerative lymphangitis that is characterized by ulcers of irregular shapes and sizes, or (ii) internally, which is characterized mainly by abscess formation in the lymph nodes and internal organs (kidney, liver, lung, and spleen) (Britz et al., 2014). In the United States, ulcerative lymphangitis outbreaks with large economic losses for horse farmers have been reported (Aleman et al., 1996; Foley et al., 2004; Spier, 2008). In addition, a recent study showed an increase in the case numbers of horses infected with *C. pseudotuberculosis* biovar *equi* during the last 10 years in the western region of the USA, which is being considered as an endemic area (Kilcoyne et al., 2014).

Whole genome sequencing of *C. pseudotuberculosis* biovar *equi* strain 258, isolated from a horse with ulcerative lymphangitis in Belgium, revealed the presence of pathogenic islands in its chromosome as well as genes that might contribute to its virulence, most of them coding for secreted proteins. Moreover, putative antigenic proteins were identified through reverse vaccinology (Soares et al., 2013a). Some studies have shown that extracellular proteins are related to the pathogenic process of *C. pseudotuberculosis* (Wilson et al., 1995; Billington et al., 2002; Pacheco et al., 2012; Seyffert et al., 2014). However, phospholipase D (Pld) exotoxin, which contributes to bacterial spread in the host, is considered the major virulence factor of this pathogen (McKean et al., 2007). In addition, some secreted factors related to the virulence of *C. pseudotuberculosis* have already been described, such as serine protease CP40 (Wilson et al., 1995) and two operons, *fagABC* and *ciuABCDEF*, that are involved in iron uptake (Billington et al., 2002; Ribeiro et al., 2014).

The study of host-bacteria interactions in natural hosts, such as horses, cattle or sheep, is difficult because of the underlying genetic variability among animals; it is also extremely expense and requires multiple replicates and control animals. Thus, several studies have used mice as a model for studying both the pathogenic process (Jolly, 1965; Zaki, 1966; Nieto et al., 2009) and vaccination testing against infection by *C. pseudotuberculosis* (Simmons et al., 1997; Lan et al., 1999; Gorman et al., 2010; Ribeiro et al., 2014; Droppa-Almeida et al., 2016). In regards to host-bacteria interactions, some work has explored the serial passage process of bacterial pathogens *in vitro* or in an *in vivo* model to identify factors that might be involved in virulence (Fernández et al., 2000, 2013; Bleich et al., 2005; Chapuis et al., 2011; Fernandez-Brando et al., 2012; Liu et al., 2015). In this study, we adopted an *in vivo* assay in which the strain 258_*equi* was experimentally inoculated in mice followed by high-throughput proteomic analysis. We screened the functional genome of 258_*equi* after experimental passage in the murine host by examining the proteins released into the culture supernatant of this strain using the three-phase partitioning (TPP) protocol for obtaining extracellular proteins (Paule et al., 2004) and a mass spectrometry, data-independent acquisition (LC-MS^E) approach to identify and quantify the proteins (Silva et al., 2006; Pacheco et al., 2011).

MATERIALS AND METHODS

Bacterial Strain and growth Conditions

Corynebacterium pseudotuberculosis biovar *equi* strain 258 was isolated from a horse in Belgium; this strain was cultivated under routine conditions in brain–heart infusion broth (BHI-HiMedia Laboratories Pvt. Ltd., India) at 37°C. When necessary, 1.5% agar was added to the medium for solid culture. For extracellular proteomic analyses, 258_*equi* was grown in a chemically defined medium (CDM) [(Na_2HPO_4_$7H_2O$ (12.93 g/L), KH_2PO_4 (2.55 g/L), NH_4Cl (1 g/L), $MgSO_4$_$7H_2O$ (0.20 g/L), $CaCl_2$ (0.02 g/L), and 0.05% (v/v) Tween 80] with 4% (v/v) MEM Vitamins Solution (Invitrogen, Gaithersburg, MD, USA), 1% (v/v) MEM Amino Acids Solution (Invitrogen), 1% (v/v) MEM Non-Essential Amino Acids Solution (Invitrogen), and 1.2% (w/v) glucose at 37°C (Moura-Costa et al., 2002).

Experimental Infection in a murine Model

The infection parameters were performed according to Moraes et al. (2014) and Ribeiro et al. (2014). In this study, female BALB/c mice between 6- and 8-weeks-old were utilized; they were provided by the Animal Care Facility at the Biological Sciences Institute at the Federal University of Minas Gerais and were handled in accordance with the CEUA guidelines of the UFMG Ethics Committee on Animal Testing (Permit Number: CETEA 103/2011). For the bacterial passage assay, three mice were infected via intraperitoneal injection with 10^6 colony forming units (CFU) of strain 258_*equi*. Thirty-six hours after infection, the animals were sacrificed, and the spleen was aseptically removed for recovering the bacteria. Each spleen was individually macerated in a sterile saline solution (0.9% $NaCl_2$) and seeded onto BHI agar plates for incubation at 37°C for 48 h. Subsequently, one bacterial colony of each BHI plate was isolated and cultured in BHI broth at 37°C with shaking (180 rpm) until the $OD_{600} = 0.8$. Three different stock cultures were generated and stored at −80°C in BHI broth and 10% glycerol. The recovered bacteria are referred to as Recovered (Rc), and bacteria with no previous host contact were used as the Control (Ct). For bacterial virulence assays, bacteria from the three individual frozen stocks of Rc and the Ct condition were centrifuged at 5,000 × g for 5 min and washed twice in saline solution, followed by resuspension in saline solution. Three groups of five mice were infected with bacteria from the Rc or Ct condition via intraperitoneal injection of a suspension containing 10^6 or 10^5 CFU. The animals' survival rates were calculated and represented in GraphPad Prism v.5.0 (GraphPad Software, San Diego, CA, USA) using the Kaplan-Meier survival function.

Preparation of extracellular proteins for proteome Analysis

For proteomic analysis, three independent control and recovered colonies from the three individual frozen stocks were grown in CDM to an OD600 = 0.8. The cultures were then centrifuged for 20 min at 2,700 × g. The supernatants were filtered using 0.22-μm filters, 30% (w/v) ammonium sulfate was added to the samples, and the pH of the mixtures was adjusted to 4.0. Next, 20 mL/L N-butanol was added to each sample. The samples were centrifuged for 10 min at 1,350 × g and 4°C. The interfacial precipitate was collected and resuspended in 1 mL of 20 mM Tris-HCl, pH 7.2 (Paule et al., 2004). Proteins were quantified using the Bradford assay. For label-free proteomic analysis, the protein extract was concentrated using a spin column with a 10 kDa threshold (Millipore, Billerica, MA, USA). The protein was denatured (0.1% *Rapi*GEST SF at 60°C for 15 min) (Waters, Milford, CA, USA), reduced (10 mM DTT), alkylated (10 mM iodoacetamide) and enzymatically digested with trypsin (Promega, Sequencing Grade Modified Trypsin, Madison, WI, USA). Glycogen phosphorylase (Waters Corporation, SwissProt P00489) was added to the digests to a final concentration of 20 fmol/μl as an internal standard for normalization prior to each replicate injection. The digestion process was stopped by adding 10 μL of 5% TFA (Fluka, Buchs, Germany) (Silva et al., 2006).

Mass Spectrometry analysis, data processing and quantification

Three independent biological replicates of each experimental condition were digested, as described above, for MS^E analysis. Qualitative and quantitative nanoUPLC tandem nanoESI-$HDMS^E$ (Nano Electrospray High Definition Mass Spectrometry) experiments were performed using a 1 h reversed-phase gradient from 7 to 40% (v/v) acetonitrile (with 0.1% v/v formic acid) at 500 nL.min^{-1} using a nanoACQUITY UPLC 2D RPxRP Technology system (Gilar et al., 2005). All analyses were performed using nano-electrospray ionization in the positive ion mode (nanoESI (+)) and a NanoLockSpray (Waters, Manchester, UK) ionization source. The mass spectrometer was calibrated with an MS/MS spectrum of human [Glu1]-Fibrinopeptide B (Glu-Fib) solution (100 fmol.mL−1) delivered through the reference sprayer of the NanoLockSpray source. The double-charged ion ($[M + 2H]^{2+} = 785.8426$) was used for initial single-point calibration, and MS/MS fragment ions of Glu-Fib were used to obtain the final instrument calibration. Multiplexed data-independent (DIA) scanning with additional specificity and selectivity for non-linear "T-wave" ion mobility ($HDMS^E$) experiments were performed using a Synapt G2-S HDMS mass spectrometer (Waters, Manchester, UK), which was constructed to automatically switch between the application of standard MS (3 eV) and elevated collision energies $HDMS^E$ (19–45 eV) to the transfer "T-wave" CID (collision-induced dissociation) cell with argon gas.

The proteins were identified, and quantitative data were packaged using dedicated algorithms (Silva et al., 2005; Geromanos et al., 2009) and searching against a database with default parameters to account for ions (Li et al., 2009). The databases used were reversed "on-the-fly" during the database queries and appended to the original database to assess the false positive rate during identification. For proper spectra processing and database searching conditions, the ProteinLynx Global SERVER v.2.5.2 (PLGS) with Identity^E and Expression^E informatics v.2.5.2 (Waters, Manchester, UK) was used. UniProtKB (release 2013_01) with manually reviewed annotations was used, and the search conditions were based on taxonomy (*Corynebacterium pseudotuberculosis*). The maximum allowed missed cleavages by trypsin was up to 1, and various modifications, including carbamidomethyl (C), N-terminal acetyl, phosphoryl (STY) and oxidation (M), were allowed. A peptide mass tolerance value of 10 ppm was used. The search threshold to accept each spectrum was the default value in the program with a false discovery rate value of 4% (Curty et al., 2014). For protein quantitation, PLGS v2.5.2 software was used with the Identity^E algorithm using Hi3 methodology and glycogen phosphorylase (muscle form; P00489) peptides were used as internal standards. The collected proteins were organized by the PLGS Expression^E tool algorithm into a statistically significant list (p-value ≤ 0.05) that corresponded to higher or lower regulation ratios between the different groups. The calculation of the log ratio and the confidence interval was based on a Gaussian distribution model, which allows for the possibility of an uncertain peptide assignment, an incorrect assignment of data to a cluster or interference. The confidence interval of 95% was used, and the probability distribution of the measured value of a log2 ratio more than a 1.2 was more symmetric than that obtained for the direct ratio, making the results interpretations more meaningful (Levin et al., 2011). For comparing pairs of experimental groups, proteins with a differential expression $\log_2$ ratio greater than or equal to 1.2 between the two conditions were considered for higher or lower abundance level determination (Levin et al., 2011).

Bioinformatics Analysis

The proteins identified in both conditions were analyzed using the following prediction tools: SurfG+ v1.0 (Barinov et al., 2009) was used to predict subcellular localization, SecretomeP 2.0 server was used to predict proteins exported from non-classical systems (positive prediction scores greater than 0.5; Bendtsen et al., 2005a), TatP was used to predict proteins with twin-arginine signal peptides (Bendtsen et al., 2005b) and the PIPs software was used to predict the proteins in pathogenicity islands (Soares et al., 2012). Gene ontology (GO) functional annotations were generated using the COG data base (Tatusov et al., 2001). Pathway enrichment analysis of significant proteins was carried out using the Kyoto encyclopedia of genes and genomes (KEGG) database. A protein-protein interaction network was generated using Cytoscape version 2.8.3 (Shannon et al., 2003) with a spring-embedded layout.

RESULTS

Evaluation of the virulence potential of 258_*equi* after passage in a murine host

In the first *in vivo* assay, BALB/c mice were infected with 10^6 CFU of bacteria that had no previous host contact and with bacteria that were recovered from mouse spleens. We observed that all of the infected animals under both the control (Ct) and recovered (Rc) conditions died within 48 h of infection (**Figure 1A**). These results reveal the virulence potential of 258_*equi*; however, in this assay, we did not observe differences in the virulence potential between the control and recovered condition. Next, the Ct and Rc condition were analyzed using a new survival assay in BALB/c mice, but a 10^5 CFU infection dose was used (**Figure 1B**). In this assay, we observed altered virulence in the Rc condition; the mice began dying in the first 10 days post-infection (40% decrease in survival rate) and mortality reached 100% in less than 20 days post-infection. For the Ct condition, while mice died during the first 10 days of infection, this early stage only resulted in 20% mortality, and mortality did not reach 100% until 23 days post-infection. Finally, we detected abscess formation in the internal organs (kidney and liver) of all animals infected with either the Ct or Rc condition in the assays using 10^5 CFU (data not shown). These results show that passage in a murine host affects the virulence potential of 258_*equi*.

Overview of the exoproteome of *C. pseudotuberculosis* strain 258 after passage in a murine Host

After passage of 258_*equi* in BALB/c mice, we detected changes in its virulence potential. To assess whether this change is reflected in its proteome, considering the importance of exported proteins in bacterial infection (Hilbi et al., 2012), we used a TPP/LC-MS^E approach (Pacheco et al., 2011) to compare the extracellular proteome of the control and recovered conditions. From our proteomic analysis, a total of 113 non-redundant 258_*equi* proteins were detected with high confidence and were identified in at least two of the three biological replicates of the two conditions tested, with an average of 17 peptides per protein and an FDR of 1%. The peptides were identified with a normal distribution of 10 ppm error for the total identified peptides (**Supplementary Figure 1A**). In addition, only the source fragments of peptides with a charge state of at least $[M + 2H]^{2+}$ and the absence of decoys were considered to increase data quality.

The absolute quantitation of proteins present within a complex protein mixture is extremely important for understating physiological adaptations in response to biological demands promoted by environmental changes (Mallick and Kuster, 2010). To estimate the absolute abundance of identified proteins in the 258_*equi* exoproteome, we utilized the Hi3 method (Silva et al., 2006) where the average MS signal responses for the three most intense tryptic peptides for each protein were determined, including those of the internal standard protein glycogen phosphorylase (muscle form; P00489). All samples were normalized prior to injection using "scouting runs," and the stoichiometry between the intensity and molarity proportion prior to the replicate runs per condition were considered. From this analysis a dynamic range of protein abundance was generated spanning three orders of magnitude (**Supplementary Figure 1B**). Lysozyme M1 was the most abundant protein detected. This protein, which is related to bacterial virulence, is localized in the pathogenic island Cp258PiCp02. Lysozyme M1 was also detected in a membrane shaving of a field isolate of *C. pseudotuberculosis* biovar *ovis* (Rees et al., 2015). Other proteins, such as hydrolase domain containing protein, trehalose corynomycolyl transferase B, which is involved in the cell wall synthesis, and FtsX, a protein related to division cellular, were among the most abundant proteins. All of the identified proteins on the protein abundance scale are listed in **Supplementary Table 1**.

For evaluating the relative differences between the core exoproteome of the Ct and Rc conditions, we used label-free quantification (Silva et al., 2005, 2014; Pacheco et al., 2011). In agreement with the PLGS analyses, 105 proteins between the Rc and Ct conditions presented significant statistical values ($p < 0.05$) using the Expression^E algorithm tool (**Supplementary Table 2**). Differential expression was considered for proteins that were significantly different ($p < 0.05$) and had $\log_2$ ratios equal to or greater than a factor of 1.2, as described by Levin et al. (2011). Based on this analysis, 39 proteins were induced and 16 were down-regulated in the Rc condition (**Table 1**). In addition, we detected proteins exclusive to the proteome of each condition; cytochrome c nitrate reductase small subunit NrfH was detected only in the Ct condition. In only the Rc condition, two multidrug resistance proteins, the cytochrome oxidase assembly, a thioredoxin-related protein and three proteins with unknown function were identified (**Supplementary Table 3**).

In silico prediction of 258_*equi* exoprotein localization

Extracellular proteins produced by prokaryotic organisms are a subset of proteins present in the extracellular milieu, which is composed of both proteins with signal peptides that are actively secreted by classical secretion systems and proteins without signal sequences that are exported by non-classical secretion systems (Bendtsen et al., 2005a; Desvaux et al., 2009). To identify proteins that contain signal peptides and to determine their subcellular localizations, we utilized the SurfG tool (Barinov et al., 2009), which enables the classification of proteins within the following categories: cytoplasmic (CYT), membrane (MEM), potentially surface-exposed (PSE) and secreted (SEC) (**Figure 2A** and **Supplementary Tables 2**, **3**). Of the total proteins identified, 66% ($n = 74$) presented positive predictions for signal peptides (**Figure 2B**). This group was composed of predicted proteins in the SEC and PSE categories. When these results were compared with the *in silico* data of the 258_*equi* genome, we had identified approximately 65% and 16% of the proteins predicted to be SEC and PSE, respectively. The proteins that did not present positive predictions for signal peptides were analyzed by SecretomeP (Bendtsen et al., 2005a) to identify proteins that could eventually be exported by non-classical secretion systems. According this analysis, nine proteins were predicted to be PSE, seven proteins were predicted to be MEM and two proteins were predicted to be CYT as they presented High SecP scores above 0.5 (**Supplementary Tables 2**, **3**), suggesting that these proteins might be exported by a non-classical secretion system. Taken together, 86% of the 258_*equi* exoproteome was

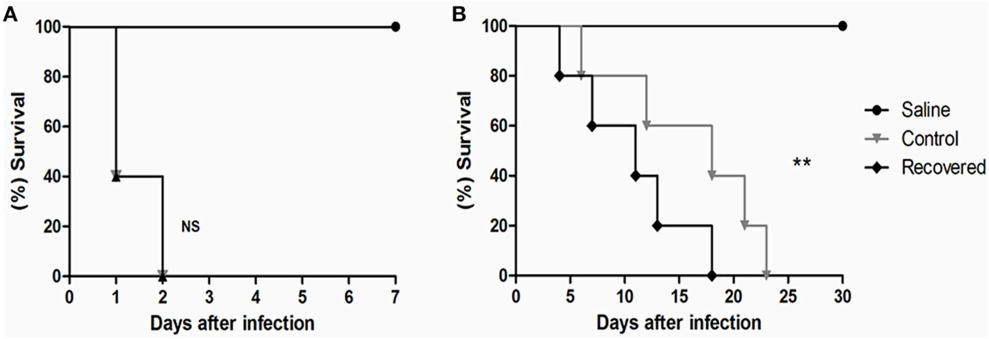


FIGURE 1 | Survival assay of Balb/C mice infected with strain 258_*equi*. **(A)** Percent survival of BALB/c mice infected with 10^6 CFU of bacteria. **(B)** Percent survival of BALB/c mice infected with 10^5 CFU of bacteria. Ct, control condition and Rc, recovered condition. The mortality rates were measured daily. The results presented in **(A,B)** represents three independent experiments. The *p*-values were calculated using the log rank test. Note that infection with 10^5 CFU of bacteria changes the potential virulence of Rc ($p = 0.0024$, log-rank test) relative to Ct. NS = $P > 0.05$; $^{**}P < 0.01$.

composed of extracellular proteins (**Figure 2B**). In addition, our proteomic analysis detected 17 proteins predicted to be lipoproteins (**Supplementary Table 2**).

Functional Classifications of the differentially Expressed proteins in strain 258_*equi* after passage in a murine Host

To evaluate the functional characteristics of the 258_*equi* exoproteome, we performed a Clusters of Orthologous Groups analysis (Tatusov et al., 2001). According GO analysis, the proteins were organized by clusters of orthologous groups (**Figure 3A**). When we evaluated each functional category, we observed that the majority of the proteins detected as induced in the Rc condition were predicted as “function unknown” and “general function only” (**Table 1** and **Figure 3B**). These results represent a lack of knowledge regarding a protein set that might play an important role in the pathogenic process of 258_*equi*, and therefore, more studies are necessary to investigate the true roles of these proteins in the virulence of this strain. When we evaluated proteins with known or predicted functions, the majority of those that were more abundant in the Rc condition were related to cellular processes and signaling (**Figures 3A,B**). According to *in silico* data of the 258_*equi* genome, this pathogen has five iron uptake systems (**Supplementary Figure 2**; Soares et al., 2013a). Interestingly, in our proteomic analysis, we identified components of each of the 5 systems as more abundant in the Rc condition, including CiuA, FhuD, FagC, HmuT, HmuV, and HtaA (**Table 1**), suggesting that iron uptake pathways may play an important role in the pathogenesis of *C. pseudotuberculosis*. Moreover, we detected several proteins related to bacterial pathogenesis that contribute to processes of adherence, intracellular growth and evasion of the immune system (**Table 1** and **Supplementary Tables 2**, **3**).

To identify the most relevant biological pathways of the proteins differentially expressed between the Ct and Rc conditions, we performed a KEGG enrichment analysis. Enrichment results revealed eight biological pathways with significant differences ($p < 0.05$). The proteins that were induced in the Rc condition are in pathways such as ABC transporters, bacterial secretion systems, peptidoglycan biosynthesis and protein export (**Figure 3C**). This finding confirms that secretion and secreted proteins are key elements in *C. pseudotuberculosis* virulence and adaptation, as suggested by previous reports that identified several secreted proteins as potential virulence factors in *C. pseudotuberculosis* (Pacheco et al., 2011; Silva et al., 2013; Rees et al., 2015). Most proteins perform their function in a context of networks by interacting with other proteins (Schleker et al., 2012). To evaluate the 258_*equi* exoproteome at the network level, we performed a protein-protein interaction analysis of the differentially expressed proteins using the Cytoscape tool. After Cytoscape analysis, the 258_*equi* exoproteome network was composed of 87 proteins (**Figure 4**). In the PPI-network, we observed enrichment clusters in heme biosynthesis and ABC transporters related to iron uptake, peptidoglycan biosynthesis and antibiotic biosynthesis. In addition, we observed that some clusters were formed by unknown proteins, which shows that these proteins may play an important role in the virulence of *C. pseudotuberculosis*.

DISCUSSION

Here, we report a comprehensive analysis of the exoproteome of an *equi* isolate of *C. pseudotuberculosis*. The 258_*equi* exoproteome was composed of a high number of extracellular proteins, and a similar result was observed in a study conducted by Pacheco et al. (2011), which characterized the extracellular proteomes of *C. pseudotuberculosis* biovar *ovis* strains. The infections caused by *C. pseudotuberculosis* are chronic in character, and due to this, post-infection disease signs may not begin to appear until after 6 months. Necropsy is the only viable way to identify abscesses, but the cost is high. Testing several strains requires many animals and would result in high economic and ethical costs. Thus, mice are used as an alternative model for studying *C. pseudotuberculosis* infection, because they are relatively resistant to experimental challenge and are able to contain infection (Jolly, 1965; Zaki, 1966). In addition, mice have been shown to be efficient for the evaluation of different vaccine compounds and of humoral and cellular immune responses

TABLE 1 | Proteins differentially expressed between the recovered (Rc) and control (Ct) conditions.

Accession	Description	Score	Gene	Rc:Ct Log(2)Ratio
ADHESION AND MOTILITY CELL				
I3QZX5_CORPS	Sortase A	13261.7	*srtA*	1.28
AMINO ACID TRANSPORT AND METABOLISM				
I3QWW3_CORPS	Diaminopimelate decarboxylase	600.78	*lysA*	−1.27
I3QXT1_CORPS	Chorismate synthase aroC	614.35	*aroC*	−1.54
I3QY54_CORPS	4-hydroxy-tetrahydrodipicolinate reductase	1212.64	*dapB*	−1.67
CELL DIVISION AND CELLULAR CYCLE				
I3QW64_CORPS	Cell division protein FtsX	2967.84	*ftsX*	1.88
I3QYI0_CORPS	Cell division protein FtsQ	4225.31	*ftsQ*	1.24
I3QYH4_CORPS	Antigen 84	31337.36	*ag84*	1.23
CELL WALL/MEMBRANE AND ENVELOPE BIOGENESIS				
I3R031_CORPS	Trehalose corynomycolyl transferase B	112067.8	*cmtB*	3.10
I3QV43_CORPS	Penicillin binding protein transpeptidase	18512.68	*pbpB*	1.80
I3QZM5_CORPS	D-alanyl-D-alanine carboxypeptidase	2988.16	*pbp4*	1.46
I3QX04_CORPS	Mycothiol acetyltransferase	5934.73	*mshD*	−4.62
COENZYME METABOLISM				
I3QVB7_CORPS	Uroporphyrinogen decarboxylase	3376.83	*hemE*	1.89
DNA METABOLISM: REPLICATION, RECOMBINATION AND REPAIR				
I3QXT3_CORPS	Amino deoxychorismate lyase	4162.99	*yceG*	1.56
GENERAL FUNCTION PREDICTION ONLY				
I3QZJ3_CORPS	Lipoprotein LpqE	44732.01	*lpqE*	3.74
I3QXX7_CORPS	Lipoprotein	36815.57	*Cp258_1221*	3.62
I3QXE1_CORPS	Hemolysin related protein	893.83	*tlyC*	1.75
I3QXC3_CORPS	Esterase	496.62	*Cp258_1017*	1.30
I3QZ50_CORPS	Peptidase S8A Subtilisin family	40174.72	*Cp258_1653*	1.23
I3QW71_CORPS	Periplasmic binding protein	11884.29	*fecB*	1.21
I3QW24_CORPS	Hydrolase domain containing protein	17234.12	*Cp258_0564*	−1.26
I3QYP5_CORPS	MutT NUDIX family protein	5870.55	*Cp258_1498*	−1.38
I3QV42_CORPS	Protein yqeY	23153.72	*yqeY*	−1.57
I3R0F7_CORPS	Anthranilate synthase component II	382.53	*trpG*	−1.63
I3QXJ1_CORPS	Prolipoprotein LppL	2671.28	*lppL*	−3.38
I3QZA3_CORPS	Protein NrdI	7211.97	*nrdI*	−4.70
INORGANIC ION TRANSPORT AND METABOLISM				
I3QVU3_CORPS	Cell surface hemin receptor HtaA	11874.87	*htaA*	2.22
I3QUS8_CORPS	Iron-regulated membrane protein	7389.79	*piuB*	1.85
I3QUW4_CORPS	ABC type metal ion transport system	1016.6	*mntA*	1.75
I3QXC5_CORPS	CiuA protein	12242.99	*ciuA*	1.64
I3QVU4_CORPS	Hemin binding periplasmic protein HmuT	14010	*hmuT*	1.47
I3QVU6_CORPS	Hemin import ATP binding protein HmuV	785.22	*hmuV*	1.34
I3QUM8_CORPS	FagC protein	392.95	*fagC*	1.23
I3QX10_CORPS	Iron(3+)-hydroxamate-binding protein FhuD	13922.33	*fhuD*	1.21
I3QUW5_CORPS	Manganese zinc iron transport system ATP-binding	391.52	*mntB*	−1.38
INTRACELLULAR TRAFFICKING, SECRETION, AND VESICULAR TRANSPORT				
I3QX59_CORPS	ABC transporter domain containing protein	1150.79	*Cp258_0956*	1.79
I3QXV8_CORPS	Protein translocase subunit SecF	1729.78	*secF*	1.66
I3R0D7_CORPS	Oligopeptide binding protein OppA	3469.47	*oppA7*	1.50
I3QWP1_CORPS	Oligopeptide binding protein OppA	41781.98	*oppA3*	1.38
I3QZC0_CORPS	ABC type antimicrobial peptide transport	1159.51	*Cp258_1740*	1.34
I3QXV9_CORPS	Protein translocase subunit SecD	2879.45	*secD*	1.24
LIPID TRANSPORT AND METABOLISM				
I3QW96_CORPS	Enoyl CoA hydratase echA6	611.75	*echA6*	−1.40
POST-TRANSLATIONAL MODIFICATION, PROTEIN TURNOVER, CHAPERONES				
I3QW38_CORPS	Lon protease	6439.26	*lon*	1.34
UNKNOWN FUNCTION				
I3QYD5_CORPS	Unknown Function	8143.98	*Cp258_1380*	4.17
I3QZP8_CORPS	Unknown Function	14676.76	*Cp258_1869*	2.93
I3QW25_CORPS	Unknown Function	1433.27	*Cp258_0565*	2.74
I3R0E2_CORPS	Unknown Function	141172.41	*Cp258_2121*	2.60
I3QW83_CORPS	Unknown Function	1589.85	*Cp258_0622*	2.39
I3QWP8_CORPS	Unknown Function	54353.63	*Cp258_0793*	2.29
I3R049_CORPS	Unknown Function	2318.32	*Cp258_2028*	2.11
I3QZK0_CORPS	Unknown Function	1030.59	*Cp258_1819*	1.37
I3QV90_CORPS	Unknown Function	12638.51	*Cp258_0263*	1.34
I3QVZ1_CORPS	Unknown Function	2256.66	*Cp258_0531*	−1.28
I3QYV3_CORPS	Unknown Function	141.46	*Cp258_1555*	−1.53
I3QWK1_CORPS	Unknown Function	221.77	*Cp258_0745*	−2.02
I3R080_CORPS	Unknown Function	2564.2	*Cp258_2060*	−2.64

(Simmons et al., 1997; Lan et al., 1999; Gorman et al., 2010; Droppa-Almeida et al., 2016). Other studies have used mice to study virulence and pathogenesis, including an evaluation of hepatic disease (Nieto et al., 2009), or to study knockout strains (Moraes et al., 2014; Ribeiro et al., 2014).

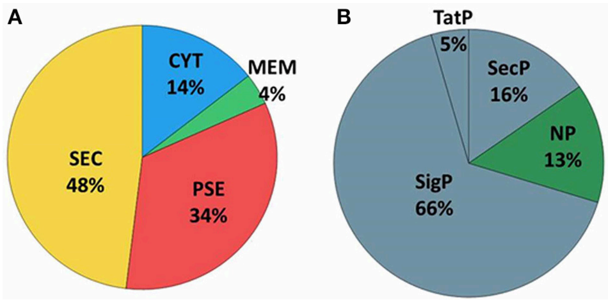


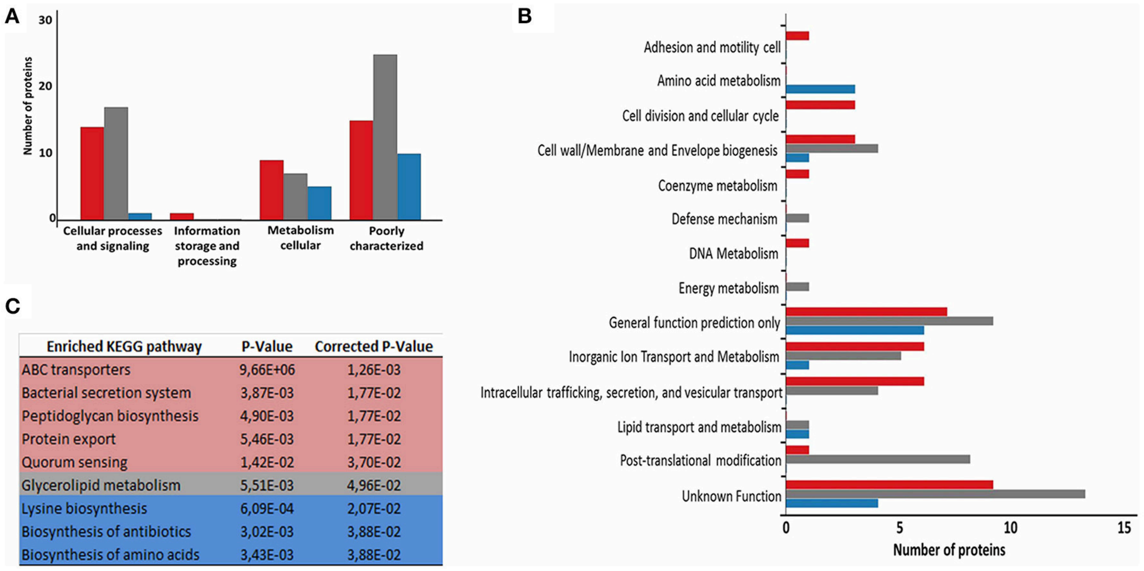
FIGURE 2 | Prediction of the subcellular localization of the 258_*equi* exoproteome. **(A)** CYT, cytoplasmic; MEM, membrane; PSE, potentially surface-exposed; and SEC, secreted. **(B)** SecP, SecretomeP prediction for non-classical pathways; TatP, Tat-pathway prediction; SigP, SignalP prediction for peptide signal prediction; and NP, no prediction for SecP, TatP, or SigP.

In our study, the serial passage process in mice was efficient to induce changed both virulence and functional genome of 258_*equi*, which was showed through of proteomic analysis. In the *in vivo* assay, we observed changes in its virulence potential in a new infection assay with a lower infection dose. Similar results from an evaluation of the virulence potential of Shiga toxin (Stx)-producing *Escherichia coli* (STEC) were also observed when mice were infected with a lower dose of the recovered bacteria that had been recovered after serial passage in a murine model (Fernandez-Brando et al., 2012). Other studies have also shown that the serial passage process through *in vitro* or *in vivo* models leads to changes in the virulence potential of pathogens, including *Helicobacter pylori*, *Escherichia coli, Xenorhabdus nematophila*, *Arcobacter butzleri*, *Salmonella enterica,* and *Shigella flexneri* (Fernández et al., 2000, 2013; Bleich et al., 2005; Chapuis et al., 2011; Fernandez-Brando et al., 2012; Liu et al., 2015). Changes in the virulence potential of these pathogens, as well as in 258_*equi*, show that this strategy promotes the activation of genes related to bacterial pathogenesis.

A proteomic study conducted with *Shigella flexneri* after passage in an *in vivo* model showed the induction of important proteins that might contribute to its adaptation process during infection (Liu et al., 2015). In our proteomic analysis, we also observed changes in the 258_*equi* exoproteome after the recuperation process, and proteins that might play an important role in the pathogenesis of *C. pseudotuberculosis* were detected (**Figure 5**). Interestingly, when compared the proteins that were differentially induced in the Rc condition, with *in silico* data of the core-genome of *C. pseudotuberculosis* biovar *equi* and biovar *ovis* strains (Soares et al., 2013b), we observed that all proteins are present in this core-genome. This result represents a set of proteins that might be important to pathogenesis of biovar *equi* and biovar *ovis* strains. Within our proteomic repertoire, we detected predicted proteins such as lipoprotein. This class of proteins is produced by several prokaryotic organisms and then translocated across the membrane through the Sec or Tat pathway (Pugsley, 1993; Shruthi et al., 2010). Different studies have shown that these peripherally anchored membrane proteins perform an important role in the physiology, virulence and immune response of different gram-negative and gram-positive pathogens. In addition, lipoproteins are recognized as excellent vaccine targets (Nguyen and Götz, 2016). In the closely related pathogen *M. tuberculosis*, lipoproteins have been shown to be extremely important for virulence, contributing directly to evasion of the immune system (Su et al., 2016).

Adhesion to host cells is a key determinant that contributes to bacteria–host interaction; this process is required for bacterial colonization and persistence. *In vitro* and *in silico* studies showed that *C. pseudotuberculosis* contains pili, and these structures play an important role in cellular adhesion (Yanagawa and Honda, 1976; Soares et al., 2013b). In *C. pseudotuberculosis*, *spaA* is a major pili gene that is encoded by the following gene cluster: *srtB-spaA-srtA-spaB-spaX-spaC* (Soares et al., 2013b). We found that sortase A (SrtA) was induced in the Rc supernatant. This cell surface anchored transpeptidase catalyzes the covalent attachment of precursor cell wall-attached proteins (LPXTG proteins) to the peptidoglycan. In gram-positive pathogens, such as *Listeria monocytogenes* (Bierne et al., 2002), *Streptococcus pneumoniae* (Kharat and Tomasz, 2003), and *Staphylococcus aureus* (Oh et al., 2006), *srtA* mutant strains had reduced virulence in animal infection models. We also detected important proteins related to the cell division and growth of *Corynebacterium* induced in the Rc condition, such as FtsQ and FtsX, which form part of the *ftsXE* cluster, and penicillin-binding proteins (PBPs) (Letek et al., 2008). In *E. coli*, the FtsEX proteins were suggested to form an ABC transporter system involved in the uptake of substrates necessary to maintain osmotic pressure during cell division (Schmidt et al., 2004; Reddy, 2007). FtsX was detected among the most abundant proteins of the 258_*equi* exoproteome, which suggests it is an important protein within the biology of this strain. PBPs proteins have an important role in cell-wall biosynthesis in *Corynebacterium* as they are essential to peptidoglycan biosynthesis. In addition, this class of proteins is a target of antibiotics (Letek et al., 2008). Antigen 84 (Ag84) was also induced in the Rc condition. Interestingly, Ag84 was also detected in a membrane shaving of an *ovis* strain isolated directly from the caseous nodes of a diseased animal (Rees et al., 2015). In *M. tuberculosis*, this protein presents antigenic characteristics and is required for growth (Sassetti et al., 2003). These proteins may have key functions in the replication and growth of *C. pseudotuberculosis* during the infection process.

Iron is an essential element for both the virulence and growth of several bacterial pathogens during the infection process. However, free iron is not available to the bacterial inside the host, thus several pathogens utilize different mechanisms to acquire both free iron and iron from host iron proteins (Brown and Holden, 2002). For *C. pseudotuberculosis*, iron acquisition is a required step in its pathogenic process (Billington et al., 2002; Ribeiro et al., 2014), and according to an *in silico* analysis of the 258_*equi* genome, this bacterium has different genetic loci associated with high-affinity iron transport systems as well as surface-associated heme-uptake pathways (Soares et al., 2012). In our proteomic analysis, we detected that specific proteins related to iron acquisition were induced in the Rc condition. Some these proteins are involved directly in the virulence of *C. pseudotuberculosis*, such as the FagC protein that is component



Enriched KEGG pathway	P-Value	Corrected P-Value
ABC transporters	9,66E+06	1,26E-03
Bacterial secretion system	3,87E-03	1,77E-02
Peptidoglycan biosynthesis	4,90E-03	1,77E-02
Protein export	5,46E-03	1,77E-02
Quorum sensing	1,42E-02	3,70E-02
Glycerolipid metabolism	5,51E-03	4,96E-02
Lysine biosynthesis	6,09E-04	2,07E-02
Biosynthesis of antibiotics	3,02E-03	3,88E-02
Biosynthesis of amino acids	3,43E-03	3,88E-02

FIGURE 3 | Functional analysis of the differentially expressed proteome between the control and recovered conditions. **(A)** Proteins classified by COG functional categories. **(B)** Categorization of differentially expressed proteins in biological processes. **(C)** KEGG pathway enrichment analysis of differentially expressed proteins. The colors are based on the Rc:Ct relation; red, up-regulated; gray, unchanged; and blue, down-regulated.

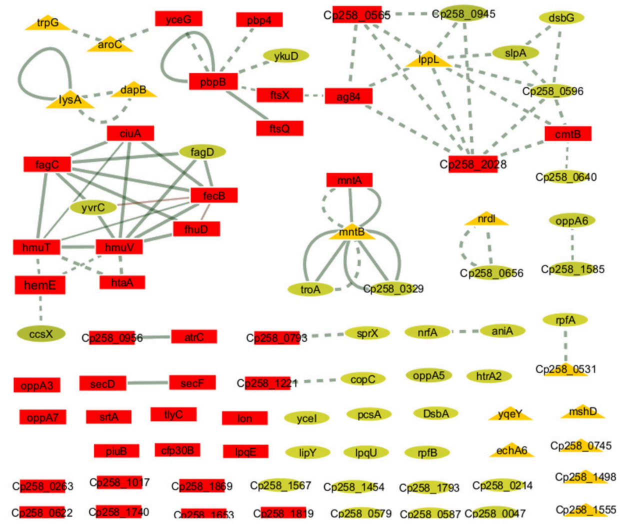


FIGURE 4 | Protein-protein interaction network. The network nodes represent proteins, and the edges represent protein-protein associations. The node size (protein) is proportional to the amount of protein interacting (degree of interaction). Dotted line = regulatory interactions (functional), solid line, physical interactions; triangles, down-regulated proteins in the Rc condition; squares, highly induced proteins in the Rc condition; and circles, unchanged proteins.

of the Fag system. A study performed with strains with defective *fagB(C)* genes showed that these strains presented reduced virulence in goats (Billington et al., 2002).

To acquire iron inside a host, bacteria synthesize and secrete siderophores, which are low-molecular-weight iron chelators that have a high and specific affinity for ferric iron (Ellermann and Arthur, 2017). We detected the CiuA siderophore, which is localized in the operon *ciuABCDE* (*Corynebacterium* iron uptake). Studies performed with the pathogen *C. diphtheria* showed that Ciu is a high-affinity iron uptake system (Kunkle and Schmitt, 2005). Additionally, in a study performed with *C. pseudotuberculosis*, a *ciuA* mutated strain showed reduced virulence, demonstrating the role of this protein in the virulence of this bacterium, and was also able to protect immunized mice when they were challenged with a virulent strain (Ribeiro et al., 2014). Another siderophore detected was the FhuD siderophore, which is part of the conserved ferric hydroxamate uptake system, Fhu. The uptake of ferric ferrichrome is described in pathogens such as *L. monocytogenes* (Jin et al., 2006; Xiao et al., 2011), *Streptococcus pyogenes* (Hanks et al., 2005) and *S. aureus* (Sebulsky and Heinrichs, 2001). In *L. monocytogenes*, FhuD contributes to the uptake of ferric hydroxamate from ferrichrome, ferrichrome A and ferrioxamine B (Jin et al., 2006; Xiao et al., 2011). In *S. aureus*, this protein was shown to contribute both to proliferation within the blood and to the formation of renal abscesses in mice (Mishra et al., 2012). Like *C. diphtheria*, the genome of 258_*equi* also has genetic loci with genes related to heme acquisition, such as the hemin-uptake (hmu) operon *HmuTUV* and cell hemin specific receptors *htaA, htaB* and *htaC*. The ABC hemin transporter HmuTUV, together with cell-surface hemin receptors, is involved in heme uptake from hemoglobin (Hb), hemoglobin/haptoglobin, and myoglobin (Mb) (Kunkle and Schmitt, 2005; Allen and Schmitt, 2015). The presence of these systems shows the versatility of 258_*equi* in acquiring iron from different sources. Interestingly, the HtaA protein was detected to be immunoreactive in an immunoproteomic study of *C. pseudotuberculosis* biovar *ovis*

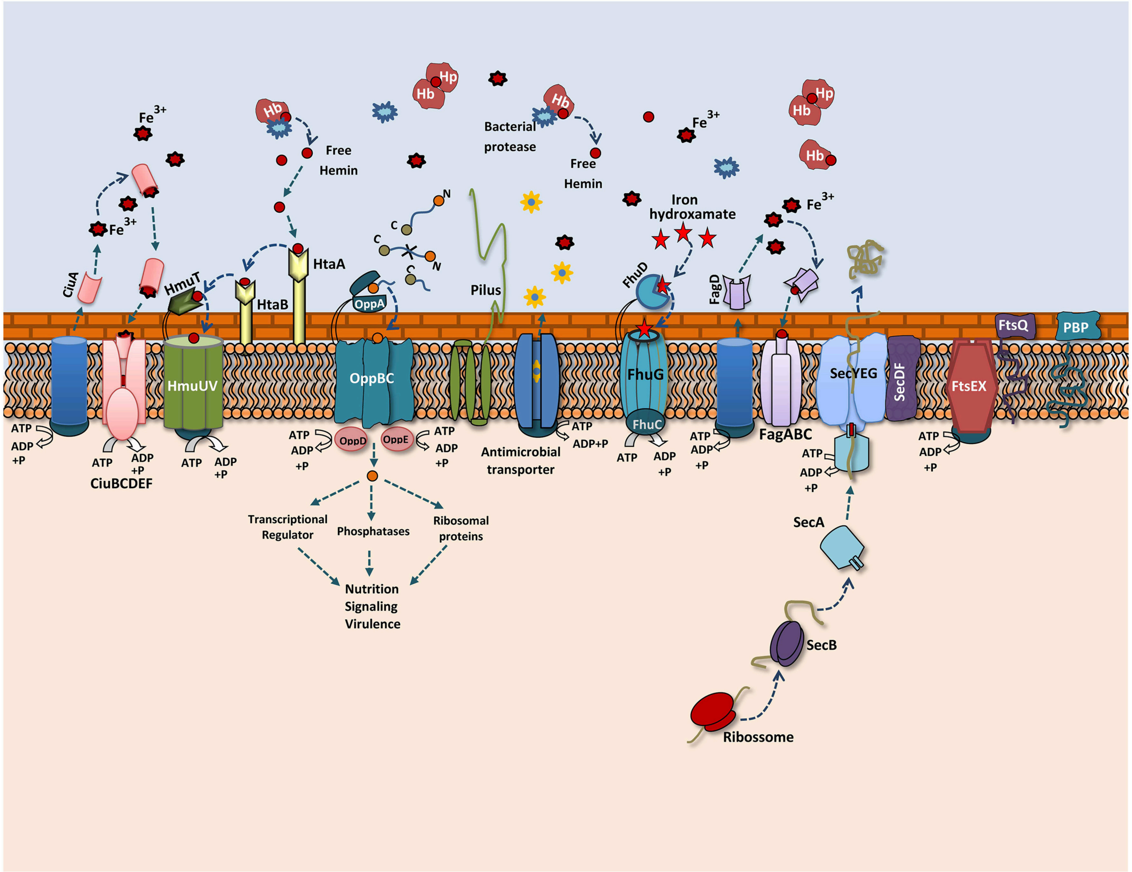


FIGURE 5 | Overview of the 258_*equi* proteome after the recuperation process. A model representing the main exoproteins induced in the recovered condition, including proteins related to biogenesis of the cell wall, cellular adhesion and different secretion pathways related to iron acquisition, bacterial nutrition, efflux pumps and the Sec pathway.

(Seyffert et al., 2014). Taken together, these results indicate that, similar to other pathogens, iron acquisition likely plays an important role in 258_*equi* virulence as this strain uses distinct iron acquisition systems during infection. This ability to acquire iron may contribute to the increase virulence observed in the 258_*equi* strain.

Opp transport systems belong to the superfamily of conserved ATP-binding cassette transporters and play an important role in bacterial nutrition, signaling and virulence (Yu et al., 2014). The OppA protein, which is responsible for the uptake of peptides from the external medium, was induced in the Rc supernatant. In *Mycobacterium avium*, the *oppA* gene contributed to infections in a mouse model as well as to its viability in macrophages (Danelishvili et al., 2014). The Sec pathway is the major secretion system in several prokaryotic pathogens, components of this system was also induced in the Rc supernatant. SecDF are accessory factors from this translocation machinery and act to increase protein translocation. Different studies show that in *S. aureus*, the role of SecDF is related to the export of several virulence factors that contribute to parts its pathogenic process, such as adhesion, invasion and immune system evasion (Sibbald et al., 2006). In addition, SecDF belong to the resistance-nodulation-cell division (RND) family of multidrug export pumps and contribute to the resistance process against the antimicrobial effects of cathelicidins, a class of antimicrobial peptides produced by the immune system (Blodkamp et al., 2017). Similarly, proteins for antimicrobial agent resistance, such as ABC-type antimicrobial peptide transporters, which are localized in the pathogenicity island Cp258PiCp14, and efflux transporters, such as NorM, which belongs to the multidrug and toxic compound extrusion (MATE) transporter, were also detected. These data are consistent with previous *in vitro* studies, which showed that *C. pseudotuberculosis* is resistant to several classes of antimicrobial agents (Judson and Songer, 1991), and that activation of these defense pathways against antimicrobial agents might contribute to survival of this pathogen.

CONCLUSION

Herein, we characterized, for the first time, the exoproteome of a *C. pseudotuberculosis equi* isolate. In addition, we showed changes in both the virulence and proteomic profiles of 258_*equi* after its recovery from murine host spleens. Through a TPP/LC-MS^E approach, we detected secreted virulence-associated proteins. The up-regulation of these proteins may account for the difference in virulence potential we observed in the Rc condition compared with the Ct condition. Altogether, our proteomic repertoire identified several extracellular proteins involved in key processes of bacterial pathogenesis that might contribute to the pathogenic process of *C. pseudotuberculosis*.

AUTHOR CONTRIBUTIONS

WS, VA, and YL Conceived and designed the experiments. WS and FD performed *in vivo* experiments. WS and RD performed microbiological analyses and sample preparation for proteomic analysis. WS and GS conducted the proteomic analysis. EF performed bioinformatics analysis. AP, YL, and HF contributed substantially to data interpretation and revisions. AS, VA, and YL participated in all steps of the project as coordinators, and critically reviewed the manuscript.

FUNDING

The work was supported by the Brazilian Federal Agency for the Support and Evaluation of Graduate Education (CAPES), Pará Research Foundation (FAPESPA), Minas Gerais Research Foundation (FAPEMIG) and the National Council for Scientific and Technological Development (CNPq). YL is the recipient of a PVE grant (71/2013) from Programa Ciências sem Fronteiras.

ACKNOWLEDGMENTS

The authors would like to thank to Pará State Genomics and Proteomics Network and Waters Corporation, Brazil.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fcimb.2017.00325/full#supplementary-material

Supplementary Figure 1 | (A) Two-dimensional nanoUPLC HDMS^E analysis showing the distribution of fragment masses and the exact mass accuracy for 90% of the precursor ions with a 10 ppm maximum error. **(B)** Dynamic range based on the absolute quantitation of the proteins identified by LC-HDMS^E analysis.

Supplementary Figure 2 | Genomic loci encoding proteins related to iron-acquisition in 258_*equi*. Red genes encode proteins that were identified in our proteomic analysis.

Supplementary Table 1 | Total list of identified proteins on the protein abundance scale.

Supplementary Table 2 | Total list of differentially expressed proteins between the recovered and control conditions of stain 258_*equi*.

Supplementary Table 3 | Proteins unique to the recovered and control conditions.

REFERENCES

Aleman, M., Spier, S. J., Wilson, W. D., and Doherr, M. (1996). *Corynebacterium pseudotuberculosis* infection in horses: 538 cases (1982-1993). *J. Am. Vet. Med. Assoc.* 15, 804–809.

Allen, C. E., and Schmitt, M. P. (2015). Utilization of host iron sources by *Corynebacterium diphtheriae*: multiple hemoglobin-binding proteins are essential for the use of iron from the hemoglobin-haptoglobin complex. *J. Bacteriol.* 197, 553–562. doi: 10.1128/JB.02413-14

Barinov, A., Loux, V., Hammani, A., Nicolas, P., Langella, P., Ehrlich, D., et al. (2009). Prediction of surface exposed proteins in *Streptococcus pyogenes*, with a potential application to other Gram-positive bacteria. *Proteomics* 9, 61–73. doi: 10.1002/pmic.200800195

Bendtsen, J. D., Kiemer, L., Fausboll, A., and Brunak, S. (2005a). Non-classical protein secretion in bacteria. *BMC Microbiol.* 5:58. doi: 10.1186/1471-2180-5-58

Bendtsen, J. D., Nielsen, H., Widdick, D., Palmer, T., and Brunak, S. (2005b). Prediction of twin-arginine signal peptides. *BMC Bioinformatics* 2:167. doi: 10.1186/1471-2105-6-167

Bierne, H., Mazmanian, S. K., Trost, M., Pucciarelli, M. G., Liu, G., Dehoux, P., et al. (2002). Inactivation of the srtA gene in *Listeria monocytogenes* inhibits anchoring of surface proteins and affects virulence. *Mol. Microbiol.* 43, 869–881. doi: 10.1046/j.1365-2958.2002.02798.x

Billington, S. J., Esmay, P. A., Songer, J. G., and Jost, B. H. (2002). Identification and role in virulence of putative iron acquisition genes from *Corynebacterium pseudotuberculosis*. *FEMS Microbiol. Lett.* 208, 41–45. doi: 10.1111/j.1574-6968.2002.tb11058.x

Bleich, A., Köhn, I., Glage, S., Beil, W., Wagner, S., and Mähler, M. (2005). Multiple *in vivo* passages enhance the ability of a clinical *Helicobacter pylori* isolate to colonize the stomach of Mongolian gerbils and to induce gastritis. *Lab. Anim.* 39, 221–229. doi: 10.1258/0023677053739800

Blodkamp, S., Kadlec, K., Gutsmann, T., Quiblier, C., Naim, H. Y., and Schwarz, S. (2017). Effects of SecDF on the antimicrobial functions of cathelicidins against *Staphylococcus aureus*. *Vet. Microbiol.* 200, 52–58. doi: 10.1016/j.vetmic.2016.03.021

Britz, E., Spier, S. J., Kass, P. H., Edman, J. M., and Foley, J. E. (2014). The relationship between *Corynebacterium pseudotuberculosis* biovar *equi* phenotype with location and extent of lesions in horses. *Vet. J.* 200, 282–286. doi: 10.1016/j.tvjl.2014.03.009

Brown, J. S., and Holden, D. W. (2002). Iron acquisition by Gram-positive bacterial pathogens. *Microb. Infect.* 4, 1149–1156. doi: 10.1016/S1286-4579(02)01640-4

Chapuis, É., Pagès, S., Emelianoff, V., Givaudan, A., and Ferdy, J. B. (2011). Virulence and pathogen multiplication: a serial passage experiment in the hypervirulent bacterial insect-pathogen *Xenorhabdus nematophila*. *PLoS ONE* 31:e15872. doi: 10.1371/journal.pone.0015872

Curty, N., Kubitschek-Barreira, P. H., Neves, G. W., Gomes, D., Pizzatti, L., Abdelhay, E., et al. (2014). Discovering the infectome of human endothelial cells challenged with *Aspergillus fumigatus* applying a mass spectrometry label-free approach. *J. Proteomics* 31, 126–140. doi: 10.1016/j.jprot.2013.07.003

Danelishvili, L., Stang, B., and Bermudez, L. E. (2014). Identification of *Mycobacterium avium* genes expressed during *in vivo* infection and the role of the oligopeptide transporter OppA in virulence. *Microb. Pathog.* 76, 67–76. doi: 10.1016/j.micpath.2014.09.010

Desvaux, M., Hébraud, M., Talon, R., and Henderson, I. R. (2009). Secretion and subcellular localizations of bacterial proteins: a semantic awareness issue. *Trends Microbiol.* 17, 139–145. doi: 10.1016/j.tim.2009.01.004

Dorella, F. A., Pacheco, L. G., Oliveira, S. C., Miyoshi, A., and Azevedo, V. (2006). *Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. *Vet. Res.* 37, 201–218. doi: 10.1051/vetres:2005056

Droppa-Almeida, D., Vivas, W. L., Silva, K. K., Rezende, A. F., Simionatto, S., Meyer, R., et al. (2016). Recombinant CP40 from *Corynebacterium pseudotuberculosis* confers protection in mice after challenge with a virulent strain. *Vaccine* 17, 1091–1096. doi: 10.1016/j.vaccine.2015.12.064

Ellermann, M., and Arthur, J. C. (2017). Siderophore-mediated iron acquisition and modulation of host-bacterial interactions. *Free Radic. Biol. Med.* 105, 68–78. doi: 10.1016/j.freeradbiomed.2016.10.489

Fernandez-Brando, R. J., Miliwebsky, E., Mejías, M. P., Baschkier, A., Panek, C. A., Abrey-Recalde, M. J., et al. (2012). Shiga toxin-producing *Escherichia coli* O157: H7 shows an increased pathogenicity in mice after the passage through the gastrointestinal tract of the same host. *J. Med. Microbiol.* 61, 852–859. doi: 10.1099/jmm.0.041251-0

Fernández, H., Flores, S., P., Villanueva, M., Medina, G. and Carrizo, M. (2013). Enhancing adherence of *Arcobacter butzleri* after serial intraperitoneal passages in mice. *Rev. Argent. Microbiol.* 45, 75–79. doi: 10.1016/s0325-7541(13)70002-6

Fernández, H., Vivanco, T., and Eller, G. (2000). *Expression of invasiveness of Campylobacter jejuni ssp.* jejuni after serial intraperitoneal passages in mice. *J. Vet. Med. B. Infect. Dis. Vet. Public Health* 47, 635–639. doi: 10.1046/j.1439-0450.2000.00392.x

Foley, J. E., Spier, S. J., Mihalyi, J., Drazenovich, N., and Leutenegger, C. M. (2004). Molecular epidemiologic features of *Corynebacterium pseudotuberculosis* isolated from horses. *Am. J. Vet. Res.* 65, 1734–1737. doi: 10.2460/ajvr.2004.65.1734

Geromanos, S. J., Vissers, J. P., Silva, J. C., Dorschel, C. A., Li, G. Z., Gorenstein, M. V., et al. (2009). The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS. *Proteomics* 9, 1683–1695. doi: 10.1002/pmic.200800562

Gilar, M., Olivova, P., Daly, A. E., and Gebler, J. C. (2005). Two-dimensional separation of peptides using RP-RP-HPLC system with different pH in first and second separation dimensions. *J. Sep. Sci.* 28, 1694–1703. doi: 10.1002/jssc.200500116

Gorman, J. K., Gabriel, M., MacLachlan, N. J., Nieto, N., Foley, J., and Spier, S. (2010). Pilot immunization of mice infected with an equine strain of *Corynebacterium pseudotuberculosis*. *Vet. Ther.* 11, E1–E8.

Hanks, T. S., Liu, M., McClure, M. J., and Lei, B. (2005). ABC transporter FtsABCD of *Streptococcus pyogenes* mediates uptake of ferric ferrichrome. *BMC Microbiol.* 5:62. doi: 10.1186/1471-2180-5-62

Hilbi, H., and Haas, A. (2012). Secretive bacterial pathogens and the secretory pathway. *Traffic* 13, 1187–1197. doi: 10.1111/j.1600-0854.2012.01344.x

Jin, B., Newton, S. M., Shao, Y., Jiang, X., Charbit, A., and Klebba, P. E. (2006). Iron acquisition systems for ferric hydroxamates, haemin and haemoglobin in *Listeria monocytogenes*. *Mol. Microbiol.* 59, 1185–1198. doi: 10.1111/j.1365-2958.2005.05015.x

Jolly, R. D. (1965). The pathogenesis of experimental *Corynebacterium* ovis infection in mice. *N.Z. Vet. J.* 13, 141–147. doi: 10.1080/00480169.1965.33618

Judson, R. and Songer, J. G. (1991). *Corynebacterium pseudotuberculosis*: *in vitro* susceptibility to 39 antimicrobial agents. *Vet. Microbiol.* 27, 145–150. doi: 10.1016/0378-1135(91)90005-Z

Kharat, A. S., and Tomasz, A. (2003). Inactivation of the srtA gene affects localization of surface proteins and decreases adhesion of *Streptococcus pneumoniae* to human pharyngeal cells *in vitro*. *Infect. Immun.* 71, 2758–2765. doi: 10.1128/IAI.71.5.2758-2765.2003

Kilcoyne, I., Spier, S. J., Carter, C. N., Smith, J. L., Swinford, A. K., and Cohen, N. D. (2014). Frequency of *Corynebacterium pseudotuberculosis* infection in horses across the United States during a 10-year period. *J. Am. Vet. Med. Assoc.* 245, 309–314. doi: 10.2460/javma.245.3.309

Kunkle, C. A., and Schmitt, M. P. (2005). Analysis of a DtxR-regulated iron transport and siderophore biosynthesis gene cluster in *Corynebacterium diphtheriae*. *J. Bacteriol.* 187, 422–433. doi: 10.1128/JB.187.2.422-433.2005

Lan, D. T., Makino, S., Shirahata, T., Yamada, M., and Nakane, A. (1999). Tumor necrosis factor alpha and gamma interferon are required for the development of protective immunity to secondary *Corynebacterium pseudotuberculosis* infection in mice. *J. Vet. Med. Sci.* 61, 1203–1208. doi: 10.1292/jvms.61.1203

Letek, M., Fiuza, M., Ordóñez, E., Villadangos, A. F., Ramos, A., Mateos, L. M., et al. (2008). Cell growth and cell division in the rod-shaped actinomycete *Corynebacterium glutamicum*. *Antonie Van Leeuwenhoek*. 94, 99–109. doi: 10.1111/j.1574-6968.2009.01679.x

Levin, Y., Hradetzky, E., and Bahn, S. (2011). Quantification of proteins using data-independent analysis (MSE) in simple and complex samples: a systematic evaluation. *Proteomics* 11, 3273–3287. doi: 10.1002/pmic.201000661

Li, G. Z., Vissers, J. P., Silva, J. C., Golick, D., Gorenstein, M. V., and Geromanos, S. J. (2009). Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures. *Proteomics* 9, 1696–1719. doi: 10.1002/pmic.200800564

Liu, X., Lu, L., Liu, X., Pan, C., Feng, E., Wang, D., et al. (2015). Comparative proteomics of *Shigella flexneri* 2a strain using a rabbit ileal loop model reveals key proteins for bacterial adaptation in host niches. *Int. J. Infect. Dis.* 40, 28–33. doi: 10.1016/j.ijid.2015.09.014

Mallick, P., and Kuster, B. (2010). Proteomics: a pragmatic perspective. *Nat. Biotechnol.* 28, 695–709. doi: 10.1038/nbt.1658

McKean, S. C., Davies, J. K., and Moore, R. J. (2007). Expression of phospholipase D, the major virulence factor of *Corynebacterium pseudotuberculosis*, is regulated by multiple environmental factors and plays a role in macrophage death. *Microbiology* 153, 2203–2211. doi: 10.1099/mic.0.2007/005926-0

Mishra, R. P., Mariotti, P., Fiaschi, L., Nosari, S., Maccari, S., Liberatori, S., et al. (2012). *Staphylococcus aureus* FhuD2 is involved in the early phase of staphylococcal dissemination and generates protective immunity in mice. *J. Infect. Dis.* 206, 1041–1049. doi: 10.1093/infdis/jis463

Moraes, P. M., Seyffert, N., Silva, W. M., Castro, T. L., Silva, R. F., and Lima, D. D. (2014). Characterization of the opp peptide transporter of *Corynebacterium pseudotuberculosis* and its role in virulence and pathogenicity. *Biomed. Res. Int.* 2014:489782. doi: 10.1155/2014/489782

Moura-Costa, L. F., Paule, B. J. A., Freire, S. M., Nascimento, I., Schaer, R., Regis, L. F., et al. (2002). Chemically defined synthetic medium for *Corynebacterium pseudotuberculosis* culture. *Rev. Bras. Saúde Prod. An.* 3, 1–9.

Nguyen, M. T., and Götz, F. (2016). Lipoproteins of gram-positive bacteria: key players in the immune response and virulence. *Microbiol. Mol. Biol. Rev.* 10, 891–903. doi: 10.1128/MMBR.00028-16

Nieto, N. C., Foley, J. E., MacLachlan, N. J., Yuan, T., and Spier, S. J. (2009). Evaluation of hepatic disease in mice following intradermal inoculation with *Corynebacterium pseudotuberculosis*. *Am. J. Vet. Res.* 70, 257–262. doi: 10.2460/ajvr.70.2.257

Oh, K. B., Oh, M. N., Kim, J. G., Shin, D. S., and Shin, J. (2006). et al. inhibition of sortase-mediated *Staphylococcus aureus* adhesion to fibronectin via fibronectin-binding protein by sortase inhibitors. *Appl. Microbiol. Biotechnol.* 70, 102–106. doi: 10.1007/s00253-005-0040-8

Pacheco, L. G., Castro, T. L., Carvalho, R. D., Moraes, P. M., Dorella, F. A., Carvalho, N. B., et al. (2012). A role for sigma factor σ(E) in *Corynebacterium pseudotuberculosis* resistance to nitric oxide/peroxide stress. *Front. Microbiol.* 3:126. doi: 10.3389/fmicb.2012.00126

Pacheco, L. G., Slade, S. E., Seyffert, N., Santos, A. R., Castro, T. L., Silva, W. M., et al. (2011). A combined approach for comparative exoproteome analysis of *Corynebacterium pseudotuberculosis*. *BMC Microbiol.* 17:12. doi: 10.1186/1471-2180-11-12

Paule, B. J., Meyer, R., Moura-Costa, L. F., Bahia, R. C., Carminati, R., Regis, L. F., et al. (2004). Three-phase partitioning as an efficient method for extraction/concentration of immunoreactive excreted-secreted proteins of *Corynebacterium pseudotuberculosis*. *Protein Expr. Purif.* 34, 311–316. doi: 10.1016/j.pep.2003.12.003

Pugsley, A. P. (1993). The complete general secretory pathway in gram-negative bacteria. *Microbiol. Rev.* 57, 50–108.

Reddy, M. (2007). Role of FtsEX in cell division of *Escherichia coli*: viability of *ftsEX* mutants is dependent on functional sufI or high osmotic strength. *J. Bacteriol.* 189, 98–108. doi: 10.1128/JB.01347-06

Rees, M. A., Kleifeld, O., Crellin, P. K., Ho, B., Stinear, T. P., Smith, A. I., et al. (2015). Proteomic characterization of a natural host–pathogen interaction: repertoire of *in vivo* expressed bacterial and host surface-associated proteins. *J. Proteome Res.* 14, 120–132. doi: 10.1021/pr5010086

Ribeiro, D., Rocha, F. S., Leite, K. M., Soares, S. C., Silva, A., and Portela, R. W. (2014). An iron-acquisition-deficient mutant of *Corynebacterium pseudotuberculosis* efficiently protects mice against challenge. *Vet. Res.* 6:28. doi: 10.1186/1297-9716-45-28

Sassetti, C. M., Boyd, D. H., and Rubin, E. J. (2003). Genes required for mycobacterial growth defined by high density mutagenesis. *Mol. Microbiol.* 48, 77–84. doi: 10.1046/j.1365-2958.2003.03425.x

Schleker, S., Sun, J., Raghavan, B., Srnec, M., Müller, N., Koepfinger, M., et al. (2012). The current Salmonella-host interactome. *Proteomics Clin. Appl.* 6, 117–133. doi: 10.1002/prca.201100083

Schmidt, K. L., Peterson, N. D., Kustusch, R. J., Wissel, M. C., Graham, B., Phillips, G. J., et al. (2004). A predicted ABC transporter, FtsEX, is needed for cell division in *Escherichia coli*. *J. Bacteriol.* 186, 785–793. doi: 10.1128/JB.186.3.785-793.2004

Sebulsky, M. T., and Heinrichs, D. E. (2001). Identification and characterization of fhuD1 and fhuD2, two genes involved in iron-hydroxamate uptake in *Staphylococcus aureus*. *J. Bacteriol.* 183, 4994–5000. doi: 10.1128/JB.183.17.4994-5000.2001

Seyffert, N., Silva, R. F., Jardin, J., Silva, W. M., Castro, T. L., and Tartaglia, N. R. (2014). Serological proteome analysis of *Corynebacterium pseudotuberculosis* isolated from different hosts reveals novel candidates for prophylactics to control caseous lymphadenitis. *Vet. Microbiol.* 7, 255–260. doi: 10.1016/j.vetmic.2014.08.024

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Shruthi, H., Babu, M. M., and Sankaran, K. (2010). TAT pathway-dependent lipoproteins as a niche-based adaptation in prokaryotes, *J. Mol. Evol.* 70, 359–370. doi: 10.1007/s00239-010-9334-2

Sibbald, M. J., Ziebandt, A. K., Engelmann, S., Hecker, M., de Jong, A., Harmsen, H. J., et al. (2006). Mapping the pathways to staphylococcal pathogenesis by comparative secretomics. *Microbiol. Mol. Biol. Rev.* 70, 755–788. doi: 10.1128/MMBR.00008-06

Silva, J. C., Denny, R., Dorschel, C. A., Gorenstein, M., Kass, I. J., Li, G. Z., et al. (2005). Quantitative proteomic analysis by accurate mass retention time pairs. *Anal. Chem.* 1, 2187–2000. doi: 10.1021/ac048455k

Silva, J. C., Gorenstein, M. V., Li, G. Z., Vissers, J. P., and Geromanos, S. J. (2006). Absolute quantification of proteins by LCMS^{E}: a virtue of parallel MS acquisition. *Mol. Cell. Proteomics* 5, 144–156. doi: 10.1074/mcp.M500230-MCP200

Silva, W. M., Carvalho, R. D., Soares, S. C., Bastos, I. F., Folador, E. L., Souza, G. H., et al. (2014). Label-free proteomic analysis to confirm the predicted proteome of *Corynebacterium pseudotuberculosis* under nitrosative stress mediated by nitric oxide. *BMC Genomics* 4:1065. doi: 10.1186/1471-2164-15-1065

Silva, W. M., Seyffert, N., Santos, A. V., Castro, T. L., Pacheco, L. G., Santos, A. R., et al. (2013). Identification of 11 new exoproteins in *Corynebacterium pseudotuberculosis* by comparative analysis of the exoproteome. *Microb. Pathog.* 62, 37–42. doi: 10.1016/j.micpath.2013.05.004

Simmons, C. P., Hodgson, A. L., and Strugnell, R. A. (1997). Attenuation and vaccine potential of aroQ mutants of *Corynebacterium pseudotuberculosis*. *Infect. Immun.* 65, 3048–3056.

Soares, S. C., Abreu, V. A., Ramos, R. T., Cerdeira, L., Silva, A., et al. (2012). PIPS: pathogenicity island prediction software. *PLoS ONE* 7:e30848. doi: 10.1371/journal.pone.0030848

Soares, S. C., Silva, A., Trost, E., Blom, J., Ramos, R., Carneiro, A., et al. (2013b). The pan-genome of the animal pathogen *Corynebacterium pseudotuberculosis* reveals differences in genome plasticity between the biovar *ovis* and *equi* strains. *PLoS ONE* 8:e53818. doi: 10.1371/journal.pone.0053818

Soares, S. C., Trost, E., Ramos, R. T., Carneiro, A. R., Santos, A. R., Pinto, A. C., et al. (2013a). Genome sequence of *Corynebacterium pseudotuberculosis* biovar *equi* strain 258 and prediction of antigenic targets to improve biotechnological vaccine production. *J. Biotechnol.* 20, 135–141. doi: 10.1016/j.jbiotec.2012.11.003

Spier, S. J. (2008). *Corynebacterium pseudotuberculosis* infection in horses: an emerging disease associated with climate change? *Equine Vet. Educ.* 20, 37–39. doi: 10.2746/095777307X260106

Su, H., Zhu, S., Zhu, L., Huang, W., Wang, H., Zhang, Z., et al. (2016). Recombinant lipoprotein Rv1016c derived from *Mycobacterium tuberculosis* Is a TLR-2 ligand that induces macrophages apoptosis and inhibits mhc ii antigen processing. *Front. Cell. Infect. Microbiol.* 18:147. doi: 10.3389/fcimb.2016.00147

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., et al. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res*. 29, 22–28. doi: 10.1093/nar/29.1.22

Wilson, M. J., Brandon, M. R., and Walker, J. (1995). Molecular and biochemical characterization of a protective 40-kilodalton antigen from *Corynebacterium pseudotuberculosis*. *Infect. Immun.* 63, 206–211.

Xiao, Q., Jiang, X., Moore, K. J., Shao, Y., Pi, H., Dubail, I., et al. (2011). Sortase independent and dependent systems for acquisition of haem and haemoglobin in *Listeria monocytogenes*. *Mol. Microbiol.* 80, 1581–1197. doi: 10.1111/j.1365-2958.2011.07667.x

Yanagawa, R., and Honda, E. (1976). Presence of pili in species of human and animal parasites and pathogens of the genus corynebacterium. *Infect. Immun.* 13, 1293–1295.

Yu, D., Pi, B., Yu, M., Wang, Y., Ruan, Z., Feng, Y., et al. (2014). Diversity and evolution of oligopeptide permease systems in staphylococcal species. *Genomics* 104, 8–13. doi: 10.1016/j.ygeno.2014.04.003

Zaki, M. M. (1966). The ability of *Corynebacterium ovis* to produce suppurative osteomyelitis and arthritis in white mice. *J. Comp. Pathol.* 76, 121–126. doi: 10.1016/0021-9975(66)90014-4

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.