Pan-genomic analyses of *Corynebacterium pseudotuberculosis* and characterization of the biovars *ovis* and *equi* through comparative genomics

ORIENTADO: Siomar de Castro Soares

SUPERVISOR: Prof. Dr. Vasco Ariston de Carvalho Azevedo

CO-SUPERVISOR: PD. Dr. Andreas Tauch

BELO HORIZONTE

August - 2013
Siomar de Castro Soares

Pan-genomic analyses of *Corynebacterium pseudotuberculosis* and characterization of the biovars *ovis* and *equi* through comparative genomics.

Thesis presented as partial requirement for the degree of Doctor of Philosophy in Genetics, to the Department of General Biology at the Institute of Biological Sciences, Federal University of Minas Gerais.

SUPERVISOR: Prof. Dr. Vasco Ariston de Carvalho Azevedo

CO-SUPERVISOR: PD. Dr. Andreas Tauch

BELO HORIZONTE

August - 2013
I dedicate this work to my family, my wife and every single person who stood by my side during the years.
ACKNOWLEDGEMENTS

I would like to thank everybody who worked with or stood by me since 2007, when I first decided to begin my research way of life.

Collaborators:

I would like to thank:

- Prof. Dr. Vasco Ariston de Carvalho Azevedo, for the supervision during the development of my work; for the non-academic advices; for the time expended during my own development; for seeking the best for all of your students, no matter who; and, also, for the friendship.

- PD. Dr. Andreas Tauch, for the receptivity when I arrived in Germany; for the time expended, and also patience, when reviewing every section of the works we have developed in Germany; for giving me the opportunity to take part of CLIB; and, also, for coming to Belo Horizonte for my Ph.D. defense.

- Prof. Dr. Artur Silva, for the great collaboration during the development of the work; and, for the all opportunities to give talks and conferences in Pará, together with the LPDNA group, which helped improving my knowledge.

- Prof. Dr. Anderson Miyoshi, for the supervision during my master's, which helped in my own development, organization and writing skills; and, for the advices in all presentations and manuscripts during the Ph.D.

- Programs of Post-graduation in Genetics and Bioinformatics, for all the disciplines; and, also, for the cooperation and support, mainly in the last steps, when trying to organize flights and accommodations.

- Laboratory of cellular and molecular Genetics, specially the long-date friends Fernanda, Thiago, Luís, Marcela, Wanderson and several other friends, for all the knowledge I have acquired from all of you; for the experience and maturity; and, of course, for all the non-academic time we have expended together.

- Laboratory of DNA polymorphism, specially Rommel, Adriana, Diego, Rafael, Hivana and Leonardo, for the receptivity in Pará, collaboration, support and friendship.

- Laboratory of Nanobiotechnology, specially Paula, Juliana, Lara, Galber, Fabiana and Professors Carlos Ueira and Luiz Goulart, for the opportunity to work with this great group; and, for all the knowledge I have acquired during my stay.
Family:

I would also like to thank:

- My parents, for the faith you have in my future, since I was a kid; and, for the comprehension, even when you do not understand why someone needs to be a student for so long.

- My brothers and sister for the support and great time I have every time I go back to my hometown. Believe me, to be with you for one weekend is like one year anti-stress therapy and it makes all the difference.

- My nephews, the simple fact that you exist is enough to give strength to everybody on the family and I work hoping that I may be a good reference for you in future the same way you are my strength.

- My wife, Letícia, only you know the backstage. In the words of Vasco, you give me stability. I avoid explaining everything you represent because there are no words that could explain all support, comprehension and peace. And, even if I would to try, half of this thesis would not be enough. I hope I can do for you, during your newly started academic life, the same you have done for me. "The important thing is to be happy".
“The only reason for time is so that everything does not happen at once.”
(Albert Einstein)

“If I have seen further it is by standing on the shoulders of Giants.”
(Isaac Newton)
Table of contents

List of Figures .................................................................i
List of Tables ......................................................................ii
Abbreviations ...................................................................iii
Abstract .............................................................................1
I. Presentation.....................................................................2
   I.1 Collaborators .............................................................3
II. Preface ..........................................................................4
   II.1 *C. pseudotuberculosis* - state of the art ....................5
      II.1.1 Biovars of *C. pseudotuberculosis* .........................5
      II.1.1.1 Biovar **ovis** ..................................................5
      II.1.1.2 Biovar **equi** ..................................................8
   II.2 Manuscript Structure and author's contributions ..........10
III. Introduction ....................................................................12
   III.1 *Corynebacterium* pathogenic species in next-generation genomic era: the use of EDGAR and PIPS software and the importance of pathogenicity islands identification in pan-genomic analyses of pathogenic species ..................13
IV. Goals ............................................................................30
   IV.1 Main goal ..................................................................31
   IV.2 Specific goals ..........................................................31
V. Research Articles ............................................................32
   V.1 Chapter I. PIPS: Pathogenicity Island Prediction Software ..................................................33
      V.1.1 Appendix S1 ..........................................................44
      V.1.2 Figure S1 ................................................................45
      V.1.3 Figure S2 ................................................................46
      V.1.4 Table S1 ...............................................................47
      V.1.5 Discussion ............................................................49
   V.2 Chapter II. Genome sequence of *Corynebacterium pseudotuberculosis* biovar **equi** strain 258 and prediction of antigenic targets to improve biotechnological vaccine production ........................................51
V.3 Chapter III. The Pan-Genome of the Animal Pathogen Corynebacterium pseudotuberculosis Reveals Differences in Genome Plasticity between the Biovar ovis and equi Strains ................................................................. 60

V.3.1 Figure S1 ............................................................................... 75

V.3.2 Discussion ............................................................................. 76

VI. General Discussion .................................................................. 77

VII. Conclusions ............................................................................ 79

VIII. Bibliography ......................................................................... 81

IX. Appendices ............................................................................. 87

IX.1 Curriculum Vitae .................................................................... 88
List of Figures

Figure 1. The whole genome of Corynebacterium pseudotuberculosis .................. 6
Figure 2. Gene regions encoding adhesive pili of C. pseudotuberculosis FRC41 .... 7
Figure 3. Model of pilus biogenesis ................................................................. 8
Figure 4. Genomic map comparing strains of Corynebacterium pseudotuberculosis, Corynebacterium ulcerans and Corynebacterium diphtheriae ........................................ 9
Figure S1. Prediction of PICD12 of C. diphtheriae with a different size than the literature prediction ................................................................. 45
Figure S2. Graphic representation of PAI features in the genome (A) and in the pathogenicity islands (B) of C. pseudotuberculosis and C. diphtheriae .................. 46
Figure 5. Heatmap showing the presence/absence of PAIs identified by PIPS in 13 strains of C. diphtheriae ................................................................. 50
Figure S1. Plasticity of PiCps 4, 5 and 9 ................................................................. 75
List of Tables

Table S1. PAI composition. .............................................................................. 47
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRIG</td>
<td>Blast Ring Image Generator</td>
</tr>
<tr>
<td>CAPES</td>
<td>Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Coordination for the Improvement of Higher Education Personnel)</td>
</tr>
<tr>
<td>CDS</td>
<td>Coding Sequence</td>
</tr>
<tr>
<td>CeBiTec</td>
<td>Center for Biotechnology</td>
</tr>
<tr>
<td>CLA</td>
<td>Caseous Lymph Adenitis</td>
</tr>
<tr>
<td>CLIB</td>
<td>Cluster Industrial Biotechnology</td>
</tr>
<tr>
<td>CMNR</td>
<td>Group composed of Corynebacterium, Mycobacterium, Nocardia and Rhodococcus</td>
</tr>
<tr>
<td>CNPq</td>
<td>Conselho Nacional de Desenvolvimento Científico e Tecnológico (National Counsel of Technological and Scientific Development)</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>Fapemig</td>
<td>Fundação de Amparo à Pesquisa do Estado de Minas Gerais (Foundation for Research Support of the State of Minas Gerais)</td>
</tr>
<tr>
<td>G+C</td>
<td>Guanine + Thymine</td>
</tr>
<tr>
<td>GEI</td>
<td>Genomic Island</td>
</tr>
<tr>
<td>LGCM</td>
<td>Laboratório de Genética Celular e Molecular (Laboratory of Cellular and Molecular Genetics)</td>
</tr>
<tr>
<td>LPDNA</td>
<td>Laboratório de Polimorfismo de DNA (Laboratory of DNA Polimorphism)</td>
</tr>
<tr>
<td>MHC</td>
<td>Major Histocompatibility Complex</td>
</tr>
<tr>
<td>PAI</td>
<td>Pathogenicity Island</td>
</tr>
<tr>
<td>PIPS</td>
<td>Pathogenicity Island Prediction Software</td>
</tr>
<tr>
<td>RGMG</td>
<td>Rede Genoma de Minas Gerais (Minas Gerais Genome Network)</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic Acid</td>
</tr>
<tr>
<td>RPGP</td>
<td>Rede Paraense de Genômica e Proteômica (The Genomics and Proteomics Network of the State of Pará)</td>
</tr>
<tr>
<td>rRNA</td>
<td>Ribosomal ribonucleic acid</td>
</tr>
<tr>
<td>tRNA</td>
<td>Transporter ribonucleic acid</td>
</tr>
<tr>
<td>UFMG</td>
<td>Universidade Federal de Minas Gerais (Federal University of Minas Gerais)</td>
</tr>
<tr>
<td>UFPA</td>
<td>Universidade Federal do Pará (Federal University of Pará)</td>
</tr>
</tbody>
</table>
Abstract

*Corynebacterium pseudotuberculosis* is the causative agent of diverse communicable diseases in small ruminants (biovar *ovis*), horses, camels, buffalo and other animals (biovar *equi*), which mainly differ in symptoms and site of infection. Additionally, the diseases present a highly important economic problem worldwide and there is still a lack of efficient treatments against *C. pseudotuberculosis*. In this work, we describe the steps from the first genome sequencing of a strain of *C. pseudotuberculosis* to the pan-genomic analyses of 15 strains isolated from different hosts and countries with diverse symptoms. Briefly, we introduce the genus *Corynebacterium* and the in silico analyses performed in pathogenic species of this genus to date. Then, we describe the implementation of a software for the prediction of pathogenicity islands (PAIs) in bacteria (PIPS), which outperformed the other available software, and identified 7 PAIs with important virulence factors in *C. pseudotuberculosis* biovar *ovis*. Moreover, we extend the analyses of PAIs to strains of *C. pseudotuberculosis* biovar *equi* and predict 49 putative vaccine targets, in silico, which are commonly shared by both biovars, *ovis* and *equi*. Finally, we present the phylogenomic, pan-genomic, core genomic, singletons and genomic plasticity analyses of the 15 strains of *C. pseudotuberculosis*, from both biovars. All the analyses performed here point for a clonal-like behavior of *C. pseudotuberculosis*, which could be the result of the facultative intracellular behavior of the species. Moreover, the biovar *equi* presents a higher variability in gene content when compared to biovar *ovis*, specially in PAI regions. Noteworthy, the strains from biovar *ovis* present a high degree of similarity in pili clusters of genes, whereas the biovar *equi* strains are very variable. The conservation of pili clusters of genes in biovar *ovis* could account for the ability of these strains to spread inside host tissues and penetrate live cells to live intracellularly, where they would have less contact to other organisms, thus, possibly explaining the clonal-like behavior of the biovar *ovis*. 


I. Presentation
I.1 Collaborators

This work was performed on the Laboratories of Molecular and Cellular Genetics (LGCM) and DNA Polimorphism (LPDNA), at Federal University of Minas Gerais (UFMG) and Federal University of Pará, respectively, and the Center for Biotechnology (CeBiTec), at the Bielefeld University, in a collaboration between the following researchers in alphabetic order:

Profª. Drª. Ana Luiza de Mattos Guaraldi, Researcher and Professor from UERJ, Brazil;

Prof. Dr. Anderson Miyoshi, Researcher and Professor from LGCM-UFMG, Brazil;

PD. Dr. Andreas Tauch, Researcher from CeBiTec and member of the Graduate Cluster Industrial Biotechnology (CLIB), Germany.

Prof. Dr. Artur Silva, Researcher and Professor from LPDNA-UFPA and member of The Genomics and Proteomics Network of the State of Pará (RPGP), Brazil;

Prof. Dr. Raphael Hirata Jr., Researcher and Professor from UERJ, Brazil;

Prof. Dr. Robert Moore, Researcher from CSIRO, Australia;

Prof. Dr. Vasco Ariston de Carvalho Azevedo, Researcher and Professor from LGCM-UFMG and member of the Minas Gerais Genomics Network (RGMG), Brazil;

The work was supported by: Fundação de Amparo à Pesquisa do Estado de Minas Gerais (Fapemig), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Graduate Cluster Industrial Biotechnology (CLIB) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).
II. Preface
II.1 **C. pseudotuberculosis** - state of the art

*Corynebacterium* species are members of the CMNR group, which also includes *Mycobacterium*, *Nocardia* and *Rhodococcus* and are mainly characterized by: (i) high G+C content and (ii) a specific cell wall structure. *Corynebacterium* genus harbours several bacteria of high biotechnological, medical and veterinary relevance (Dorella *et al.*, 2006). *C. pseudotuberculosis*, the main subject of this work, is closely related to the pathogenic species *C. diphtheriae* and *C. ulcerans*, which share several virulence genes and present a high degree of genomic synteny (Buck *et al.*, 1985; Groman *et al.*, 1984; Ruiz *et al.*, 2011).

II.1.1 Biovars of *C. pseudotuberculosis*

II.1.1.1 Biovar ovis

*C. pseudotuberculosis* presents two biovars, *ovis* (nitrate negative reduction) and *equi* (nitrate positive reduction), where the former is mainly associated to the worldwide distributed disease Caseous Lymph Adenitis (CLA), which affects lymph nodes and visceral organs of goat and sheep and causes several economic losses by compromising the animal skin, weight, milk and meat production, and causing carcass condemnation and death (Biberstein *et al.*, 1971). Finally, although many vaccines do exist, they are mainly intended to sheep and goat and provide variable protection levels (Williamson, 2001).

In order to better understand the pathogenic mechanisms underlying CLA, the genome sequencing of *C. pseudotuberculosis* 1002 biovar *ovis*, isolated from goat in Bahia, was initially proposed by our group in 2006. The genome sequencing was finished alongside with another biovar *ovis* strains, C231, which was firstly sequenced in Australia, by Prof. Robert Moore, and later finished and analyzed by 3 collaborating groups from UFMG, UFPA and CeBiTec (Ruiz *et al.*, 2011). Concomitantly, the genome sequence of the strain FRC41 isolated from human, biovar *ovis*, was also finished by the group of PD. Dr. Andreas Tauch (CeBiTec) in collaboration with the Brazilian’s groups (UFMG and UFPA) (Trost *et al.*, 2010). Finally, all genome sequences were deposited to Genbank and further analyses on gene content and synteny were made to achieve a global view of the pathogen and its virulence factors.
A common feature of virulence factors is their high concentration inside Pathogenicity Islands (PAIs), a class of Genomic Islands (GEIs). PAIs are large genomic regions acquired through horizontal gene transfer, which have in common: deviations in G+C content and codon usage; the presence of transposase and virulence factors; flanking insertion sequences and/or tRNA genes; and the absence in non-pathogenic organism of the same genus or related species (Azevedo et al., 2011). In order to predict PAIs in the genome sequences of *C. pseudotuberculosis*, our group has developed a software named PIPS (Pathogenicity Island Prediction Software), which predicts PAIs taking into account the concentration of the before mentioned features along the genome sequence (Soares et al., 2012). In analyses of *C. pseudotuberculosis* 1002 and C231, PIPS has identified 7 PAIs (Figure 1), which harbour: the *pld* gene that codes for the exotoxin Phospholipase D; the *fagABC* operon and *fagD* gene that codes for iron uptake proteins; and, several other virulence factors and hypothetical proteins (Ruiz et al., 2011; Soares et al., 2012).

**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. pseudotuberculosis* and its location in the genomes. (Figure from doi:10.1371/journal.pone.0018551.g001).
Additionally, in analyses of *C. pseudotuberculosis* FRC41, it was reported the presence of 2 clusters of pili genes (Figure 2), which could contribute to the facultative intracellular behavior of this species by coding proteins with roles in adhesion and internalization mechanisms. The pili clusters of genes are named accordingly to their major pilin gene as follow: 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 cell wall (Trost *et al.*, 2010).

![Figure 2. Gene regions encoding adhesive pili of *C. pseudotuberculosis* FRC41.](image)

The gene clusters involved in the synthesis of adhesive (Spa-like) pili of *C. pseudotuberculosis* FRC41 are shown. The gene clusters encode sortases required for the assembly of the pilus (blue), major pilins (red), minor pilins (yellow), pilus tip proteins (green), and proteins of unknown function (grey). The detected sorting (LPxTG) signals are indicated. Specifically marked in the major pilin proteins are the characteristic pilin boxes (blue) and E-boxes (white). The predicted binding of the transcription regulator GlxR in the spaA-srtB intergenic region is shown. (Figure from doi: 10.1186/1471-2164-11-728).

The polymerization of pili structures on *C. pseudotuberculosis* has not been deeply studied yet, however, there are several studies on the closely related species, *C. diphtheriae* (Mandlik *et al.*, 2007). Briefly, in *C. diphtheriae*, the housekeeping sortase forms intermediates with the precursor proteins (the products of the major, minor and tip pilin, designated as A, B and C on Figure 3, respectively) and the specific sortase catalyzes pilus polymerization and transfer the pilus polymer to lipid II (Figure 3). Noteworthy, the polymerization of the complete structure requires all related genes and also, the polymerization of minor and tip pilin (B and C on Figure 3, respectively) depends on the presence of the major pilin (A on Figure 3). In the absence of a pilus-specific sortase, the major, minor and tip pilin are attached to the cell wall as monomers and the same is true for the minor and tip pilin when the pilus-specific sortase is present and the major pilin is absent (Mandlik *et al.*, 2008).
II.1.1.2 Biovar *equi*

After achieving a better view of the genome sequences of *C. pseudotuberculosis* biovar *ovis* strains isolated from sheep, human and goat, our groups began a great effort to sequence other biovar *ovis* strains isolated from other hosts and also biovar *equi* strains from different hosts and countries. *C. pseudotuberculosis* biovar *equi* strains were isolated from horses, camels and buffalos where the disease symptoms are very variable and visceral commitment is rare (Cerdeira *et al.*, 2011; Lopes *et al.*, 2012; Pethick *et al.*, 2012; Pethick *et al.*, 2012; Ramos *et al.*, 2012; Ramos *et al.*, 2013; Silva *et al.*, 2011; Silva *et al.*, 2012; Soares *et al.*, 2012).
After finishing the sequences of biovar *equi* genomes, we were able to identify 4 additional PAIs in *C. pseudotuberculosis* 316 and 258 (PICPs 8-11), both isolated from horses (Figure 4) (Ramos *et al.*, 2013; Soares *et al.*, 2012). Moreover, further reverse vaccinology based analyses were performed in *C. pseudotuberculosis* 258, biovar *equi*, in order to find new vaccine candidates that could possibly elicit immune response against this organism. Finally, we have accomplished the genome sequencing of 15 strains of *C. pseudotuberculosis* from both biovars, isolated from different countries and hosts, and performed pan-genomics analyses on the whole species aiming to correlate regions of genome plasticity with the disease patterns and host-preference (Soares *et al.*, 2013).

![Figure 4. Genomic map comparing strains of Corynebacterium pseudotuberculosis, Corynebacterium ulcerans and Corynebacterium diphtheriae.](image)

Comparative genomic analyses between: Corynebacterium pseudotuberculosis strains 1002, C231, CIP52.97 and 316; Corynebacterium ulcerans strains BR-AD22 and 809; Corynebacterium diphtheriae NCTC 13129; Corynebacterium glutamicum ATCC 13032; and pathogenicity islands identified in *C. pseudotuberculosis*. The figure shows the presence/absence of the pathogenicity islands of *C. pseudotuberculosis* 1002, strain which was also used as reference to create the figure, on the other strains and species. *(Adapted from 10.1111/1751-7915.12006)*
II.2 Manuscript Structure and author’s contributions

The thesis is divided into Introduction and 3 chapters based on 1 book chapter and 3 research articles, as follow:

a. The Introduction, presented as a book chapter, shows an overview of *in silico* studies performed in pathogenic *Corynebacterium* species to date, showing the importance of PAIs, reverse vaccinology and pan-genomics analyses in pathogenic species and providing tables with putative PAIs and vaccine targets. In this work, I have written the whole manuscript with scientific support from all the co-authors. Additionally, I have also performed the identification of PAIs of *C. ulcerans* and created the tables and figures;

b. The first chapter presents a research article showing the implementation of the software "PIPS: Pathogenicity Island Prediction Software" and a comparison between this software and other previously available programs. For this matter, analyses were performed using previously described PAIs of *C. diphtheriae* NCTC 13129 and *Escherichia coli* CFT 073. Finally, the article shows data on analyses performed in *C. pseudotuberculosis* 1002 and C231 as a case study. In this work, I have created all the scripts, except for two of them that were kindly provided and one that was implemented by Dr. Rommel Ramos (the credits were added to the specific scripts). I also had support from the co-authors in news ideas for predicting transposases and writing the manuscript;

c. The article in the second chapter describes the identification of PAIs in *C. pseudotuberculosis* 258, biovar *equi*, using PIPS, in order to identify regions of plasticity between both biovars. Furthermore, we applied the reverse vaccinology's theory in *C. pseudotuberculosis* 258 in comparison with other *C. pseudotuberculosis* strains, aiming to identify new putative vaccine targets that could elicit immune response against both biovars. In this work, I have made the identification of pathogenicity islands and all the reverse vaccinology analyses, except for the cell wall measurement and prediction of subcellular location that were performed by Dr. Anderson Santos;
d. The third chapter presents the pan-genomics article, where all 15 genome sequences of different hosts and strains were used. In this article, we review basic concepts about the biovar *ovis* and *equi* and create a phylogenomics tree to find the evolutionary relationship between species of the genus *Corynebacterium*. Moreover, we assess the pan-genome, core genome and singletons subsets of *C. pseudotuberculosis* and perform comparisons between both biovars according to these datasets and the PAIs content. In order to accomplish this work, I have participated in specific tasks on all previous steps of genome sequencing, annotation and comparative analyses of all 15 strains. In this work, the retrieving of data from Genbank by GeneDB, the incorporation of data into EDGAR and the statistical analyses performed by R package were triggered by Dr. Jochen Blom. Finally, I have performed all analyses, data interpretation, figure creation and manuscript writing with support from the other authors.

Additional files and discussions for the research articles are appended after the referred article, in the same section. After the chapters, we present the general conclusions of the work. Finally, after bibliography, there is an "appendices" section, where one can find the curriculum vitae.
III. Introduction
III.1 *Corynebacterium* pathogenic species in next-generation genomic era: the use of EDGAR and PIPS software and the importance of pathogenicity islands identification in pan-genomic analyses of pathogenic species


Recently, our group has been invited to write a book chapter for "Microbial pathogens and strategies for combating them: science, technology and education". In this book chapter, we review the pathogenic species *C. pseudotuberculosis*, *C. diphtheriae* and *C. ulcerans*, highlighting the *in silico* studies performed in these organisms. Additionally, we also review Edgar, PIPS and other software that were used in our previous works. Furthermore, we summarize potential vaccine targets and PAIs identified in those analyses in a compendium-like work. The description of such data will be helpful in driving future *in vitro* studies performed by our group and the analyses and software may also be easily applied by other groups.
Corynebacterium pathogenic species in next-generation genomic era: the use of EDGAR and PIPS software and the importance of pathogenicity islands identification in pan-genomic analyses of pathogenic species

S. C. Soares ¹, R. T. J. Ramos ², W. M. Silva ¹, L. C. Oliveira ¹, L. G. Amorim ¹, R. Hirata Jr ³, A. L. Mattos-Gueraldì ³, A. Miyoshi ³, A. Silva ³, V. Azevedo ³, 4

¹ Laboratory of Cellular and Molecular Genetics, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
² Microbiology and Immunology Discipline, Medical Sciences Faculty, State University of Rio de Janeiro, Rio de Janeiro, Brazil
³ Department of Genetics, Federal University of Pará, Belém, Pará, Brazil

Corynebacterium genus presents several opportunistic, non-pathogenic, and pathogenic species of high industrial, medical and veterinary importance. Between Corynebacterium pathogenic species, 3 highly virulent organisms deserve higher attention as the causative agents of the worldwide distributed and communicable diseases diphtheria, caseous lymphadenitis and diphtheria-like, caused by Corynebacterium diphtheriae, Corynebacterium pseudotuberculosis and Corynebacterium ulcerans, respectively. In order to better understand the virulence mechanisms underlying the diseases caused by those organisms, several in silico studies have been performed, focusing in: phylogenetics analyses and how those species correlate with each other; pan-genomics analyses and the degree of variability within the species; pathogenicity island identification, commonly shared virulence factors and how genome plasticity may influence the genomes of those species; and, reverse vaccinology and the identification of new candidate targets for future vaccine developments. In this chapter, we review the disease patterns of each species according to their hosts, the high potential of the methodologies and their resulting data, and the putative pathogenicity islands and candidate targets identified in C. diphtheriae, C. pseudotuberculosis and C. ulcerans to date.

Keywords: PIPS, EDGAR, Pan-genomics, Phylogenomics, Reverse Vaccinology, Subtractive Genomics, Pan-Exoproteome, Pathogenicity Islands, Vaccine Targets.

1 Corynebacterium genus

Corynebacterium genus is part of the CMNR group, a suprageneric group of the Actinomycetales family, which includes several genera with high medical, veterinary and biotechnological importance, like: Corynebacterium, Mycobacterium, Nocardia and Rhodococcus. Bacteria from the CMNR group have in common: (i) high G+C content and (ii) a specific cell wall structure composed of mycolic acid, peptidoglycan and arabinogalactan [1]. Corynebacterium genus was first created to harbour Corynebacterium diphtheriae and other pathogenic species [2]. Later on, other species were included, which differed in shape, pathogenicity and sporulation [3]. Nowadays, the genus is mainly composed of: the non-pathogenic species Corynebacterium glutamicum and Corynebacterium efficiens, which are of great biotechnological interest in amino acid production [4,5], and Corynebacterium variabile, a bacterium isolated from the microflora contributing to the development of flavour and texture in cheese ripening [6]; the opportunistic species Corynebacterium jeikeium, Corynebacterium urealyticum and Corynebacterium resitsens, which are frequently associated with nosocomial infections [7-9], and the opportunistic and potentially pathogenic Corynebacterium aurimucosum, which is mainly isolated from women with urogenital infections and appears associated with complications in pregnancy [10]; and, the pathogenic species Corynebacterium pseudotuberculosis, C. diphtheriae and Corynebacterium ulcerans, of high veterinary and medical relevance, and a low pathogenic potential bacterium, Corynebacterium kroppenstedtii, which is associated with pulmonary disease and cases of mastitis [9,11-14].

1.1 Corynebacterium pseudotuberculosis

C. pseudotuberculosis is a gram-positive, non-motile, facultative anaerobic, pleomorphic, and intracellular facultative pathogen that proliferates inside macrophages [15]. C. pseudotuberculosis presents two biovars, ovis and equi, which are mainly distinguished by their ability to reduce nitrate and by their host preference [16]. The biovar equi strains (positive nitrate reduction) of C. pseudotuberculosis may be isolated from buffalo, camels, cows and horses; the biovar ovis strains (negative nitrate reduction), on the other hand, are mainly isolated from small ruminants, like sheep and goats, causing a contagious chronic disease, named Caseous Lymphadenitis (CLA), but may also be found in llamas, antelopes, cows and even humans [16-20]. Additionally, Cows are the only hosts from which both biovars, ovis and equi, have been isolated to date, causing a broad range of symptoms that vary from pyogranulomatous reactions, mastitis and ulcerative dermatitis to abscess formation and visceral commitment [21,22]. In horses, there are reports of C. pseudotuberculosis associated with external abscesses (pigeon fever), ulcerative lymphangitis and, more rarely, a visceral form affecting internal organs [23,24]. However, except for CLA, there is still a lack of information and studies
about the diseases caused by *C. pseudotuberculosis* and the underlying pathogenic mechanisms and virulence factors [1,25,26].

CLA is characterized by the presence of caseous necrosis in lymphatic glands or abscess formation in superficial lymph nodes and subcutaneous tissues of sheep and goats [27], compromising the animal skin, weight, milk and meat production, and causing carcass condemnation and death [1]. The disease has a worldwide distribution and was already reported in several countries, like Australia, New Zealand, South Africa, United States, Canada and Brazil, where sheep and goat farming are very intense [1,28-32]. In Brazil, epidemiologic studies report that a high number of the animals are infected, where the states from the North-East region are the most affected and the underlying losses in this region are highly significant [33,34]. Besides, in the state of Minas Gerais, 78.9% of goats are seropositive for *C. pseudotuberculosis* infection [35]. The treatment of CLA infected animals is normally performed by draining infected superficial lymph nodes, however, this treatment does not eliminate 100% of the bacteria, it is not viable when visceral organs have been affected, and it may also contaminate the soil [1]. Moreover, although *C. pseudotuberculosis* is susceptible to a broad range of antibiotics *in vitro*, the inefficacy of antibiotics in penetrating the abscess capsule and the highly expensive treatment make the antibiotic therapy not applicable [36]. Finally, the licensed vaccines intended for use in sheep present variable efficacy in goat immunization [21] and this scenario is much worst when other hosts and diseases are considered.

1.2 *Corynebacterium diphtheriae*

*C. diphtheriae* is a gram-positive, aerobic, non-motile, rod-shaped and pathogenic bacterium [11]. This bacterium is mainly isolated from humans, although other hosts have already been reported, like horses [37], domestic cats [38] and dogs [39]. In humans, *C. diphtheriae* is responsible for causing the diphtheria disease, an acute upper respiratory tract communicable disease [40,41], and, based on the severity of the infection along with the biochemical profile, the strains are classified under 4 biovars: *mitis*, *gravis*, *intermedius* and *belfanti* [42]. The cases of diphtheria over the world have decreased drastically since the development of a vaccine based on the inactivation of diphtheria toxin (DT), coded by the tox gene [41]. However, despite the existence of DTP vaccine (diphtheria-tetanus-pertussis) and the decrease in cases worldwide, the disease remains endemic in several regions including Africa, Bangladesh, Vietnam, the tropics and areas of South America, including Brazil [43]. Moreover, more than 150,000 cases have been reported in the former Soviet Union in the 1990s and there are several reports of either re-emergence or persistence of diphtheria in Indian states from 1998-2008 [41,43-46].

The reasons for the reemergence of diphtheria remain to be fully elucidated, however, factors mainly point to: an increased susceptibility of both children and adults; and to the inefficacy of control measures due to shortages of vaccine and deteriorating health infrastructure [41,43,45,47]. Besides, the fact that the tox gene is harboured by a pathogenicity island (PAI), which was horizontally acquired from corynephage, accounts for the emergence of new non-toxigenic strains to which the immune response elicited by the toxoid-based vaccine is not effective [11,48]. The non-toxigenic strains of *C. diphtheriae* cause infectious diseases varying from cutaneous lesions and pharyngitis to severe invasive commitments, which are characterized by bacteraemia and endocarditis in the absence of toxin mediated lesions [49].

1.3 *Corynebacterium ulcerans*

The emergence of *C. ulcerans* strains causing diphtheria-like diseases are of major concern in industrialized countries, like United Kingdom, France and Germany [50]. The symptoms of the diphtheria-like disease caused by *C. ulcerans* vary from skin ulcers to pharyngitis, sinusitis, tonsillitis and pulmonary nodules [51,52]. Although the virulence of *C. ulcerans* is not necessarily dependent on the production of DT, there are reports of strains producing a potent toxin and causing severe diphtheria-like symptoms [53,54]. Interestingly, the amino acid sequence of the DT harboured by *C. ulcerans* presents 95% similarity to the one of *C. diphtheriae* [55], which could account to differences in immune response by vaccinated individuals. The infection route of *C. ulcerans* producing DT is not fully understood and person-to-person transmission was not yet reported [52]. However, toxigenic *C. ulcerans* were isolated from domestic animals, like cats with nasal discharge [38] and dogs [56,57], pointing these animals as potential reservoirs.
2. Comparative genomics in *Corynebacterium* pathogenic species

2.1 Phylogenomics - *Corynebacterium* genus

In past, evolutionary reconstructions of the tree of life were mainly performed based on identification of the point of divergence between species solely based in shared homologous characters. However, this methodology could be very trick due to convergent and divergent evolution. With the advent of molecular techniques, phylogenetics was greatly improved by the use of nucleotide differences in universal reference markers, creating the area of phylogenomics [58]. In the post-genomic era, a second wave of changes brought new approaches to phylogenomics, which now infers the evolutionary divergence by taking advantage of whole-genome data, like: gene content and gene order; orthology; and, DNA string or DNA signature [58,59]. In this sense, phylogenomics may be defined as the junction of phylogenetics and genomics for reconstructing reliable species trees, analysing the distribution and spread of bacterial pathogenicity and predicting orthologous and paralogous genes [60,61].

An approach for the reconstruction of phylogenomics trees and inference of evolutionary divergences is Gegenees, a software that splits the genome data of a group of strains or species in small sequences using pre-defined sizes, performs similarity searches using BLAST, identifies genes commonly shared between the genomes and creates a distance matrix based on the percentage of similarity between the variable contents of the underlying genomes [62]. From the heatmap and phylogenomics tree generated by Gegenees, although *C. ulcerans* appears more related to *C. pseudotuberculosis* than to *C. diphtheriae*, all 3 pathogenic species, *C. ulcerans*, *C. pseudotuberculosis* and *C. diphtheriae*, cluster together, whereas the non-pathogenic and opportunistic species appear separately [63]. This close relationship was already described in previous works [64] and is probably due to the presence of commonly shared virulence factors between those 3 species.

2.2 Pan-genomics analyses using EDGAR - *C. diphtheriae* and *C. pseudotuberculosis*

A new methodology to achieve a broad genome view of a species or genus is the pan-genomics approach. The Pan-genome idea was initially introduced by Tettelin and collaborators in 2005 [65] using the genome sequences of 8 strains of *Streptococcus agalactiae*. Pan-genome is defined as the complete and non-redundant repertoire of genes from a species or genus and is composed of three subsets: the core genome, which harbours all commonly shared genes of the studied dataset; the extended core, which consists of genes that are shared by two or more strains but are not present in all strains; and, the singletons, which are strain-specific genes [65,66]. According to those datasets, the pan-genomics studies, to date, are based in three main steps: identification of orthologous and paralogous genes; classification of each gene into the subsets; and, curve fitting to achieve a prospect of the pan-genome state and development [67-69].

The identification of orthologous and paralogous genes in pan-genomics approaches may be performed by using all-versus-all sequence similarity searches (BBH, Best Blast Hit) or identification of clusters of orthologous groups (OrthoMCL) [63,69-72]. The classification of each gene as part of the pan-genome, core genome and singletons groups is performed based on the information generated on the analysis of orthologous and paralogous genes by the addition of 1 genome at a time. For instance, considering 2 genomes *A* and *B* with 1500 and 1400 genes, respectively (Figure 1 - A1), and commonly sharing 1000 genes (Figure 1 - A2): the pan-genome will be represented by 1900 genes, i.e., 400 singletons from genome *B* will be added to genome *A* (Figure 1 - A3); the core genome will have 1000 genes (Figure 1 - A4); and, the singletons will consist of 400 genes (Figure 1 - A5). For ease of representation, in the second round, the “virtual” pan-genome of genomes *A* and *B* (PanAB) is used for comparison (Figure 1 - B1), where: 800 genes are shared by all strains; 1100 genes in the PanAB and 400 genes in genome *C* are part of the extend core, i.e., genes which are shared by *A* and *B*, *A* and *C* or *B* and *C* (Figure 1 - B2); and, 100 genes are newly added as singletons (Figure 1 - B1). In this scenario, the subsets will develop in the following manner: the pan-genome will now increase to 2000 genes, i.e., 100 singletons from genome *C* will be added to PanAB (Figure 1 - B3); the core genome will consist of 800 genes (Figure 1 - B4); and, the singletons will decrease to 100 (Figure 1 - B5). Finally, those steps are repeated for each newly added genome, creating a development curve for each subset.
Fig. 1 Schematic representation of classification of genes into pan-genome, core genome and singleton subsets.
In the last step of pan-genomics approach, a curve fitting is performed to correct the whole curve and also the initial number of genes in pan-genome, core genome and singletons (1500 genes in the example). In order to perform this step, a permutation of all genomes in each position is made and the mean or median number of genes is used for curve fitting of the pan-genome, using Power Law or Heap's Law, and the core genome and singletons, using least-squares fit of the exponential regression decay [65-67,72]. The resulting fitted curves are represented by the formulas $n = k \cdot N^\alpha$ and $n = k \cdot \exp(-a \cdot N)$ for the Heap' Law and least-squares fit of the exponential regression decay, respectively, where $n$ is the number of genes for a given number of genomes, $N$ is the number of genomes in pan-genome extrapolations, $x$ is the number of genomes in core genome and singletons extrapolations, exp is Euler's number and the other terms are constants defined to fit the specific curve. Interestingly, an $\alpha \leq 1$ in Heap's Law represents an open pan-genome and singletons, whereas an $\alpha > 1$ is representative of a closed pan-genome, i.e., no substantial change will be seen in the number of genes in the pan-genome with the addition of newly sequenced genomes [63]. Moreover, the formula for least-squares fit of the exponential regression decay may also be used to achieve a development prospect, where $tg(\theta)$ in core genome analysis is representative of the number of genes found in the stabilized core genome after a given number of genomes are sequenced and added to the analysis, whereas in singletons analysis, it is the approximate number of genes which will be added to the pan-genome by each newly sequenced genome [66].

In *C. pseudotuberculosis* and *C. diphtheriae*, the software EDGAR was used to perform pan-genome analysis. This software identifies orthologous and paralogous genes by performing all-versus-all BLAST searches and using the score from the alignments to define orthology. As the BLAST score is a very variable metric, EDGAR normalizes the value by using the score rate value (SRV). SRV is calculated as the division of the bit score of a protein B against a protein A by the highest bit score against protein A, i.e., the bit score of protein A against itself. The resulting value, in the range from 0 to 1, is then rounded and multiplied by 100 to represent the percentage of homology [67]. The next steps in pan-genome analysis were performed as described above, using Heap's Law and least-squares fit to the exponential regression decay, and the $\alpha$ and $tg(\theta)$ were calculated for both species, *C. pseudotuberculosis* and *C. diphtheriae*, and for both biovars of *C. pseudotuberculosis*, ovis and equi. According to the results, *C. diphtheriae* presents an $\alpha$ of 0.69, whereas the $\alpha$ value of *C. pseudotuberculosis* is 0.89 [63,69]. Besides, in singletons analysis, the $tg(\theta)$ of *C. diphtheriae* and *C. pseudotuberculosis* are ~65 and ~19 genes, respectively. Altogether, the findings show that both pan-genomes are open, although the pan-genome of *C. pseudotuberculosis* is growing at slower rates when compared with the pan-genome of *C. diphtheriae*, given that *C. pseudotuberculosis* presents a $\alpha$ value that is closer to 1 and a lower number of genes will be added to the pan-genome for each newly sequenced genome. Finally, according to $\alpha$ value of the pan-genome and $tg(\theta)$ of the singletons of *C. pseudotuberculosis*, the pan-genome of the biovar ovis strains ($\alpha = 0.94$ and $tg(\theta) = 16.811$) is also growing at slower rates when compared with the pan-genome of the biovar equi strains ($\alpha = 0.89$ and $tg(\theta) = 34.533$) [63,69]. These examples illustrate how powerful the use of pan-genome may be in comparative analysis of bacterial species and give new directions on possible targets for genome sequencing, e.g., the choice of a higher number of strains from *C. pseudotuberculosis* biovar equi, opposing to biovar ovis, for future sequencing projects due to its higher variability.

2.3 Pathogenicity Islands identification using PIPS - *C. diphtheriae*, *C. ulcerans* and *C. pseudotuberculosis*

Prokaryotes are very promiscuous organisms, compared with Eukaryotes, in the sense they may achieve new environmental fitness through incorporation of incoming DNA from different organisms via horizontal gene transfer (HGT) [73]. Several different mechanisms may be involved in HGT events, playing a pivotal role in evolution by leaps through the incorporation of: plasmids, bacteriophages, transposons, insertion elements and genomic islands. Due to specific features of the source, and the mechanisms used in genome incorporation, horizontally acquired regions have in common: a deviant genomic signature (G+C content and codon usage), which reflects the genomic signature from the donor organism; the presence of flaking insertion sequences (IS) and/or tRNAs, which, in turn, may present a specific IS in their 3’ end region; and, the presence of transposases [74]. Moreover, genomic islands, large regions acquired by HGT events, may be absent from organisms of the same genus or related species and harbour high concentrations of specific genes, which classifies them in: resistant islands, with high concentrations of antibiotic resistance genes [75]; symbiotic islands, which may be related to the association of bacteria to host plants from the family leguminosae, for example [76]; metabolic islands, with several genes associated with the biosynthesis of secondary metabolites [77]; and, pathogenicity islands (PAIs), which have high concentrations of virulence factors, are present in pathogenic bacteria and absent from non-pathogenic species from the same genus and/or related species, and are involved in the re-emergence of several pathogenic organisms due to the insertion of new virulence determinants [78]. In view of its high instability and the incorporation of several genes in block, GEIs are very interesting for comparative genomics analysis, mainly for pan-genomics, where the information for several strains is available. The pan-genomics analysis associated with GEIs prediction may shade a light in genome plasticity on the whole species/genus and help in correlating the presence/absence of different regions with host- or environmental-adaptability.
In pathogenic bacteria, the prediction of PAIs may be performed by the identification of the common features described earlier, such as: deviant genomic signature, i.e., G+C content and codon usage deviation; presence of transposases and flanking tRNA and/or IS; high concentration of virulence factors; and, absence in non-pathogenic organism of the same genus or related species. In *C. pseudotuberculosis*, *C. diphtheriae* and *C. ulcerans*, those analyses were performed by the software PIPS: pathogenicity island prediction software [79] using *C. glutamicum* ATCC 13032 as non-pathogenic organism of the same genus. In *C. diphtheriae*, 13 PAIs (PICDs1-13) were firstly described in the strain NCTC 13129 [11]. Later, PIPS has identified 11 new putative PAIs (PICDs14-24)(Figure 2A) [79,80]. Finally, using information from the pan-genome of the species, 57 putative GEIs were identified through the whole species [69].

In *C. pseudotuberculosis*, PIPS has firstly identified 7 putative PAIs (PICPs 1-7) [12]. In a more recent work, 4 additional putative PAIs were identified (PICPs8-11), which presented variations in gene content in comparisons using biovar *ovis* and *equi* strains [81,82]. Finally, in pan-genomics analysis, 5 additional putative PAIs were identified, where several PAIs had already been predicted on the first work, but were discarded as the prediction force was low, and revalidated after manual curation during pan-genomics analyses (Figure 2B) [63]. In the last case, the higher number of genomes (15), isolated from different hosts, biovars and locations, gave a better view of regions of plasticity inside the PAIs, which helped in a better classification.

Finally, in *C. ulcerans*, PIPS has identified 16 putative PAIs (Figure 2C)(this work).

Interesting, most of genome plasticity in all 3 species occur in PAIs, like: PICDs 1, 4, 6, 7, 8, 9, 12 and 13 in *C. diphtheriae* NCTC 13129 (Figure 2A); PICPs 4, 5 and 9 in *C. pseudotuberculosis* 1002 (Figure 2B); and, PICUs 8 and 12 in *C. ulcerans* BR-AD22 (Figure 2C). Besides, *C. diphtheriae* shows a higher variability than *C. pseudotuberculosis*, in PAI content and variation, and all the biovar *equi* strains present specific deletions in the same PAIs. Additionally, *C. diphtheriae* NCTC 13129, *C. pseudotuberculosis* 1002 and *C. ulcerans* BR-AD22 share similarities in PAI content. For instance, the *tox* gene is harboured by PICD 1 that is acquired through HGT from corynephage [11], which rises the potential of other *Corynebacterium* to harbour this same region and, thus, have the potential to code DT. In fact, *C. pseudotuberculosis* 31, isolated from bufallo, and *C. ulcerans* 0102 were already showed to present the *tox* gene [63,83]. Additionally, *pld* gene, which is present inside a PAI in *C. pseudotuberculosis* 1002 (PICP 1), is also harboured by *C. ulcerans* BR-AD22 (PICU 1). Other prominent functions of genes harboured by PAIs in those species are: ABC-type transport systems, like the *fag* (Fe acquisition gene) cluster of genes and *ciu* (*Corynebacterium* iron uptake) operon; CRISPR loci; phage-related genes; and, pili clusters of genes (Table 1). The *fagABC* operon, *fagD* gene and *ciu* operon render bacteria able to acquire iron from low iron environments inside host, thus, contributing to virulence [84,85]. Additionally, the CRISPR loci are involved in a bacterial immunity-like system, or self-defence system, and play an pivotal role in degrading DNA from viruses and other organisms, preventing unbridled acquirement of DNA [86]. Finally, the pili clusters of genes code for proteins that are responsible for polymerization and attachment of pili structures to the cell wall and play an important role in adhesion and internalization [87]. In work performed by Zasada in 2012 [88], all invasive strains of *C. diphtheriae* were shown to present every tested pilus genes, which points for their important role in bacterial spread inside the host. Finally, in pan-genomics analyses of *C. pseudotuberculosis*, it was reported a high variability in pili clusters of genes in biovar *equi* strains, which may be co-related with the characteristic superficial diseases caused by those strains, where visceral commitment is rarely reported [63]. Finally, the study of PAIs may be very interesting for elucidating not only host-preference and disease patterns, but also for the identification of new vaccine candidates, like in reverse vaccinology approaches.
Fig. 2 Circular genome comparison between C. ulcerans, C. pseudotuberculosis, C. diphtheriae and C. glutamicum. Reference genomes used were: A, C. diphtheriae NCTC 13129; B, C. pseudotuberculosis 1002; and, C, C. ulcerans BR-AD22. The circular genome comparisons show the positions of putative pathogenicity islands in C. diphtheriae NCTC 13129 (A), C. pseudotuberculosis 1002 (B) and C. ulcerans BR-AD22 (C) and the presence/absence of the pathogenicity islands in other strains of the same species and species of Corynebacterium. GC content, G+C profile in the genome region; Cu, Corynebacterium ulcerans; Cd, Corynebacterium diphtheriae; Cp, Corynebacterium pseudotuberculosis; Cg, Corynebacterium glutamicum; PICD, putative pathogenicity island of C. diphtheriae; PICP, putative pathogenicity island of C. pseudotuberculosis; PICU, putative pathogenicity island of C. ulcerans.
<table>
<thead>
<tr>
<th>Strain*</th>
<th>Name**</th>
<th>Begin CDS</th>
<th>End CDS</th>
<th>Prominent function of island genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD</td>
<td>PICD 1</td>
<td>DIP0179</td>
<td>DIP0222</td>
<td>diphtheria toxin encoding corynephage</td>
</tr>
<tr>
<td>CD</td>
<td>PICD 2</td>
<td>DIP0223</td>
<td>DIP0247</td>
<td>adhesive pilus SpaDEF</td>
</tr>
<tr>
<td>CD</td>
<td>PICD 3</td>
<td>DIP0282</td>
<td>DIP0290</td>
<td>iron transport system</td>
</tr>
<tr>
<td>CD</td>
<td>PICD 4</td>
<td>DIP0334</td>
<td>DIP0359</td>
<td>secreted proteins, including polysaccharide degradation enzyme</td>
</tr>
<tr>
<td>CD</td>
<td>PICD 5</td>
<td>DIP0438</td>
<td>DIP0445</td>
<td>metal transport system and secreted proteins</td>
</tr>
<tr>
<td>CD</td>
<td>PICD 6</td>
<td>DIP0750</td>
<td>DIP0766</td>
<td>lantibiotic biosynthesis proteins</td>
</tr>
<tr>
<td>CD</td>
<td>PICD 7</td>
<td>DIP0794</td>
<td>DIP0823</td>
<td>phage-related proteins</td>
</tr>
<tr>
<td>CD</td>
<td>PICD 8</td>
<td>DIP1645</td>
<td>DIP1664</td>
<td>secreted proteins, including extracellular matrix-binding protein</td>
</tr>
<tr>
<td>CD</td>
<td>PICD 9</td>
<td>DIP1817</td>
<td>DIP1843</td>
<td>phage-related proteins</td>
</tr>
<tr>
<td>CD</td>
<td>PICD 10</td>
<td>DIP2010</td>
<td>DIP2015</td>
<td>adhesive pilus SpaABC</td>
</tr>
<tr>
<td>CD</td>
<td>PICD 11</td>
<td>DIP2064</td>
<td>DIP2093</td>
<td>fimbrial-associated protein and surface-anchored protein</td>
</tr>
<tr>
<td>CD</td>
<td>PICD 12</td>
<td>DIP2143</td>
<td>DIP2170</td>
<td>siderophore biosynthesis and transport proteins</td>
</tr>
<tr>
<td>CD</td>
<td>PICD 13</td>
<td>DIP2208</td>
<td>DIP2234</td>
<td>CRISPR locus</td>
</tr>
<tr>
<td>CD</td>
<td>PICD 14</td>
<td>DIP0028</td>
<td>DIP0051</td>
<td>CRISPR locus</td>
</tr>
<tr>
<td>CD</td>
<td>PICD 15</td>
<td>DIP0071</td>
<td>DIP0115</td>
<td>iron transport system</td>
</tr>
<tr>
<td>CD</td>
<td>PICD 16</td>
<td>DIP0267</td>
<td>DIP0275</td>
<td>antibiotic resistance protein</td>
</tr>
<tr>
<td>CD</td>
<td>PICD 17</td>
<td>DIP0320</td>
<td>DIP0326</td>
<td>transport system with unknown function</td>
</tr>
<tr>
<td>CD</td>
<td>PICD 18</td>
<td>DIP0448</td>
<td>DIP0466</td>
<td>two-component system and transport system with unknown function</td>
</tr>
<tr>
<td>CD</td>
<td>PICD 19</td>
<td>DIP0582</td>
<td>DIP0607</td>
<td>siderophore biosynthesis and transport proteins</td>
</tr>
<tr>
<td>CD</td>
<td>PICD 20</td>
<td>DIP1891</td>
<td>DIP1901</td>
<td>transport system with unknown function</td>
</tr>
<tr>
<td>CD</td>
<td>PICD 21</td>
<td>DIP1944</td>
<td>DIP1971</td>
<td>diverse functions and Proteins with unknown function</td>
</tr>
<tr>
<td>CD</td>
<td>PICD 22</td>
<td>DIP2021</td>
<td>DIP2049</td>
<td>secreted proteins, including secretory lipases</td>
</tr>
<tr>
<td>CD</td>
<td>PICD 23</td>
<td>DIP2123</td>
<td>DIP2135</td>
<td>transport system with unknown function</td>
</tr>
<tr>
<td>CD</td>
<td>PICD 24</td>
<td>DIP2302</td>
<td>DIP2345</td>
<td>two-component system and transport system with unknown function</td>
</tr>
<tr>
<td>CP</td>
<td>PICP1</td>
<td>Cp1002_0022</td>
<td>Cp1002_0031</td>
<td>phospholipase D and Fe acquisition genes (fag operon)</td>
</tr>
<tr>
<td>CP</td>
<td>PICP2</td>
<td>Cp1002_0040</td>
<td>Cp1002_0067</td>
<td>Fe and Choline transport system and transcriptional regulators</td>
</tr>
<tr>
<td>CP</td>
<td>PICP3</td>
<td>Cp1002_0173</td>
<td>Cp1002_0185</td>
<td>iron transport system</td>
</tr>
<tr>
<td>CP</td>
<td>PICP4</td>
<td>Cp1002_0980</td>
<td>Cp1002_0992</td>
<td>iron transport system</td>
</tr>
<tr>
<td>CP</td>
<td>PICP5</td>
<td>Cp1002_1445</td>
<td>Cp1002_1472</td>
<td>iron transport system and transcriptional regulators</td>
</tr>
<tr>
<td>CP</td>
<td>PICP6</td>
<td>Cp1002_1553</td>
<td>Cp1002_1565</td>
<td>transport system</td>
</tr>
<tr>
<td>CP</td>
<td>PICP7</td>
<td>Cp1002_1903</td>
<td>Cp1002_1932</td>
<td>diverse functions and proteins with unknown function</td>
</tr>
<tr>
<td>CP</td>
<td>PICP8</td>
<td>Cp1002_0159</td>
<td>Cp1002_0167</td>
<td>diverse functions and Proteins with unknown function</td>
</tr>
<tr>
<td>CP</td>
<td>PICP9</td>
<td>Cp1002_0553</td>
<td>Cp1002_0573</td>
<td>diverse functions</td>
</tr>
<tr>
<td>CP</td>
<td>PICP10</td>
<td>Cp1002_1617</td>
<td>Cp1002_1633</td>
<td>diverse functions</td>
</tr>
<tr>
<td>CP</td>
<td>PICP11</td>
<td>Cp1002_2069</td>
<td>Cp1002_2080</td>
<td>diverse functions and proteins with unknown function</td>
</tr>
<tr>
<td>CP</td>
<td>PICP12</td>
<td>Cp1002_0120</td>
<td>Cp1002_0140</td>
<td>diverse functions</td>
</tr>
<tr>
<td>CP</td>
<td>PICP13</td>
<td>Cp1002_0194</td>
<td>Cp1002_0196</td>
<td>ABC transport system</td>
</tr>
<tr>
<td>CP</td>
<td>PICP14</td>
<td>Cp1002_1681</td>
<td>Cp1002_1694</td>
<td>ABC transport system and diverse functions</td>
</tr>
<tr>
<td>CP</td>
<td>PICP15</td>
<td>Cp1002_1866</td>
<td>Cp1002_1883</td>
<td>spaA cluster of pili genes</td>
</tr>
<tr>
<td>CP</td>
<td>PICP16</td>
<td>Cp1002_2014a</td>
<td>Cp1002_2026</td>
<td>ABC transport system, transcriptional regulator, phage-related protein and diverse functions</td>
</tr>
<tr>
<td>CU</td>
<td>PICU1</td>
<td>CULC22_00019</td>
<td>CULC22_00042</td>
<td>CRISPR locus, phospholipase D and Fe acquisition genes (fag operon)</td>
</tr>
<tr>
<td>CU</td>
<td>PICU2</td>
<td>CULC22_00051</td>
<td>CULC22_00112</td>
<td>CRISPR locus, ABC transport systems, two-component</td>
</tr>
</tbody>
</table>
The idea of reverse vaccinology has been initially proposed by Rappuoli in 2000 [89] and relies on the identification of putative vaccine targets using the genome sequence of the pathogen and assaying the chosen targets in vitro, rather than cultivating the pathogen and isolating putative targets one by one, like in conventional vaccine development methodology [90,91]. The reverse vaccinology approach considers proteins that are somehow exposed to the host as putative targets for vaccine development as they are promptly recognized by the immune system. In view of this, proteins from the exoproteome and membrane proteins are considered good targets to elicit immune response [91]; where the exoproteome of an organism is defined as the total repertoire of exported proteins, consisting of 2 main classes: secreted proteins that, after cleavage, remain anchored to the cell wall, releasing the mature portion into the extracellular space; and, surface exposed proteins that, after cleavage, remain anchored to the cell wall [92].

Subtractive genomics may take advantage of the exoproteome, the idea of comparative genomics, or pan-genomics, and the main concepts of reverse vaccinology to search for vaccine candidates from genome sequences in a subtractive way [93]. Firstly, all gene sequences are considered for their presence in all genome sequences of a given species, because commonly shared genes (core genome) are better targets for vaccine development as they could possibly elicit immune response against all strains. Secondly, the commonly shared genes are analysed for their putative subcellular location (pan-exoproteome and membrane proteins). Next, adhesion probabilities and MHC I and II binding properties analyses are performed in the resulting dataset. Finally, a search for the presence of virulence factors or pathogenicity associated proteins (e.g., genes harbouring by pathogenicity islands) may be performed in order to find better targets, although it does not exclude the targets from the previous step [82]. In C. pseudotuberculosis and C. diphtheriae, several in silico strategies have been performed in order to identify new virulence factors and candidate vaccine targets using the ideas of reverse vaccinology [80,82], pan-exoproteome [92] and subtractive genomics [93], summarized at Table 2. In those works, the most recurrent functions are related to maltose transport system (malE and malL), penicillin binding proteins (pbpA, pbpB and pbpC) and resuscitation-promoting factors (rpfA, rpfB and rpfI). Briefly, penicillin-binding proteins are the primary targets of β-lactam antibiotics and play a pivotal role in bacterial cell elongation, septation and modulation of cellular morphology; maltose transport system genes code for carbohydrate-binding proteins, which were already reported to elicit host immune response; and, resuscitation-promoting factors were shown to restore the culturability of dormant mycobacteria and are also important for bacterial growth [82,94,95]. However, although many attractive vaccine candidates may be identified through those in silico strategies, experimental assaying is still a major requirement in order to validate those targets. Finally, the reverse vaccinology allied with pan-genomics, PAI analyses and subtractive genomics, as exposed here, have a great potential for the identification of new targets.

### Abbreviations:
- CD, C. diphtheriae NCTC 13129; CP, C. pseudotuberculosis 1002; and, CU, C. ulcerans BR-AD22.
- PICD, putative PAI of C. diphtheriae; PICP, putative PAI of C. pseudotuberculosis; and, PICU, putative PAI of C. ulcerans.

<table>
<thead>
<tr>
<th>CU</th>
<th>PICU</th>
<th>CULC22_00166</th>
<th>CULC22_00182</th>
<th>diverse functions and proteins with unknown function</th>
</tr>
</thead>
<tbody>
<tr>
<td>CU</td>
<td>PICU7</td>
<td>CULC22_01155</td>
<td>CULC22_01200</td>
<td>phage-related proteins</td>
</tr>
<tr>
<td>CU</td>
<td>PICU8</td>
<td>CULC22_01654</td>
<td>CULC22_01724</td>
<td>diverse functions and proteins with unknown function</td>
</tr>
<tr>
<td>CU</td>
<td>PICU9</td>
<td>CULC22_01773</td>
<td>CULC22_01788</td>
<td>secreted proteins and proteins with unknown function</td>
</tr>
<tr>
<td>CU</td>
<td>PICU10</td>
<td>CULC22_01794</td>
<td>CULC22_01816</td>
<td>diverse functions and proteins with unknown function</td>
</tr>
<tr>
<td>CU</td>
<td>PICU11</td>
<td>CULC22_01853</td>
<td>CULC22_01866</td>
<td>diverse functions</td>
</tr>
<tr>
<td>CU</td>
<td>PICU12</td>
<td>CULC22_01921</td>
<td>CULC22_01985</td>
<td>diverse functions; proteins with unknown function</td>
</tr>
<tr>
<td>CU</td>
<td>PICU13</td>
<td>CULC22_02033</td>
<td>CULC22_02044</td>
<td>chaperone and proteins with unknown function</td>
</tr>
<tr>
<td>CU</td>
<td>PICU14</td>
<td>CULC22_02071</td>
<td>CULC22_02085</td>
<td>cytochrome C biosynthesis and proteins with unknown function</td>
</tr>
<tr>
<td>CU</td>
<td>PICU15</td>
<td>CULC22_02101</td>
<td>CULC22_02116</td>
<td>spaDEF cluster of pili genes</td>
</tr>
<tr>
<td>CU</td>
<td>PICU16</td>
<td>CULC22_02130</td>
<td>CULC22_02168</td>
<td>spaBC cluster of pili genes; iron and oligopeptide transport system; urease operon; diverse functions</td>
</tr>
<tr>
<td>CU</td>
<td>PICU17</td>
<td>CULC22_02256</td>
<td>CULC22_02267</td>
<td>ABC transport system; transcriptional regulator; diverse functions</td>
</tr>
<tr>
<td>CU</td>
<td>PICU18</td>
<td>CULC22_02307</td>
<td>CULC22_02325</td>
<td>diverse functions; proteins with unknown function</td>
</tr>
</tbody>
</table>

2.4 Reverse vaccinology, pan-exoproteome and subtractive genomics analyses for identifying vaccine targets - C. pseudotuberculosis
<table>
<thead>
<tr>
<th>Species*</th>
<th>Gene name</th>
<th>Locus tag</th>
<th>Product/function</th>
<th>Methodology**</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD</td>
<td>CRISPR I</td>
<td>DIP0036-DIP0038</td>
<td>CRISPR-associated proteins</td>
<td>PAI</td>
<td>[80]</td>
</tr>
<tr>
<td>CD</td>
<td>CRISPR II</td>
<td>DIP208-DIP2215</td>
<td>CRISPR-associated proteins</td>
<td>PAI</td>
<td>[80]</td>
</tr>
<tr>
<td>CD</td>
<td>pbpB</td>
<td>DIP0298</td>
<td>Penicillin-binding protein 1B</td>
<td>RV</td>
<td>[80]</td>
</tr>
<tr>
<td>CD</td>
<td>slpA</td>
<td>DIP0365</td>
<td>Surface layer protein A</td>
<td>RV</td>
<td>[80]</td>
</tr>
<tr>
<td>CD</td>
<td>essT</td>
<td>DIP0559</td>
<td>ESAT-6-like protein</td>
<td>RV</td>
<td>[80]</td>
</tr>
<tr>
<td>CD</td>
<td>mepA</td>
<td>DIP0836</td>
<td>Secreted metalloendopeptidase</td>
<td>RV</td>
<td>[80]</td>
</tr>
<tr>
<td>CD</td>
<td>pbpC</td>
<td>DIP2294</td>
<td>Penicillin-binding protein</td>
<td>RV</td>
<td>[80]</td>
</tr>
<tr>
<td>CD</td>
<td>oppA</td>
<td>DIP0515</td>
<td>ABC-type oligopeptide transport system</td>
<td>RV</td>
<td>[80]</td>
</tr>
<tr>
<td>CD</td>
<td>oppA</td>
<td>DIP0956</td>
<td>ABC-type oligopeptide transport system</td>
<td>RV</td>
<td>[80]</td>
</tr>
<tr>
<td>CD</td>
<td>ciu operon</td>
<td>DIP0582-DIP0586</td>
<td>ABC-type iron transport system</td>
<td>PG, PAI</td>
<td>[69,80]</td>
</tr>
<tr>
<td>CD</td>
<td>cid operon</td>
<td>DIP1898-DIP1901</td>
<td>Cytochrome d ubiquinol oxidase</td>
<td>PAI, PAI</td>
<td>[80]</td>
</tr>
<tr>
<td>CD</td>
<td>dha operon</td>
<td>DIP2334-DIP2336</td>
<td>Dihydroxyacetone kinase</td>
<td>PAI</td>
<td>[80]</td>
</tr>
<tr>
<td>CD</td>
<td>pdx operon</td>
<td>DIP2022-DIP2022</td>
<td>Pyridoxine biosynthesis</td>
<td>PAI</td>
<td>[80]</td>
</tr>
<tr>
<td>CD</td>
<td>spaA cluster</td>
<td>DIP2010-DIP2013</td>
<td>Pili polymerization and attachment proteins</td>
<td>PG, PAI</td>
<td>[69]</td>
</tr>
<tr>
<td>CD</td>
<td>spaD cluster</td>
<td>DIP2033-DIP2038</td>
<td>Pili polymerization and attachment proteins</td>
<td>PG, RV, PAI</td>
<td>[69,80]</td>
</tr>
<tr>
<td>CD</td>
<td>spaH cluster</td>
<td>DIP2223-DIP2227</td>
<td>Pili polymerization and attachment proteins</td>
<td>PG, RV, PAI</td>
<td>[69,80]</td>
</tr>
<tr>
<td>CP</td>
<td>pld</td>
<td>Cp1002_0027</td>
<td>Phospholipase D</td>
<td>GS, 2GC, PAI, PE</td>
<td>[12,13,92]</td>
</tr>
<tr>
<td>CP</td>
<td>fagABC operon</td>
<td>Cp1002_0028-0030</td>
<td>ABC-type iron transport system</td>
<td>2GC, PAI</td>
<td>[12]</td>
</tr>
<tr>
<td>CP</td>
<td>fagD</td>
<td>Cp1002_0031</td>
<td>Iron siderophore binding protein</td>
<td>2GC, PAI</td>
<td>[12]</td>
</tr>
<tr>
<td>CP</td>
<td>mgtE</td>
<td>Cp1002_0046</td>
<td>Mg(2+) transporter</td>
<td>2GC, PAI</td>
<td>[12]</td>
</tr>
<tr>
<td>CP</td>
<td>mllL</td>
<td>Cp1002_0047</td>
<td>Oligo-1,6-glucosidase</td>
<td>2GC, PAI</td>
<td>[12]</td>
</tr>
<tr>
<td>CP</td>
<td>tetA</td>
<td>Cp1002_0049</td>
<td>Tetraclitin-efflux transporter</td>
<td>2GC, PAI</td>
<td>[12]</td>
</tr>
<tr>
<td>CP</td>
<td>sigK</td>
<td>Cp1002_0051</td>
<td>ECF family sigma factor K</td>
<td>2GC, PAI</td>
<td>[12]</td>
</tr>
<tr>
<td>CP</td>
<td>cskE</td>
<td>Cp1002_0050a</td>
<td>Anti-sigma factor</td>
<td>2GC, PAI</td>
<td>[12]</td>
</tr>
<tr>
<td>CP</td>
<td>dipZ</td>
<td>Cp1002_0053</td>
<td>C-type cytochrome biogenesis protein</td>
<td>2GC, PAI</td>
<td>[12]</td>
</tr>
<tr>
<td>CP</td>
<td>potG</td>
<td>Cp1002_0180</td>
<td>Putrescine ABC transport system</td>
<td>2GC, PAI</td>
<td>[12]</td>
</tr>
<tr>
<td>CP</td>
<td>gtpT</td>
<td>Cp1002_0183</td>
<td>Glycerol-3-phosphate transporter</td>
<td>2GC, PAI</td>
<td>[12]</td>
</tr>
<tr>
<td>CP</td>
<td>phoB</td>
<td>Cp1002_0184</td>
<td>Two-component regulatory protein</td>
<td>2GC, PAI</td>
<td>[12]</td>
</tr>
<tr>
<td>CP</td>
<td>lcoS</td>
<td>Cp1002_0185</td>
<td>Two-component sensor protein</td>
<td>2GC, PAI</td>
<td>[12]</td>
</tr>
<tr>
<td>CP</td>
<td>ciu operon</td>
<td>Cp1002_0981-0985</td>
<td>Corynebacterium iron uptake system</td>
<td>2GC, PAI, PE</td>
<td>[12,92]</td>
</tr>
<tr>
<td>CP</td>
<td>pfoS</td>
<td>Cp1002_1468</td>
<td>Regulatory protein</td>
<td>2GC, PAI</td>
<td>[12]</td>
</tr>
<tr>
<td>CP</td>
<td>spaA cluster</td>
<td>Cp1002_1867-1874</td>
<td>Pili polymerization and attachment proteins</td>
<td>GS, PG, PAI</td>
<td>[13,63]</td>
</tr>
<tr>
<td>CP</td>
<td>spaD cluster</td>
<td>Cp1002_1899-1901</td>
<td>Pili polymerization and attachment proteins</td>
<td>GS, PG, PAI</td>
<td>[13,63]</td>
</tr>
<tr>
<td>CP</td>
<td>pbdA</td>
<td>Cp1002_0035</td>
<td>Penicillin-binding protein A</td>
<td>RV, PE</td>
<td>[82]</td>
</tr>
<tr>
<td>CP</td>
<td>pbxB</td>
<td>Cp1002_0200</td>
<td>Penicillin-binding protein B</td>
<td>RV, PE</td>
<td>[82]</td>
</tr>
<tr>
<td>CP</td>
<td>malE</td>
<td>Cp1002_0377</td>
<td>Maltotriose-binding protein</td>
<td>RV, PE</td>
<td>[82]</td>
</tr>
<tr>
<td>CP</td>
<td>htaC</td>
<td>Cp1002_0454</td>
<td>Hypothetical protein with HtaA family domain</td>
<td>RV</td>
<td>[82]</td>
</tr>
<tr>
<td>CP</td>
<td>rpfA</td>
<td>Cp1002_0594</td>
<td>Resuscitation-promoting factor A</td>
<td>RV, GS</td>
<td>[13,82]</td>
</tr>
<tr>
<td>CP</td>
<td>gltB</td>
<td>Cp1002_0648</td>
<td>Glutamate ABC transporter</td>
<td>RV</td>
<td>[82]</td>
</tr>
<tr>
<td>CP</td>
<td>fhuD</td>
<td>Cp1002_0876</td>
<td>Iron(3+)-hydroxamate-binding protein FhuD</td>
<td>RV</td>
<td>[82]</td>
</tr>
<tr>
<td>CP</td>
<td>yceI</td>
<td>Cp1002_1013</td>
<td>Hypothetical protein YceI</td>
<td>RV, PE</td>
<td>[82]</td>
</tr>
<tr>
<td>CP</td>
<td>rnuA</td>
<td>Cp1002_1173</td>
<td>Holliday junction ATP-dependent DNA helicase</td>
<td>RV</td>
<td>[82]</td>
</tr>
<tr>
<td>CP</td>
<td>copC</td>
<td>Cp1002_1189</td>
<td>Copper resistance protein CopC</td>
<td>RV</td>
<td>[82]</td>
</tr>
<tr>
<td>Gene</td>
<td>Description</td>
<td>Reference</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>--------------------------------------------------</td>
<td>-----------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP thIX</td>
<td>Thiamine biosynthesis protein ThiX</td>
<td>RV</td>
<td>[82]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP lpgE</td>
<td>Lipoprotein LpqE</td>
<td>RV</td>
<td>[82]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP nrdC</td>
<td>Cytochrome c nitrite reductase</td>
<td>RV</td>
<td>[82]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP sfpA</td>
<td>Surface layer protein A</td>
<td>PE</td>
<td>[92]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP maIE</td>
<td>Maltose transport system</td>
<td>PE</td>
<td>[92]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP sprT</td>
<td>Trypsin</td>
<td>PE, PAI</td>
<td>[92]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP cynT</td>
<td>Carbonic anhydrase</td>
<td>PE</td>
<td>[92]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP rpfB</td>
<td>Resuscitation-promoting factor B</td>
<td>PE, GS</td>
<td>[13, 92]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP yceG</td>
<td>Amino deoxychorismate lyase</td>
<td>PE</td>
<td>[92]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP ctaC</td>
<td>Cytochrome c oxidase</td>
<td>PE</td>
<td>[92]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP lipY</td>
<td>Secretory lipase</td>
<td>PE</td>
<td>[92]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP oppA1</td>
<td>Oligopeptide binding protein</td>
<td>PE</td>
<td>[92]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP ltrR</td>
<td>Transcriptional regulator</td>
<td>PE</td>
<td>[92]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP pknL</td>
<td>Serine/Threonine protein kinase</td>
<td>PE</td>
<td>[92]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP gipQ</td>
<td>Glycerocephosphoryl diester phosphodiesterase</td>
<td>PE</td>
<td>[92]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP hemE</td>
<td>Uroporphyrinogen decarboxylase</td>
<td>PE</td>
<td>[92]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP accD5</td>
<td>Propionyl-CoA carboxylase</td>
<td>PE</td>
<td>[92]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP manB</td>
<td>D-alpha-D-mannose-1-phosphate guanylyltransferase</td>
<td>PE</td>
<td>[92]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP glmU</td>
<td>Bifunctional N-acetylglucosamine-1-phosphate uridylyltransferase/glucosamine-1-phosphate acetyltransferase</td>
<td>PE</td>
<td>[92]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP rpmI</td>
<td>50S ribosomal protein L35</td>
<td>PE</td>
<td>[92]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP ndhA</td>
<td>Membrane NADH dehydrogenase</td>
<td>PE</td>
<td>[92]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP dhp</td>
<td>Bifunctional phosphopantothenoylcysteine decarboxylase/phosphopantothenate synthase</td>
<td>PE</td>
<td>[92]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP qcrA</td>
<td>Rieske iron-sulfur protein</td>
<td>PE</td>
<td>[92]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP dca</td>
<td>C4-dicarboxylate-transport protein</td>
<td>PE</td>
<td>[92]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP ponA1</td>
<td>Penicillin-binding protein</td>
<td>PE</td>
<td>[92]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP cwlM</td>
<td>Hydrolase</td>
<td>PE</td>
<td>[92]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP dcd</td>
<td>Deoxycytidine triphosphate deaminase</td>
<td>2CG, SG, PAI</td>
<td>[12, 96]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP folP1</td>
<td>Dihydropteroate synthase</td>
<td>SG</td>
<td>[96]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP nrdH</td>
<td>Glutaredoxin-like protein</td>
<td>SG</td>
<td>[96]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP nrdI</td>
<td>Ribonucleotide-diphosphate reductase</td>
<td>SG</td>
<td>[96]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP murA</td>
<td>UDP-N-acetyl glucosamine 1-carboxyvinyltransferase</td>
<td>SG</td>
<td>[96]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP murE</td>
<td>UDP-N-acetyl muramoyl alanyl-D-glutamate--2,6-diamino pimelate ligase</td>
<td>SG</td>
<td>[96]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP nanH</td>
<td>Neurominidase H (sialidase)</td>
<td>GS</td>
<td>[13]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP rpfI</td>
<td>Resuscitation-promoting factor interacting protein</td>
<td>GS, PAI</td>
<td>[13]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP nor</td>
<td>Nitric oxide reductase</td>
<td>GS</td>
<td>[13]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP nps</td>
<td>Nonribosomal peptide synthetase 1</td>
<td>GS, PAI</td>
<td>[13]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP dtsR1</td>
<td>Acetyl-CoA carboxylase β-subunit</td>
<td>GS</td>
<td>[13]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP dtsR2</td>
<td>Acetyl-CoA carboxylase β-subunit</td>
<td>GS</td>
<td>[13]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP accD3</td>
<td>Acetyl-CoA carboxylase β-subunit</td>
<td>GS</td>
<td>[13]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CU cpp/ndoE</td>
<td>Corynebacterial protease</td>
<td>2GC</td>
<td>[14]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CU plb</td>
<td>Phospholipase D</td>
<td>2GC, PAI</td>
<td>[14]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CU spaBC</td>
<td>Pili polymerization and attachment proteins</td>
<td>2GC, PAI</td>
<td>[14]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CU spaDEF</td>
<td>Pili polymerization and attachment</td>
<td>2GC, PAI</td>
<td>[14]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster</td>
<td>CULC22_02106</td>
<td>Proteins</td>
<td>Abbreviation(s)</td>
<td>Reference</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
<td>----------</td>
<td>----------------</td>
<td>-----------</td>
<td></td>
</tr>
<tr>
<td>CU</td>
<td>rpfI</td>
<td>CULC22_01148</td>
<td>Rpf interacting protein</td>
<td>2GC</td>
<td>[14]</td>
</tr>
<tr>
<td>CU</td>
<td>cwlH</td>
<td>CULC22_01537</td>
<td>Cell wall-associated hydrolase</td>
<td>2GC</td>
<td>[14]</td>
</tr>
<tr>
<td>CU</td>
<td>nanH</td>
<td>CULC22_00437</td>
<td>Sialidase precursor (neuraminidase H)</td>
<td>2GC</td>
<td>[14]</td>
</tr>
<tr>
<td>CU</td>
<td>vspI</td>
<td>CULC22_00515</td>
<td>Venome serine protease</td>
<td>2GC</td>
<td>[14]</td>
</tr>
<tr>
<td>CU</td>
<td>tspA</td>
<td>CULC22_02007</td>
<td>Trypsin-like serine protease</td>
<td>2GC</td>
<td>[14]</td>
</tr>
<tr>
<td>CU</td>
<td>CRISPR I</td>
<td>CULC22_00029-CULC22_00032</td>
<td>CRISPR-associated proteins</td>
<td>2GC, PAI</td>
<td>[14]</td>
</tr>
<tr>
<td>CU</td>
<td>CRISPR II</td>
<td>CULC22_00106-CULC22_00111</td>
<td>CRISPR-associated proteins</td>
<td>2GC, PAI</td>
<td>[14]</td>
</tr>
</tbody>
</table>

**Abbreviations:**
* CD, *C. diphtheriae* NCTC 13129; CP, *C. pseudotuberculosis* 1002; and, CU, *C. ulcerans* BR-AD22.
** PAI, pathogenicity islands identification; GS, genome sequence and annotation; 2GC, 2 genomes comparison; PG, pan-genomics; SG, subtractive genomics; and, RV, reverse vaccinology.

### 3. Conclusions

*C. pseudotuberculosis*, *C. diphtheriae* and *C. ulcerans* are highly relevant pathogens for medical and veterinary research and have been extensively studied using genome sequencing and in silico strategies. In general, *C. ulcerans* is closely related to *C. pseudotuberculosis* and all 3 pathogenic species cluster together, regardless of the other *Corynebacteria*. *C. pseudotuberculosis* genome shows a more clonal-like behaviour in comparison with *C. diphtheriae*, which is shown on the pan-genome analyses and also in the higher number of PAIs of *C. diphtheriae* (PICDs 1-24) when compared with *C. pseudotuberculosis* (PICP1-16) and *C. ulcerans* (PICU 1-18). Regarding the methodologies, pan-genomic analyses data may be used to infer the development of the gene repertoire of the target organism and also to define conserved genes and variable content, where the variable content may arise from incorporation of GEIs, like PAIs. Analyses of PAIs are very interesting on the identification of genes related with new strain-specific features and host-adaptation, e.g., the presence of complete clusters of pili genes in invasive strains of *C. diphtheriae* and the underlying disease pattern. Moreover, subtractive genomics may take advantage of the conserved genes from the core genome (pan-genomic analyses), virulence factors (PAI analyses), reverse vaccinology and pan-exoproteome to identify putative vaccine and drug targets. Concluding, all those methodologies presented here are powerful tools for in silico analyses of pathogenic organisms and are very helpful in driving in vitro experimentation, thus, saving time and money.
References


IV. Goals
IV.1 Main goal

The main goal of this thesis was to perform pan-genomic analyses of *Corynebacterium pseudotuberculosis* and characterize the biovars *ovis* and *equi* through comparative genomics.

IV.2 Specific goals

The specific goals of this thesis were:

- to develop a new software for the identification of pathogenicity islands in pathogenic bacteria;
- to validate this software, to compare it with other gold standard programs and to assess its performance;
- to predict putative pathogenicity islands of *C. pseudotuberculosis* biovar *ovis*;
- to predict putative pathogenicity islands of *C. pseudotuberculosis* biovar *equi*;
- to predict new putative vaccine targets that could possibly elicit host immune response against both biovars of *C. pseudotuberculosis*, *ovis* and *equi*;
- to create a phylogenomic tree to find the evolutionary relationship between species of the genus *Corynebacterium*;
- to access the pan-genome, core genome and singleton subsets of *C. pseudotuberculosis* and to perform comparisons between both biovars according to these datasets;
- to predict additional putative pathogenicity islands of *Corynebacterium pseudotuberculosis* and to compare the genome plasticity between both biovars.
V. Research Articles
V.1 Chapter I. PIPS: Pathogenicity Island Prediction Software


In a first attempt to predict virulence factors in *C. pseudotuberculosis* 1002, we have tried to use several software intended for the identification of PAIs. However, most of the software are limited to the analysis of specific features of PAIs, instead of considering the whole scenario. PredictBias and IslandViewer are the only exception to those software as they predict PAIs in a multi-pronged way. However, both software presented other limitations, like: computationally expensive processes and unsolved dependencies that impaired the installation of IslandViewer; and, an online based architecture that required the submission of the genome sequence to PredictBias, which is a problem when the genome sequence is not allowed for submission before publication. In order to circumvent those problems, we have created a new software for the prediction of PAIs, named PIPS, that is publicly available for installation in personal computers and, which also outperformed the other previously available software in terms of performance. The following paper describes the implementation of PIPS and also presents comparisons between this software with PredictBias and IslandViewer.
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.

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 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-culturab le 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 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 fimbral 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 platforms 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-graph, 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]. Pandir 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 efficiency 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.
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],

---

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
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

![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](image-url)
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 corynebacterium-acquired region and is responsible for the pathognomonic symptoms of diphtheria [41].

*C. glutamicum* strain ATCC 13092 [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 pyelonephriotic uropathogenic *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.

PredictBias showed good specificity (88.7%), at the cost of sensitivity (2.4%), when using only predicted PAIs (PredictBias_PAIs) as a positive dataset for the test (Table 1). The sensitivity was higher (30.8%) when GEs 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 14.0%, 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 1.** Comparison between the software used to identify pathogenicity islands in the *C. diphtheriae* strain NCTC 13129.

<table>
<thead>
<tr>
<th>Software</th>
<th>Sensitivity (%)</th>
<th>Specificity(%)</th>
<th>Accuracy(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IslandPath_DIMOB</td>
<td>13.6</td>
<td>98.3</td>
<td>89.2</td>
</tr>
<tr>
<td>IslandPick</td>
<td>65.2</td>
<td>81.9</td>
<td>80.1</td>
</tr>
<tr>
<td>SIGI_HMM</td>
<td>14.0</td>
<td>94.9</td>
<td>86.2</td>
</tr>
<tr>
<td>IslandViewer</td>
<td>74.4</td>
<td>76.4</td>
<td>76.2</td>
</tr>
<tr>
<td>PredictBias_GE</td>
<td>30.8</td>
<td>84.4</td>
<td>78.6</td>
</tr>
<tr>
<td>PredictBias_PAIs</td>
<td>2.4</td>
<td>88.7</td>
<td>79.4</td>
</tr>
<tr>
<td>PIPS_Auto</td>
<td>86.4</td>
<td>85.0</td>
<td>85.1</td>
</tr>
<tr>
<td>PIPS_Manual</td>
<td>96.8</td>
<td>87.1</td>
<td>88.1</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0030848.t001

![Page 38](image-url)
Table 2. Comparison between the software used to identify pathogenicity islands in the uropathogenic E. coli strain CFT 073.

<table>
<thead>
<tr>
<th>Software</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IslandPath_DIMOB</td>
<td>44.5</td>
<td>99.3</td>
<td>90.2</td>
</tr>
<tr>
<td>IslandPick</td>
<td>7.5</td>
<td>99.7</td>
<td>84.5</td>
</tr>
<tr>
<td>SIGl_HMM</td>
<td>21.9</td>
<td>96.9</td>
<td>84.5</td>
</tr>
<tr>
<td>IslandViewer</td>
<td>55.8</td>
<td>96.2</td>
<td>89.5</td>
</tr>
<tr>
<td>PredictBias_GEI</td>
<td>60.0</td>
<td>93.7</td>
<td>88.1</td>
</tr>
<tr>
<td>PredictBias_PAI</td>
<td>39.2</td>
<td>96.2</td>
<td>86.8</td>
</tr>
<tr>
<td>PIPS_Auto</td>
<td>94.8</td>
<td>93.7</td>
<td>93.9</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Codon usage deviation (%)</th>
<th>GC content deviation (%)</th>
<th>Virulence factors (%)</th>
<th>Hypothetical proteins (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. pseudotuberculosis NCTC 13129 PICDs</td>
<td>20.80</td>
<td>18.40</td>
<td>39.20</td>
</tr>
<tr>
<td>C. diphtheriae NCTC 13129 genome</td>
<td>26.89</td>
<td>9.52</td>
<td>17.45</td>
</tr>
<tr>
<td>C. pseudotuberculosis 1002 PICPs</td>
<td>14.79</td>
<td>23.08</td>
<td>30.77</td>
</tr>
<tr>
<td>C. pseudotuberculosis 1002 genome</td>
<td>3.52</td>
<td>11.65</td>
<td>17.27</td>
</tr>
<tr>
<td>C. pseudotuberculosis 231 PICPs</td>
<td>19.62</td>
<td>20.25</td>
<td>32.91</td>
</tr>
<tr>
<td>C. pseudotuberculosis 231 genome</td>
<td>3.80</td>
<td>10.76</td>
<td>17.77</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0030848.t003

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 (75.5–60%). 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.
In a comparison of the frequency of CDSs with codon usage deviation, \textit{C. pseudotuberculosis} PICPs 3, 5, 6 and 7 had higher frequencies than those found in the whole genome of \textit{C. pseudotuberculosis} 1002. In \textit{C. pseudotuberculosis} C231, together with the previously described PAIs (PICPs 3, 5, 6 and 7), \textit{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 \textit{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 \textit{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 \textit{C. pseudotuberculosis} PICP5. However, this island, along with \textit{C. pseudotuberculosis} PICPs 3 and 6, had higher frequencies of hypothetical proteins.

No single characteristic was consistent throughout all \textit{C. pseudotuberculosis} PICPs. However, the absence of \textit{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 \textit{C. pseudotuberculosis} and \textit{C. diphtheriae}

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

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

Moreover, the flanking regions of the PICP5 of \textit{C. pseudotuberculosis} are the same as those of PICD8 of \textit{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 \textit{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 \textit{C. pseudotuberculosis} strains 1002 and C231. Y-axis: frequency in percentage; X-axis: PICPs and genomes of \textit{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](https://www.plosone.org/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.

**Figure S2** Graphic representation of PAI features in the genome (A) and in the pathogenicity islands (B) of *C. diphtheriae* and *C. pseudotuberculosis*. Y-axis: frequency; 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.

**References**


**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.

**Appendix S1** Availability and Requirements.

**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 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.
V.1.1 Appendix S1

Availability and requirements

- **Project name:** PIPS
- **Project home page:** http://www.genoma.ufpa.br/lgcm/pips
- **Link on bioinformatics.org:** http://www.bioinformatics.org/groups/?group_id=1063
- **Operating system(s):** UNIX Platform
- **Programming language:** Perl
- **Other requirements:** Java Virtual Machine v1.6.0_20, HMMER3, PERL v5.10.1, COLOMBO/SIGI-HMM v3.8 or higher
- **License:** GNU GPL
- **Restrictions for use by non-academics:** None
- **Run time:** Varies from 20 min to 1 h for 2.48-5.23 Mb genomes using a computer with two 2.20 GHz processors and 4 GB of RAM.
V.1.2 Figure S1

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. doi:10.1371/journal.pone.0030848.s001
V.1.3 Figure S2

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. doi:10.1371/journal.pone.0030848.s002
### V.1.4 Table S1

Table S1. PAI composition.

<table>
<thead>
<tr>
<th>PAI</th>
<th>Literature</th>
<th>PIPS</th>
<th>IslandViewer</th>
<th>PredictBias</th>
<th>Classification attributed by PredictBias</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Colombo SIGI-HMM</td>
<td>IslandPath DIMOB</td>
<td>IslandPick</td>
</tr>
<tr>
<td>PiCd 1</td>
<td>DIP0180-DIP0222</td>
<td>DIP0179-DIP0222</td>
<td>DIP0177-DIP0179</td>
<td>Not identified</td>
<td>DIP0180-DIP0222</td>
</tr>
<tr>
<td>PiCd 2</td>
<td>DIP0223-DIP0244</td>
<td>DIP0223-DIP0247</td>
<td>DIP0223-DIP0226</td>
<td>Not identified</td>
<td>DIP0227-DIP0241</td>
</tr>
<tr>
<td>PiCd 3</td>
<td>DIP0282-</td>
<td>DIP0282-</td>
<td>DIP0281-DIP0290</td>
<td>Not identified</td>
<td>DIP0279-DIP0289</td>
</tr>
<tr>
<td>PiCd 4</td>
<td>DIP0334-DIP0357</td>
<td>DIP0334-DIP0359</td>
<td>Not identified</td>
<td>Not identified</td>
<td>Not identified</td>
</tr>
<tr>
<td>PiCd 5</td>
<td>DIP0438-</td>
<td>Not identified</td>
<td>Not identified</td>
<td>Not identified</td>
<td>Not identified</td>
</tr>
<tr>
<td>PiCd 6</td>
<td>DIP0752-DIP0820</td>
<td>DIP0750-DIP0823</td>
<td>DIP0806-DIP0821</td>
<td>Not identified</td>
<td>DIP0795-DIP0804</td>
</tr>
<tr>
<td>PiCd 7</td>
<td>DIP1645-DIP1663</td>
<td>DIP1645-DIP1664</td>
<td>Not identified</td>
<td>Not identified</td>
<td>Not identified</td>
</tr>
<tr>
<td>PiCd 9</td>
<td>DIP1817-</td>
<td>DIP1817-</td>
<td>Not identified</td>
<td>DIP1817-DIP1817</td>
<td>Not identified</td>
</tr>
<tr>
<td>PiCd 10</td>
<td>DIP2010-</td>
<td>DIP2010-</td>
<td>DIP2010-2015</td>
<td>Not identified</td>
<td>Not identified</td>
</tr>
<tr>
<td>PiCd 11</td>
<td>DIP2066-DIP2093</td>
<td>DIP2064-DIP2093</td>
<td>Not identified</td>
<td>Not identified</td>
<td>DIP2063-DIP2081</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>---------------</td>
<td>---------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>PiCd 12</td>
<td>DIP2148-DIP2143</td>
<td>DIP2143-DIP2168</td>
<td>Not identified</td>
<td>Not identified</td>
<td>DIP2143-DIP2167</td>
</tr>
<tr>
<td>PiCd 13</td>
<td>DIP2208-DIP2234</td>
<td>DIP2208-DIP2234</td>
<td>Not identified</td>
<td>Not identified</td>
<td>DIP2207-DIP2227</td>
</tr>
</tbody>
</table>

The PAIs composition of the *C. diphtheriae* strain NCTC 13129, as described in the literature and as identified by PIPS, IslandViewer and PredicBias. doi:10.1371/journal.pone.0030848.s003
V.1.5 Discussion

In this section, we reported the implementation of the software PIPS: Pathogenicity Island Prediction Software and showed the better performance of this software in identifying pathogenicity islands in *C. diphtheriae* and *E. coli* when compared with other gold standard approaches. However, PIPS has only identified 12 out of 13 PAIs from *C. diphtheriae* (PICDs 1-13), where the PICD 5 was disregarded as a true PAI due to the absence of the following features: virulence factors, flanking tRNA genes, transposase genes and deviation in genomic signature. Besides, PIPS has also predicted 11 additional PAIs (PICDs 14-24) that were not previously described in literature. Taking into mind that the previously identified PAIs from literature were used as gold standard to assess the reliability of the software, the disagreement in prediction of PICD 5 was responsible for the loss in sensitivity, whereas the 11 additional PAIs have decreased the specificity of the software. However, in further pan-genomic studies with 13 strains of *C. diphtheriae* (Trost et al., 2012), PICD 5 did not present any sign of deletion in any of the strains, whereas 10 of the 11 additional PAIs presented regions of variability in at least one strain (Figure 5). Those patterns show that PICD 5 is probably not a true PAI, as previously anticipated by the absence of PAI features, and the 11 additional PAIs are true positive predictions, which corroborates the high efficiency of PIPS in predicting PAIs.

Furthermore, in comparison of PAIs of *E. coli* CFT 073, although PIPS had the best performance in predicting PAIs (93.9%) when compared to other software, its specificity was the lowest one (93.7%). However, the 13 PAIs available from literature, and used as gold standard, were only representative of PAIs larger than 30Kb (Lloyd et al., 2007). In this scenario, the additionally predicted PAIs in the range from 6kb up to 30kb were erroneously considered as false-positive due to the lack of gold standard data describing PAIs smaller than 30Kb, what decreased the overall specificity of PIPS. In view of this, one could anticipate that the additionally predicted PAIs are true-positive data. However, further studies have to be performed to assess whether they are true PAIs or not.

Since its development, PIPS has already been used by our group and collaborators in the prediction of PAIs of *C. pseudotuberculosis* (Ramos et al., 2013; Ruiz et al., 2011; Soares et al., 2012; Soares et al., 2013), *C. diphtheriae* (D’Afonseca et al., 2012; Trost et al., 2012), *C. ulcerans* (Introduction - book chapter), *Corynebacterium kroppenstedtii* (Ali et al., 2013), *Helicobacter pylori* (Ali, et al. - submitted article) and *Campylobacter fetus* subspecies (Ali et al., 2012). Additionally, the software has also been cited by other groups (Busby et al., 2013; Mebrhatu et al., 2013; Zhu et al., 2013); it has been visualized approximately 3200 times, where 26.21% of the visualizations lead to the download of the article; and, there are approximately 150 users from all around the world registered to the website http://www.genoma.ufpa.br/lgcm/pips. In view of the good response from
academics and taking in mind the previously proposed updates in the paper, we recently began to develop a new graphical version of PIPS in java language that may be used by researchers without any additional expertise in computer language. Furthermore, the software will be improved for the identification of all classes of genomics islands, i.e., symbiotic islands, resistance islands, metabolic islands and pathogenicity islands, using the following: ARDB - Antibiotic Resistance Genes Database (Liu & Pop, 2009) and Comprehensive Antibiotic Resistance Database (McMaster University, Canada, arpcard.mcmaster.ca); NodMutDB (Mao et al., 2005); classification under the category metabolism of COG - Cluster of Orthologous Genes; and, the database of virulence factors mVIRdb (Zhou et al., 2007). Finally, we intend to implement a graphical window for comparative analyses of genomic islands between different strains and species and, also, heatmap comparisons and phylogenomics analyses based in regions of plasticity. The initial steps in file parsing and identification of transposase and virulence factors are already implemented and we plan to finish and release the new version of the software and the underlying manuscript at the first semester of 2014.

Figura 5. Heatmap showing the presence/absence of PAIs identified by PIPS in 13 strains of C. diphtheriae. The numbers in the heatmap represent the percentage of similarity of a PAI in a given strain in a range from 0-10 (0-100%). The number 20 represents 200%, meaning that the referred PAI is duplicated on the underlying genome. Adapted from: doi: 10.1128/JB.00183-12.
V.2 Chapter II. Genome sequence of *Corynebacterium pseudotuberculosis* biovar *equi* strain 258 and prediction of antigenic targets to improve biotechnological vaccine production


As highlighted in the previous section, we have predicted 7 PAIs in *C. pseudotuberculosis* 1002 and C231, both from biovar *ovis*. Then, in the process of sequencing the 15 genomes of *C. pseudotuberculosis*, we have been confronted with the need for better characterizing the biovar *equi* strains and finding new PAIs and vaccine targets that were suitable to elicit immune response against both biovars, *ovis* and *equi*. In the following article, we have used PIPS to predict 4 additional PAIs (PICP 8-11) in *C. pseudotuberculosis* 258, biovar *equi*, which were in agreement with what we found in a parallel work in *C. pseudotuberculosis* 316, also from biovar *equi*. Furthermore, we have seen specific patterns of deletions in PAIs of both strains from biovar *equi* when compared to *C. pseudotuberculosis* 1002. Finally, we have applied the reverse vaccinology strategy, in a subtractive genomics approach, in order to identify conserved proteins between the biovars *ovis* and *equi* that could be promptly recognized by the immune system.
Genome sequence of Corynebacterium pseudotuberculosis biovar equi strain 258 and prediction of antigenic targets to improve biotechnological vaccine production


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 Pard, Guamá, Belém, PA, Brazil

A R T I C L E   I N F O

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 2088 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 vaccination strategy identified 49 putative antigenic proteins, which can be used as candidate vaccine targets in future works.

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 the macrophages (Dorella et al., 2006). The taxonomic identification of C. pseudotuberculosis is mainly performed by taking into account the 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 lymphadenitis (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 (Alemán 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. pseudotuberculosis* 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. Materials 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 assembling with the assemblers Velvet (Zerbino and Birney, 2008) and Edena (Hernandez et al., 2008), which perform data processing based on Euclidian 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); IntergroScan (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 *Corynebacterium glutamicum* ATCC 13032 (Soares et al., 2012). The plasticity comparison between strain 258 and *C. pseudotuberculosis* 1002 (CP001809), *C. pseudotuberculosis* CIP52.97 (CP003061), *Corynebacterium diphtheriae* NCTC 13129 (BX248353), *Corynebacterium 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); and (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 potential surface exposed proteins is the use of cell wall structures, 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 48 h 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 8 °C and washed 3 times with 0.1 M sodium cacodylate buffer (pH 7.2). After buffer washing, the
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 assembling with Velvet and Edena, generating 8004 contigs. Redundant sequences were then removed, reducing the number of contigs to 2289. 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 BIOL, 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 2088 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 their 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

**Table 1** Genomic features of *C. pseudotuberculosis* (Cp) strains.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Cp 1002</th>
<th>Cp CIP 52.97</th>
<th>Cp 258</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biovar</td>
<td>equi</td>
<td>equi</td>
<td>equi</td>
</tr>
<tr>
<td>Isolation</td>
<td>Goat (Brazil)</td>
<td>Horse (Kenya)</td>
<td>Horse (Belgium)</td>
</tr>
<tr>
<td>Size</td>
<td>2,335,113 bp</td>
<td>2,320,595 bp</td>
<td>2,314,404 bp</td>
</tr>
<tr>
<td>C+C content</td>
<td>52.19%</td>
<td>52.14%</td>
<td>52.15%</td>
</tr>
<tr>
<td>Proteins</td>
<td>2,090</td>
<td>2060</td>
<td>2088</td>
</tr>
<tr>
<td>rRNAs</td>
<td>12</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>tRNAs</td>
<td>48</td>
<td>47</td>
<td>49</td>
</tr>
<tr>
<td>Genes</td>
<td>2203</td>
<td>2194</td>
<td>2195</td>
</tr>
<tr>
<td>Pseudogenes</td>
<td>53</td>
<td>75</td>
<td>46</td>
</tr>
</tbody>
</table>
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).

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.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Cp 1002</th>
<th>Cp CIP 52.97</th>
<th>Cp 258</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell wall size (nm)</td>
<td>24.54</td>
<td>19.89</td>
<td>24.11</td>
</tr>
<tr>
<td>Cytoplasmic proteins</td>
<td>1417</td>
<td>1411</td>
<td>1428</td>
</tr>
<tr>
<td>Membrane proteins</td>
<td>370</td>
<td>368</td>
<td>370</td>
</tr>
<tr>
<td>PSE&lt;sup&gt;a&lt;/sup&gt; proteins</td>
<td>211</td>
<td>194</td>
<td>201</td>
</tr>
<tr>
<td>Secreted proteins</td>
<td>99</td>
<td>84</td>
<td>89</td>
</tr>
</tbody>
</table>

<sup>a</sup> Putative surface-exposed.
in SurfG+, we have classified 646–680 gene products as secreted proteins, putative surface-exposed (PSE) proteins or membrane proteins (Tables 2 and 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 (Tables 4 and 3 [rule III]).

Table 3

<table>
<thead>
<tr>
<th>Rule</th>
<th>Cp 1002</th>
<th>Cp CIP 52.97</th>
<th>Cp 258</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule I</td>
<td>680</td>
<td>646</td>
<td>660</td>
</tr>
<tr>
<td>Rule II</td>
<td>71</td>
<td>64</td>
<td>63</td>
</tr>
<tr>
<td>Rule III</td>
<td>49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rule IV</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

After searching for antigenic proteins, which are encoded by shared pathogenicity islands (Table 3 [rule IV]), we found one candidate protein, Cp528,1473 (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; Anzalda and Skaar, 2010). *C. pseudotuberculosis* also presents another antigenic protein involved in iron acquisition, represented by the *fluD* 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 *pbbB* 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).

Table 4

<table>
<thead>
<tr>
<th>Cp 1002</th>
<th>Cp CIP 52.97</th>
<th>Cp 258</th>
<th>Gene name</th>
<th>Subcellular location</th>
<th>Gene product</th>
</tr>
</thead>
<tbody>
<tr>
<td>CpCIP5297,0016</td>
<td>CpCIP5297,0016</td>
<td>Cp258,0017</td>
<td>–</td>
<td>PSE⁴</td>
<td>ABC transporter substrate-binding protein</td>
</tr>
<tr>
<td>CpCIP5297,0037</td>
<td>Cp258,0039</td>
<td>pbpA</td>
<td>secreted</td>
<td>Penicillin-binding protein A</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,0090</td>
<td>Cp258,0093</td>
<td>–</td>
<td>PSE</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,0120</td>
<td>Cp258,0139</td>
<td>–</td>
<td>secreted</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,0202</td>
<td>Cp258,0202</td>
<td>–</td>
<td>PSE</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,0207</td>
<td>Cp258,0206</td>
<td>pbbB</td>
<td>secreted</td>
<td>Penicillin-binding protein B</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,0219</td>
<td>Cp258,0218</td>
<td>–</td>
<td>membrane</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,0226</td>
<td>Cp258,0225</td>
<td>–</td>
<td>membrane</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,0321</td>
<td>Cp258,0318</td>
<td>–</td>
<td>PSE</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,0326</td>
<td>Cp258,0323</td>
<td>–</td>
<td>PSE</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,0388</td>
<td>Cp258,0385</td>
<td>maLE</td>
<td>secreted</td>
<td>L̂-Transpeptidase</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,0399</td>
<td>Cp258,0397</td>
<td>–</td>
<td>secreted</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,0420</td>
<td>Cp258,0424</td>
<td>–</td>
<td>PSE</td>
<td>Manganese ABC transporter, substrate-binding protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,0452</td>
<td>Cp258,0450</td>
<td>–</td>
<td>secreted</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,0489</td>
<td>Cp258,0467</td>
<td>htcAC</td>
<td>PSE</td>
<td>Hypothetical protein with HtAA family domain</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,0548</td>
<td>Cp258,0542</td>
<td>–</td>
<td>secreted</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,0563</td>
<td>Cp258,0557</td>
<td>–</td>
<td>PSE</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,0603</td>
<td>Cp258,0599</td>
<td>rpfA</td>
<td>secreted</td>
<td>Heme-binding protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,0654</td>
<td>Cp258,0654</td>
<td>–</td>
<td>PSE</td>
<td>Uncharacterized metalloprotease</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,0659</td>
<td>Cp258,0653</td>
<td>glbB</td>
<td>PSE</td>
<td>Glutamate ABC transporter, substrate-binding protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,0701</td>
<td>Cp258,0690</td>
<td>–</td>
<td>secreted</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,0782</td>
<td>Cp258,0771</td>
<td>–</td>
<td>secreted</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,0876</td>
<td>Cp258,0884</td>
<td>hupD</td>
<td>PSE</td>
<td>Iron(3+) hydroxamate-binding protein hupD</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,0894</td>
<td>Cp258,0892</td>
<td>–</td>
<td>membrane</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,1002</td>
<td>Cp258,0998</td>
<td>–</td>
<td>PSE</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,1017</td>
<td>Cp258,1014</td>
<td>–</td>
<td>secreted</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,1030</td>
<td>Cp258,1027</td>
<td>yceL</td>
<td>secreted</td>
<td>Hypothetical protein yceL</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,1102</td>
<td>Cp258,1100</td>
<td>–</td>
<td>PSE</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,1194</td>
<td>Cp258,1192</td>
<td>ruvA</td>
<td>PSE</td>
<td>Holiday junction ATP-dependent DNA helicase</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,1210</td>
<td>Cp258,1208</td>
<td>copC</td>
<td>PSE</td>
<td>Copper resistance protein CopC</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,1304</td>
<td>Cp258,1301</td>
<td>–</td>
<td>PSE</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,1380</td>
<td>Cp258,1380</td>
<td>–</td>
<td>membrane</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,1386</td>
<td>Cp258,1386</td>
<td>–</td>
<td>PSE</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,1404</td>
<td>Cp258,1403</td>
<td>–</td>
<td>PSE</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,1478</td>
<td>Cp258,1473</td>
<td>–</td>
<td>PSE</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,1517</td>
<td>Cp258,1510</td>
<td>thiX</td>
<td>PSE</td>
<td>Thiamine biosynthesis protein thiX</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,1520</td>
<td>Cp258,1514</td>
<td>–</td>
<td>secreted</td>
<td>Guanylyl-specific ribonuclease</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,1544</td>
<td>Cp258,1548</td>
<td>–</td>
<td>PSE</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,1617</td>
<td>Cp258,1606</td>
<td>–</td>
<td>PSE</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,1700</td>
<td>Cp258,1697</td>
<td>–</td>
<td>PSE</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,1781</td>
<td>Cp258,1780</td>
<td>lqpe</td>
<td>secreted</td>
<td>Lipoprotein lqpe</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,1785</td>
<td>Cp258,1783</td>
<td>–</td>
<td>PSE</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,1840</td>
<td>Cp258,1836</td>
<td>–</td>
<td>secreted</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,1869</td>
<td>Cp258,1864</td>
<td>nrfC</td>
<td>membrane</td>
<td>Cytochrome c nitrite reductase, small subunit</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,1920</td>
<td>Cp258,1910</td>
<td>–</td>
<td>secreted</td>
<td>Membrane protein</td>
<td></td>
</tr>
<tr>
<td>CpCIP5297,1983</td>
<td>Cp258,1972</td>
<td>–</td>
<td>PSE</td>
<td>Hypothetical protein</td>
<td></td>
</tr>
</tbody>
</table>

⁴ Putative surface-exposed.
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 malle 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 rnuA 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 control 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 equi 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.

Acknowledgments

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


Jones, M., Melo, C.R., Cole, S.M., Clarke, S.C., transfer Immuno-

Djoko, K.Y., Xiao, Z., Huffman, D.L., Wedd, A.G., Collens, A., immu-


terial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Research 18, 802–809.


Rappuoli, R., 2001. Reverse vaccinology, a genome-based approach to vaccine develop-

tors as antigens in novel tuberculosis sub-units vaccines. Microbes and Infection 14, 86–95.


58
V.2.1 Discussion

In this work, we have used PIPS to predict 11 PAIs in *C. pseudotuberculosis* 258, where we found specific patterns of deletion in PICPs 4, 5 and 9, on biovar *equi* strains when compared with *C. pseudotuberculosis* 1002, biovar *ovis*. Furthermore, we have applied reverse vaccinology strategy, by using the softwares SurfG+ and Vaxign (Barinov *et al.*,2009; He *et al.*,2010), to predict 49 putative vaccine targets that could possibly elicit immune response against both biovars, *ovis* and *equi*. From those 49 proteins, Cp258_1473 deserves further attention because it is located inside a PAI that is commonly shared by all studied strains, independently of biovar. Besides, maltotriose-binding proteins, penicillin-binding protein and resuscitation-promoting factors were shown to be very recurrent in other studies of reverse vaccinology, pan-exoproteome and genome sequence and annotation (Introduction - book chapter), thus, deserving a higher attention. Finally, the strategy used here was further applied for the identification of putative vaccine targets in *Streptococcus agalactiae* strains isolated from fish, bovine and human (Pereira *et al.*,2013) and may also be used by other researchers.
V.3 Chapter III. The Pan-Genome of the Animal Pathogen *Corynebacterium pseudotuberculosis* Reveals Differences in Genome Plasticity between the Biovar *ovis* and *equi* Strains


The genome sequencing era of *C. pseudotuberculosis* has finally culminated in the pan-genomics analyses of the whole species and both biovars separately. In the following research article, we describe the phylogenomics of the genus *Corynebacterium* using the software Gegenees and we further compare the generated tree with the ones from literature. Then, we analyse the pan-genome, core genome and singletons of the whole species and the ones from both biovars. From the extrapolation of each subset, we achieve a better view on the pan-genome evolution. Finally, we predict PAIs and correlate the plasticity in pili clusters of genes with the intracellular facultative behavior of *C. pseudotuberculosis*. 
The Pan-Genome of the Animal Pathogen Corynebacterium pseudotuberculosis Reveals Differences in Genome Plasticity between the Biovar ovis and equi Strains

Siomar C. Soares1,2,3, Artur Silva4, Eva Trost2,3, Jochen Blom2, Rommel Ramos4, Adriana Carneiro4, Amjad Ali1, Anderson R. Santos1, Anne C. Pinto1, Carlos Diniz1, Eudes G. V. Barbosa1, Fernanda A. Dorella1, Flávia Aburjaille1, Flávia S. Rocha1, Karina K. F. Nascimento1, Luis C. Guimarães1,2,3, Sintia Almeida1, Syed S. Hassan1, Syeda M. Bakhtiar1, Ulisses P. Pereira5, Vinicius A. C. Abreu1, Maria P. C. Schneider4, Anderson Miyoshi1, Andreas Tauch2,9, Vasco Azevedo1*

1Department of General Biology, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, 2Center for Biotechnology, Bielefeld University, Bielefeld, Nordrhein-Westfalen, Germany, 3CIB Graduate Cluster Industrial Biotechnology, Center for Biotechnology, Bielefeld University, Bielefeld, Nordrhein-Westfalen, Germany, 4Department of Genetics, Federal University of Para, Belem, Para, Brazil, 5Department 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.

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 [3]; 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 fimbrae, 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 *fagB* 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 symptoms 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 acquisition 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/]. 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),
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** strains, using strain 1002 as a reference; B) the biovar **equi** strains, using **C. pseudotuberculosis** strain CIP52.97 as a reference; and C) the biovar **ovis** strains, using **C. pseudotuberculosis** strain C231 and C258 as a reference. Initially, the genome sequences of all the complete **Corynebacterium** genomes, resulting in the variable content, which was calculated as genes that were present in only one strain and thus did not present orthologs in any other **C. pseudotuberculosis** strains. The gene set of subject strain A was compared with the gene set of query strain B, presenting orthologs in all the selected strains. 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 genes Singletons of **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), multi-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 cutoff 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

<table>
<thead>
<tr>
<th>Strains</th>
<th>Biovar</th>
<th>Host</th>
<th>Country of isolation</th>
<th>Clinical description</th>
<th>Genome size</th>
<th>Number of genes</th>
<th>Singletons</th>
<th>GenBank accession</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1002</td>
<td>ovis</td>
<td>Goat</td>
<td>Brazil</td>
<td>CLA abscess</td>
<td>2,335,113</td>
<td>2,203</td>
<td>0</td>
<td>CP001809</td>
<td>[43]</td>
</tr>
<tr>
<td>C231</td>
<td>ovis</td>
<td>Sheep</td>
<td>Australia</td>
<td>CLA abscess</td>
<td>2,328,208</td>
<td>2,204</td>
<td>3</td>
<td>CP001829</td>
<td>[43]</td>
</tr>
<tr>
<td>42/02-A</td>
<td>ovis</td>
<td>Sheep</td>
<td>Australia</td>
<td>CLA abscess</td>
<td>2,337,606</td>
<td>2,164</td>
<td>5</td>
<td>CP003062</td>
<td>[49]</td>
</tr>
<tr>
<td>PAT10</td>
<td>ovis</td>
<td>Sheep</td>
<td>Argentina</td>
<td>Lung abscess</td>
<td>2,335,323</td>
<td>2,200</td>
<td>1</td>
<td>CP002924</td>
<td>[50]</td>
</tr>
<tr>
<td>3/99-5</td>
<td>ovis</td>
<td>Sheep</td>
<td>Scotland</td>
<td>CLA</td>
<td>2,337,938</td>
<td>2,239</td>
<td>39</td>
<td>CP003152</td>
<td>[49]</td>
</tr>
<tr>
<td>267</td>
<td>ovis</td>
<td>Llama</td>
<td>USA</td>
<td>CLA abscess</td>
<td>2,337,628</td>
<td>2,249</td>
<td>8</td>
<td>CP003407</td>
<td>[51]</td>
</tr>
<tr>
<td>PS4896</td>
<td>ovis</td>
<td>Antelope</td>
<td>South Africa</td>
<td>CLA abscess</td>
<td>2,337,657</td>
<td>2,205</td>
<td>2</td>
<td>CP003385</td>
<td>–</td>
</tr>
<tr>
<td>I19</td>
<td>ovis</td>
<td>Cow</td>
<td>Israel</td>
<td>Bovine mastitis abscess</td>
<td>2,337,730</td>
<td>2,213</td>
<td>0</td>
<td>CP002251</td>
<td>[52]</td>
</tr>
<tr>
<td>FR4C41</td>
<td>ovis</td>
<td>Human</td>
<td>France</td>
<td>Necrotizing lymphadenitis</td>
<td>2,337,913</td>
<td>2,171</td>
<td>12</td>
<td>CP002097</td>
<td>[46]</td>
</tr>
<tr>
<td>CIP52.97</td>
<td>equi</td>
<td>Horse</td>
<td>Kenya</td>
<td>Ulcerative lymphangitis</td>
<td>2,320,595</td>
<td>2,194</td>
<td>30</td>
<td>CP003061</td>
<td>[53]</td>
</tr>
<tr>
<td>316</td>
<td>equi</td>
<td>Horse</td>
<td>USA</td>
<td>Abscess</td>
<td>2,310,415</td>
<td>2,234</td>
<td>25</td>
<td>CP003077</td>
<td>[54,55]</td>
</tr>
<tr>
<td>258</td>
<td>equi</td>
<td>Horse</td>
<td>Belgium</td>
<td>Ulcerative lymphangitis</td>
<td>2,314,404</td>
<td>2,195</td>
<td>29</td>
<td>CP003540</td>
<td>[56]</td>
</tr>
<tr>
<td>1/06-A</td>
<td>equi</td>
<td>Horse</td>
<td>USA</td>
<td>Abscess</td>
<td>2,279,118</td>
<td>2,127</td>
<td>20</td>
<td>CP003082</td>
<td>[57]</td>
</tr>
<tr>
<td>Cp162</td>
<td>equi</td>
<td>Camel</td>
<td>UK</td>
<td>Neck abscess</td>
<td>2,293,464</td>
<td>2,150</td>
<td>13</td>
<td>CP003652</td>
<td>[58]</td>
</tr>
<tr>
<td>31</td>
<td>equi</td>
<td>Buffalo</td>
<td>Egypt</td>
<td>Abscess</td>
<td>2,297,010</td>
<td>2,170</td>
<td>50</td>
<td>CP003421</td>
<td>[59]</td>
</tr>
</tbody>
</table>
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 $\kappa$ and $\gamma$ using the nonlinear least-squares curve fit to the mean values.

The core genes of all the strains, including the biovar avis 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 15032 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 30x.

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 Gegenes. 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 pseudodurum [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 avis, 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 ranging 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. avis) and C. pseudotuberculosis strain CIP52.97 (bv. equi) as references. Figure 2A shows specific patterns of deletions in all the biovar avis 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 avis had a pan-genome of 2,405 genes, 1.14-fold the average total number of genes in each biovar avis 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 $n = \kappa \cdot N^{\gamma}$, 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 $\kappa$ and $\gamma$ were determined to be 2,043.06 and 0.11, respectively, by using the statistical computing language R. According to Heaps’ Law, 1 an $\alpha \geq 1$ 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 $\alpha$ of 0.89, indicating that it has an open pan-genome. The extrapolation of the pan-genome was also separately calculated for both biovars, avis and equi. Although the biovar equi had the same $\alpha$ as the entire pan-genome (0.89), the biovar avis had a much-higher $\alpha$ 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, CpP54896, Cp267, CpPAT10, Cp19, 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, CpP54896, Cp19, 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
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 $n = \kappa \cdot \exp\left(\frac{-x}{\tau}\right) + t g(0)$, 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 $\kappa \cdot \exp\left(\frac{-x}{\tau}\right)$ term will tend toward 0, where $tg(0)$ 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 $n = \kappa \cdot \exp\left(\frac{-x}{\tau}\right) + t g(0)$, we calculated the $tg(0)$ (Figure 4) for the three datasets: A) all the genomes, B) the biovar *ovis* genomes, and C) the biovar *equi* genomes. The $tg(0)$ 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...
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
separation of the *ovis* and *equi* biovar strains, similar to the phylogenetic 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 phylogenetic 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 13 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 disulphide 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 catalyzes the transfer of adenine 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* strain 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 *corynebacterium* 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...
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 hypothetical 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 pili 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 (spaA–spaB–spaX–spaC) and spaD (spaD–spaD–spaY–spaE–spaF) clusters, where spaD 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, spaD and spaD 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 (spaD) is likely responsible for anchoring the pili to the cell wall [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 spaA–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 *spaB* 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 *spaB* 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

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>96</td>
<td>96</td>
<td>96</td>
<td>96</td>
<td>90</td>
<td>96</td>
<td>88</td>
<td>88</td>
<td>88</td>
<td>88</td>
<td>88</td>
<td>88</td>
<td>84</td>
<td>91</td>
</tr>
<tr>
<td>95</td>
<td>100</td>
<td>99</td>
<td>98</td>
<td>98</td>
<td>97</td>
<td>90</td>
<td>94</td>
<td>86</td>
<td>88</td>
<td>86</td>
<td>84</td>
<td>88</td>
<td>89</td>
<td>83</td>
</tr>
<tr>
<td>94</td>
<td>98</td>
<td>100</td>
<td>98</td>
<td>98</td>
<td>97</td>
<td>89</td>
<td>93</td>
<td>86</td>
<td>88</td>
<td>86</td>
<td>84</td>
<td>88</td>
<td>88</td>
<td>91</td>
</tr>
<tr>
<td>94</td>
<td>96</td>
<td>97</td>
<td>100</td>
<td>96</td>
<td>88</td>
<td>92</td>
<td>84</td>
<td>87</td>
<td>85</td>
<td>83</td>
<td>87</td>
<td>87</td>
<td>81</td>
<td>89</td>
</tr>
<tr>
<td>92</td>
<td>93</td>
<td>94</td>
<td>94</td>
<td>100</td>
<td>87</td>
<td>90</td>
<td>82</td>
<td>85</td>
<td>82</td>
<td>81</td>
<td>84</td>
<td>83</td>
<td>79</td>
<td>87</td>
</tr>
<tr>
<td>95</td>
<td>94</td>
<td>94</td>
<td>94</td>
<td>96</td>
<td>100</td>
<td>93</td>
<td>89</td>
<td>90</td>
<td>85</td>
<td>83</td>
<td>88</td>
<td>87</td>
<td>82</td>
<td>90</td>
</tr>
<tr>
<td>91</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>84</td>
<td>100</td>
<td>87</td>
<td>84</td>
<td>84</td>
<td>83</td>
<td>85</td>
<td>87</td>
<td>82</td>
<td>91</td>
</tr>
<tr>
<td>86</td>
<td>86</td>
<td>86</td>
<td>86</td>
<td>84</td>
<td>83</td>
<td>80</td>
<td>86</td>
<td>86</td>
<td>100</td>
<td>83</td>
<td>83</td>
<td>84</td>
<td>84</td>
<td>85</td>
</tr>
<tr>
<td>85</td>
<td>85</td>
<td>86</td>
<td>86</td>
<td>83</td>
<td>82</td>
<td>86</td>
<td>86</td>
<td>100</td>
<td>83</td>
<td>83</td>
<td>84</td>
<td>86</td>
<td>78</td>
<td>87</td>
</tr>
<tr>
<td>65</td>
<td>66</td>
<td>67</td>
<td>67</td>
<td>65</td>
<td>63</td>
<td>68</td>
<td>63</td>
<td>67</td>
<td>100</td>
<td>83</td>
<td>80</td>
<td>84</td>
<td>78</td>
<td>85</td>
</tr>
<tr>
<td>66</td>
<td>68</td>
<td>68</td>
<td>68</td>
<td>67</td>
<td>64</td>
<td>71</td>
<td>67</td>
<td>70</td>
<td>85</td>
<td>100</td>
<td>85</td>
<td>87</td>
<td>82</td>
<td>85</td>
</tr>
<tr>
<td>70</td>
<td>72</td>
<td>72</td>
<td>72</td>
<td>70</td>
<td>69</td>
<td>74</td>
<td>67</td>
<td>72</td>
<td>87</td>
<td>87</td>
<td>100</td>
<td>88</td>
<td>84</td>
<td>88</td>
</tr>
<tr>
<td>68</td>
<td>69</td>
<td>70</td>
<td>70</td>
<td>66</td>
<td>66</td>
<td>73</td>
<td>66</td>
<td>71</td>
<td>87</td>
<td>85</td>
<td>84</td>
<td>100</td>
<td>79</td>
<td>86</td>
</tr>
<tr>
<td>68</td>
<td>69</td>
<td>69</td>
<td>69</td>
<td>69</td>
<td>65</td>
<td>72</td>
<td>64</td>
<td>68</td>
<td>85</td>
<td>85</td>
<td>86</td>
<td>85</td>
<td>100</td>
<td>87</td>
</tr>
<tr>
<td>65</td>
<td>67</td>
<td>67</td>
<td>66</td>
<td>65</td>
<td>62</td>
<td>71</td>
<td>65</td>
<td>67</td>
<td>81</td>
<td>79</td>
<td>79</td>
<td>81</td>
<td>77</td>
<td>100</td>
</tr>
</tbody>
</table>

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 phylogenetic 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.
virulence genes and the low pathogenic potential characteristic of C. kroppenstedii [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 ~63 genes to the entire pan-genome [57]. Moreover, the slow increase and higher $z$ 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 $z$ 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 $z$ value of 0.89, the pan-genome of the biovar *ovis* had a higher $z$ 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 $z$ 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 (1,010 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 Gegenes (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 PAs, 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. Puls 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, ortB*, 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 spaD clusters of the biovar ovis strains, the spaD clusters presented differences in three strains of this biovar. As C. pseudotuberculosis P54B96 and PAT10, a frameshift 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 Cip162 are merged into the same reading frame. Overall, these results suggest that although C. pseudotuberculosis 258, 316, 1/06-A and Cip162 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 spaD 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 spaD 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, PiCp5; 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, Cp119, Cp42/02-A, Cp3/99-5, CpFR4C1 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, Cp119, Cp42/02-A, CpFR4C1, Cp3/99-5, Cp1/06-A; and, the biovar equi strains Cp31, CpCp162, Cp316, Cp258 and CIP52.97. CDS, coding sequences; tRNA, transfer RNA; rRNA, ribosomal RNA; and PAI, pathogenicity island. (TIF)

Acknowledgments

The authors thank the CAPES/DAAD international cooperation for financing an exchange scholarship (http://www.capes.gov.br/cooperacao-internacional) (grant number 51171119) and CLIB - Graduate Cluster Industrial Biotechnology (http://www.graduclustertc.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. AIP 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


72
40. McKean SC, Davies JK, Moore RJ (2007) Expression of phospholipase D, the
in axenic purulent exudate on common barnyard foraines. Am J Vet Res 47:
713–715.
30. Peel MM, Palmer GG, Stacpoole AM, Kerr TG (1997) Caseous lymphadenitis due to Corynebacterium pseudotuberculosis: report of ten cases from Australia and
40. McKean SC, Davies JK, Moore RJ (2007) Expression of phospholipase D, the
in axenic purulent exudate on common barnyard foraines. Am J Vet Res 47:
713–715.
30. Peel MM, Palmer GG, Stacpoole AM, Kerr TG (1997) Caseous lymphadenitis due to Corynebacterium pseudotuberculosis: report of ten cases from Australia and


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. doi:10.1371/journal.pone.0053818.s001.
V.3.2 Discussion

In view of the high variability of *C. pseudotuberculosis* biovar *equi*, our group is currently performing new sequencings of other biovar *equi* strains and reviewing all genome assemblages performed to date. Interestingly, in a recently published work (Ramos et al., 2013), it was discovered the presence of a complete Corynephage region inside *C. pseudotuberculosis* 31, which was responsible for the incorporation of the *tox* gene (coding for diphtheria toxin), and other additional regions of insertion along the genome. However, based on the high coverage of next-generation sequencers and also on the reference- and de novo-based assembling processes used so far, one could anticipate that the new genomes will change only by insertions instead of deletions. In this scenario, one could anticipate no changes in biovar *ovis* pan-genome, core genome and singletons, where the biovar *equi* would be represented by a lower α and small changes in all subsets. Finally, the pan-genomes would still be regarded as open and no variation would be expected on the pili clusters of genes.
VI. General Discussion
Throughout the works presented here, it has been shown an improvement in the performance of the prediction of PAIs by PIPS, which was mainly the result of adding new strains to the dataset. Although some PAIs were initially predicted by PIPS in *C. pseudotuberculosis* 1002, they were disregarded as true positive PAIs as they presented a "weak" prediction force. However, after the addition of new strains from biovar *equi*, like *C. pseudotuberculosis* 258 and 316, we have seen a high degree of plasticity in the regions where those PAIs were predicted and we have further classified them as true positive results. These findings are in agreement with the higher variability inside the biovar *equi* strains and show the importance of performing comparisons between different strains of the same species in order to achieve a better prediction of PAIs. Besides, those results also show the importance of implementing comparative analyses of predicted GEIs in the next version of PIPS, as highlighted on the Discussion of Chapter I.

On Chapter II, we have reported the prediction of the additional PAIs from biovar *equi* and we also showed an integrative reverse vaccinology's approach for the analyses of new vaccine candidates with the use of PAIs, exoproteome and MHC binding properties. Those newly identified targets will be very helpful for *in vitro* studies performed by our group and the methodology will also assist other researchers in identifying vaccine targets to elicit immune response against other pathogens.

Finally, on Chapter III, we have correlated the clonal-like behavior of *C. pseudotuberculosis* with: its facultative intracellular characteristic; and, the development of its pan-genome, core genome and singletons and also the ones from the underlying subsets from biovars *ovis* and *equi*. In view of the lower number of strains from biovar *equi* (6 strains) when compared to biovar *ovis* (9 strains), one could argue that this difference could account for a bias in pan-genome, core genome and singletons development. However, considering a higher number of biovar *ovis* strains, we would expect to see a lower number of core genes in biovar *ovis* and a higher number of genes in pan-genome and singletons when compared with the biovar *equi* ones, if the analyses would to be biased. However, the final scenario is the complete opposite, as a result of the higher variability of biovar *equi*, where the deletions inside the PAIs accounted for a smaller core-genome, and the high number of different insertions in all strains were responsible for a bigger pan-genome and singletons. This scenario corroborates for the higher variability of biovar *equi* and shed a light in the need for sequencing new strains of biovar *equi* as previously discussed. Finally, the high variability in pili clusters of genes in biovar *equi* strains, opposing to the conservative behavior of this same cluster in biovar *ovis*, has to be further studied *in vitro* in order to assess its putative correlation with the patterns of the diseases caused by *C. pseudotuberculosis* and the above mentioned clonal-like characteristic.
VII. Conclusions
Since the beginning of *C. pseudotuberculosis* genomic era, we have continuously explored *in silico* analyses in order to achieve a better view of the whole species and both biovars. In this process, we have:

a. implemented a new software for the prediction of PAIs with a better performance when compared with other available strategies. PIPS has predicted 16 PAIs in *C. pseudotuberculosis* and was also applied to several other organisms, like: *C. diphtheriae*, *C. ulcerans*, *C. kroppenstedtii*, *C. fetus* subspecies and *Helycobacter pylori*. Additionally, a new version of PIPS implemented in java is under development. The new software will be able to perform predictions of different classes of GEIs (resistance islands, metabolic islands, symbiotic islands and pathogenicity islands) and compare them between different strains;

b. predicted additional PAIs in biovar *ovis* strains and new vaccine targets that are present in *C. pseudotuberculosis* 1002, CIP 52.97 and 258 and could possibly elicit immune response against both biovars;

c. performed pan-genomics analyses on the whole species and separated biovars, predicting the pan-genome, core genome and singletons, along with their underlying extrapolations, for each of the subsets. From the heatmap created in phylogenomics analyses and also from the pan-genome, core genome and singleton analyses, we could affirm that *C. pseudotuberculosis* presents a clonal-like genome, which may be the result of the intracellular facultative behavior of the species, and the biovar *equi* strains are more variable in genome composition than biovar *ovis* strains. Finally, in PAI analyses, we have proposed a putative relationship between the plasticity of pili clusters of genes and the intracellular facultative behavior of *C. pseudotuberculosis* and its biovars.
VIII. Bibliography


IX. Appendices
IX.1 Curriculum Vitae

Address to this CV:  http://lattes.cnpq.br/4393381414254469

Full name:  Siomar de Castro Soares


Bibliographic Citations:  Soares, Siomar C.; Soares, S.C.; Soares, Siomar C;

Parental information:  Nilo Josias Soares and Cacilda de Fátima Soares

Birth information:  12/08/1983 - Uberaba/MG - Brazil

Identification document:  11942204 SSP - MG - 06/08/1998

CPF Number:  056.951.826-11

Passport:  FD191821

Professional Address:  Universidade Federal de Minas Gerais
                     Instituto de Ciências Biológicas
                     Departamento de Biologia Geral
                     Av. Antonio Carlos. ICB bloco Q3 259
                     Belo Horizonte
                     31270-215, MG - Brazil

Phone number:  +55 31 34092873

Eletronic Adress:  siomars@gmail.com

Formal Education

2009-current date

Doctorate in Genetics.
Universidade Federal de Minas Gerais, UFMG, Belo Horizonte, Brazil
with Sandwich Doctorate in Universität Bielefeld (Advisor : Andreas Tauch)
Title:  Pan-genomic analyses of Corynebacterium pseudotuberculosis and
characterization of the biovars ovis and equi through comparative genomics.
Advisor: Vasco Ariston de Carvalho Azevedo

2008 - 2009

Master's in Genetics.
Universidade Federal de Minas Gerais, UFMG, Belo Horizonte, Brazil
Title:  Validação de um método computacional para Identificação,
Caracterização e Comparação in silico de Ilhas de Patogenicidade no
Gênero Corynebacterium e aplicação no genoma de duas linhagens de
Corynebacterium pseudotuberculosis. Year of degree: 2009
Advisor: Anderson Miyoshi

2007 - 2008

Improvement Course in Bioinformatics (Aperfeiçoamento em Bioinformática).
Universidade Federal de Minas Gerais, UFMG, Belo Horizonte, Brazil
Title: Bioinformática
Advisor: Vasco Ariston de Carvalho Azevedo

2002 - 2007

Bachelor's in Biomedicina.
Universidade de Uberaba, UNIUBE, Uberaba, Brazil
Title: Pesquisa de Incidência de Isospora belli em pacientes portadores de
HIV.
Advisor: Elaine Grava Japaulo
Areas of Expertise
1. Molecular Genetics of Microorganisms
2. Genomics
3. Bioinformatics

Languages
German  Understanding Functional , Speaking Functional , Writing Functional , Reading Functional
English  Understanding Fluent , Speaking Fluent , Writing Fluent , Reading Fluent
Español Understanding Functional , Speaking Functional , Writing Functional , Reading Functional
Français Understanding Fluent , Speaking Functional , Writing Functional , Reading Fluent

Bibliographic Production

Articles Published in Scientific Journals

1. **Soares, SC**; Abreu, VAC; Ramos, RTJ; Cerdeira, L; Silva, A; Baumbach, J; Trost, E; Tauch, A; Hirata, R; Mattos-Guaraldi, AL; Miyoshi, A; Azevedo, V. PIPS: Pathogenicity Island Prediction Software. Plos One. v.7, p.e30848, 2012.

2. **Soares, SC**; Silva, A; Trost, E; Blom, J; Ramos, RTJ; Carneiro,AR; Ali, A; Santos, AR; Pinto, AC; Diniz, CAA; Barbosa, E; Dorella, FA; Aburjaile, FF; Rocha, FS; Fiaux, K; Guimarães, LC; Almeida, SS; Hassan, S; 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. v.8, p.e53818, 2013.

3. **Soares, SC**; Trost, E; Ramos, RTJ; Carneiro, AR; Santos, AR; Pinto, AC; Barbosa, E; Aburjaile, F; Ali, A; Diniz, CAA; Hassan, SS; Fiaux, K; Guimarães, LC; Bakhtiar, SM; Pereira, U; Almeida, SS; Abreu, VAC; Rocha, FS; Dorella, FA; 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. In press, 2012.

4. **Soares, SC**; Dorella, FA; Pacheco, LGC; Hirata Jr, R; Mattos-Guaraldi, AL; Azevedo, V; Miyoshi, A. Plasticity of Corynebacterium diphtheriae pathogenicity islands revealed by PCR. Genetics and Molecular Research. v.10, p.1290-1294, 2011.

5. Ramos, RTJ; Carneiro, AR; **Soares, SC**; 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. Journal of Microbiological Methods. In press, 2013.
6. Pereira, UP; Santos, AR; Hassan, SS; Aburjaile, FF; Soares, SC; Ramos, RTJ; Carneiro, AR; Guimarães, LC; Almeida, SS; Diniz, C; Mattos, SVM; Gomes de Sá, PHC; Ali, A; Bakhtiar, SM; Dorella, FA; Zerlotini, A; Araújo, FMG; Leite, LR; Oliveira, GC; Miyoshi, A; Silva, A; Azevedo, V; Figueiredo, HCP. Complete genome sequence of Streptococcus agalactiae strain SA20-06, a fish pathogen associated to meningoencephalitis outbreaks. STAND GENOMIC SCI. v.8(2), 2013.

7. Guimarães, LC; Soares, SC; Albersmeier, A; Blom, J; Jaenicke, S; Azevedo, V; Soriano, F; Tauch, A; Trost, E. Complete Genome Sequence of Corynebacterium urealyticum Strain DSM 7111, Isolated from a 9-Year-Old Patient with Alkaline-Encrusted Cystitis. Genome Announcements. v.1, p.e00264-13, 2013.

8. Ali, A; Soares, SC; Barbosa, EGV; Santos, AR; Barh, D; Bakhtiar, SM; Hassan, SS; Ussery, DW; Silva, A; Miyoshi, A; Azevedo, V. Microbial Comparative Genomics: An Overview of Tools and Insights Into The Genus Corynebacterium. Journal of Bacteriology & Parasitology. v.4, p.100067-100094, 2013.

9. Barh, D; Barve, N; Gupta, K; Chandra, S; Jain, N; Tiwari, S; Leon-Sicairos, N; Canizalez-Roman, A; Santos, AR; Hassan, SS; Almeida, S; Ramos, RTJ; Abreu, VAC; Carneiro, AR; Soares, SC; Castro, TLP; Miyoshi, A; Silva, A; Kumar, A; Amarendra, NM; Blum, K; Braverman, ER; Azevedo V. Exoproteome and Secretome Derived Broad Spectrum Novel Drug and Vaccine Candidates in Vibrio cholerae Targeted by Piper betel Derived Compounds. Plos One. Fator de Impacto(2011 JCR): 4,0920, v.8, p.e52773, 2013.

10. Ramos, RTJ; Carneiro, AR; Soares, SC; Santos, AR; Almeida, S; Guimarães, L, Figueira, F; Barbosa, E; Tauch, A; Azevedo, V; Silva A. Tips and tricks for the assembly of a genome using a semiconductor sequencer. Microbial Biotechnology (Online). v.6(2), p.150-6, 2013.


12. Santos, AR; Carneiro, A; Gala-García, A; Pinto, A; Barh, D; Barbosa, E; Aburjaile, F; Dorella, F; Rocha, F; Guimarães, LC; Zurita-Turk, M; Ramos, R; Almeida, S; Soares, S; Pereira, U; Abreu, VC; Silva, A; Miyoshi, A; Azevedo, V. The Corynebacterium pseudotuberculosis in silico predicted pan-exoproteome. BMC Genomics. v.13, p.S6, 2012.
13. Hassan, S; Guimarães, LC; Pereira, UP; Islam, A; Ali, A; Bakhtiar, SM; Ribeiro, D; Santos, AR; **Soares, SC**; Dorella, FA; Pinto, AC; Schneider, MPC; Barbosa, MSR; Almeida, SS; Abreu, VAC; Aburjaile, F; Carneiro, AR; Cerdeira, LT; Fiaux, K; Barbosa, E; Diniz, CAA; Rocha, FS; Ramos, RTJ; Neha, J; Tiwari, S; Barh, D; Miyoshi, A; Muller, B; Silva, A; Azevedo, V. Complete genome sequence of Corynebacterium pseudotuberculosis biovar ovis strain P54B96 isolated from antelope in South Africa obtained by rapid next generation sequencing technology. STAND GENOMIC SCI. v.7, p.189-199, 2012.

14. Pethick, FE; Lainson, AF; Yaga, R; Flockhart, A; Smith, DGE; Donachie, W; Cerdeira, LT; Silva, A; Bol, E; Lopes, TS; Barbosa, MS; Pinto, AC; Santos, AR; **Soares, SC**; Almeida, SS; Guimarães, LC; Aburjaile, FF; Abreus, VAC; Ribeiro, D; Fiaux, KK; Diniz, CAA; Barbosa, EGV; Pereira, UP; Hassan, SS; Ali, A; Bakhtiar, SM; Dorella, FA; Carneiro, AR; Ramos, RTJ; Rocha, FS; Schneider, MPC; Miyoshi, A; Azevedo, V; Fontaine, MC. Complete Genome Sequences of Corynebacterium pseudotuberculosis Strains 3/99-5 and 42/02-A, Isolated from Sheep in Scotland and Australia, Respectively. Journal of Bacteriology (Print). v.194, p.4736-4737, 2012.

15. Silva, A; Ramos, RTJ; Carneiro, AR; Pinto, AC; **Soares, SC**; Santos, AR; Almeida, SS; Guimarães, LC; Aburjaile, FF; Barbosa, EGV; Dorella, FA; Rocha, FS; Lopes, TS; Kawasaki, R; Sá, PG; Coimbra, NAR; Cerdeira, LT; Barbosa, MS; Schneider, MPC; Miyoshi, A; Selim, SAK; Moawad, MS; Azevedo, V. Complete Genome Sequence of Corynebacterium pseudotuberculosis Cp31, Isolated from an Egyptian Buffalo. Journal of Bacteriology (Print). v.194, p.6663-6664, 2012.

16. Lopes, T; Silva, A; Ramos, R; Carneiro, A; Dorella, FA; Rocha, FS; Santos, AR; Lima, ARJ; Guimarães, LC; Barbosa, EGV; Ribeiro, D; Fiaux, KK; Diniz, CAA; Abreus, VAC; Almeida, SS; Hassan, SS; Ali, A; Bakhtiar, SM; Aburjaile, FF; Pinto, AC; **Soares, SC**; Pereira, UP; Schneider, MPC; Miyoshi, A; Edman, J; Spier, S; Azevedo, V. Complete Genome Sequence of Corynebacterium pseudotuberculosis Strain Cp267, Isolated from a Llama. Journal of Bacteriology (Print). v.194, p.3567-3568, 2012.

17. Pethick, FE; Lainson, AF; Yaga, R; Flockhart, A; Smith, DGE; Donachie, W; Cerdeira, LT; Silva, A; Bol, E; Lopes, TS; Barbosa, MS; Pinto, AC; **Soares, SC**; Santos, AR; Almeida, SS; Guimarães, LC; Aburjaile, FF; Abreus, VAC; Ribeiro, D; Fiaux, KK; Diniz, CAA; Barbosa, EGV; Pereira, UP; Hassan, SS; Ali, A; Bakhtiar, SM; Dorella, FA; Carneiro, AR; Ramos, RTJ; Rocha, FS; Schneider, MPC; Miyoshi, A; Azevedo, V; Fontaine, MC. Complete Genome Sequence of Corynebacterium pseudotuberculosis Strain 1/06-A, Isolated from a Horse in North America. Journal of Bacteriology (Print). v.194, p.4476-4476, 2012.
18. Carneiro, AR; Ramos, RTJ; Dall'Agnol, H; Pinto, AC; Soares, SC; Santos, AR; Guimarães, LC; Almeida, SS; Barauna, RA; Graças, DA; Franco, LC; Ali, A; Hassan, SS; Nunes, CIP; Barbosa, MS; Fiaux, KK; Aburjaile, FF; Barbosa, EGV; Bakhtiar, SM; Vilela, D; Nobrega, F; Santos, AL; Carepo, MSP; Azevedo, V; Schneider, MPC; Pellizari, VA; Silva, A. Genome Sequence of Exiguobacterium antarcticum B7, Isolated from a Biofilm in Ginger Lake, King George Island, Antarctica. Journal of Bacteriology (Print). v.194, p.6689-6690, 2012.

19. Ramos, RTJ; Silva, A; Carneiro, AR; Pinto, AC; Soares, SC; Santos, AR; Almeida, SS; Guimarães, LC; Aburjaile, FF; Barbosa, EGV; Dorella, FA; Rocha, FS; Cerdeira, LT; Barbosa, MS; Tauch, A; Edman, J; Spier, S; Miyoshi, A; Schneider, MPC; Azevedo, V. Genome Sequence of the Corynebacterium pseudotuberculosis Cp316 Strain, Isolated from the Abscess of a Californian Horse. Journal of Bacteriology (Print). v.194, p.6620-6621, 2012.

20. Ali, A; Soares, SC; Santos, AR; Guimarães, LC; Barbosa, E; Almeida, SS; Abreu, VA; Carneiro, AR; Ramos, RT; Bakhtiar, SM; Hassan, SS; Ussery, DW; On, S; Silva, A; Schneider, MP; Lage, AP; Miyoshi, A; Azevedo, V. Campylobacter fetus subspecies: Comparative genomics and prediction of potential virulence targets. Gene (Amsterdam). v.508, p.145-156, 2012.

21. Viguetti, SZ; Pacheco, LGC; Santos, LS; Soares, SC; Bolt, F; Baldwin, A; Dowson, CG; Rosso, ML; Guiso, N; Miyoshi, A; Hirata, R; Mattos-Guaraldi, AL; Azevedo, V. Multilocus sequence types of invasive Corynebacterium diphtheriae isolated in the Rio de Janeiro urban area, Brazil. Epidemiology and Infection (Print). v.140, p.617-620, 2012.

22. Trost, E; Blom, J; Soares, SC; Huang, IH; Al-Dilaimi, A; Schroder, J; Jaenicke, S; Dorella, FA; Rocha, FS; Miyoshi, A; Azevedo, V; Schneider, MP; Silva, A; Camello, TC; Sabbadini, PS; Santos, CS; Santos, LS; Hirata, R; Mattos-Guaraldi, AL; Efstratiou, A; Schmitt, MP; Ton-That, H; Tauch, A. Pan-genomics of Corynebacterium diphtheriae: Insights into the genomic diversity of pathogenic isolates from cases of classical diphtheria, endocarditis and pneumonia. Journal of Bacteriology (Print). v.194, p.3199-3215, 2012.

23. D'Afonseca, V; Soares, SC; Ali, A; Santos, AR; Pinto, AC; Magalhães, A; Faria, CJ; Barbosa, EGV; Guimarães, LC; Almeida, SS; Abreu, VAC; Zerlotini, A; Carneiro, AR; Cerdeira, LT; Ramos, RTJ; Hirata Jr, R; Mattos-Guaraldi, AL; Trost, E; Tauch, A; Silva, A; Schneider, MPC; Miyoshi, A; Azevedo, V. Reannotation of the Corynebacterium diphtheriae NCTC13129 genome as a new approach to studying gene targets connected to virulence and pathogenicity in diphtheria. Open Access Bioinformatics. v.2012, p.1, 2012.
24. Hassan, SS; Schneider, MPC; Ramos, RTJ; Carneiro, AR; Ranieri, A; Guimarães, LC; Ali, A; Bakhtiar, SM; Pereira, UP; Santos, AR; Soares, SC; Dorella, F; Pinto, AC; Ribeiro, D; Barbosa, MS; Almeida, S; Abreu, V; Aburjaile, F; Fiaux, K; Barbosa, E; Diniz, C; Rocha, FS; Saxena, R; Tiwari, S; Zambare, V; Ghosh, P; Pacheco, LGC; Dowson, CG; Kumar, A; Barh, D; Miyoshi, A; Azevedo, V; Silva, A. Whole-Genome Sequence of Corynebacterium pseudotuberculosis Strain Cp162, Isolated from Camel. Journal of Bacteriology (Print). v.194, p.5718-5719, 2012.

25. Barh, D; Jain, N; Tiwari, S; Parida, BP; D'Afonseca, V; Li, L; Ali, A; Santos, AR; Guimarães, LC; Soares, SC; Miyoshi, A; Bhattacharjee, A; Misra, AN; Silva, A; Kumar, A; Azevedo, V. 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 (Print). v.78, p.73-84, 2011.

26. Trost, E; Al-Dilaimi, A; Papavasiliou, P; Schneider, J; Viehoever, P; Burkovski, A; Soares, SC; Almeida, SS; Dorella, FA; Miyoshi, A; Azevedo, V; Schneider, MP; Silva, A; Santos, CS; Santos, LS; Sabbadini, P; Dias, AA; Hirata, R; Mattos-Guaraldi, AL; Tauch, A. Comparative analysis of two complete Corynebacterium ulcerans genomes and detection of candidate virulence factors. BMC Genomics. v.12, p.383, 2011.

27. Cerdeira, LT; Schneider, MPC; Pinto, AC; Almeida, SS; Santos, AR; Barbosa, EGV; Ali, A; Aburjaile, FF; Abreu, VAC; Guimarães, LC; Soares, SC; Dorella, FA; Rocha, FS; Bol, E; Sá, PG; Lopes, TS; Barbosa, MS; Carneiro, AR; Ramos, RTJ; Coimbra, NAR; Lima, ARJ; Barh, D; Jain, N; Tiwari, S; Raja, R; Zambare, V; Ghosh, P; Trost, E; Tauch, A; Miyoshi, A; Azevedo, V; Silva, A. Complete Genome Sequence of Corynebacterium pseudotuberculosis Strain CIP 52.97, Isolated from a Horse in Kenya. Journal of Bacteriology (Print). v.193, p.7025-7026, 2011.

28. Stynen, APR; Lage, AP; Moore, RJ; Rezende, AM; Resende, VDS; Ruy, PC; Daher, N; Resende, DM; Almeida, SS; Soares, SC; Abreu, VAC; Rocha, AACM; Santos, AR; Barbosa, EGV; Costa, DF; Dorella, FA; Miyoshi, A; Lima, ARJ; Campos, FDS; Sá, PG; Lopes, TS; Rodrigues, RMA; Carneiro, AR; Leão, T; Cerdeira, LT; Ramos, RTJ; Silva, A; Azevedo, V; Ruiz, JC. Complete Genome Sequence of Type Strain Campylobacter fetus subsp. venerealis NCTC 10354T. Journal of Bacteriology (Print). v.193, p.5871-5872, 2011.
29. 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, ÉM; 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. v.6, p.e18551, 2011.

30. Hirata, R; Pacheco, LG; Soares, SC; Santos, LS; Moreira, LO; Sabbadini, PS; Santos, CS; Miyoshi, A; Azevedo, VA; Mattos-Guaraldi, AL. Similarity of rpoB gene sequences of sucrose-fermenting and non-fermenting Corynebacterium diphtheriae strains. Antonie Van Leeuwenhoek (Dordrecht. Online). v.99, p.733-737, 2011.

31. Almeida, SS; Magalhães, AAC; Soares, SC; Zurita-Turk, M; Goulart, LR; Miyoshi, A; Azevedo, V. The Phage Display Technique: Advantages and Recent Patents. Recent Patents on DNA & Gene Sequences. v.5, p.136-148, 2011.

32. Cerdeira, LT; Pinto, AC; Schneider, MPC; Almeida, SS; Santos, AR; Barbosa, EGV; Ali, A; Barbosa, MS; Carneiro, AR; Ramos, RTJ; Oliveira, RS; Barh, D; Barve, N; Zambare, V; Belchior, SE; Guimarães, LC; Soares, SC; Dorella, FA; Rocha, FS; Abreu, VAC; Tauch, A; Trost, E; Miyoshi, A; Azevedo, V; Silva, A. Whole-Genome Sequence of Corynebacterium pseudotuberculosis PAT10 Strain Isolated from Sheep in Patagonia, Argentina. Journal of Bacteriology (Print). v.193, p.6420-6421, 2011.

33. 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; Abreu, VAC; Shpigel, N; Miyoshi, A; Azevedo, V. Complete Genome Sequence of Corynebacterium pseudotuberculosis I-19, strain isolated from Israel Bovine mastitis. Journal of Bacteriology (Print). v.193, p.323-324, 2010.
34. Trost, E; Ott, L; Schneider, J; Schröder, J; Jaenicke, S; Goesmann, A; Husemann, P; Stoye, J; Dorella, F; Rocha, F; **Soares, SC**; 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**. v.11, p.728 - , 2010.

**Book chapters published**

1. Silva, A; Ramos, RTJ; Carneiro, AR; Almeida, SS; Abreu, VAC; Santos, AR; **Soares, SC**; Pinto, AC; Guimarães, LC; Barbosa, E; Schneider, MPC; Zambare, V; Barh, D; Miyoshi, A; Azevedo, V. Next-Generation Sequencing and Assembly of Bacterial Genomes. In: **OMICS: Applications in Biomedical, Agricultural, and Environmental Sciences.** 1ed, p. 1-713, 2013.

2. Azevedo, V; Abreu, VAC; Almeida, SS; Santos, AR; **Soares, SC**; Ali, A; Pinto, AC; Magalhães, A; Barbosa, EGV; Ramos, RTJ; Cerdeira, L; Carneiro, AR; Schneider, MPC; Silva, A; Miyoshi, A. Whole Genome Annotation: In Silico Analysis. In: **Bioinformatics - Trends and Methodologies.** 1 ed. Rijeka - Croatia : InTech - Open Access Publisher, p. 679-704, 2011.

**Articles in Magazines**

1. **Soares, SC**; Silva, A; Ramos, RTJ; Cerdeira, L; Ali, A; Santos, AR; Pinto, AC; Cassiano, AAM; Aburjaile, FF; Carneiro, AR; Guimarães, LC; Barbosa, EGV; Almeida, SS; Abreu, VAC; Miyoshi, A; Azevedo, V. Plasticidade Genômica e Evolução Bacteriana. Microbiologia in Foco, 26º CBM - Foz do Iguaçu, v. 16, p. 31-8, 2011.

2. Ruiz, JC; Santos, AR; Pinto, AC; Resende, DM; Cerdeira, L; Ramos, RTJ; Cuadros-Orellana, S; Almeida, SS; **Soares, SC**; D'Afonseca, V; AZEVEDO, V; Silva, A. Second Genomic Revolution: the use of Next-Generation Sequencers. Microbiologia in Foco, 25º CBM - Porto de Galinhas, v.9, p.15 - 18, 2009.