

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
DEPARTAMENTO DE BIOLOGIA GERAL
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA



PhD Thesis

**Comparative genomics and positive selection analysis in
*Corynebacterium pseudotuberculosis***

PhD STUDENT: Marcus Vinicius Canário Viana

SUPERVISOR: Prof. Dr. Vasco Ariston de Carvalho Azevedo

CO-SUPERVISOR: Dr. Alice Rebecca Wattam

BELO HORIZONTE

March – 2018

Marcus Vinicius Canário Viana

**Comparative genomics and positive selection analysis in
*Corynebacterium pseudotuberculosis***

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: Dr. Alice Rebecca Wattam

BELO HORIZONTE

March – 2018

043

Viana, Marcus Vinicius Canário.

Comparative genomics and positive selection analysis in *Corynebacterium pseudotuberculosis* [manuscrito] / Marcus Vinicius Canário Viana. – 2018.

91 f. : il. ; 29,5 cm.

Orientador: Prof. Dr. Vasco Ariston de Carvalho Azevedo. Co-orientadora: Dr. Alice Rebecca Wattam.

Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas.

1. Genômica - Teses. 2. *Corynebacterium* - Teses. 3. Seleção - Biologia. I. Azevedo, Vasco Ariston de Carvalho. II. Wattam, Alice Rebecca. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 575




**"Comparative genomics and positive selection analysis in
Corynebacterium pseudotuberculosis"**

Marcus Vinicius Canário Viana


Tese aprovada pela banca examinadora constituída pelos Professores:


Vasco Ariston de Carvalho Azevedo - Orientador
UFMG


Francisco Pereira Lobo
UFMG


Gabriel da Rocha Fernandes
FIOCRUZ


Rogerio Margis
UFRGS


Georgios Joannis Pappas Júnior
UnB

Belo Horizonte, 27 de março de 2018.

I dedicate this work to my father, Aduino da Silva Viana, who was and still is a role model to me; and my mother, Sonia Maria Canário Viana, who raised me and supported my decisions of professional carrier.

ACKNOWLEDGEMENTS

I would like to thank everyone who is working with me during my PhD.

- Prof. Dr. Vasco Ariston de Carvalho Azevedo, for the supervision, for giving the opportunity to work in LGCM, the trust and investment in my carrier.
- Prof. Dr. Alice Rebecca Wattam, for the co-supervision, for giving the opportunity to work in Biocomplexity Institute of Virginia Tech, for the receptivity and orientation in United States, and for being the kindest professor I ever had.
- The Post-Graduation Programs in Genetics and Bioinformatics of UFMG, for the disciplines and support.
- Laboratory of Cellular and Molecular Genetics, specially Siomar, Diego, and Luis for the knowledge I have acquired from; and Edgar, Thiago, Mariana and Douglas for the friendship and making all that time better.
- Laboratory of DNA Polymorphism, specially Rommel, for the collaboration.
- Biocomplexity Institute of Virginia Tech, for receiving me as a foreign student, and giving the necessary support and structure to work.
- To my former professors in undergraduation and master's degree, from who I got the knowledge and inspiration to continue the academic life, specially Ana Maria, Paulo Affonso and Ana Karina.

Table of contents

List of figures	i
Abbreviations	ii
Abstract	1
Resumo	2
I. Presentation	3
I.1 <i>Corynebacterium pseudotuberculosis</i> genomics research	4
I.2 Collaborators	4
I.3 Thesis structure	5
II. Introduction.....	6
II.1 <i>Corynebacterium pseudotuberculosis</i>	7
II.1.1 Taxonomy	7
II.1.2 Biovar Ovis and Equi.....	9
II.1.3 Virulence factors	10
II.1.4 Difficulties in control measures.....	11
II.2 Comparative genomics of <i>C. pseudotuberculosis</i>	12
II.2.1. Genome variation and adaptation	12
II.2.1.1 Genome synteny	13
II.2.1.2 Pangenomics	13
II.2.1.3 Pathogenicity islands.....	15
II.2.5 Phylogenomics	15
II.3 Pathogens Resource Integration Center	16
III. Goals	17
III.1 Main goal.....	18
III.2 Specific goals	18
IV. Articles	19
IV.1 Chapter I. Comparative genomic analysis between <i>Corynebacterium pseudotuberculosis</i> strains isolated from buffalo	20
II.2 Chapter II. – Positive selection analysis in bacteria: methods and findings	21
IV.3 Chapter III. Rapidly evolving changes and gene loss associated with host switching in <i>Corynebacterium pseudotuberculosis</i>	22
V. Discussion.....	23
VI. Conclusion and perspectives.....	26
VII. Bibliography	28

VIII. Appendix	37
A. Published, accepted and submitted research articles	37
VIII. Appendix	69
B. Book chapter and review	69
VIII. Appendix	75
C. National and international workshop, course and poster presentations.....	75

List of figures

Figure 1. Phylogeny of CMNR group. The tree was built based on the core proteome with progressive refinement and the Maximum Likelihood method implemented in the pipeline PEPR..... 8

Abbreviations

AA	Amino Acids
ACDH	Fatty Acyl-CoA Dehydrogenase
ADP	Adenosine Diphosphate
AQUACEN	Laboratório Oficial Central do Ministério da Pesca e Aquicultura (National Reference Laboratory for Aquatic Animal Diseases)
ATP	Adenosine Triphosphate
BRIG	Blast Ring Image Generator
BLASTp	Basic Local Alignment Search Tool (protein)
CDS	Coding Sequence
CLA	Caseous lymphadenitis
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico (“National Counsel of Technological and Scientific Development”)
COGs	Protein Database of Clusters of Orthologous Groups
Cp	<i>Corynebacterium pseudotuberculosis</i>
DNA	Deoxyribonucleic Acid
FAPEMIG	Fundação de Amparo à Pesquisa do Estado de Minas Gerais (Research Support Foundation of the State of Minas Gerais)
G+C	Guanine + Cytosine
GEI	Genomic Island
LGCM	Laboratório de Genética Celular e Molecular (Laboratory of Cellular and Molecular Genetics)
LBMCF	Laboratório de Biologia Molecular e Computacional de Fungos
LPDNA	Laboratório de Polimorfismo de DNA (Laboratory of DNA Polimorphism)
LRT	Likelihood Ratio Test
NCBI	National Center of Biotechnology Information
PAI	Pathogenicity Island
PATRIC	Pathosystems Resource Integration Center
Pb	Base Pairs
PLD	Phospholipase D Protein
PLfam	PATRIC Genus-Specific Family
PRPq	Pró-Reitoria de Pesquisa
RNA	Ribonucleic Acid
rRNA	Ribosomal Ribonucleic Acid
tRNA	Transporter ribonucleic Acid

UFMG	Universidade Federal de Minas Gerais (Federal University of Minas Gerais)
UFPA	Universidade Federal do Pará (Federal University of Pará)
UniProt	Universal Protein Resource

Abstract

Corynebacterium pseudotuberculosis is a Gram-positive, facultative intracellular bacterium of veterinary and medical importance with a global distribution. It infects a variety of mammals, such as sheep, goats, horses, buffalo and humans, and causes economic losses in animal production. Biovar Equi is nitrate positive and has horse and buffalo as exclusive hosts. Biovar Ovis is nitrate negative and has preference for sheep and goats. Antibiotic treatments have reduced efficiency due to the protection provided by bacteria sequestered in abscesses, and current vaccines have different protection efficiency in different hosts. In order to know more about the mechanisms of pathogenicity as well as the evolution of the species, we performed comparative genome and genome-scale positive selection analysis. Genomes isolated from buffalo hosts in the same disease outbreak were compared to each other and other *C. pseudotuberculosis* strains. The characteristic that differentiated buffalo isolates from others was a prophage containing the diphtheria toxin inserted into a pathogenicity island, suggesting this as a requirement for this host infection. Phylogenomic analysis of 29 genomes of different hosts and biovars suggested that Ovis is a monophyletic group derived from Equi. Positive selection analysis identified 27 genes involved in adaptations of specific branches of *C. pseudotuberculosis*, some of them are drug or vaccine targets. The results were combined with data from literature to explain the genetic structure and evolution of *C. pseudotuberculosis* based on an ecological diversification model.

Keywords: Comparative genomics, positive selection, *Corynebacterium*

Resumo

Corynebacterium pseudotuberculosis é uma bactéria Gram-positiva e intracelular facultativa de importância médica e veterinária. A espécie infecta uma variedade de mamíferos, como ovelhas, cabras, cavalos, búfalos e humanos, e causa perdas econômicas na produção animal. O biovar Equi é nitrato positivo e tem o cavalo e o búfalo como hospedeiros exclusivos. O biovar Ovis é nitrato negativo e tem preferência por ovinos e caprinos. Os tratamentos com antibióticos reduziram a eficiência, devido à proteção proporcionada pelos abscessos, e as vacinas atuais apresentam eficiência de proteção diferente em diferentes hospedeiros. Com o objetivo de conhecer melhor os mecanismos de patogenicidade e a evolução das espécies, realizamos análises de genômica comparativa e seleção positiva em escala genômica. Genomas isolados de hospedeiros bubalinos no mesmo surto da doença foram comparados entre si e com outras cepas de *C. pseudotuberculosis*. A característica que diferenciou isolados de búfalos de outros isolados foi um prófago contendo a toxina da difteria inserida em uma ilha de patogenicidade, sugerindo este prófago como um requisito para essa infecção deste hospedeiro. Análise de análise filogenômica de 29 genomas de diferentes hospedeiros e biovars sugeriu que Ovis é um grupo monofilético derivado de Equi. A análise de seleção positiva identificou 27 genes envolvidos em adaptações de ramos específicos de *C. pseudotuberculosis*, alguns deles são alvos de drogas ou vacinas. Os resultados foram combinados com a literatura de forma de dados para explicar a estrutura genética e evolução de *C. pseudotuberculosis* com base em um modelo de diversificação ecológica.

Palavras-chave: Genômica comparativa, seleção positiva, *Corynebacterium*

I. Presentation

I.1 *Corynebacterium pseudotuberculosis* genomics research

The Laboratory of Cellular and Molecular Genetics and collaborators research the pathogenic mechanisms, diagnostics and vaccines of *Corynebacterium pseudotuberculosis* by structural and functional genomics. The first complete genome sequenced by our group was strain 1002 biovar Ovis, isolated from a goat, released in 2010. The comparative analysis done by our group identified virulence factors, pathogenicity islands, predicted vaccine targets (BARAÚNA et al., 2017; RUIZ et al., 2011; SOARES et al., 2013a, 2013b), and genes related to stress response (PINTO et al., 2014). More information about the species is becoming available, as more genomes are sequenced. Nowadays, 75 genomes of this species are available in GeneBank, 69 of them deposited by our research group and collaborators.

I.2 Collaborators

This work was performed on the Laboratories of Molecular and Cellular Genetics (LGCM) and National Reference Laboratory for Aquatic Animal Diseases (AQUACEN), at Federal University of Minas Gerais (UFMG), and the Biocomplexity Institute of Virginia Tech, at the Virginia Polytechnic Institute and State University, in a collaboration between the following researchers in alphabetic order:

Dr. Alice Rebecca Wattam, Researcher from Biocomplexity Institute - Virginia Tech, USA;
Msc. Andrew Warren, Researcher from Biocomplexity Institute - Virginia Tech, USA;
Dr. Arne Sahm, Researcher from Leibniz Institute on Aging - Fritz Lipmann Institute, Germany;
Dr. Artur Silva, Researcher and Professor from LPDNA - UFPA, Brazil;
Dr. Rommel Ramos, Researcher and Professor from LPDNA - UFPA, Brazil;
Msc. Felipe Luiz Pereira, Information technology manager from AQUACEN - UFMG, Brazil;
Dr. Fernanda Alves Dorella, Researcher from AQUACEN - UFMG, Brazil;
Dr. Henrique Figueiredo, Researcher and Professor from AQUACEN - UFMG, Brazil;
Dr. Aristóteles Góes Neto, Researcher and Professor from LBMCF – UFMG, Brazil;
Dr. Mohammad Salaheldean, Researcher from Faculty of Veterinary Medicine - Cairo University, Egypt;
Dr. Salah Abdel Karim Selim, Researcher from Faculty of Veterinary Medicine - Cairo University, Egypt;
Dr. Vasco Azevedo, Researcher and Professor from LGCM - UFMG, 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), and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

I.3 Thesis structure

This manuscript is divided into Introduction and two chapters based in one research article and one ongoing project.

- a. The Introduction shows relevant characteristics of *C. pseudotuberculosis*, such as taxonomy, biovars, economic impact, and current findings of genomic studies;
- b. The first chapter is a review about positive selection in bacteria, submitted to PLOS Computational Biology, describing the current methodologies and their contribution to studies of bacterial biology;
- c. The second chapter presents a research article showing comparative genomics analysis, published in PLOS One, with focus on 11 strains isolated from buffalo hosts;
- d. The third paper is an analysis of genes under positive selection in the species, submitted to Nature Genetics, where 28 genes involved in adaptations of specific branches of *C. pseudotuberculosis* were identified;
- e. The Appendix presents other scientific productions, including research articles, book chapter and reviews, participation in national and international workshop, courses and poster presentations.

After bibliography, there is an "appendix" section, where one can find other works developed during the PhD, and the certification of courses, events, and work presentations.

II. Introduction

II.1 *Corynebacterium pseudotuberculosis*

Corynebacterium pseudotuberculosis is a Gram-positive bacterium and a facultative intracellular pathogen. It is non-sporulating, non-capsulated, non-motile and has fimbriae. The cells are small irregular rods of 0.5–0.6 × 1.0–3.0 µm, with club forms and metachromatic granules. Colonies on blood agar are yellowish-white, opaque, and convex with a matt surface. It is positive for glucose, fructose, galactose, mannose, and maltose. Cell walls contain arabinose, galactose, glucose, and mannose, meso-Diaminopimelic acid and mycolic acids. Biovar Ovis is nitrate negative, while biovar Equi is nitrate positive. Infections occur in sheep, goats, horses and other warm blooded animals, occasionally in humans (BERNARD; FUNKE, 2015). The species has a worldwide distribution (BAIRD; FONTAINE, 2007).

II.1.1 Taxonomy

The species is classified under the genus *Corynebacterium*, which belongs to a suprageneric group of Actinomycetes that also includes the genera *Mycobacterium*, *Nocardia* and *Rhodococcus*, known as the CMNR group. This group contains species of medical, veterinary and biotechnological importance, and has in common characteristics that include high DNA G+C content and a cell wall composed mainly of peptidoglycans, arabinogalactans and mycolic acids (DORELLA et al., 2006). The genus *Corynebacterium* is composed of pathogenic, opportunistic and nonpathogenic species as *C. diphtheria*, *C. jeikeium*, and *C. glutamicum*, respectively (CERDEÑO-TÁRRAGA et al., 2003; IKEDA; NAKAGAWA, 2003; SING et al., 2015; TAUCH et al., 2005) (Figure 1). Phylogenetic analyses using 16S rRNA and *rpoB* (β-subunit of RNA polymerase) markers and phylogenomic analysis showed that *C. pseudotuberculosis* to *C. ulcerans* and *C. diphtheriae* form a cluster, with *C. diphtheriae* as the cluster most external branch (KHAMIS; RAOULT; LA SCOLA, 2005; SOARES et al., 2013a; TAKAHASHI et al., 1997).

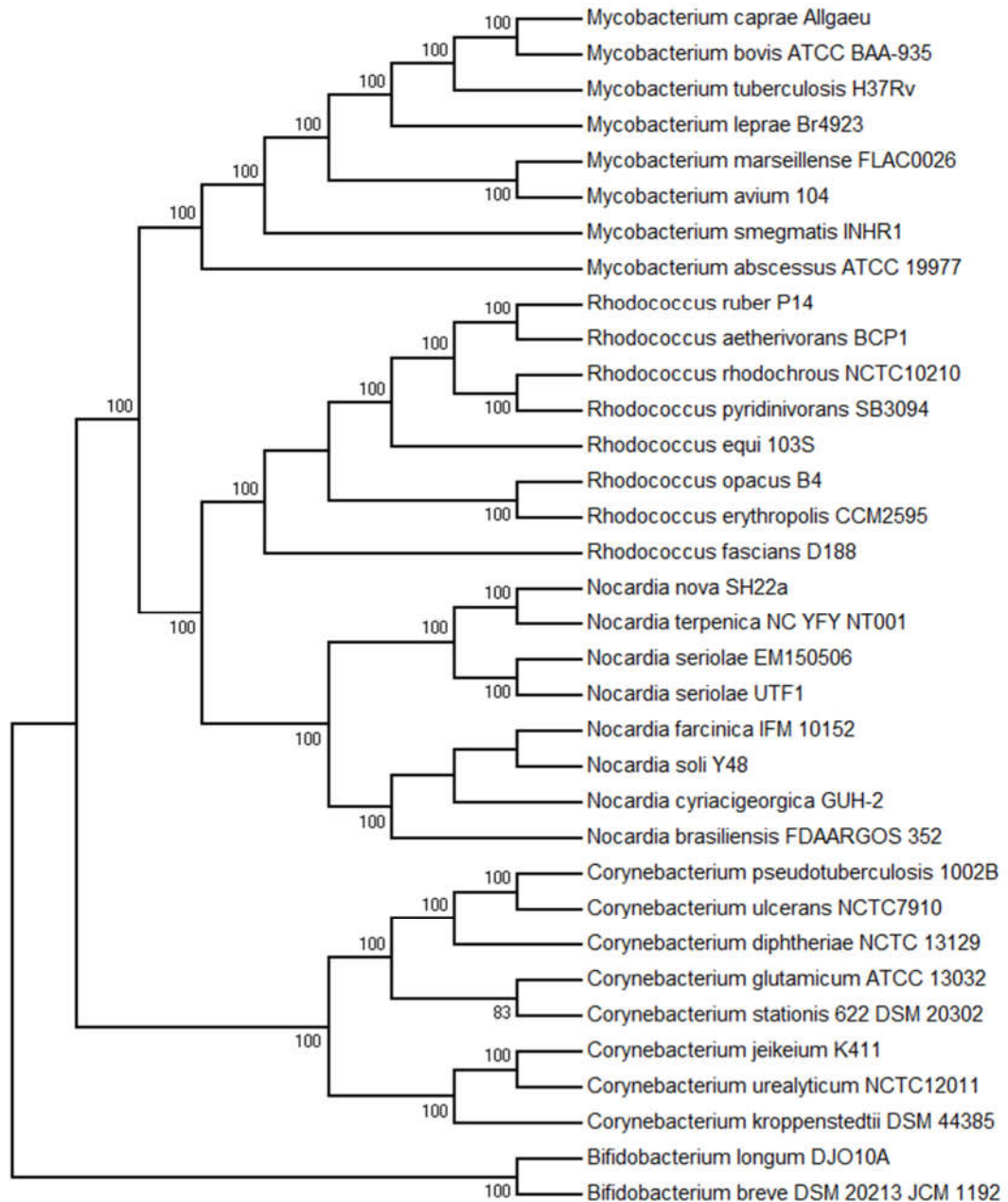


Figure 1. Phylogeny of CMNR group. The tree was built based on the core proteome with progressive refinement and the Maximum Likelihood method implemented in the pipeline PEPR.

II.1.2 Biovar Ovis and Equi

Two biotypes were identified initially by nitrate reductase activity and by the hosts they infected. Biotype 1 is nitrate reductase negative and infected mainly sheep and goats, while biotype 2, nitrate positive, was found in larger hosts, such as horses and buffalo (BIBERSTEIN; KNIGHT; JANG, 1971; SONGER et al., 1988). The groups were later renamed to serotype 1 and 2 (BARAKAT et al., 1984), and biovar Ovis and Equi (SONGER et al., 1988), respectively. This classification was supported by 16S-rRNA ribotyping (COSTA; SPIER; HIRSH, 1998; SUTHERLAND; HART; BULLER, 1996), amplified rDNA restriction analysis (ARDRA) (VANEECHOUTTE et al., 1995), and genomic analysis (ALMEIDA et al., 2017; RUIZ et al., 2011; SOARES et al., 2013b).

The two biovars also have different disease phenotypes that varies depending upon the host. Ovis strains causes Caseous Lymphadenitis (CLA) in goats and sheep, characterized mainly by abscess formation in the superficial and internal lymph nodes (WILLIAMSON, 2001), with visceral and external forms, affecting lungs, liver, kidney, and other organs (BAIRD; FONTAINE, 2007). The range of CLA reflects distribution of their hosts, being identified in Europe, Australia, North and South America, Africa and the Middle East (BAIRD; FONTAINE, 2007). The disease causes significant economic losses, with reduced wool, meat and milk yields, decreased reproductive efficiency and also in the condemnation of carcasses and skins in abattoirs (DORELLA et al., 2006; WILLIAMSON, 2001). Cows infected with Ovis strains have skin lesions in the form of single or multiple ulcerative granulomatous lesions on the forehead, neck, shoulder, tail base and hip, and a mastitic form (YERUHAM et al., 2004). In camels, the infection manifests as lymphadenitis with abscesses predominantly subcutaneous and muscular (HAWARI, 2008). Humans are occasionally infected by occupational exposure and consumption of raw goat and cow milk (PEEL et al., 1997). The described manifestations are chronic infections, localized suppurative granulomatous lymphadenitis affecting the axillary, inguinal or cervical lymph node, necrotizing lymphadenitis and pneumonia (BAIRD; FONTAINE, 2007; HEGGELUND et al., 2015; TROST et al., 2010).

Biovar Equi causes ulcerative granulomatous lesions in cattle (YERUHAM et al., 1996, 2004). Horses can have different disease patterns, as external abscesses; ulcerative lymphangitis of the limbs, also known as pigeon fever due to the swelling in the chest; and a visceral form that affects the internal organs (BARAÚNA et al., 2017; FOLEY et al., 2004; SPIER; AZEVEDO, 2016). In camel, it causes enlargement in ventral cervical superficial lymph nodes (TEJEDOR-JUNCO et al., 2008).

Oedematous skin disease (OSD) caused by *C. pseudotuberculosis* biovar Equi in buffaloes is manifested by a large cutaneous oedematous swellings in the dewlap, whole hind or forelimbs, and the belly. The swelling in the initial site of infection usually involves drainage of lymph nodes which can be enlarged to the size of a small watermelon. The second stage of

the disease is the oozing of serous fluid from cutaneous lesions, leading to secondary infections and larger abscesses. The third stage is an extensive dermal necrosis, respiratory manifestations, and hemoglobinuria, which can be fatal if not treated promptly (HUSSEIN, 2012; SELIM, 2001). OSD is a local disease of buffalo in governorates of Lower Egypt (Nile Delta) and surrounding Cairo, associated with particular climatic conditions of high temperatures and relative humidity (HAFEZ; HILALI, 1978). The disease is of considerable importance to the livestock industry in Egypt, as it causes economic loss due to reduced milk and meat production, reduced work efficiency of the animals, and expensive treatment that may extend to months (AHMED; EL-TAHAWY, 2012; SELIM, 2001). Besides cows and camels, *Ovis* cannot infect other Equi hosts, but experimental infection of a buffalo isolate could cause CLA in sheep (MOUSSA et al., 2016).

II.1.3 Virulence factors

During infection, *C. pseudotuberculosis* provokes both the innate immune response and the phagocytosis process. It can survive within the macrophages. After multiplying and lysing these cells, it migrates to the lymph, forming necrotic lesions (JONES; HUNT; KING, 2000). The infective process depends upon the presence of genes and structures described as niche and/or virulence factors, which the pathogen needs to survive and reproduce in the host (HILL, 2012; WEBB; KAHLER, 2008)(TAUCH; BURKOVSKI, 2015). Some of the virulence factors of the species have been described, such as pili, toxic cell wall lipids, and the exotoxins phospholipase D and diphtheria toxin.

Pili are bacterial adhesion structures that play an important role in the initiation of extracellular and intracellular invasion and proliferation (MANDLIK et al., 2008; YANAGAWA; HONDA, 1976). Two pili operons were described in *C. pseudotuberculosis* (YANAGAWA; HONDA, 1976), *spaA* (*spaA-spaB-spaC*) and *spaD* (*spaD-spaE-spaF*), where SpaA and SpaD are shaft pilins, SpaB and SpaE are base pilins, and SpaC and spaF are tip pilins (TROST et al., 2010).

The toxic cell wall lipids, specifically mycolic acids, cause dermonecrotic lesion, and permit intracellular survival and lethal injury in the macrophages (CARNE; WICKHAM; KATER, 1956; HARD, 1975). The exotoxin phospholipase D is the most well characterized virulence factor of *C. pseudotuberculosis*. It promotes the hydrolysis and degradation of sphingomyelin in endothelial cell membranes, contributes to the spread of bacteria to secondary sites within the host by increasing the vascular permeability, and plays a role in macrophage death (D'AFONSECA et al., 2008; MCKEAN; DAVIES; MOORE, 2007).

Iron is an essential nutrient for most bacteria, involved in process as DNA biosynthesis, respiration, and control of gene expression, and its access is limited inside the host (ANDREWS; ROBINSON; RODRÍGUEZ-QUIÑONES, 2003; HOOD; SKAAR, 2012). Two siderophores were described, one encoded by the *fagABC* operon and *fagD* (BILLINGTON et al., 2002), and the other encoded by the operon *ciuABCDE* (RAYNAL et al., 2018).

The diphtheria toxin is found in *C. pseudotuberculosis* in buffalo isolates (SELIM et al., 2015). This toxin leads to the inactivation of the host protein synthesis, and is acquired through a lysogenic conversion during an infection by a β corynephage (HOLMES, 2000). The toxin precursor domain B binds to the host cell surface receptor to mediate the internalization of domain A. Domain A causes inactivation of protein synthesis and the consequent cell apoptosis by NAD⁺-dependent ADP-ribosylation of a Elongation Factor 2 (EF-2) (HOLMES, 2000; MURPHY, 2011).

II.1.4 Difficulties in control measures

C. pseudotuberculosis can be transmitted by contact with superficial wounds (YERUHAM et al., 2004) and was shown to survive up to 55 days in fomites (AUGUSTINE; RENSHAW, 1986), and up to 8 months in soil (BAIRD; FONTAINE, 2007). The transmission can be as simple as skin contact between animals (YERUHAM et al., 2004), or can be the result of common procedures like shearing, castration and ear-tagging (WILLIAMSON, 2001). Transmission can also occur through blood sucking flies *Musca domestica* (BARBA et al., 2015; BRAVERMAN et al., 1999) and *Hippobosca* (GHONEIM et al., 2001). For humans, it is associated with occupational exposure of infected animals and consumption of raw goat and cow milk (PEEL et al., 1997).

For buffalo, vaccines against *C. pseudotuberculosis* were developed using formalin killed bacterin (SYAME; EL-HEWAIRY; SELIM, 2008), and phospholipase D inactivated by formalin (SYAME; EL-HEWAIRY; SELIM, 2008) or by a mutation of one amino acid in its active site (rPLD) (MOHAMED MOUSSA et al., 2014). The vaccines caused an immune response, but could not protect the host. A recent vaccine based on rPLD and formaline killed bacterin from a buffalo isolated strain didn't protect this host, but protected sheep. An experimental infection of non-vaccinated buffalos demonstrated that this host is immune to biovar *Ovis* strains. Therefore, it was proposed that OSD is caused by toxin (s) other than PLD (MOUSSA et al., 2016). In previous vaccines, the DT was not used as antigen in wild, formaldehyde inactivated, nor recombinant form (MOHAMED MOUSSA et al., 2014; MOUSSA et al., 2016; SYAME; EL-HEWAIRY; SELIM, 2008). As all reported buffalo isolated strains have the DT, a vaccine based on an inactivated form of this toxin was proposed (VIANA et al., 2017).

Low detection rates (YERUHAM et al., 2004), inefficacy of antibiotic therapies due to intra-macrophagic lifestyle and abscess formation (COLLETT; BATH; CAMERON, 1994) and variability in the efficiency of the vaccines between host species (DORELLA et al., 2009), also contribute to persistence and the spread of disease, making this pathogen very difficult to control. These difficulties require research for new diagnosis, treatments and vaccines against the pathogen, which can be acquired by genomics studies.

II.2 Comparative genomics of *C. pseudotuberculosis*

II.2.1. Genome variation and adaptation

Bacterial genome sequencing has significantly increased the knowledge of genome structure, function, variation and diversity, phylogeny and their relation to lifestyle, with many practical applications, such as genome-scale metabolic modeling, biosurveillance, bioforensics and epidemiology (LAND et al., 2015; LOMAN; PALLEEN, 2015). The analysis of bacterial genome architecture has been shown a high correlation between genome size and gene number, and that variation in size mostly due to the acquisition (insertion) and loss (deletion) of functional accessory genes, but also due to duplications, inversions, and translocations (BOBAY; OCHMAN, 2017). There is a mutation bias toward deletion of superfluous sequences after niche change, most intense in clonal species (BOLOTIN; HERSHBERG, 2016).

Neutral markers and multilocus sequence typing (MLST) data have been used to differentiate species of bacteria, but these approaches have the problem of deciding the cut-off level that separates phylogenetic clusters as separated species (VOS, 2011). To solve this problem, different bacterial speciation models and tests were developed based on adaptation to ecological niches. Mutations in gene sequence or horizontal gene transfer can provide adaptations that allow the mutant to explore a new resource or the same resource in a different way. This niche change leads to clonal expansion of the founder mutant and creates a new ecotype. As the new ecotype diverges genetically from the original population, the genomic cohesion (recombination) eventually disappears and the new ecotype becomes a new species (HALL; BROCKHURST; HARRISON, 2017; LASSALLE; MULLER; NESME, 2015). In this approach, a new species is identified when a monophyletic group diverges from another group by adaptive divergence, with significance for a statistical test based on the difference in the proportion of non-synonymous to synonymous mutations between the groups (VOS, 2011).

II.2.1.1 Genome synteny

During evolution genomes undergo structural variations that can span from a few kilobases to millions of base pairs that can be classified as deletions, duplications, insertions, inversions and translocations. The variations are caused by horizontal gene transfer, homologous recombination, imprecise DNA repair, transposable elements and horizontal gene transfer. These processes can contribute to evolution by disrupting or creating new chimeric genes and acquisition of new genes that can have divergent phenotypic consequences such as higher fitness, reduced fitness or lethality (PERIWAL; SCARIA, 2015). Genome synteny is the conservation of genomic blocks order within genomes. Using genome sequencing and assembly data, it can be estimated by mapping of paired reads onto a reference genome or by alignment of entire chromosomes (DARLING, 2004; DARLING; MAU; PERNA, 2010; PERIWAL; SCARIA, 2015).

The first synteny analysis in *C. pseudotuberculosis* showed a high synteny of 97% between the biovar Ovis strains 1002 and C231, isolated from goat and sheep, respectively (RUIZ et al., 2011). However, the use of optical map in genome assembly identified a large inversion in the strain 1002 genome, containing more than half of the chromosome (MARIANO et al., 2016). High synteny was also observed when compared to *C. diphtheriae*, as expected from species of the genus (NAKAMURA et al., 2003; TAUCH et al., 2005).

II.2.1.2 Pangenomics

The identification of homologous genes found within different genomes is useful for genome annotation, studies on gene evolution, comparative genomics, and for understanding taxonomically restricted sequences (HAGGERTY et al., 2014). Among the different homology groups, orthologs evolve from vertical descent from a single ancestral gene, paralogs are the result of duplication events within the same genome, and xenologs are orthologs from a distant lineage (KOONIN, 2005). Different methods are used to identify gene homology (ALTENHOFF; DESSIMOZ, 2009; CHEN et al., 2007; PEARSON, 2013), and these are usually based on BLASTp bidirectional best hits (BBH) and Markov Clustering Algorithm (MCL) (WOLF; KOONIN, 2012).

The pangenome is the complete gene inventory of a species, while the core genome is the orthologous genes that are shared by all genomes. The accessory genome is that subset of genes present in more than one, but not all genomes. Singletons are genes that are present in a single genome (MEDINI et al., 2005). A pangenome is defined as open when the addition of more genome to the analysis will contribute to the identification of new genes in the species repertory (MEDINI et al., 2005; TETTELIN et al., 2008). For microbiology, the study of the

pangenomes have applications in medicine and biotechnology. It is possible to detect genome plasticity events as lateral transfer, genes involved in metabolism, regulation, pathogenicity, drug resistance, and vaccine targets (MUZZI; MASIGNANI; RAPPUOLI, 2007; ROULI et al., 2015; TETTELIN et al., 2008).

The first sequenced genomes of *C. pseudotuberculosis* were the biovar Ovis strains 1002 from goat and C231 from sheep. The comparative analysis of these genomes showed that *C. pseudotuberculosis* lost numerous genes and has one of the smallest genomes within the genus, with approximately 2.3 Mb and 2,200 genes, representing a loss of approximately 1,220 genes. Those two specific strains were very similar, with at least 95% of amino acid sequence similarity, the same mean G+C content (52.19%), gene length, operon composition and gene density (RUIZ et al., 2011). The classification of the genes by function showed that the most abundantly represented categories are linked to metabolic processes in the two strains (cellular metabolic, biosynthetic, primary and macromolecule processes). In addition, a low proportion of the proteins were linked to metabolism of secondary metabolites, which confirms that the species is a facultative intracellular pathogen.

A analysis performed on 15 genomes from both biovars identified an open pangenome with 2,782 total genes, with a core genome that included 54% of all the genes (1,504 genes) (SOARES et al., 2013a). In comparison to Equi biovar, Ovis showed more clonal behavior by having a larger core genome (1,818 against 1,599), and a smaller pangenome (2,403 against 2,521). In addition, a statistical estimation predicted that the Ovis core and pangenome will stabilize closer than Equi, with the addition of more genomes, what means that Equi has a higher genome plasticity. A large number of hypothetical proteins were found when the core genomes of Ovis and Equi were compared and it was suggested to be related to adaptations to the host ranges of each biovar. Another interesting finding was the first identification of the diphtheria toxin gene in strain 31 from buffalo, which could be essential for host infection (SOARES et al., 2013a). The pangenome of the species has been characterized as open, which justifies the sequencing and analysis of more genomes.

A study of 12 strains isolated from horses (Equi biovar) had the goal to determine whether there was a relationship between the genetic content and different diseases manifestations: ulcerative lymphangitis and external or internal abscesses (BARAÚNA et al., 2017). The pangenome of the strains was estimated as 3,183 and the core genome as 1,355 genes (42,56%). Among genomic islands, the pathogenicity islands were largely conserved, while resistance islands were more similar within strains that caused internal abscesses. No genotypic differences were observed between the strains that caused the different diseases manifestations.

An analysis focused on Ovis and Equi strains isolated in Mexico showed that strains in the Ovis biovar had an exclusive Type III restriction-modification system, while Equi had an exclusive CRISPR-Cas system. The phylogeny suggests that Equi strains were clustered by host, while the genetic structure of Ovis is influenced by hosts transport within farms than host species (PARISE et al., 2018, accepted).

II.2.1.3 Pathogenicity islands

Genomic islands are elements acquired by horizontal gene transfer and can be detected by analyzing deviations in G+C content and codon usage, identification of transposases and tRNA genes, and class specific factors, such as virulence, metabolism and symbiosis genes (SOARES et al., 2016). Pathogenicity islands (PAI) include virulence genes that mediate mechanisms of adhesion, invasion, colonization, proliferation into the host and evasion of the immune system (KARAOLIS et al., 1998).

Seven pathogenicity islands (PiCp1-7) were first described in *C. pseudotuberculosis* Ovis strains 1002 and C231 (RUIZ et al., 2011) and identified as harboring classical virulence factors, including genes for fimbrial subunits, adhesion factors, iron uptake and secreted toxins. This was followed by the identification of four new PAIs (Cp8-11) after comparison with Equi strains 258 and 316 (SOARES et al., 2013b), followed by four more (Cp12-16) (SOARES et al., 2013a). The analysis of the variability in size, gene content and deletion patterns of the 16 PAIs explained most of the differences between biovars, with Ovis having their PAIs more tightly conserved across the biovar (SOARES et al., 2013a).

II.2.5 Phylogenomics

Phylogenomics is the application of genome-scale data to infer evolutionary relations, improving previous studies done with one or a few genes (CHAN; RAGAN, 2013). Phylogenomics of the genus *Corynebacterium* using nucleotide sequences of the accessory genome showed a cluster of pathogenic and another of non-pathogenic species, with the pathogens *C. pseudotuberculosis*, *C. ulcerans* and *C. diphtheriae* all branching together (SOARES et al., 2013a). A closer examination of the relationships within the *C. pseudotuberculosis* strains showed that a cluster of Equi strains with 95 to 100% similarity, and a cluster of biovar Ovis genomes with 99 to 100% of similarity, suggesting a more clonal behavior. *C. pseudotuberculosis* is more clonal than *C. ulcerans* and *C. diphtheriae*, with 90 to 94% and 82 to 100% of similarity, respectively (SOARES et al., 2013a).

II.3 Pathogens Resource Integration Center

In this work, we will use the tools available in the Pathogens Resource Integration Center (PATRIC). PATRIC is a web based information system that provides integrated genome-scale data, metadata, and analysis tools for all publically available bacterial genomes (GILLESPIE et al., 2011; WATTAM et al., 2014a). It is one of the Bioinformatics Resource Centers (BRCs) funded by the National Institute of Allergy and Infectious Diseases (NIAID) of the US National Institutes of Health (NIH), originally for A, B and C priority pathogens (SNYDER et al., 2007) but now contains all bacterial and archaeal genomes (WATTAM et al., 2014a). PATRIC integrates services of genome assembly and annotation, proteome comparison, differential expression and RNA-Seq analysis, reconstruction and comparison of metabolic models, and sequence variation analysis (<http://www.patricbrc.org>).

The genome annotation is performed by the Rapid Annotation using Subsystems Technology tool kit (RASTtk) pipeline (BRETTIN et al., 2015). Its functional annotation is based on subsystems, which are sets of genes with specific functional roles that work together to implement a specific biological process or structural complex. These subsystems are curated by expert researchers (OVERBEEK, 2005; OVERBEEK et al., 2014). The Specialty Genes Search tool allows the search for antibiotic resistance, human homolog genes, and virulence factors. For visualization, it provides circular maps and genome browsing.

For proteome analysis, the comparisons can be performed by bidirectional BLASTp best hits against a reference (Proteome Comparison tool), or by the Protein Family Sorter, where the proteins distribution across the genomes can be analyzed interactively and visualized by a heatmap. The protein families are classified and clustered by isofunctional families (FIGfams) (MEYER; OVERBEEK; RODRIGUEZ, 2009), or orthologs in genus level (PLFams) or cross-genus level (PGFams) (DAVIS et al., 2016). The Comparative Pathway Tool allows comparison of metabolic pathways across closely related or diverse groups of genomes and visualization using interactive KEGG maps and heatmap viewer (WATTAM et al., 2014b).

III. Goals

III.1 Main goal

The main goal of this thesis is to perform comparative genome analysis between *Corynebacterium pseudotuberculosis* genomes isolated from buffalo hosts, and to identify genes under positive selection in this species.

III.2 Specific goals

The specific goals of this thesis are:

- to sequence, assembly, and annotate 10 *C. pseudotuberculosis* genomes isolated from buffalo hosts;
- to perform phylogenomic analysis to identify the relationship between the buffalo isolated strains;
- to estimate the genomes plasticity using synteny analysis, pangenome analysis, and prediction of pathogenicity islands and prophages;
- to compare buffalo isolated strains with other strains of the same species, by pangenome, phylogenomics and functional analyses;
- to identify genes under positive selection in *C. pseudotuberculosis* and their importance in species biology.

IV. Articles

IV.1 Chapter I. Comparative genomic analysis between *Corynebacterium pseudotuberculosis* strains isolated from buffalo

Marcus Vinicius Canário Viana, Henrique Figueiredo, Rommel Ramos, Luis Carlos Guimarães, Felipe Luiz Pereira, Fernanda Alves Dorella, Salah Abdel Karim Selim, Mohammad Salaheldean, Artur Silva, Alice Rebecca Wattam, Vasco Azevedo. **PLoS One, v. 12, p. e0176347, 2017.**

One of the main goals of this thesis was to perform a comparative genome analysis of *Corynebacterium pseudotuberculosis* genomes isolated from buffalo hosts. We wanted to better understand the pathogenic mechanisms of Oedematous Skin Disease and to provide information that could be used for the development of control methods. This specific chapter is a comparative genomic analysis of 44 different *Corynebacterium pseudotuberculosis* strains, 11 from buffalo, making this the first to include different isolates from this host. The results identified an overall genomic synteny, a core genome of 94.75%, a prophage containing the diptheriae toxin gene, and conservation of pilus genes cluster among buffalo isolates. In addition, phylogenomics and sequence analysis showed that the strains tended to cluster phylogenetically by host, and that the cow isolate from the Equi biovar showed similarities with the strains from the Ovis biovar. This study could play a role in directing future studies on host adaptation and provide a framework for a broader examination of pathogenicity across the *Corynebacterium* genus.

RESEARCH ARTICLE

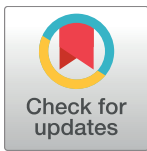
Comparative genomic analysis between *Corynebacterium pseudotuberculosis* strains isolated from buffalo

Marcus Vinicius Canário Viana^{1,2}, Henrique Figueiredo³, Rommel Ramos⁴, Luis Carlos Guimarães⁴, Felipe Luiz Pereira³, Fernanda Alves Dorella³, Salah Abdel Karim Selim⁵, Mohammad Salehdean⁵, Artur Silva⁴, Alice R. Wattam²✉, Vasco Azevedo¹✉*

1 Department of General Biology, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, **2** Biocomplexity Institute of Virginia Tech, Virginia Tech, Blacksburg, Virginia, United States of America, **3** AQUACEN, National Reference Laboratory for Aquatic Animal Diseases, Ministry of Fisheries and Aquaculture, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, **4** Center of Genomic and System Biology, Federal University of Pará, Belém, Pará, Brazil, **5** Department of Microbiology, Faculty of Veterinary Medicine, Cairo University, Giza, Egypt

✉ These authors contributed equally to this work.

* vasco@icb.ufmg.br



OPEN ACCESS

Citation: Viana MVC, Figueiredo H, Ramos R, Guimarães LC, Pereira FL, Dorella FA, et al. (2017) Comparative genomic analysis between *Corynebacterium pseudotuberculosis* strains isolated from buffalo. PLoS ONE 12(4): e0176347. <https://doi.org/10.1371/journal.pone.0176347>

Editor: Baochuan Lin, Defense Threat Reduction Agency, UNITED STATES

Received: November 21, 2016

Accepted: April 10, 2017

Published: April 26, 2017

Copyright: This is an open access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the [Creative Commons CC0](https://creativecommons.org/licenses/by/4.0/) public domain dedication.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Ministério da Pesca e Agricultura. A.R. Wattam was supported in part by federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department

Abstract

Corynebacterium pseudotuberculosis is a Gram-positive, pleomorphic, facultative intracellular pathogen that causes Oedematous Skin Disease (OSD) in buffalo. To better understand the pathogenic mechanisms of OSD, we performed a comparative genomic analysis of 11 strains of *C. pseudotuberculosis* isolated from different buffalo found to be infected in Egypt during an outbreak that occurred in 2008. Sixteen previously described pathogenicity islands (PiCp) were present in all of the new buffalo strains, but one of them, PiCp12, had an insertion that contained both a corynephage and a diphtheria toxin gene, both of which may play a role in the adaptation of *C. pseudotuberculosis* to this new host. Synteny analysis showed variations in the site of insertion of the corynephage during the same outbreak. A gene functional comparison showed the presence of a nitrate reductase operon that included genes involved in molybdenum cofactor biosynthesis, which is necessary for a positive nitrate reductase phenotype and is a possible adaptation for intracellular survival. Genomes from the buffalo strains also had fusions in minor pilin genes in the *spaA* and *spaD* gene cluster (*spaCX* and *spaYEF*), which could suggest either an adaptation to this particular host, or mutation events in the immediate ancestor before this particular epidemic. A phylogenomic analysis confirmed a clear separation between the Ovis and Equi biovars, but also showed what appears to be a clustering by host species within the Equi strains.

Introduction

Corynebacterium, *Mycobacterium*, *Nocardia*, and *Rhodococcus*, collectively known as the CMNR group, are all members of the order Corynebacteriales. This group contains species of medical, veterinary and biotechnological importance with shared common characteristics that

of Health and Human Services, under contract no. HHSN272201400027C.

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

include a high GC content and a cell wall composed mainly of peptidoglycans, arabinogalactans and mycolic acids [1]. *Corynebacterium* species are Gram-positive, rod-shaped, non-motile and non-spore forming bacteria [2], and include both pathogens like *C. diphtheria*, *C. jeikeium* and *C. pseudotuberculosis*, and non-pathogenic strains like *C. glutamicum* [3]. *Corynebacterium pseudotuberculosis* are facultative intracellular pathogens that are divided into two distinct biovars, each of which can infect a range of mammalian hosts. Nitrate reductase activity seems to correlate with the type of host that they infect, and thus the biovar that each strain belongs to. Isolates belonging to the Ovis biovar infect sheep and goats, and they are nitrate reductase negative. Members of the Equi biovar infect larger hosts like the horses, cattle, and buffalo [4], and they have positive nitrate reductase activity [5,6]. The separation into two biovars is also supported by molecular markers [6,7] and by distinct differences in gene content that has been observed across genomes [8].

C. pseudotuberculosis causes different diseases and symptoms in these various hosts. Infection in goats and sheep results in a manifestation called Caseous Lymphadenitis [9]. Infected cattle have ulcerative granulomatous lesions and mastitis [10,11], while horses are diagnosed with ulcerative lymphangitis, or with pigeon fever, so named due to a swelling in the chest of infected animals [12]. Oedematous Skin Disease (OSD) is the manifestation seen in buffalo. OSD is characterized by an extensive cutaneous, oedematous swelling in the dewlap, whole hind or forelimbs, and belly [13]. The disease causes economic loss due to reduced milk and meat production as well as reduced work efficiency of the animals. Moreover, the treatment is expensive and can last for months [13,14]. OSD is endemic in regions of lower Egypt and the areas surrounding Cairo, with sporadic cases throughout the year and periodic epidemics seen in the summer [13]. Unlike some of the other hosts that are infected by *C. pseudotuberculosis*, there is no documented transmission directly between infected buffalo. Instead, it has been suggested that insects are involved with transmitting the disease between these animals. Specifically, biting flies of the genus *Hippobosca* have been associated with transmission, the suggestion being that they inject the bacteria intradermally with each bite [15]. Further proof of involvement is the close correlation between the breeding season of these blood-sucking flies with outbreaks of OSD in buffalo [13].

All strains that have been isolated in Egypt from infected buffalo belonged to the Equi biovar, and all those sequenced genomes contain a gene that produces the diphtheria toxin [16]. This toxin is found in other *Corynebacterium* species, but within *C. pseudotuberculosis* it has only been found in isolates from buffalo to date. It is unclear if the presence of this gene indicates an adaptation that allows it to specifically infect buffalo, or if it is present due to a horizontal transfer event that occurred recently and spread across the specific geographic region where the animals were found [8].

To fully characterize these isolates and determine additional genomic features that not only distinguish them, but might explain the pathogenic mechanisms of OSD, a comparative analysis of 11 strains of *C. pseudotuberculosis* isolated from different buffalo in Egypt was conducted. These genomes were compared to other *C. pseudotuberculosis* strains from both Ovis and Equi biovars. Here we describe those differences, explore the species-wide pangenome, and define the differences related to not only the biovars, but also to the hosts they have been isolated from.

Materials and methods

Strains isolation

Each of the 11 *C. pseudotuberculosis* strains were isolated from individual water buffalo (*Bubalus bubalis*) in Egypt during an outbreak of OSD that occurred during the summer of 2008

Table 1. Strains of *Corynebacterium pseudotuberculosis* isolated from buffalo diagnosed with Oedematous Skin Disease in Egypt during the summer of 2008.

Strain	Genome size (bp)	Mean coverage depth	De novo assembler	CDS	tRNA	rRNA	Accession number
31 [19]	2,402,956	550x	MIRA 4.0.2; SPADES 3.1.1	2248	47	12	CP003421.3
32	2,403,533	56.61x	Newbler 2.9	2272	52	12	CP015183
33	2,403,550	104.11x	Newbler 2.9	2281	52	12	CP015184
34	2,403,454	58.94x	MIRA 3.9.18	2286	49	12	CP015192
35	2,403,502	112.68x	Newbler 2.9	2276	52	12	CP015185
36	2,403,412	172.02x	SPAdes 3.6.0	2279	49	12	CP015186
38	2,403,515	160.27x	SPAdes 3.6.0	2281	49	12	CP015187
39	2,403,579	241.79x	SPAdes 3.6.0	2281	49	12	CP015188
43	2,365,075	218.07x	Newbler 2.9	2218	51	12	CP015189
46	2,366,565	249.57x	SPAdes 3.6.0	2218	48	12	CP015190
48	2,403,301	306.54x	SPAdes 3.6.0	2281	49	12	CP015191

<https://doi.org/10.1371/journal.pone.0176347.t001>

(Table 1). The infected animals were all located in separate farms either in the Menofia (located 89 km north of Cairo) or El-Fayoum (50 km west Cairo) regions. The samples were collected following permission from the owners of the individual animals. No additional permission was required as Egyptian buffalo are not protected, and are considered to be domesticated livestock. Samples were collected from internal lesions that were located within the brisket, fore or hind limbs of these animals. The samples were plated on selective media (Brain Heart Agar with Fosfomycin). A single colony from each of the samples was selected for species identification, and identification of *C. pseudotuberculosis* was determined by API Coryne (BioMerieux, France) [17] according to the manufacturer’s instructions. An aliquot of each sample was lyophilized and sent to the Laboratory of Cellular and Molecular Genetics in Brazil where they were grown again, and a second confirmation showed them to be *C. pseudotuberculosis*. Each strain was demonstrated to produce diphtheria toxin by both API Coryne System and multiplex PCR tests [18]. Colonies from these pure cultures were plated on agar, and one colony from each of these was randomly selected for sequencing.

Genome sequencing and assembly

The *in silico* analysis workflow is presented in Fig 1. Sequencing, assembly and annotation of individual genomes was conducted at one of three laboratories. Two of these (Laboratory of Cellular and Molecular Genetics and the National Reference Laboratory for Aquatic Animal Diseases) are part of the Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil. The third (Center of Genomic and System Biology) is located at the Federal University of Pará in Belém, Pará, Brazil. All genomes were sequenced by Ion Personal Genome Machine (PGM) with the Ion 318™ chip, using Ion PGM Template OT2 400 Kit and Ion PGM Hi-Q Sequencing Kit. The quality of the reads for each genome was examined using FastQC 0.10.1 (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>).

Simba 1.2.1 [20] was used to assemble the genomes, unite them into a single scaffold, and establish the start of the chromosome. Simba includes three different assembly packages: SPAdes 3.6.0 [21]; MIRA 3.9.18 (<http://sourceforge.net/projects/mira-assembler/>); and Newbler 2.9 (http://swes.cals.arizona.edu/maier_lab/kartchner/documentation/index.php/home/docs/newbler). A “best” assembly was selected for each isolate based on having the highest N50, the smallest number of contigs, and higher values of maximum and minimum contigs size [22]. The contigs of this “best” assembly were then united into a single scaffold using CONTIGuator [23], with *C. pseudotuberculosis* 31 (CP003421.2) [19] used as a reference. The

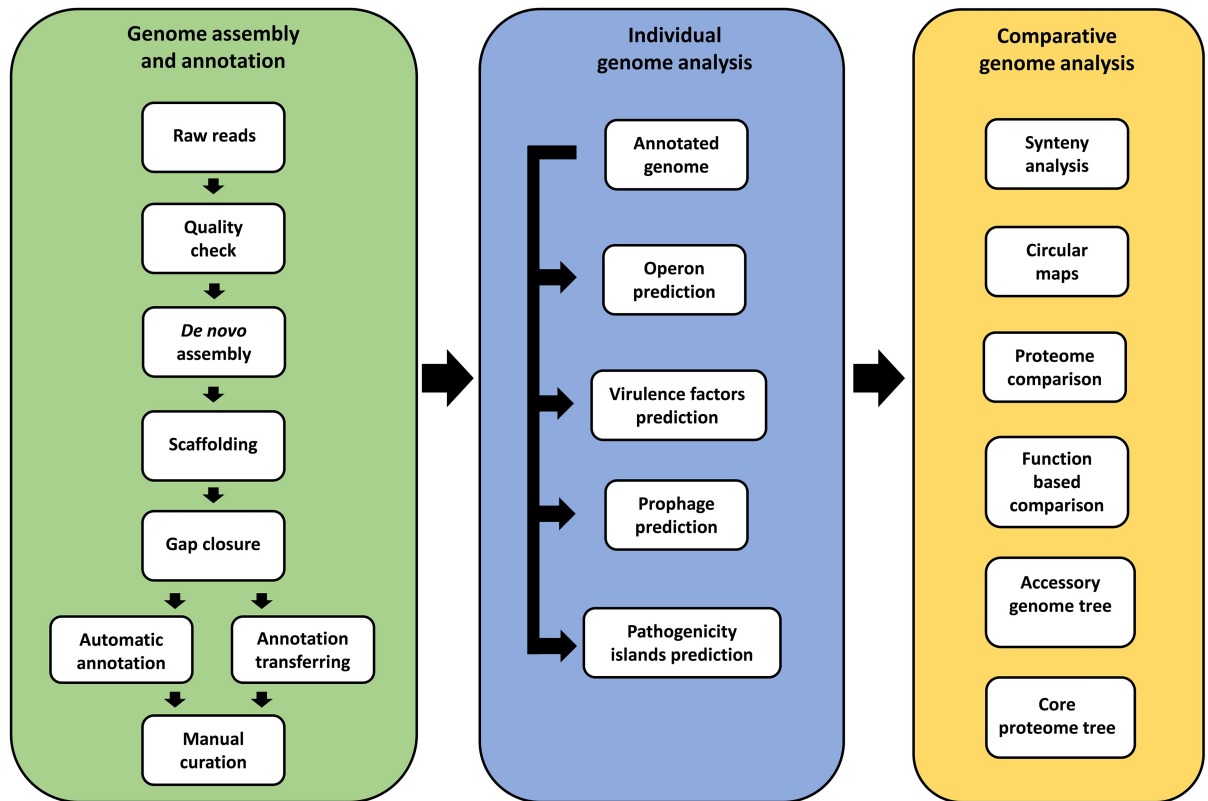


Fig 1. Workflow of the genome assembly and annotation of 11 *Corynebacterium pseudotuberculosis* strains isolated from buffalo, and comparative genomics analysis with other strains.

<https://doi.org/10.1371/journal.pone.0176347.g001>

overlap between contigs was verified by BLASTN, and an in-house script was used to establish the beginning of the chromosome at the *dnaA* gene. The gaps were closed using CLC Genomics Workbench 6.5 (<http://www.clcbio.com/products/clc-main-workbench/>) in the following way. First, the reads were mapped to the reference genome to establish a consensus sequence. Next, the sequences of the flanking regions around each gap were identified in the reference genome. Then, the consensus sequence that had been generated from the reads was used to fill the gap when it was inserted between the flanking sequences.

Genome annotation

All genomes, including strain 31 (CP003421.2) [24], were annotated consistently using the RASTtk (Rapid Annotation Using Subsystem Technology) [25] annotation service in PATRIC (Pathosystems Resource Integration Center) [26]. To verify the authenticity of the identified frameshifts, a specific curation process that involved manual examination of potential frameshifts was performed. Insertion/deletion (indel) errors are associated to a certain extent with any sequencing platform, and those associated with homopolymers in the Ion Torrent sequencing platform are well defined [27]. Our curation process involved the following steps. First, a manually curated annotation of the reference genome was transferred to each of the other 10 genomes using an in-house script to identify the pseudogenes. Frameshifts were examined in Artemis v16.0.0 [28]. To identify the frameshift location and see if the mutation was conserved across other genomes, the genes with potential frameshifts were compared to other complete genes by BLASTN against the NR database at NCBI (National Center for

Biotechnology Information). Second, we checked for indels among the reads at the location of the frameshift by mapping them against the assembled genome using the CLC Genomics Workbench v6.5. When an examination of the reads showed that an identified frameshift was a sequencing artifact, the sequence was adjusted. The new sequence, translated into a protein sequence, was verified by BLASTP against the Uniprot database [29]. The genome with the corrected sequences, which included the fixed frameshifts, was then re-annotated using RASTtk.

Genome plasticity

Pathogenicity islands (PAIs) were predicted by GIPSY [30]. All 11 genomes were individually compared to the *C. glutamicum* ATCC1302 genome (NC_006958.1), a non-pathogenic strain. DoubleACT v2 (http://www.hpa-bioinfotools.org.uk/pise/double_act.html) was used to align the pairs of genomes, and ACT (Artemis Comparison Tool) v13.0.0 [28] was used to visualize the alignments and check if the sequence of a predicted island in one genome was present but not annotated in the second genome. Comparison maps of the genomes were generated using BRIG (Blast Ring Image Generator) v0.95 [31].

PHAST, which predicts and annotates prophages from raw DNA sequence data or GenBank files, was used to predict prophages in the genomes by BLASTing against the NCBI and the prophage databases [32]. Syntenic and unique regions were both verified in the genomes by the progressiveMauve algorithm [33].

Phylogenomics

Gegenees v2.2.1 [34] was used to create a matrix of similarity across the genomes. This matrix was exported as a nexus file and used to generate a phylogenomic tree using Splitstree v4.14.2 [35] with the UPGMA method.

The PEPR (Phylogenomic Estimation with Progressive Refinement) program (<https://github.com/enordber/pepr.git>) was used to generate a phylogenomic tree with the publicly available and complete *C. pseudotuberculosis* genomes, and also included the new buffalo isolates. This is an automated system for generation of phylogenomic trees from amino acid sequences by a maximum likelihood algorithm. It identifies the common orthologs among all genomes, filters out genes that have been transferred horizontally, aligns and concatenates sequences, and generates a tree. Subtrees with low bootstrap values are refined by subsequent steps of addition genes that are shared across smaller clusters. This method is appropriated for uneven sampling in databases, where taxons of interest have a denser sampling. The resulting Newick tree file was visualized using Mega 6 [36].

Pangenomics

We compared the 11 buffalo isolates with the 33 other genomes publicly available in NCBI (Table 2). The Protein Family Sorter [26] was used to examine the pan-, core- and accessory genomes using the *Corynebacterium* genus-specific protein families (PLfams) that are available in PATRIC [37]. In addition, PATRIC's Proteome Comparison tool (<https://www.patricbrc.org/app/SeqComparison>), an adaptation of RAST's Sequence Based Comparison Tool [38], was used to generate a matrix of the bidirectional BLASTP hits across the proteins annotated in the genomes, with each genome used in a separate comparison as the reference genome to which the other 10 were compared. RAST's Function based Comparison tool (<http://rast.nmpdr.org/seedviewer.cgi>) was used to assess similarities and differences in the presence of functional roles among the genomes. FgenesB (<http://www.softberry.com/berry.phtml?topic=fgenesb>) was used to identify operons in regions of interest.

Table 2. List of the 33 strains of *Corynebacterium pseudotuberculosis* used in this study that were isolated from hosts other than buffalo.

Strain	Biovar	Host	Country	Accession number
Cp162	Equi	Camel	UK	CP003652.1
262	Equi	Cow	Belgium	CP012022.1
258	Equi	Horse	Belgium	CP003540.2
E19	Equi	Horse	Chile	CP012136.1
CIP52.97	Equi	Horse	Kenya	CP003061.1
1/06-A	Equi	Horse	USA	CP003082.1
316	Equi	Horse	USA	CP003077.1
MB11	Equi	Horse	USA	CP013260.1
MB14	Equi	Horse	USA	CP013261.1
MB30	Equi	Horse	USA	CP013262.1
MB66	Equi	Horse	USA	CP013263.1
29156	Ovis	Cow	Israel	CP010795.1
I19	Ovis	Cow	Israel	CP002251.1
1002B	Ovis	Goat	Brazil	CP012837.1
VD57	Ovis	Goat	Brazil	CP009927.1
CS_10	Ovis	Goat	Norway	CP008923.1
Ft_2193/67	Ovis	Goat	Norway	CP008924.1
PO222/4-1	Ovis	Goat	Portugal	CP013698.1
PO269-5	Ovis	Goat	Portugal	CP012695.1
226	Ovis	Goat	USA	CP010889.1
FRC41	Ovis	Human	France	CP002097.1
48252	Ovis	Human	Norway	CP008922.1
267	Ovis	Llama	USA	CP003407.1
N1	Ovis	Sheep	Equatorial Guinea	CP013146.1
PAT10	Ovis	Sheep	Argentina	CP002924.1
C231	Ovis	Sheep	Australia	CP001829.1
42/02-A	Ovis	Sheep	Australia	CP003062.1
12C	Ovis	Sheep	Brazil	CP011474.1
PA01	Ovis	Sheep	Brazil	CP013327.1
E56	Ovis	Sheep	Egypt	CP013699.1
MEX25	Ovis	Sheep	Mexico	CP013697.1
3/99-5	Ovis	Sheep	Scotland	CP003152.1
P54B96	Ovis	Wildebeest	South Africa	CP003385.1

<https://doi.org/10.1371/journal.pone.0176347.t002>

Specialty genes search

PATRIC’s Specialty Genes Search tool [39] was used to identify some of the virulence genes that were found in the 11 genomes. This tool BLASTs all genes in the annotated genome against a virulence factor database that includes VFDB [40] and manually curated virulence genes [39].

Results and discussion

Genome sequencing, assembly and annotation

All the 10 new genomes were closed, with all reads united into a single chromosome. For each genome, the theoretical mean coverage depth varied between 56.72 and 307.02x, while mean coverage depth genome varied between 56.61 and 306.54x. The 11 strains (including strain 31) have a genome size of approximately 2.4 Mb, a GC content that ranges between 52.07 and

52.1%, 2218 to 2,286 coding sequences (CDSs), 44 to 52 tRNA and 12 rRNA genes (Table 1). These values are within the ranges seen among other strains of *C. pseudotuberculosis*, which is considered to be more clonal compared to its closest relatives *C. diphtheriae* and *C. ulcerans* [8].

Genome plasticity

Genome plasticity analysis can identify areas of the genome, known as PAIs, which have been acquired by horizontal transfer and frequently include virulence genes [30]. Sixteen PAIs (PiCp1-16) were previously described in *C. pseudotuberculosis* and are shared across the biovars [8,41–43]. The variability in the size, gene content and deletion patterns in these islands explain most of the genomic differences seen between the two biovars. The content of sixteen PAIs are highly conserved in all the genomes belonging to Ovis, while biovar Equi has a greater variability, specifically with regard to deletions within the pilus genes [8].

These same sixteen PAIs were found in all the genomes isolated from buffalo (Fig 2). Interestingly, we discovered three regions that were missing in the second assembly of strain 31 (CP003421.2, re-sequenced by Ion PGM) when it was compared to other buffalo strains, but found that these regions were present in the first assembly of the same genome (CP003421.1, sequenced by Solid v3) (S1 Fig, S1 File, S1 Table). To fix the second assembly, we mapped the reads of the Ion PGM to the first assembly to verify whether those regions were represented within the reads, and to generate a consensus sequence. The second assembly was updated to include these regions, verified using an optical map, and deposited at GenBank under the accession number CP003421.3.

The Ovis and Equi strains share a genomic island known as PiCp12. All the buffalo isolates except strains 43 and 46 have a large insertion in PiCp12 that the other isolates from other hosts lack. This insertion is 36.6 Kb in length and includes 48 CDS and which are flanked by tRNA-Arg genes (Table 3). This insertion includes an intact β -corynephage, a prophage that was predicted by PHAST and estimated to be 30.4Kb and contains 39 CDSs (Fig 3) (Table 3, CDSs 3 to 41). A search of the NCBI database showed that the 36.6 Kb insertion sequence is unique and probably a new corynephage. This insertion also includes a tyrosine integrase gene, which codes for an enzyme known for site specific recombination [44]. tRNA-Arg genes, known as integration sites of β -corynephages [45], flank this region. This provides a possible explanation for the insertion of this unique 36.6 Kb region inside of the PiCp12 genomic island.

The sequence that is most similar to this new corynephage is BFK20 (NC_009799), a non-toxicogenic dsDNA virus that infects *Brevibacterium flavum*. BFK20 has structural and lytic genes closely related to known phages found in the Corynebacterinae and to *C. diphtheriae* prophages [46]. The sequence of the new β -corynephage of strain 31 was deposited in Genbank under accession number KY566218. Additional accession numbers for the sequences of all of the buffalo strains are provided in S2 Table.

The 3' end of the 36.6 Kb insertion into the PiCp12 island includes a gene of particular interest that produces the diphtheria toxin. This gene is only found in *C. pseudotuberculosis* strains that were isolated from buffalo, and is found 3,266 bp from the end of the predicted prophage in our isolates. The sequence is identical in all isolates, and differs from the toxin genes that are found in *C. diphtheria* and *C. ulcerans* (Fig 4). Maximescu et al. [47] have reported the presence of a diphtheria toxin from two separate strains isolated from Egyptian buffalo, indicating that this gene has been associated with infection in these animals since at least 1974. We cannot be certain that the entire insert, including the prophage, was also present in these previous infections as we do not have genomic sequences from these animals.

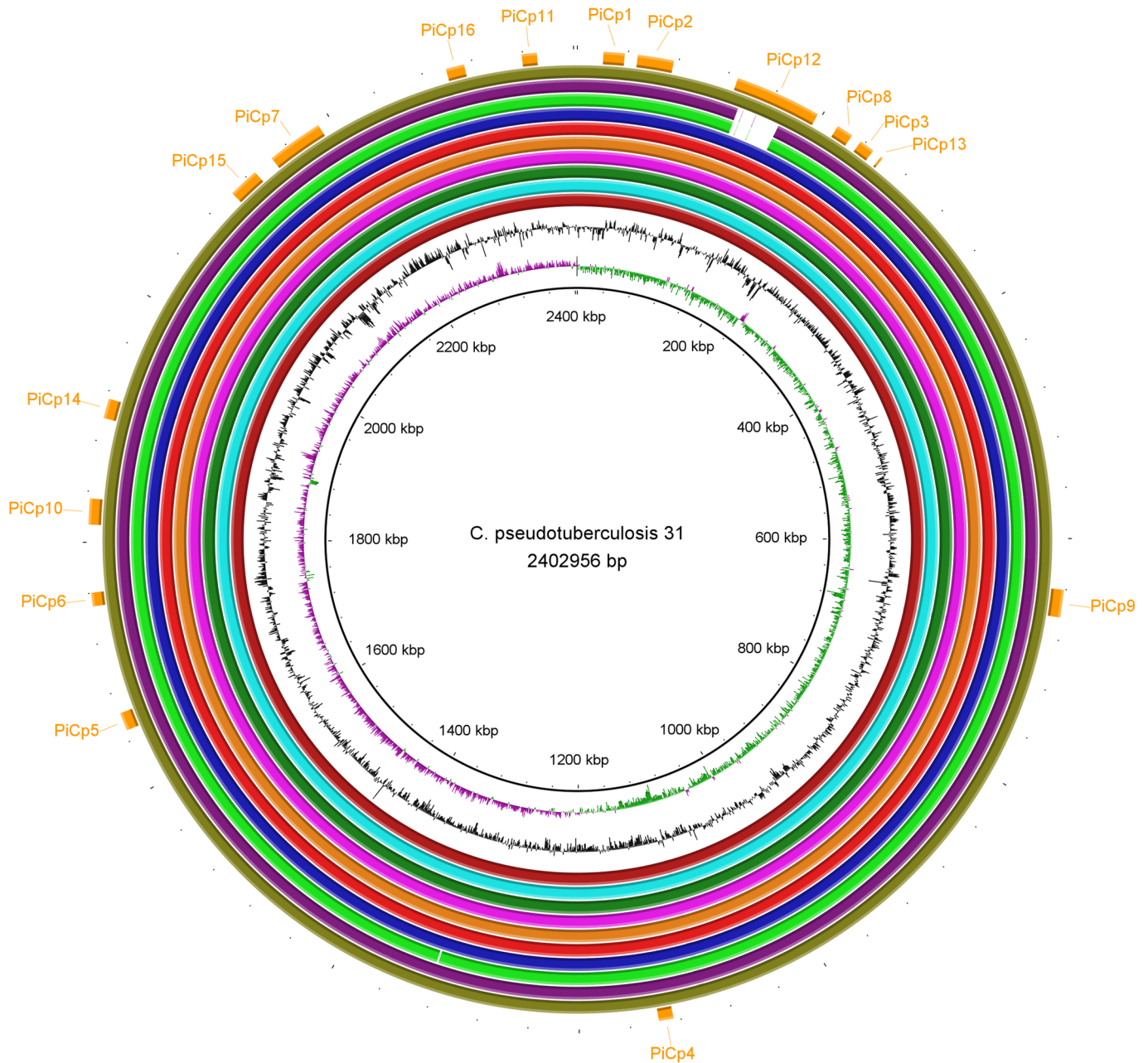


Fig 2. A circular genomic map that compares 11 *Corynebacterium pseudotuberculosis* strains isolated from Egyptian buffalo. The rings, from the inner to outer circle, are: strain 31, GC skew, GC content, strains 32, 33, 34, 35, 36, 38, 39, 43, 46, and 48, and pathogenicity islands.

<https://doi.org/10.1371/journal.pone.0176347.g002>

However, previous studies have suggested that toxins can be acquired through a lysogenic conversion during an infection by a β -corynephage [2,48], and this gene in the buffalo isolates is downstream from a corynephage in the same insertion sequence. If these studies are correct, then the proximity of the corynephage and the toxin gene supports the premise that the acquisition of the two together is linked in buffalo, but it is certainly not definitive proof.

Toxins can cause disease. When injected intradermally into susceptible animals, the toxin from *C. diphtheriae* resulted in erythema, induration, and dermonecrosis [48]. Parenteral injection caused myocarditis, polyneuritis, and focal necrosis in organs that included the adrenal

Table 3. Gene content of pathogenicity island PiCp12, a 36.6 Kb insertion sequence found in the buffalo isolates.

CDS	Position	Strand	Product	Refseq locus tag
1	136439..137053	+	Hypothetical protein	CP31_RS00695
2	137233..137475	+	Hypothetical protein	CP31_RS00700
	137418..137430	+	attL GCTAAAAGGGGC	CP31_RS00715
3	137506..138771	-	PHAGE_Mycoba_Ariel_NC_028876: tyrosine integrase; phage(gi971751237)	CP31_RS00705
4	138901..139740	-	Hypothetical protein	CP31_RS00710
5	139742..140548	-	Hypothetical protein	CP31_RS00715
6	140658..140882	-	PHAGE_Rhodoc_REQ2_NC_016652: Hypothetical protein; phage(gi372449852)	CP31_RS00720
7	140885..141031	-	Hypothetical protein	None
8	141106..141219	-	Hypothetical protein	None
9	141313..141720	-	No significant database matches	CP31_RS00725
10	141730..141981	-	No significant database matches	CP31_RS0073
11	142339..142575	+	Transcriptional regulator	CP31_RS00735
12	142624..142884	+	Hypothetical protein	None
13	143334..143807	+	Hypothetical protein	CP31_RS00745
14	143834..144655	+	PHAGE_Bacter_Lily_NC_028841: antirepressor; phage(gi971748300)	CP31_RS00750
15	144870..145103	+	Hypothetical protein	CP31_RS00760
16	145127..145351	+	Hypothetical protein	CP31_RS00765
17	145612..145725	-	Hypothetical protein	None
18	145779..146525	+	Hypothetical protein	CP31_RS00775
19	146683..147729	+	Hypothetical protein	CP31_RS00780
20	148215..148577	+	No significant database matches	CP31_RS00785
21	148665..148946	+	PHAGE_Coryne_BFK20_NC_009799: gp55, HNH endonuclease; phage(gi157168428)	None
22	149063..149455	+	Hypothetical protein	CP31_RS00790
23	149445..150839	+	PHAGE_Coryne_BFK20_NC_009799: gp2, terminase; phage(gi157168375)	CP31_RS00795
24	150875..151048	+	Hypothetical protein	CP31_RS00795
25	151061..152302	+	PHAGE_Coryne_BFK20_NC_009799: gp3, phage portal protein; phage(gi157168376)	CP31_RS00800
26	152299..153345	+	PHAGE_Coryne_BFK20_NC_009799: gp5, head maturation protease; phage(gi157168378)	CP31_RS00805
27	153342..154592	+	PHAGE_Coryne_BFK20_NC_009799: gp6, major capsid protein; phage(gi157168379)	CP31_RS00810
28	154592..154762	+	No significant database matches	None
29	154785..155267	+	PHAGE_Coryne_BFK20_NC_009799: gp8; phage(gi157168381)	CP31_RS00815
30	155264..155626	+	PHAGE_Coryne_BFK20_NC_009799: gp9; phage(gi157168382)	CP31_RS00820
31	155619..155885	+	PHAGE_Coryne_BFK20_NC_009799: gp10; phage(gi157168383)	CP31_RS00825
32	155875..156252	+	PHAGE_Coryne_BFK20_NC_009799: gp11; phage(gi157168384)	CP31_RS00830
33	156281..157228	+	PHAGE_Coryne_BFK20_NC_009799: gp12, major tail protein; phage(gi157168385)	CP31_RS00835
34	157323..157700	+	PHAGE_Coryne_BFK20_NC_009799: gp13; phage(gi157168386)	CP31_RS00840
35	157862..158071	+	No significant database matches	CP31_RS00845
36	158081..158770	+	PHAGE_Coryne_BFK20_NC_009799: gp15, minor tail protein; phage(gi157168388)	CP31_RS00850
37	158733..163745	+	PHAGE_Coryne_P1201_NC_009816: putative tail measure protein; phage(gi157310951)	CP31_RS00850
	162051..162063	+	attR GCTAAAAGGGGC	None
38	163755..164546	+	Immunity-specific protein Beta201	CP31_RS00855
39	164587..165369	+	Immunity-specific protein Beta286	CP31_RS00860
40	165370..166449	+	Immunity-specific protein Beta371	CP31_RS00865
41	166449..167825	+	PHAGE_Salmon_Fels_1_NC_010391: putative bacteriophage tail fiber protein; Lambda gpN homolog; phage(gi169257204)	CP31_RS00870
42	167829..168005	+	Hypothetical protein	None
43	168307..168861	+	Hypothetical protein	CP31_RS00875
44	168969..169706	+	Teichoic acid phosphorylcholine esterase/choline binding protein E (cbpE)	CP31_RS00880

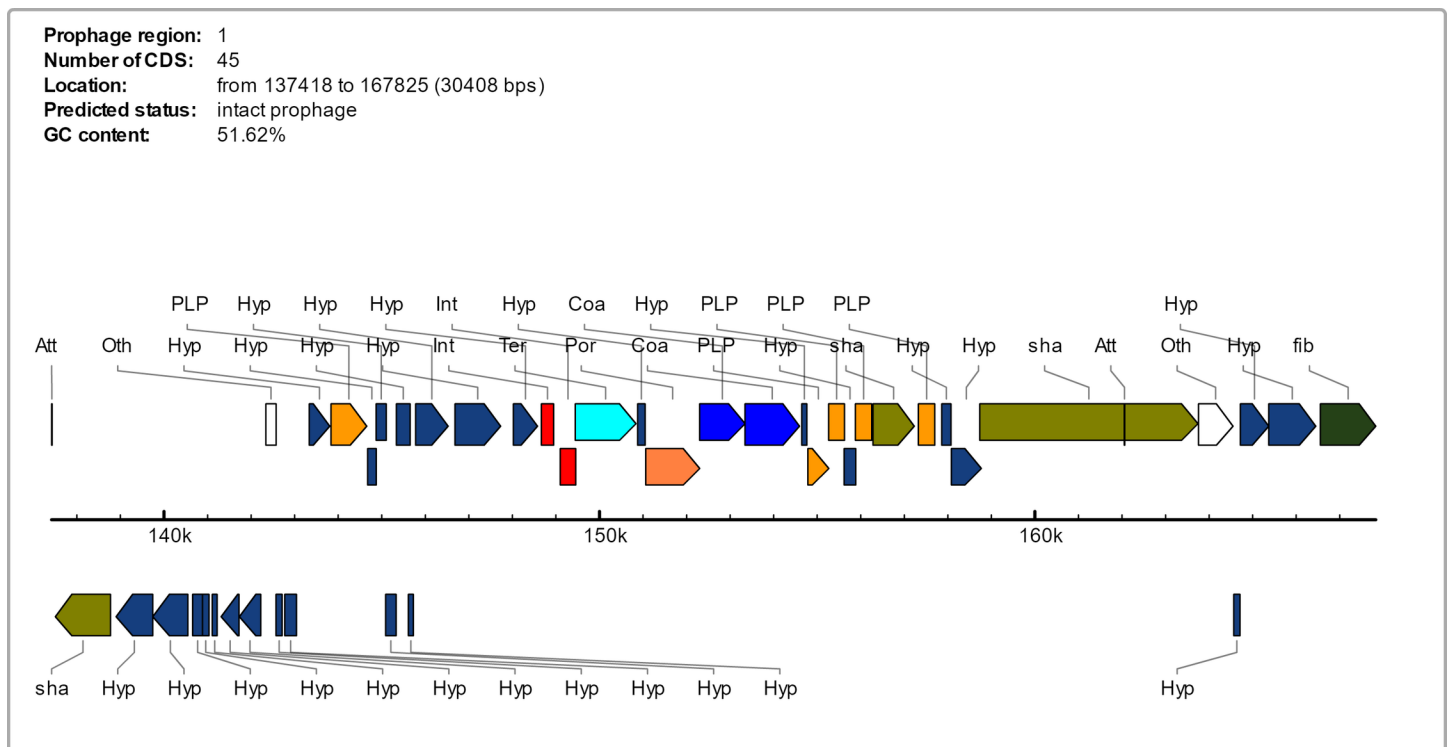
(Continued)

Table 3. (Continued)

CDS	Position	Strand	Product	Refseq locus tag
45	169703..169945	+	Hypothetical protein	CP31_RS00885
46	169948..170412	+	Putative membrane protein	CP31_RS00890
47	170409..170747	+	Putative secreted protein	CP31_RS00895
48	171092..172774	+	Diphtheria toxin	CP31_RS00900

<https://doi.org/10.1371/journal.pone.0176347.t003>

glands, kidneys, and liver [48]. The toxin gene described in this study was detected in *C. pseudotuberculosis* isolated from buffalo with OSD [16]. The presence of this toxin may be responsible for the unique disease manifestations seen in the infected buffaloes. These animals have huge abscesses in the draining lymph nodes of dewlap, belly, and limbs that can extend across the whole limb, giving an aspect of elephantiasis. Other manifestations include skin eruptions and hair loss around the eruptions, extensive dermal necrosis, spontaneous bleeding



Identified CDS types:

- | | | |
|--|---|--|
| 1 Lysis | 2 Terminase | 3 Portal |
| 4 Protease | 5 Coat | 6 Tail shaft |
| 7 Attachment site | 8 Integrase | 9 Other phage-like protein |
| 10 Hypothetical protein | 11 Other | 12 Transposase |
| 13 Tail fiber | 14 Plate | 15 tRNA |

Fig 3. Gene content of the intact prophage predicted by PHAST. The prophage is inserted in pathogenicity island PiCp12 of the strains isolated from buffalo.

<https://doi.org/10.1371/journal.pone.0176347.g003>

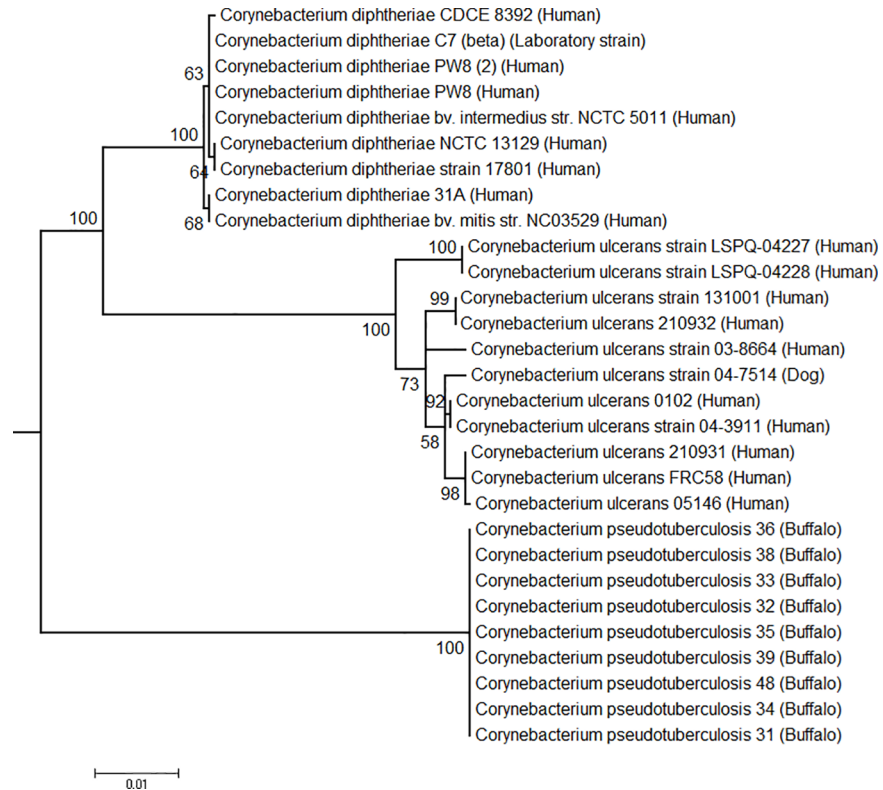


Fig 4. Phylogeny of diphtheria toxin gene (*tox*) from *Corynebacterium pseudotuberculosis*, *C. diphtheriae* and *C. ulcerans* inferred by using the maximum likelihood method based on the Tamura 3-parameter model, on Mega v6.

<https://doi.org/10.1371/journal.pone.0176347.g004>

from the skin, rapid respiration, coughing, dyspnea, and haemoglobinuria [13,16]. In contrast, horses infected by *C. pseudotuberculosis* (biovar Equi), which lack the toxin, have abscesses in different locations that include the pectoral or ventral abdomen, or in internal organs, or ulcerative lymphangitis of the limbs [12]. As buffalo are immune to any strains from the Ovis biovar [49], and there has never been a report of buffalo infected with any other strain within the Equi biovar, a possible preventative measure could be to use an inactivated form of the *C. pseudotuberculosis* diphtheria toxin as an antigen in a vaccine.

A synteny graph including all the *C. pseudotuberculosis* buffalo genomes showed two large regions (green blocks) that had the same relative position, and two smaller regions (brown and red blocks) that were close to each other, but had a variable position or were missing (Fig 5). These variable regions are found within the same pathogenicity island PiCp12 and are flanked by tRNA-Arg-ACG genes. The brown block shows the 36.6 kb insertion that harbors the prophage and the diphtheria toxin (*tox*), and is absent in strains 43 and 46. The red block is also present in the genomes of the Ovis and Equi strains, and contains the Nitric-oxide reductase (*norZ*) gene. This enzyme is a defense against the cytotoxic actions of Nitrous-oxide (NO) that the host uses as a defensive mechanism [50]. As stated previously, the tRNA-Arg genes are known to be a hotspot for phage integration [45] and the fact that we see these rearrangements within genomes isolated during the same outbreak, in the same region and from the same host species confirms that this is a volatile region in these genomes. As if in confirmation of this, strains 43 and 46 were found to be positive for the toxin gene by PCR when originally isolated from infected animals. By the time they were sequenced, this region was missing, indicating

C. pseudotuberculosis 31, 32, 33, 36 and 48



C. pseudotuberculosis 34, 35, 38 and 39



C. pseudotuberculosis 43 and 46

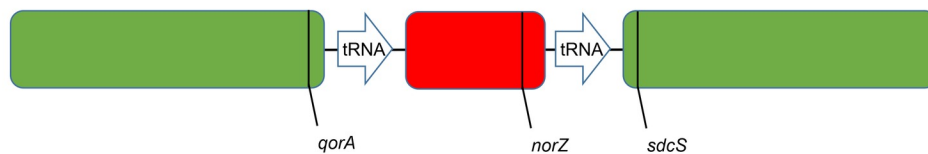


Fig 5. Synteny graph of 11 *Corynebacterium pseudotuberculosis* strains isolated from Egyptian buffalo.

<https://doi.org/10.1371/journal.pone.0176347.g005>

that there was an excision of the 36.6 Kb insert that occurred between isolation and sequencing. Another indication of volatility is the fact that we observed not only a prophage inserted in a copy of a tRNA-Arg gene, but also saw rearrangements within this same region in other genomes that were collected during a short period of time during the summer when this disease outbreak occurred.

A previous study showed that all Equi strains shared the same deletion pattern when compared to strains in biovar Ovis [8]. When we compared the Equi strains with a single Ovis strain, 1002B, we, too, found that they all shared the same deletion pattern except for strain 262 (Fig 6). However, the comparison between Equi strain 31 and the other Equi strains showed that E19, 258, MB11, MB14, MB30 and MB66 had no gaps, besides the 36.6 Kb insertion exclusive of buffalo isolated strains. Also, nitrate reductase genes are missing in the strains that have gaps (Fig 7). This suggests that the gaps in those Equi strains are assembly issues instead of genetic differences.

Phylogenomics

The phylogenomic tree generated by PEPR has two clusters representing the two biovars, with support values of 100 (Fig 8). In the Equi cluster, the buffalo were clearly separated from the horse isolates, which were collected in different countries. In the Ovis cluster, a group containing most of the goat isolates were separated from other hosts. These results suggest that *C. pseudotuberculosis* strains are grouped by host, at least in the Equi biovar. The fact that Equi strains 262, isolated from a cow (*Bos taurus*), and Cp162, isolated from a camel are outside of the two biovar clusters, with support values of 100, corroborates this hypothesis. Also, strain 262 was closer to biovar Ovis, as suggested by the circular map (Fig 6). It is possible that isolations from additional hosts like camels will create additional subclusters based on these hosts.

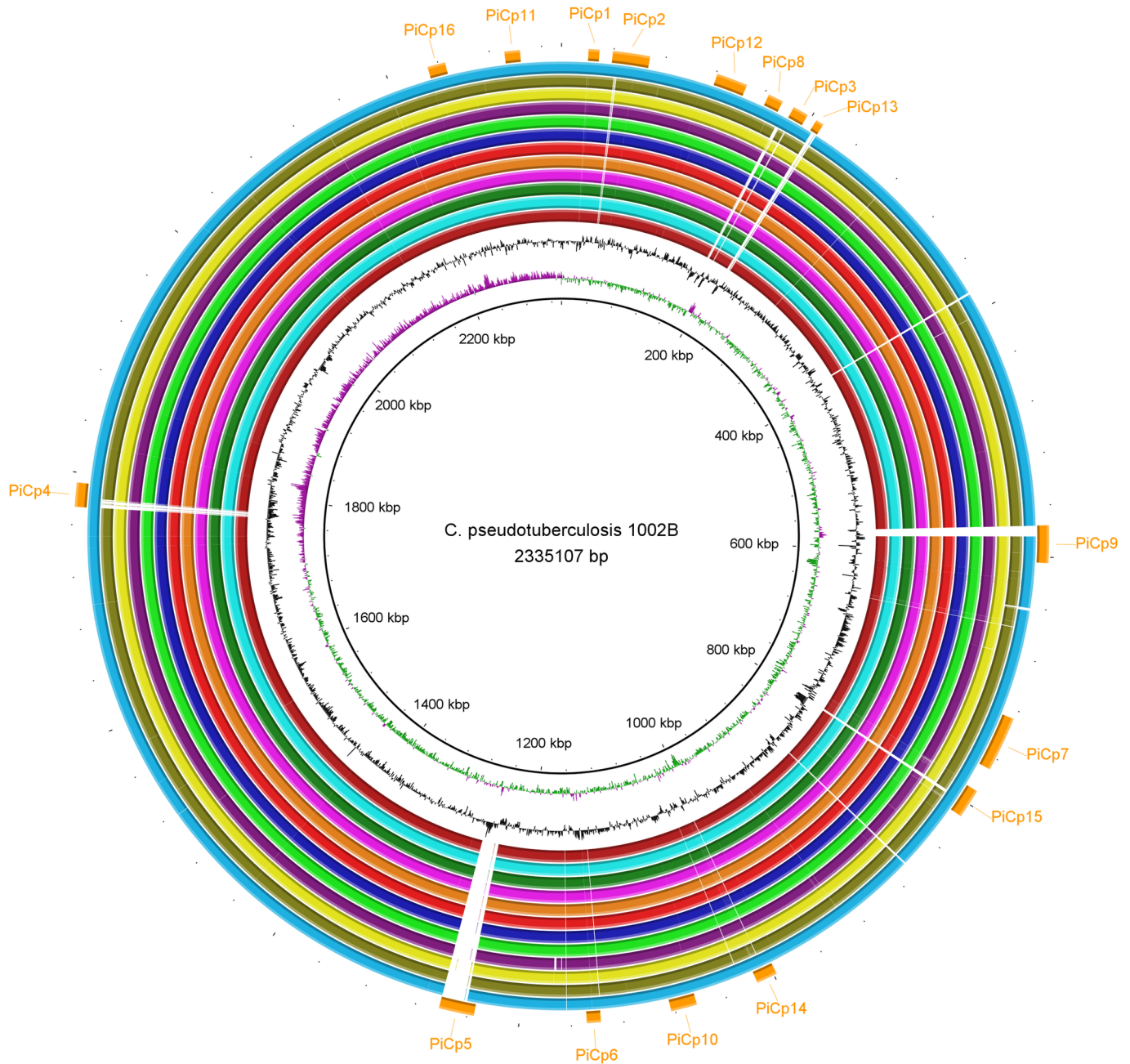


Fig 6. Circular genomic maps comparing *Corynebacterium pseudotuberculosis* 1002B (biovar Ovis) with other Equi strains. The rings, from the inner to outer circle, are strain 1002B, CG Skew, CG Content, strains 31, 258, E19, MB11, MB14, MB30, MB66, 316, CIP52.97, 1/06-A, Cp162 and 262, and pathogenicity islands.

<https://doi.org/10.1371/journal.pone.0176347.g006>

The phylogenomic tree produced by Gegenees showed two clusters representing the two biovars, with similarity percentages within Equi strains (92–99%) being lower than within Ovis strains (99–100%) (Fig 9), a result similar to what was found previously [8]. Here, the clustering of strains by host is also apparent. Buffalo are grouped with the horse and camel isolates, with strains 43 and 46 separated from the rest of the buffalo isolates, probably due to the absence of the 36.6 Kb insertion in PiCp12. The heatmap shows higher values of similarity

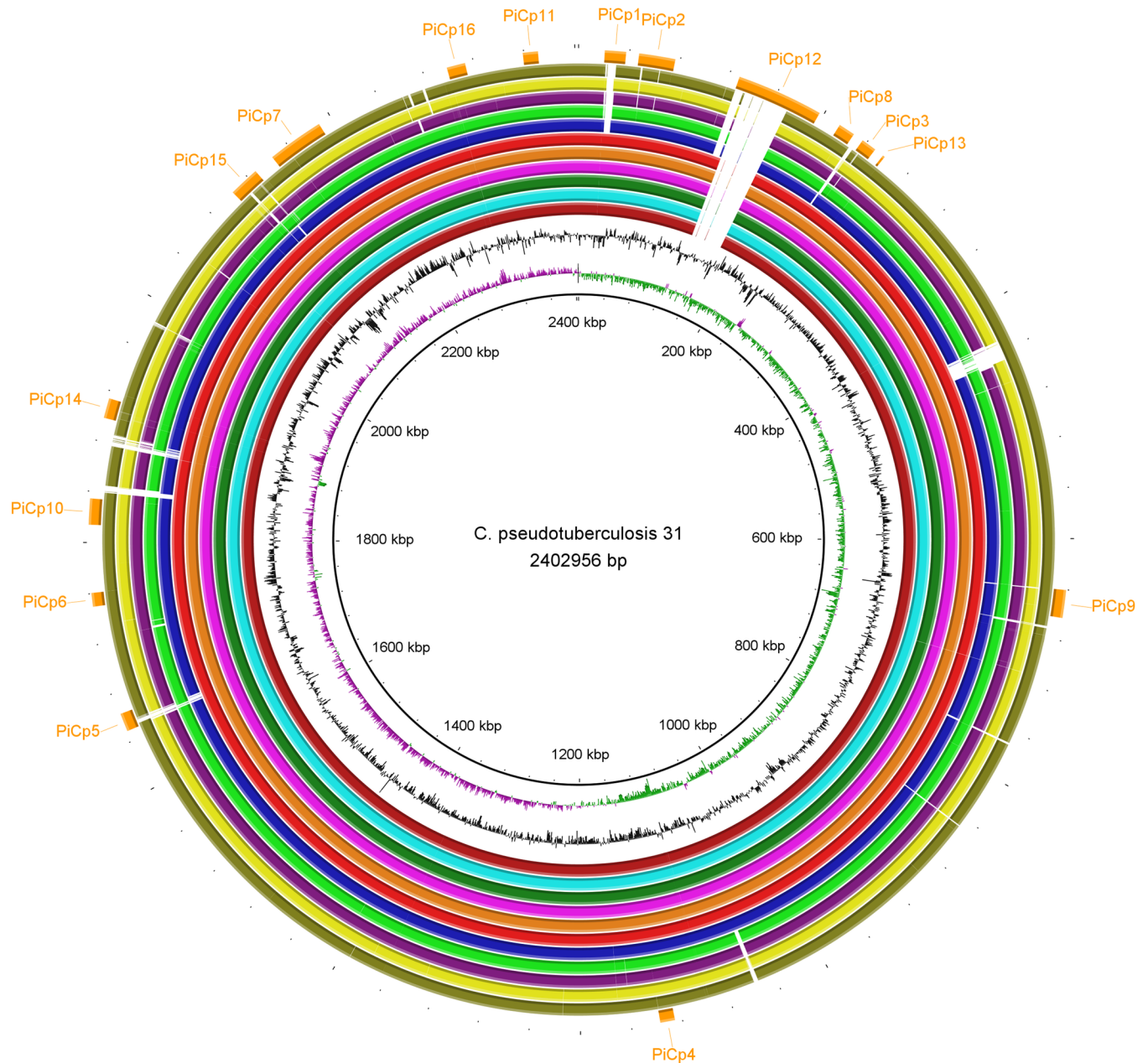


Fig 7. Circular genomic maps comparing *Corynebacterium pseudotuberculosis* 31 with other Equi strains. The rings, from the inner to outer circle, are strain 31, CG Skew, CG Content, strains 258, E19, MB11, MB14, MB30, MB66, 316, CIP52.97, 1/06-A, Cp162 and 262, and pathogenicity islands.

<https://doi.org/10.1371/journal.pone.0176347.g007>

within isolates of the same type of host, with 262 (the cow isolate) and 162 (the camel isolate) having the lowest values of similarity when compared to the other Equi strains.

Pangenomics

The “pangenome” is the complete gene inventory of a species. The “core genome” is the subset of orthologous genes present in all genomes, the “accessory genome” is a subset present in more than one genome, but not all genomes. “Singletons” are genes that are present in only

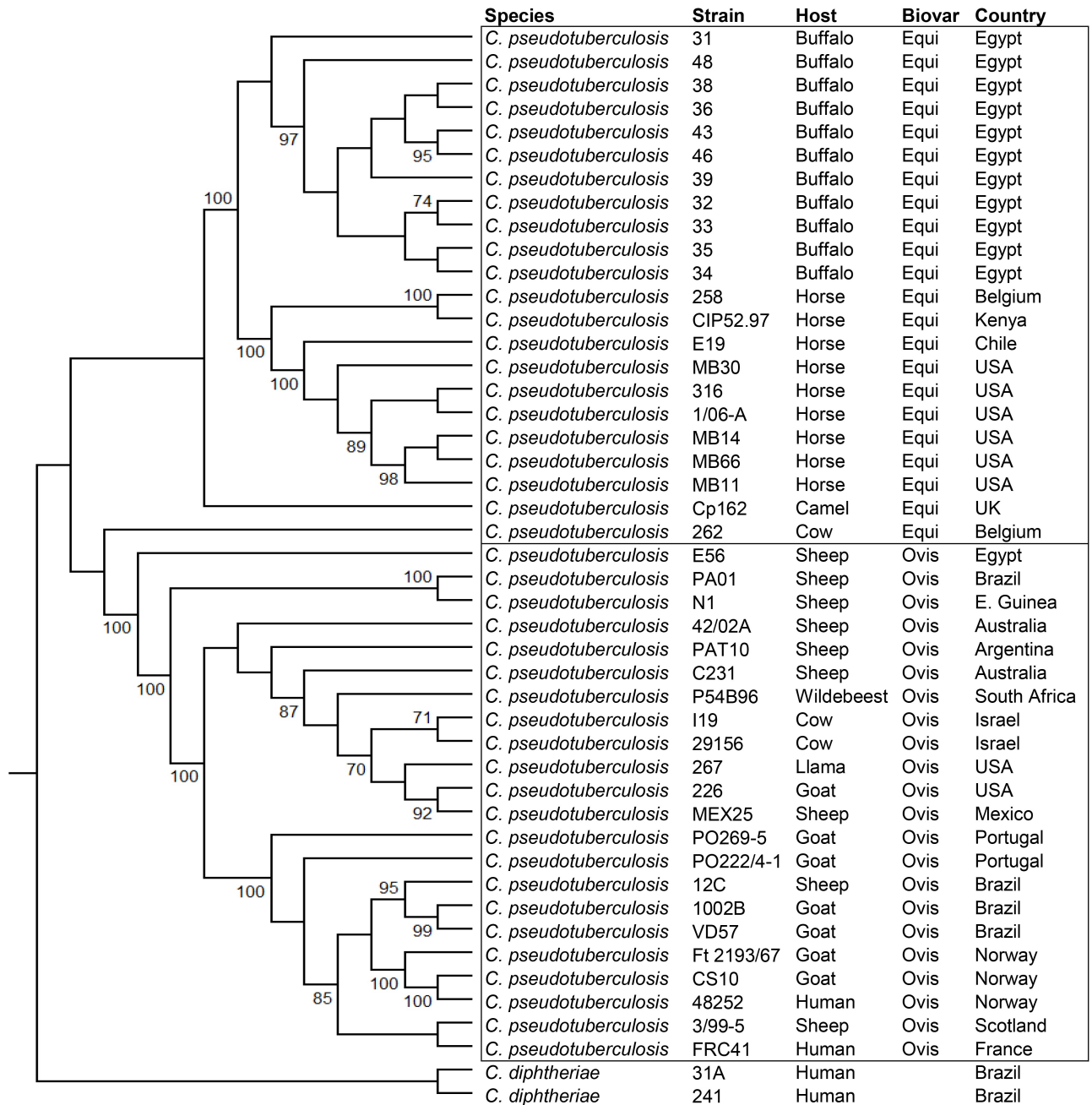


Fig 8. Phylogenomic tree of *Corynebacterium pseudotuberculosis* genomes based on the proteome of 44 complete genomes, generated by PEPR. The core proteins were used to produce a tree, and additional protein families were added to refine subtrees with low bootstrap values.

<https://doi.org/10.1371/journal.pone.0176347.g008>

one genome [51]. A previous *C. pseudotuberculosis* study identified 1,504 genes in the core genome and a pangenome with 2,782 genes. The pangenome was characterized as open [8], meaning that sequencing new genomes should significantly contribute to the identification of new genes and thus better characterize the genetic repertoire of the species [52].

Using PATRIC's Protein Family Sorter, we identified a pangenome that included 2,172 protein families among those genomes isolated from buffalo. Most of these same protein families

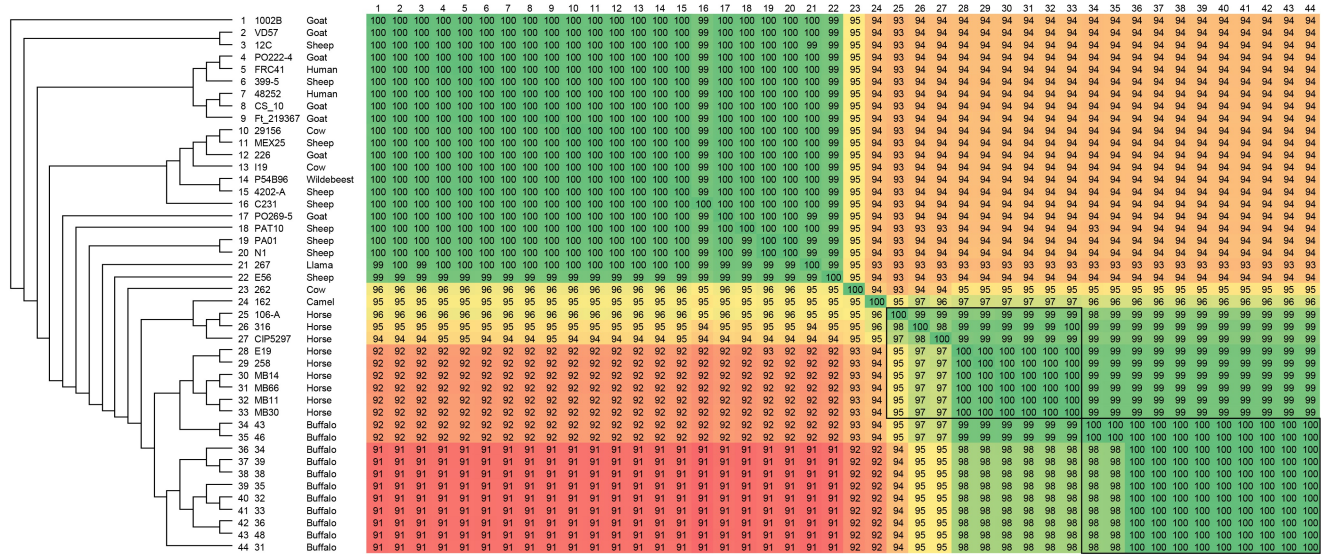


Fig 9. Phylogenomic tree of *Corynebacterium pseudotuberculosis* genomes based on the variable content of 44 complete genomes. The percentages of similarity were plotted on a heatmap generated by Gegenees 2.2.1, and then used to produce a phylogenomic tree using Splitstree v4.14.2 with the UPGMA method.

<https://doi.org/10.1371/journal.pone.0176347.g009>

were also conserved as part of the core genome, which included 2,058 families. The accessory genome was limited to 91 protein families and 13 singletons, indicating that these genomes are all very similar. Analyzing the 11 buffalo with the other 33 strains public available showed that the *C. pseudotuberculosis* pangenome has 3,067 genes, and 1,541 in its core genome. In addition, we used the Proteome Comparison tool, a bidirectional BLASTP analysis, to examine the genomes and found 48 genes that were unique to the buffalo isolates. All of these were part of the 36.6 kb prophage that is located in the PiCp12 island.

When *C. pseudotuberculosis* strain 31 was used as the reference to compare all the strains isolated from buffalo, a strong homology was shared across the genes within these genomes. All genes had a highly conserved sequence identity of at least 90.2%. Also, we identified regions absent in the Ovis biovar that were present in most of the genomes of the Equi strains (S2 Table). A comparison of the functionality of these genes was conducted to look for metabolic and functional changes across the different groups. As expected, the buffalo genomes were all consistently similar, but when compared to genomes from the Ovis biovar, we saw that the buffalo genomes had some unique functionality. The buffalo genomes contained CRISPR and phage genes that were absent in Ovis. They also included the genes involved in nitrate reduction, and additional genes that are part of molybdenum cofactor biosynthesis. The presence of nitrate reductase is one characteristic that differentiates the biovar Equi from Ovis [5,8].

The use of nitrogen oxides as alternative electron acceptors, providing energy in an anaerobic environment and contributing to the persistence of the bacteria in their host, has been suggested as an advantage in adaptation to an intracellular lifestyle [50]. When we examined the region containing nitrate reduction and molybdenum cofactor biosynthesis genes, we found 13 of 16 genes were in operons: *moeBR-moaE*, *narKGHJI-modAB*, and *mobA-moaC-moeA-moaA* (Fig 10). The respiratory nitrate reductase enzyme has three subunits that are the product of the genes *narGHI*. NarI, the gamma subunit, is a transmembrane peptide that oxidizes quinol, liberates protons in the periplasm, and transports the electrons to NarH. NarH is the beta subunit, and it transports electrons from NarI to NarG. The alpha subunit NarG contains the catalytic domain that requires a molybdenum cofactor Mo-(bis-MGD) for activity. This

subunit uses nitrate as the electron acceptor, reducing it to nitrite [53]. NarJ assists in the insertion of molybdenum cofactor in NarG, which must happen before the interaction between NarGH and NarI [54]. NarK is a transporter that performs nitrate/nitrite exchange [55]. NarT is probably involved in nitrate/nitrite transport as it has 80% identity with *narK*, and also includes the same Major Facilitator Superfamily (MFS) domain. MFS transporters are single-polypeptide secondary carriers that transport small solutes in response to a chemiosmotic ion gradients [56]. The nitrate reductase has been suggested as a drug target in *Mycobacterium tuberculosis* [53], and could be used in a similar fashion in *C. pseudotuberculosis*.

Several studies have attempted to define the roles that the *nar* genes play in survival within the host. NarGH nitrate reductase provides resistance from acid stress and reactive nitrogen species in mycobacteria [57]. A *narG* mutant of *M. tuberculosis* was unable to persist in the lung, kidney and liver of immunocompetent mice, but it was able to grow and persist in the spleen as the wild-type strain, suggesting that the role of nitrate reduction in virulence is tissue specific [58]. However, another study showed that there no difference in persistence in the mice lungs between a *narG* mutant and wild-type *M. tuberculosis*, probably because mouse granulomas are not sufficiently hypoxic to affect the growth and survival of these bacteria [59]. The benefit, if any, that these genes give to the Equi strains has not been determined, but could be a target of further experimentation to see if they are important in either host specificity or disease manifestation. A genetic manipulation of *narG* in *C. pseudotuberculosis* could shed light on the different disease manifestations seen between hosts infected with either of the two biovars.

NarG requires molybdenum to function, and, in the buffalo genomes, the molybdenum transport genes are adjacent to the ones needed for nitrate reduction. The molybdenum cofactor biosynthesis operon *modABC* encodes a molybdate transmembrane transporter. ModA binds molybdate, ModB is a transmembrane subunit, and ModC provides the energy required for transportation by ATPase activity [60]. MoaA and MoaC convert a guanine nucleotide (GTP) to cyclic pyranopterin monophosphate (cPMP). MoaD and MoaE are subunits of the molybdopterin (MPT) synthase, which converts cPMP to MPT. MoeBR (a sulfurtransferase) activates MPT synthase by transferring sulfur groups to MoaD subunit. MogA (an adenylyltransferase) adenylylates MPT, resulting in MPT-AMP. Molybdate is added by MoeA (MPT Mo-transferase), resulting in Mo-MPT (molybdenum cofactor). MobA (guanylyltransferase) converts Mo-MPT to a bis-Mo-MPT intermediary and attaches two guanines (GMP) to its phosphate groups, converting it to bis-MGD (molybdopterin guanine dinucleotide), which is the required cofactor for dimethylsulfoxide reductase (DMSO) enzymes, including nitrate reductase [60,61]. The insertion of the molybdenum cofactor in nitrate reductase NarGHI is performed by NarJ [62]. We found that the genes coding the transmembrane and ATP binding subunits of molybdate transporter (*modB* and *modC*) are fused in *C. pseudotuberculosis* and are transcribed with the nitrate reductase genes *narKGHJI* in the same operon, as predicted by FgenesB. The fusion between *modB* and *modC* does not appear to affect the phenotype, as all Equi strains that have been sequenced are positive for nitrate reduction, and they all share this fusion.

Specialty genes search

Specific genes that function as virulence factors can determine bacterial adhesion, invasion, colonization, dissemination within the host and evasion of the immune system [63]. The diphtheria toxin gene was the only virulence factor found that was unique to the buffalo isolates. An additional virulence factor, inositol-1-phosphate synthase, a product of the *ino1* gene, was found in the buffalo strains. This gene is broadly shared across the *Corynebacterium*. Ino1 catalyzes the first step in the synthesis of inositol [64], a compound required for the production of

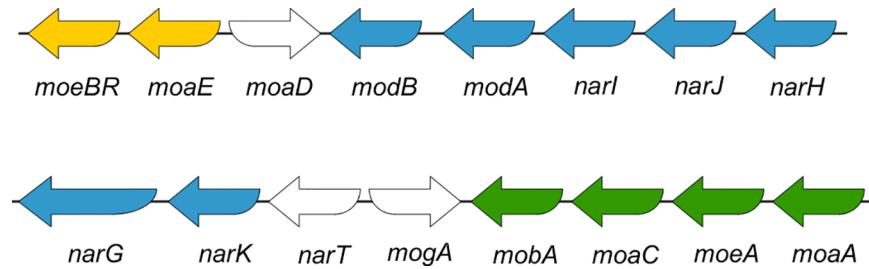


Fig 10. Organization of nitrate reductase and molybdenum cofactor biosynthesis genes in *Corynebacterium pseudotuberculosis* 31. Genes with the same color are transcribed in the same operon.

<https://doi.org/10.1371/journal.pone.0176347.g010>

essential cell wall lipoglycans in *Mycobacterium* [65], and is a major thiol that plays a role in protection from oxidative stress [66,67]. In an experiment using *M. tuberculosis*, the role that *ino1* plays in virulence was demonstrated when the CFUs of *ino1* mutants fell sharply, and the bacteria were virtually cleared in seven days by infected macrophages, while they remained constant in wild type strains [68]. Furthermore, mice infected with wild type *ino1* strains died in 38 days, while mice infected by the *ino1* mutant were alive and healthy when the experiment concluded at 56 days post infection [68]. The role, if any, that this gene plays in the virulence of *C. pseudotuberculosis* has not yet been determined.

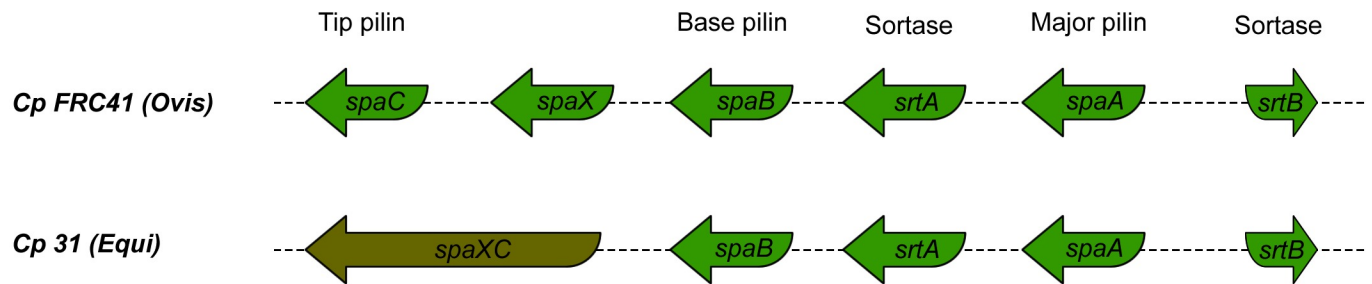
Another virulence factor identified was the exotoxin phospholipase D (*pld* gene), which promotes bacterial dissemination by degradation of sphingomyelin in endothelial cell membranes, and also plays a role in macrophage death [69,70]. In strain 31, *pld* was found to have a frameshift mutation near the 3', and it was suggested that this could decrease the ability of this strain to spread throughout the host, and this, along with the *tox* gene, have been proposed as either a requirement for infection or a geographic variant [8]. This mutation was not seen after resequencing strain 31, however, nor did we find it any other genomes used in this study. This suggests that the original finding was probably a sequencing artifact.

Pili are structures responsible for bacterial adhesion and play a major role in the initiation of extracellular and intracellular invasion and proliferation [71]. Pathogenicity islands PiCp7 and PiCp15 harbor the pilus gene cluster *spaA* (*srtB-spaA-srtA-spaB-spaX-spaC*) and *spaD* (*srtC-spaD-spaY-spaE-spaF*), respectively [8]. The genes *srtABC* are specific pilus sortases, while *spaAD*, *spaBE* and *spaCF* are major, base, and tip proteins, respectively. The genes *spaX* and *spaY* have unknown functions [72]. Specific sortases cleave the LPTxG motif of the pilin proteins and polymerize them to assembly the pilus, while the housekeeping sortase incorporates the final structure to the cell wall [73]. When we compared the Ovis and Equi strains, we saw some unique differences in these regions.

A comparison of pilus gene clusters *spaD* and *spaA* showed conservation among buffalo isolates (Fig 11) and polymorphisms when compared to other Equi strains. The *spaD* pilus genes from the buffalo isolates had a fusion between three genes, including the base and tip pilin (genes *spaE-spaF-spaY*), which is also seen in the Equi strains 258, E19 (isolated from a horse) and Cp162 (camel isolate). In all of the buffalo isolates except for strain 31, the major pilin gene *spaA* has a frameshift. A similar frameshift is also seen in all the other Equi isolates except for strains 316 (isolated from a horse), and 262 (isolated from a cow). Fusion is also seen between with the *spaC* (tip pilin) and *spaX* genes in the buffalo isolates. Not all members of Equi share this fusion, but it is also found in 262 (cow), Cp162 (camel), E19 and 258 (horse isolates). The major pilin (*spaA*) gene is frameshifted in most of the Equi strains.

In *C. diphtheria*, the three pili structures SpaA, SpaD and SpaH have been found to be necessary for adhesion to pharyngeal, laryngeal and lung human epithelial cells, respectively. In

spaA pilus gene cluster



spaD pilus gene cluster

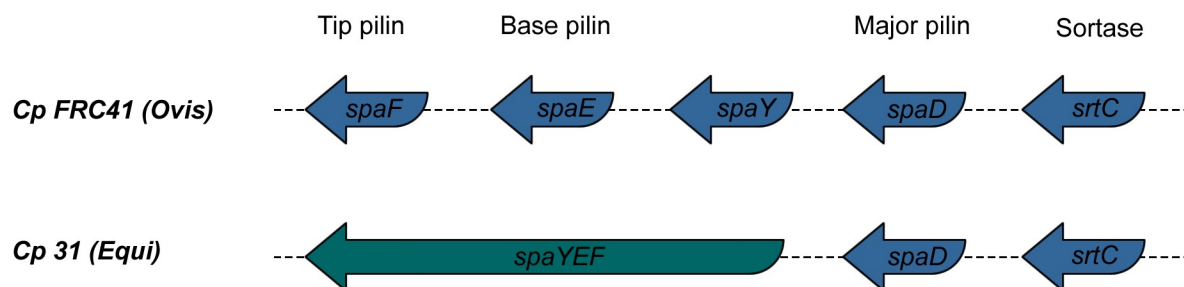


Fig 11. Comparison of pilus gene cluster *spaA* and *spaD* in *Corynebacterium pseudotuberculosis* isolated from buffalo, represented by strain 31, and biovar Ovis, represented by strain FRC41.

<https://doi.org/10.1371/journal.pone.0176347.g011>

this same species, a deletion of the major pilin gene showed that the tip and base pilin were the most important proteins for adhesion on the host cell wall [74]. The frameshift in *spaA* (major pilin) found in *C. pseudotuberculosis* may be not critical to the adhesion. Although, given the importance of base and tip pilin, the fusions of these genes and the conservation within strains may represent an adaptation to the buffalo host, or to mutation events in the immediate ancestor this particular outbreak. More data, perhaps from isolates of future outbreaks of OSD before could determine if this an anomaly or a true adaptation. The polymorphisms in the Equi pilus genes could be due to the higher variability of host species in this biovar, leading to specificity for tissues of different host species [74].

The cow isolate 262 is classified as Equi and has the same nitrate reductase genes, but the sequences of two pilus genes (*srtB* and *spaA*) are identical to those found in the Ovis biovar. Also, the phylogenetic tree (Fig 8) and the circular map (Fig 7) showed the similarity of this strain with the Ovis isolates, yet it could be viewed as ancestral to that clade. Perhaps other Equi isolates, including those from *B. taurus* cattle, will help determine if this is an anomaly, or perhaps the first of a new biovar of *C. pseudotuberculosis*.

Conclusion

C. pseudotuberculosis strains isolated from buffalo showed an overall synteny, conservation in pilus genes, and a unique insertion that contained both a corynephage and the diphtheria toxin gene. This insertion could explain the expansion of the known host range for *C. pseudotuberculosis* to include buffalo, as these genes may play a role in adaptation to this new host. The tRNA-Arg gene was identified as a hotspot of phage insertion and rearrangements, events

observed during in this outbreak of OSD. By comparison with Ovis strains, we identified and described the known nitrate reductase (*narGHI*) genes and identified genes involved molybdenum cofactor biosynthesis, which are necessary for the action of the *nar* genes. A phylogenomic tree confirmed a clear separation between the Ovis and Equi biovars, and indicated that Equi strains are clustered depending on the host they infected.

Supporting information

S1 Fig. Comparison of three versions of *Corynebacterium pseudotuberculosis* 31 genome assembly and strain 32. The rings, from the inner to outer circle, are strain 31 v3 (CP003421.3), GC skew, GC content, strains 31 v1 (CP003421.1), 31 v2 (CP003421.2), and 32 (CP015183.1), and pathogenicity islands.

(TIF)

S1 File. Differences in the assembly versions of *Corynebacterium pseudotuberculosis* strain 31.

(DOCX)

S1 Table. Gene content in the three genome sequences present in the first version of *Corynebacterium pseudotuberculosis* 31 (CP003421.1) and absent in the second version (CP003421.2).

(DOCX)

S2 Table. Accession number of the prophage sequences in *Corynebacterium pseudotuberculosis* strains isolated from buffalo.

(DOCX)

Author Contributions

Conceptualization: MVCV HF RR LCG FLP FAD SAKS MS AS ARW VA.

Data curation: ARW VA.

Formal analysis: MVCV ARW VA.

Funding acquisition: VA.

Methodology: MVCV HF RR FAD AS ARW VA.

Project administration: ARW VA.

Resources: HF SAKS MS ARW VA.

Supervision: ARW VA.

Writing – original draft: MVCV ARW VA.

Writing – review & editing: MVCV HF RR LCG FLP FAD SAKS MS AS ARW VA.

References

1. Dorella FA, Carvalho Pacheco L, Oliveira SC, Miyoshi A, Azevedo V. *Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. *Vet Res*. 2006; 37: 201–218. <https://doi.org/10.1051/vetres:2005056> PMID: 16472520
2. Guaraldi AL de M, Júnior RH, Azevedo CVA de. *Corynebacterium diphtheriae*, *Corynebacterium ulcerans* and *Corynebacterium pseudotuberculosis*—General Aspects. In: Burkovski A, editor. *Corynebacterium diphtheriae* and Related Toxigenic Species. London, New York; 2014. pp. 15–37.

3. Bernard K. The genus *Corynebacterium* and other medically relevant coryneform-like bacteria. *J Clin Microbiol.* 2012; 50: 3152–8. <https://doi.org/10.1128/JCM.00796-12> PMID: 22837327
4. Oliveira A, Teixeira P, Azevedo M, Jamal SB, Tiwari S, Almeida S, et al. *Corynebacterium pseudotuberculosis* may be under anagenesis and biovar Equi forms biovar Ovis: a phylogenetic inference from sequence and structural analysis. *BMC Microbiol. BMC Microbiology;* 2016; 16: 100. <https://doi.org/10.1186/s12866-016-0717-4> PMID: 27251711
5. Biberstein EL, Knight HD, Jang S. Two biotypes of *Corynebacterium pseudotuberculosis*. *Vet Rec.* 1971; 89: 691–692. PMID: 5168555
6. Songer JG, Beckenbach K, Marshall MM, Olson GB, Kelley L. Biochemical and genetic characterization of *Corynebacterium pseudotuberculosis*. *Am J Vet Res.* 1988; 49: 223–226. PMID: 2831763
7. Sutherland SS, Hart RA, Buller NB. Genetic differences between nitrate-negative and nitrate-positive *C. pseudotuberculosis* strains using restriction fragment length polymorphisms. *Vet Microbiol.* 1996; 49: 1–9. PMID: 8861638
8. Soares SC, Silva A, Trost E, Blom J, Ramos R, Carneiro A, et al. The Pan-Genome of the Animal Pathogen *Corynebacterium pseudotuberculosis* Reveals Differences in Genome Plasticity between the Biovar ovis and equi Strains. *PLoS One.* 2013; 8.
9. Windsor PA, Bush RD. Caseous lymphadenitis: Present and near forgotten from persistent vaccination? *Small Rumin Res.* 2016; 142: 6–10.
10. Silva A, Schneider MPC, Cerdeira L, Barbosa MS, Ramos RTJ, Carneiro AR, et al. Complete genome sequence of *Corynebacterium pseudotuberculosis* I19, a strain isolated from a cow in Israel with bovine mastitis. *J Bacteriol.* 2011; 193: 323–324. <https://doi.org/10.1128/JB.01211-10> PMID: 21037006
11. Yeruham I, Friedman S, Perl S, Elad D, Berkovich Y, Kalgard Y. A herd level analysis of a *Corynebacterium pseudotuberculosis* outbreak in a dairy cattle herd. *Vet Dermatol.* 2004; 15: 315–320. <https://doi.org/10.1111/j.1365-3164.2004.00388.x> PMID: 15500484
12. Spier SJ, Azevedo V. *Corynebacterium pseudotuberculosis* infection in horses: Increasing frequency and spread to new regions of North America. *Equine Vet Educ.* 2016;
13. Selim SA. Oedematous skin disease of buffalo in Egypt. *Journal of Veterinary Medicine, Series B.* 2001. pp. 241–258.
14. Ahmed IM, El-Tahawy AS. Prevalence of So-called Oedematous Skin Disease in Egyptian buffaloes with particular study on its economic influence. *Alexandria J Vet Sci.* 2012; 37: 129–133.
15. Ghoneim MA, Mousa AW, Ibrahim AK, Amin AS, Khafagy A, Selim SA. Role of *Hippobosca equina* as a transmitter of *C. pseudotuberculosis* among buffaloes as revealed by PCR and dot blot hybridization. *J Egypt Vet Med Assoc.* 2001; 61: 165–176.
16. Selim SA, Mohamed FH, Hessain AM, Moussa IM. Immunological characterization of diphtheria toxin recovered from *Corynebacterium pseudotuberculosis*. *Saudi J Biol Sci. King Saud University;* 2015; 0–5.
17. Soto a, Zapardiel J, Soriano F. Evaluation of API Coryne system for identifying coryneform bacteria. *J Clin Pathol.* 1994; 47: 756–759. PMID: 7962633
18. Costa Torres L de F, Ribeiro D, Hirata R, Pacheco LGC, Souza MC, dos Santos LS, et al. Multiplex polymerase chain reaction to identify and determine the toxigenicity of *Corynebacterium* spp with zoonotic potential and an overview of human and animal infections. *Mem Inst Oswaldo Cruz.* 2013; 108: 272–279.
19. Silva A, Ramos RTJ, Carneiro AR, Pinto AC, Soares SDC, Santos AR, et al. Complete genome sequence of *Corynebacterium pseudotuberculosis* Cp31, isolated from an Egyptian buffalo. *J Bacteriol.* 2012; 194: 6663–6664. <https://doi.org/10.1128/JB.01782-12> PMID: 23144408
20. Mariano DCB, Pereira FL, Aguiar EL, Oliveira LC, Benevides L, Guimarães LC, et al. SIMBA: a web tool for managing bacterial genome assembly generated by Ion PGM sequencing technology. *BMC Bioinformatics.* 2016; 17: 456. <https://doi.org/10.1186/s12859-016-1344-7> PMID: 28105921
21. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol.* 2012; 19: 455–477. <https://doi.org/10.1089/cmb.2012.0021> PMID: 22506599
22. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics.* 2013; 29: 1072–1075. <https://doi.org/10.1093/bioinformatics/btt086> PMID: 23422339
23. Galardini M, Biondi EG, Bazzicalupo M, Mengoni A. CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. *Source Code Biol Med.* 2011; 6: 11. <https://doi.org/10.1186/1751-0473-6-11> PMID: 21693004
24. Ramos RTJ, Carneiro AR, de Castro Soares S, Barbosa S, Varuzza L, Orabona G, et al. High efficiency application of a mate-paired library from next-generation sequencing to postlight sequencing:

Corynebacterium pseudotuberculosis as a case study for microbial de novo genome assembly. *J Microbiol Methods*. 2013; 95: 441–447. <https://doi.org/10.1016/j.mimet.2013.06.006> PMID: 23792707

25. Brettin T, Davis JJ, Disz T, Edwards R a, Gerdes S, Olsen GJ, et al. RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci Rep*. 2015; 5: 8365. <https://doi.org/10.1038/srep08365> PMID: 25666585
26. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, et al. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res*. 2014; 42: D581–91. <https://doi.org/10.1093/nar/gkt1099> PMID: 24225323
27. Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW. Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data. *PLoS Comput Biol*. 2013; 9: e1003031. <https://doi.org/10.1371/journal.pcbi.1003031> PMID: 23592973
28. Carver T, Berriman M, Tivey A, Patel C, Böhme U, Barrell BG, et al. Artemis and ACT: Viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics*. 2008; 24: 2672–2676. <https://doi.org/10.1093/bioinformatics/btn529> PMID: 18845581
29. Wasmuth E V, Lima CD. UniProt: the universal protein knowledgebase. *Nucleic Acids Res*. 2017; 45: D158–D169. <https://doi.org/10.1093/nar/gkw1099> PMID: 27899622
30. Soares SC, Geyik H, Ramos RTJ, de Sá PHCG, Barbosa EGV, Baumbach J, et al. GIPSy: Genomic island prediction software. *J Biotechnol*. 2016; 232: 2–11. <https://doi.org/10.1016/j.jbiotec.2015.09.008> PMID: 26376473
31. Alikhan N- F, Petty NK, Ben Zakour NL, Beatson SA. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics*. 2011; 12: 402. <https://doi.org/10.1186/1471-2164-12-402> PMID: 21824423
32. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. PHAST: a fast phage search tool. *Nucleic Acids Res*. 2011; 39: W347–52. <https://doi.org/10.1093/nar/gkr485> PMID: 21672955
33. Darling AE, Mau B, Perna NT. Progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS One*. 2010; 5.
34. Ågren J, Sundström A, Håfström T, Segerman B. Gegenees: Fragmented Alignment of Multiple Genomes for Determining Phylogenomic Distances and Genetic Signatures Unique for Specified Target Groups. *PLoS One*. 2012; 7: e39107. <https://doi.org/10.1371/journal.pone.0039107> PMID: 22723939
35. Huson DH. Application of Phylogenetic Networks in Evolutionary Studies. *Mol Biol Evol*. 2005; 23: 254–267. <https://doi.org/10.1093/molbev/msj030> PMID: 16221896
36. Tamura K, Stecher G, Peterson D, Filipinski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol*. 2013; 30: 2725–9. <https://doi.org/10.1093/molbev/mst197> PMID: 24132122
37. Davis JJ, Gerdes S, Olsen GJ, Olson R, Pusch GD, Shukla M, et al. PATtyFams: Protein Families for the Microbial Genomes in the PATRIC Database. *Front Microbiol*. 2016; 7: 1–12.
38. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res*. 2014; 42: D206–14. <https://doi.org/10.1093/nar/gkt1226> PMID: 24293654
39. Mao C, Abraham D, Wattam AR, Wilson MJC, Shukla M, Yoo HS, et al. Curation, integration and visualization of bacterial virulence factors in PATRIC. *Bioinformatics*. 2015; 31: 252–258. <https://doi.org/10.1093/bioinformatics/btu631> PMID: 25273106
40. Chen F, Ding X, Ding Y, Xiang Z, Li X, Ghosh D, et al. Proinflammatory caspase-2-mediated macrophage cell death induced by a rough attenuated *Brucella suis* strain. *Infect Immun*. 2011; 79: 2460–2469. <https://doi.org/10.1128/IAI.00050-11> PMID: 21464087
41. Ruiz JC, D’Afonseca V, Silva A, Ali A, Pinto AC, Santos AR, et al. Evidence for reductive genome evolution and lateral acquisition of virulence functions in two *Corynebacterium pseudotuberculosis* strains. *PLoS One*. 2011; 6.
42. Ramos RTJ, Carneiro AR, Soares SDC, Santos AR Dos, Almeida S, Guimarães L, et al. Tips and tricks for the assembly of a *Corynebacterium pseudotuberculosis* genome using a semiconductor sequencer. *Microb Biotechnol*. 2013; 6: 150–156. <https://doi.org/10.1111/1751-7915.12006> PMID: 23199210
43. Soares SC, Trost E, Ramos RTJ, Carneiro AR, Santos AR, Pinto AC, et al. Genome sequence of *Corynebacterium pseudotuberculosis* biovar equi strain 258 and prediction of antigenic targets to improve biotechnological vaccine production. *J Biotechnol*. 2013; 167: 135–141. <https://doi.org/10.1016/j.jbiotec.2012.11.003> PMID: 23201561
44. Fogg PCM, Colloms S, Rosser S, Stark M, Smith MCM. New applications for phage integrases. *Journal of Molecular Biology*. 2014. pp. 2703–2716. <https://doi.org/10.1016/j.jmb.2014.05.014> PMID: 24857859

45. Sekizuka T, Yamamoto A, Komiya T, Kenri T, Takeuchi F, Shibayama K, et al. *Corynebacterium ulcerans* 0102 carries the gene encoding diphtheria toxin on a prophage different from the *C. diphtheriae* NCTC 13129 prophage. *BMC Microbiol.* 2012; 12: 72. <https://doi.org/10.1186/1471-2180-12-72> PMID: 22583953
46. Bukovska G, Klucar L, Vlcek C, Adamovic J, Turna J, Timko J. Complete nucleotide sequence and genome analysis of bacteriophage BFK20—a lytic phage of the industrial producer *Brevibacterium flavum*. *Virology.* 2006; 348: 57–71. <https://doi.org/10.1016/j.virol.2005.12.010> PMID: 16457869
47. Maximescu P, Oprişan A, Pop A, Potorac E. Further studies on *Corynebacterium* species capable of producing diphtheria toxin (*C. diphtheriae*, *C. ulcerans*, *C. ovis*). *J Gen Microbiol.* 1974; 82: 49–56. <https://doi.org/10.1099/00221287-82-1-49> PMID: 4212024
48. Holmes RK. Biology and molecular epidemiology of diphtheria toxin and the *tox* gene. *J Infect Dis.* 2000; 181: S156–S167. <https://doi.org/10.1086/315554> PMID: 10657208
49. Moussa IM, Ali MS, Hessain AM, Kabli SA, Hemeg HA, Selim SA. Vaccination against *Corynebacterium pseudotuberculosis* infections controlling caseous lymphadenitis (CLA) and oedematous skin disease. *Saudi J Biol Sci.* 2016; 23: 718–723. <https://doi.org/10.1016/j.sjbs.2016.06.005> PMID: 27872567
50. Vázquez-Torres A, Baumler AJ. Nitrate, nitrite and nitric oxide reductases: From the last universal common ancestor to modern bacterial pathogens. *Curr Opin Microbiol.* 2016; 29: 1–8. <https://doi.org/10.1016/j.mib.2015.09.002> PMID: 26426528
51. Rouli L, Merhej V, Fournier P-E, Raoult D. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes New Infect.* 2015; 7: 72–85. <https://doi.org/10.1016/j.nmni.2015.06.005> PMID: 26442149
52. Medini D, Donati C, Tettelin H, Massignani V, Rappuoli R. The microbial pan-genome. *Curr Opin Genet Dev.* 2005; 15: 589–594. <https://doi.org/10.1016/j.gde.2005.09.006> PMID: 16185861
53. Huang Q, Abdalla AE, Xie J. Phylogenomics of *Mycobacterium* Nitrate Reductase Operon. *Curr Microbiol.* 2015; 71: 121–128. <https://doi.org/10.1007/s00284-015-0838-2> PMID: 25980349
54. Zakian S, Lafitte D, Vergnes A, Pimentel C, Sebban-Kreuzer C, Toci R, et al. Basis of recognition between the NarJ chaperone and the N-terminus of the NarG subunit from *Escherichia coli* nitrate reductase. *FEBS J.* 2010; 277: 1886–1895. <https://doi.org/10.1111/j.1742-4658.2010.07611.x> PMID: 20236317
55. Zheng H, Wisedchaisri G, Gonen T. Crystal structure of a nitrate/nitrite exchanger. *Nature.* 2013; 497: 647–651. <https://doi.org/10.1038/nature12139> PMID: 23665960
56. Yan N. Structural Biology of the Major Facilitator Superfamily Transporters. *Annu Rev Biophys.* 2015; 44: 257–283. <https://doi.org/10.1146/annurev-biophys-060414-033901> PMID: 26098515
57. Tan MP, Sequeira P, Lin WW, Phong WY, Cliff P, Ng SH, et al. Nitrate respiration protects hypoxic *Mycobacterium tuberculosis* against acid- and reactive nitrogen species stresses. *PLoS One.* 2010; 5: 1–8.
58. Fritz C, Maass S, Kreft A, Bange F. Dependence of *Mycobacterium bovis* BCG on Anaerobic Nitrate Reductase for Persistence Is Tissue Specific. *Infect Immun.* 2002; 70: 286–291. <https://doi.org/10.1128/IAI.70.1.286-291.2002> PMID: 11748194
59. Aly S, Wagner K, Keller C, Malm S, Malzan A, Brandau S, et al. Oxygen status of lung granulomas in *Mycobacterium tuberculosis*-infected mice. *J Pathol.* 2006; 210: 298–305. <https://doi.org/10.1002/path.2055> PMID: 17001607
60. Williams M, Mizrahi V, Kana BD. Molybdenum cofactor: A key component of *Mycobacterium tuberculosis* pathogenesis? *Crit Rev Microbiol.* 2014; 40: 18–29. <https://doi.org/10.3109/1040841X.2012.749211> PMID: 23317461
61. Leimkühler S. The Biosynthesis of the Molybdenum Cofactor in *Escherichia coli* and Its Connection to FeS Cluster Assembly and the Thiolation of tRNA. *Adv Biol.* 2014; 2014: 1–21.
62. Iobbi-Nivol C, Leimkühler S. Molybdenum enzymes, their maturation and molybdenum cofactor biosynthesis in *Escherichia coli*. *Biochim Biophys Acta—Bioenerg.* 2013; 1827: 1086–1101.
63. Tauch A, Burkovski A. Molecular armory or niche factors: virulence determinants of *Corynebacterium* species. *FEMS Microbiol Lett.* 2015; 67: fnv185.
64. Bachhawat N, Mande SC. Identification of the INO1 gene of *Mycobacterium tuberculosis* H37Rv reveals a novel class of inositol-1-phosphate synthase enzyme. *J Mol Biol.* 1999; 291: 531–536. <https://doi.org/10.1006/jmbi.1999.2980> PMID: 10448034
65. Jackson M, Crick DC, Brennan PJ. Phosphatidylinositol is an essential phospholipid of mycobacteria. *J Biol Chem.* 2000; 275: 30092–30099. <https://doi.org/10.1074/jbc.M004658200> PMID: 10889206
66. Fahey RC. Novel thiols of prokaryotes. *Annu Rev Microbiol.* 2001; 55: 333–356. <https://doi.org/10.1146/annurev.micro.55.1.333> PMID: 11544359

67. Newton GL, Buchmeier N, Fahey RC. Biosynthesis and functions of mycothiol, the unique protective thiol of Actinobacteria. *Microbiol Mol Biol Rev.* 2008; 72: 471–94. <https://doi.org/10.1128/MMBR.00008-08> PMID: [18772286](https://pubmed.ncbi.nlm.nih.gov/18772286/)
68. Movahedzadeh F, Smith D a., Norman RA, Dinadayala P, Murray-Rust J, Russell DG, et al. The *Mycobacterium tuberculosis ino1* gene is essential for growth and virulence. *Mol Microbiol.* 2004; 51: 1003–14. PMID: [14763976](https://pubmed.ncbi.nlm.nih.gov/14763976/)
69. McKean SC, Davies JK, Moore RJ. Expression of phospholipase D, the major virulence factor of *Corynebacterium pseudotuberculosis* is regulated by multiple environmental factors and plays a role in macrophage death. *Microbiology.* 2007; 153: 2203–2211. <https://doi.org/10.1099/mic.0.2007/005926-0> PMID: [17600064](https://pubmed.ncbi.nlm.nih.gov/17600064/)
70. D'Afonseca V, Moraes PM, Dorella FA, Pacheco LGC, Meyer R, Portela RW, et al. A description of genes of *Corynebacterium pseudotuberculosis* useful in diagnostics and vaccine applications. *Genet Mol Res.* 2008; 7: 252–260. PMID: [18551390](https://pubmed.ncbi.nlm.nih.gov/18551390/)
71. Rogers EA, Das A, Ton-That H. Adhesion by Pathogenic Corynebacteria. 2011. pp. 91–103.
72. Trost E, Ott L, Schneider J, Schröder J, Jaenicke S, Goesmann A, et al. The complete genome sequence of *Corynebacterium pseudotuberculosis* FRC41 isolated from a 12-year-old girl with necrotizing lymphadenitis reveals insights into gene-regulatory networks contributing to virulence. *BMC Genomics.* BioMed Central Ltd; 2010; 11: 728.
73. Mandlik A, Swierczynski A, Das A, Ton-That H. Pili in Gram-positive bacteria: assembly, involvement in colonization and biofilm development. *Trends Microbiol.* 2008; 16: 33–40. <https://doi.org/10.1016/j.tim.2007.10.010> PMID: [18083568](https://pubmed.ncbi.nlm.nih.gov/18083568/)
74. Mandlik A, Swierczynski A, Das A, Ton-That H. *Corynebacterium diphtheriae* employs specific minor pilins to target human pharyngeal epithelial cells. *Mol Microbiol.* 2007; 64: 111–124. <https://doi.org/10.1111/j.1365-2958.2007.05630.x> PMID: [17376076](https://pubmed.ncbi.nlm.nih.gov/17376076/)

II.2 Chapter II. – Positive selection analysis in bacteria: methods and findings

Marcus Vinicius Canário Viana, Arne Sahm, Alessandra Lima da Silva, Rodrigo Profeta, Aristóteles Góes Neto, Henrique Cesar Pereira Figueiredo, Alice Rebecca Wattam, Vasco Azevedo. **Formatted according to Frontiers in Microbiology.**

One of main goals of this thesis is to identify genes under positively selection in *C. pseudotuberculosis*, which would provide information about the adaptations required for its survival, and also identify probable drug and vaccine targets. This paper describes the methods used to detect positive selection in bacteria, and the contributions these genes contribute to the biology of this organism. The information presented here could play a role in directing future studies with the goal of detecting adaptations at molecular level related to ecological niche, which includes host preference and virulence, speciation and possible applications in the development of disease control methods.

Positive selection analysis in bacteria: methods and findings

Marcus Vinicius Canário Viana¹, Arne Sahm², Alessandra Lima da Silva¹, Rodrigo Profeta¹, Aristóteles Góes Neto³, Henrique Cesar Pereira Figueiredo³, Alice Rebecca Wattam⁴, Vasco Azevedo⁵.

¹Laboratory of Cellular and Molecular Genetics, Institute of Biological Sciences, Department of General Biology, Federal University of Minas Gerais, Belo Horizonte, Brazil.

²Leibniz Institute on Aging, Fritz Lipmann Institute, Jena, Germany.

³Laboratory of Molecular and Computational Biology of Fungi, Institute of Biological Sciences, Department of General Biology, Federal University of Minas Gerais, Belo Horizonte, Brazil.

⁴AQUACEN, National Reference Laboratory for Aquatic Animal Diseases, Ministry of Fisheries and Aquaculture, Federal University of Minas Gerais, Belo Horizonte, Brazil.

⁵Biocomplexity Institute of Virginia Tech, Virginia Tech, Blacksburg, VA, USA.

Abstract

Molecular genetics techniques, population genetics and genomics can associate genetic variants to individual fitness. A mutated allele can increase its frequency when it is adaptive (positive or adaptive selection), decrease in frequency when it is deleterious (negative or purifying selection) or have its frequency influenced by mutation rate and genetic drift when neutral. Detecting the selective pressures on a bacterial species is useful to understand mechanisms related to a specific ecological niche, such as host preference and virulence, to speciation and to the development control methods. In this review, we describe methods to detect positive selection in bacteria and their contributions bacterial biology.

Keywords: Positive selection, genomics, evolution, ecological niche, bacteria.

36 **1. Detecting Darwinian selection at molecular level**

37 Natural selection can be classified as positive selection (Darwinian, adaptive, or directional
38 selection) if advantageous mutations are fixed by increasing allele frequency. Selection is
39 negative, or purifying, if it acts against deleterious mutations, decreasing their allele
40 frequency. The frequency of alleles that emerged from neutral mutations is influenced by the
41 mutation rate and random genetic drift if these alleles are not linked to genome regions under
42 selection (genetic hitchhiking) (Casillas and Barbadilla, 2017; Kimura, 1968).

43 Molecular genetics techniques and sequencing technologies identify genetic variants that
44 can be related to organismal fitness. Molecular population genetics and genomics approaches
45 analyze genome-wide patterns of DNA variation, within and between species, that are being
46 used to understand the contribution of natural selection to molecular evolution (Casillas and
47 Barbadilla, 2017; Kryazhimskiy and Plotkin, 2008; Stephan, 2010; Thurman and Barrett,
48 2016). Selection tests were developed based on intra-species polymorphisms, e.g. on
49 observations of reduced genetic variation or elevated levels of linkage disequilibrium
50 (Casillas and Barbadilla, 2017; Pavlidis and Alachiotis, 2017), inter-species divergence, e.g.
51 the ratio of non-synonymous to synonymous substitution rates (d_N/d_S) (Biswas and Akey,
52 2006; Miyata and Yasunaga, 1980) as well as mixtures of the both aforementioned categories
53 such as the McDonald-Kreitmann test (McDonald and Kreitman, 1991). Identifying genes
54 under positive selection is useful to generate biological hypothesis for mutation studies and
55 functional analyses (Anisimova et al., 2002; Yang and Dos Reis, 2011).

56 The most representative categories of positively selected genes are involved in arms-race
57 adaptation between parasites and hosts. These genes are on perpetual adaptation against drugs
58 and hosts immune systems (Studer and Robinson-Rechavi, 2009). In bacteria, positively
59 selected genes can determine the adaptation of bacteria to new ecological niches and lead to
60 speciation (Kopac et al., 2014; Lassalle et al., 2015). Positive selection analysis have been
61 performed on bacterial species of medical and veterinary relevance to understand its
62 pathogenicity mechanisms (Xu et al., 2016; Yang et al., 2016) and to the development of
63 control methods (Wang et al., 2017).

64

65 **2. Analysis of protein-coding sequences**

66 Nowadays, analyses of protein-coding sequences evolution usually involve codon based
67 substitution models since they are more realistic and offer a greater robustness of derived
68 evolutionary inferences in comparison to empirical amino acid models (Arenas, 2015). These
69 codon models consider, e.g., the knowledge of genetic code, transition/transversion bias,
70 codon usage, the information of synonymous and non-synonymous nucleotide substitutions,
71 and the difference between amino acids (Goldman and Yang, 1994). Mutations that change
72 the amino acid (non-synonymous) can be adaptive, deleterious or neutral, while mutations
73 that do not change the amino acid (synonymous) can be considered neutral (Miyata and
74 Yasunaga, 1980), although synonymous mutations may impact fitness due to codon usage
75 bias (Mahajan and Agashe, 2018). Thus, positive selection can be inferred if an excess of the
76 nonsynonymous (d_N) over the synonymous codon substitution rate (d_S) was detected ($\omega = d_N /$
77 $d_S > 1$) (Miyata and Yasunaga, 1980). Negative selection (purifying) and neutral evolution
78 can be inferred, vice versa, in case that $\omega < 1$ and $\omega = 1$, respectively (Yang and Bielawski,
79 2000).

80 Initial codon substitution models assumed a single ω for all lineages and sites (Goldman
81 and Yang, 1994) and were limited to detect positive selection if the average value is above 1.
82 The lineage-specific models (branch models) are suitable for detecting positive selection

83 along lineages. However, as they assume no variation in ω among sites, positive selection is
84 detected only if the average dN over all sites is higher than the average dS (Yang, 1998; Yang
85 and Nielsen, 1998). Those models are useful for genes that were duplicated and one copy may
86 have acquired a new function and probably evolved at accelerated rates (Yang, 2007).

87 As few sites are responsible for molecular adaptation, codon substitution models with
88 heterogeneous ω among sites (site models) (Nielsen and Yang, 1998; Yang, 2005; Yang et al.,
89 2000), and among sites and among lineages (branch-site models) (Yang and Nielsen, 2002;
90 Zhang, 2005) were implemented. The site models are more specific to detect amino acids that
91 make rapid adaptation to external factors, and the branch-site models, that can be used to
92 detect ancient or specific adaptive mutational events (Studer and Robinson-Rechavi, 2009).

93

94 **3. Positive selection tests using codon-based methods**

95 The above-mentioned codon substitution models allow to calculate a likelihood for a tree
96 topology given the observed alignment data. A p-value for the inference of positive selection
97 based on tree and alignment can be calculated by comparing the likelihood of a general model
98 that allows positive selection ($\omega > 1$) against the likelihood of null model that does not allow
99 positive selection via a likelihood ratio test (LRT) (Yang and Bielawski, 2000).

100 Site and branch-site models allow for the application of a Bayesian method to calculate the
101 posterior probability of each codon belonging to a site of positive selection. The Bayes
102 empirical Bayes (BEB) method (Yang, 2005) has at least the same power of the Naive
103 Empirical Bayes (NEB) method in big data sets and does not generate as many false positives
104 in small data sets (Yang, 2005). The used of LRT and BEB provide a robust and trustworthy
105 framework for inference of positive selection, but there are limitations. The LRT has
106 sufficient power to detect positive selection only if multiple substitutions have occurred at the
107 same codon site throughout the phylogeny (in branch-site tests). However, the assumptions
108 made in the test can be violated in real data, as in case of intragenic recombination (Yang,
109 2005). In addition, the test is not is not sensitive and reliable for analysis of a single
110 population (Kryazhimskiy and Plotkin, 2008).

111

112 **4. Workflow for positive selection analysis**

113 Positive selection analyses of protein coding sequences require the identification of ortholog
114 groups, followed by codon based alignments, phylogenetic tree reconstruction, and testing
115 whether the data fits a sectional model significantly better than a null model (Moretti et al.,
116 2012; Sahm et al., 2017; Yang, 2007).

117 This analysis is sensitive to data quality. Errors in sequencing and assembly cause false
118 polymorphisms, while errors in annotation, ortholog assignment, alignment and intragenic
119 recombination can lead to the comparison of codons from different phylogenetic origins
120 (Mallick et al., 2009; Markova-Raina and Petrov, 2011; Sahm et al., 2017; Schneider et al.,
121 2009; Studer and Robinson-Rechavi, 2009). This requires strategies for quality filtering such
122 as using of isoforms of similar size, filtering unreliable alignment regions, filtering sequences
123 with evidence of recombination and significance value correction for multiple testing (Hongo
124 et al., 2015; Privman et al., 2012; Sahm et al., 2017; Villanueva-Cañas et al., 2013).

125

126 **5. Methodology for positive selection analysis**

127 **5.1. Sampling and identification of homologous genes**

128 It is important that positive selection tests are made using samples of different populations,
129 because d_N / d_S is not sensitive and reliable to measure the strength of selection within the
130 same population (Kryazhimskiy and Plotkin, 2008). The positive selection test has enough
131 power and accuracy with a minimum number of six sequences that are divergent enough to
132 contain enough information for reliable estimation of d_N and d_S . Sequences that are too
133 divergent can lead to alignment errors (Studer and Robinson-Rechavi, 2009). In the beginning
134 of the workflow, a quality filter could remove sequences with absence of valid start and/or
135 stop codons; presence of non-standard nucleotides; length that is not a multiple of three and
136 size out of a specified lower and upper bounds for sequence length (Hongo et al., 2015).

137 Within homologous genes, orthologs derive from the same ancestor after a speciation
138 event, while paralogs are genes derived from duplication events in the same genome (Koonin,
139 2005). Different methodologies can be used to identify gene homology across genomes
140 (Altenhoff and Dessimoz, 2009; Chen et al., 2007). Most positive selection analyses in
141 bacteria used strategies based on BLASTp bidirectional best hit (BBH) method. This method
142 identifies orthologs by blasting the proteome A against the proteome B and vice-versa, and
143 subsequent identification of protein that are the best hits of each other. At least in prokaryotes,
144 genes known to be orthologs typically form BBH due to orthologs being more similar to each
145 other than to any other sequences, and BBH can serve as a strong indication of gene orthology
146 (Kristensen et al., 2011; Wolf and Koonin, 2012). OrthoMCL uses BLASTp BBH approach
147 with cutoff of e^{-5} and 50% coverage, normalizes the species distance to distinguish orthologs
148 from in-paralogs (derived from recent duplication) and co-orthologs (derived from recent
149 descent and duplication) and used Markov clustering to generate groups (Fischer et al., 2011;
150 Li et al., 2003). After the groups of homologous genes are defined, the quality filter strategy
151 can remove groups and/or sequences. Groups can be removed based on homology
152 relationships requirement for the analysis, such as orthologs or paralogs, based on number of
153 sequences and species count relative to user-defined ranges or based in its the presence on a
154 user-defined anchor genome. Sequences can be removed based on relative and absolute size
155 and identity (Hongo et al., 2015).

156 **5.2. Alignment of sequences of each homolog group**

157 After the groups of homologous genes are defined and filtered, the sequences of each group
158 must be aligned for further analysis. The alignment and alignment filtering software have
159 impact on the detection of positively selected genes (Fletcher and Yang, 2010; Jordan and
160 Goldman, 2012; Markova-Raina and Petrov, 2011; Privman et al., 2012; Wong et al., 2008).

162 Simulations an real data analysis were performed to quantify the influence of aligners and
163 alignment filters on false positives and false negatives in positive selection analyses using site
164 models (Jordan and Goldman, 2012). The used the aligners were ClustalW (Thompson et al.,
165 1994), MAFFT (Kato et al., 2005), ProbCons (Do et al., 2005), T-Coffe (Notredame et al.,
166 2000) and two variations of PRANK (Loytynoja and Goldman, 2008) based on an amino acid
167 model or an empirical codon model. The alignment filters were GUIDANCE (Penn et al.,
168 2010), T-Coffe (Notredame et al., 2000) and Gblocks (Castresana, 2000). PRANK based on
169 empirical codon model performed the best and ClustalW performed the worst alignments,
170 while GUIDANCE performed the best and Gblocks performed the worst alignment filtering
171 (Jordan and Goldman, 2012). PRANK was also shown to outperform MUSCLE (Edgar,
172 2004) and AMAP (Schwartz and Pachter, 2007), as well as ClustalW, ProbCons and T-Coffe,
173 in positive selection tests using site models (Markova-Raina and Petrov, 2011).

174

175 **5.3. Phylogenetic trees**

176 The application of codon substitution models in order to identify positively selected genes
177 requires a phylogenetic tree (Goldman and Yang, 1994). Common methods for reconstruction
178 of phylogenetic trees are, e.g., maximum likelihood (ML), maximum parsimony (MP),
179 Bayesian approaches (BA) and distance-based methods such as neighbor-joining (NJ) (Saitou
180 and Nei, 1987). Frequently used tools for tree reconstruction comprise, e.g., RAxML (ML)
181 (Stamatakis, 2014), Phylip (ML and MP) (Felsenstein, J., 2005, Phylogeny Inference Package
182 version 3.6.), and BEAST (BA) (Bouckaert et al., 2014). A common strategy to reconstruct
183 phylogenetic species trees is involves concatenation of the alignments of either all core genes
184 or all orthologs (Joseph et al., 2011; Kopac et al., 2014; Sahm et al., 2017).

185

186 **5.4. Recombination tests**

187 Intragenic recombination can cause diversification and adaptation that contribute to bacterial
188 evolution (Lefébure and Stanhope, 2007; Orsi et al., 2007, 2008; Soyer et al., 2009; Tan et al.,
189 2017). However, recombination causes the alignment of codons from different phylogenetic
190 origins and lead to false positive results in positive selection analysis (Anisimova et al., 2003;
191 Shriner et al., 2003). The most used recombination test used in positive selection analysis are
192 Pairwise Homoplasy Index (PHI) (Bruen et al., 2006), Neighbor Similarity Score (NSS)
193 (Jakobsen and Easteal, 1996), Maximum Chi-Square (Smith, 1992) and Geneconv (Sawyer,
194 1989). As no single method perform optimally under all scenarios, the best strategy is to use a
195 combination of them (Posada et al., 2002).

196

197 **5.5. Correction of significance values**

198 Genome scale analyses test thousands of features against a null hypothesis (multiple testing)
199 and requires correction to avoid increasing the proportion of false positive results (family-
200 wise error rate) (Anisimova and Yang, 2007; Storey and Tibshirani, 2003). The Bonferroni
201 method (Noble, 2009) is too strict and lead to many false negative results, while Benjamini-
202 Hochberg False Discovery Rate (FDR) (Benjamini and Hochberg, 1995) provides a sensible
203 balance between the number of true and false positives. False positive rate measures the rate
204 that truly null features are called significant and provide little information about the features
205 called significant. FDR measures the rate that significant features are truly null and provides a
206 direct measure about the features called significant (Storey and Tibshirani, 2003). For each
207 tested feature in the genome, the p-value measures the false positive rate and the q-value
208 measures the FDR (Storey and Tibshirani, 2003). The FDR method must be used in
209 recombination test and LRT for positive selection.

210

211 **5.6. Estimation of evolutionary rate by maximum likelihood**

212 Given two sequences of equal length, d_S and d_N can be estimated by performing the following
213 steps: i) counting the numbers of synonymous (S_d) and non-synonymous (N_d) substitutions
214 between the two sequences, ii) calculating the fractions of potential synonymous as well as
215 non-synonymous substitutions multiplied by sequence length in each sequence and building
216 the averages across both sequences (S and N), (iii) calculating the ratios S_d/S as well as N_d/N ,
217 and (iv) correcting these ratios for multiple substitutions at the same site to determine d_S and
218 d_N (Miyata and Yasunaga, 1980; Nei and Gojoborit, 1986). This approach simplistically
219 assumes that all substitutions occur with an equal rate which typically leads to a biased

220 estimation of $\omega = d_N / d_S$ (Yang and Nielsen, 2000). Therefore, maximum likelihood methods
 221 were developed that are based on explicit codon substitution models that allow, e.g., different
 222 substitution rates for transitions and transversions. Model parameters such as ω or the
 223 transition/transversion rate ratio κ are estimated by maximum likelihood. The basic codon
 224 substitution model that is implemented in the software CodeML specifies the substitution rate
 225 Q_{ij} from sense codon i to sense codon j as follows:

226

$$Q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at more than one position} \\ \pi_j, & \text{for synonymous transversions} \\ \kappa\pi_j, & \text{for synonymous transition} \\ \omega\pi_j, & \text{for nonsynonymous transversions} \\ \omega\kappa\pi_j, & \text{for nonsynonymous transitions} \end{cases}$$

228

229 where π_j is the equilibrium frequency of codon j (Yang and Bielawski, 2000). The
 230 transition probability rate matrix P of time (or branch length) t can be determined through
 231 $P(t) = e^{Qt}$, where $Q = \{Q_{ij}\}$. The likelihood of a given phylogenetic tree can then be
 232 calculated from the probabilities $P(t)$ following (Felsenstein, 1981; Goldman and Yang,
 233 1994). There exist several extensions of the basic codon substitution model that allow ω to
 234 vary between alignment sites or branches of a phylogeny (see below).

235

236 5.7. Positive selection tests using site models

237 Site models allow heterogeneous ω among sites (Nielsen and Yang, 1998; Yang, 2005; Yang
 238 et al., 2000). This allows detection of positive selection on a protein coding sequence
 239 alignment even if only a small proportion of sites are affected which is the usual case
 240 (Kosakovsky Pond and Frost, 2005; Nielsen, 2001). These tests are useful to detect amino
 241 acids that are continuously under positive selection, as the ones involved in the arms race
 242 between pathogens and hosts (Studer and Robinson-Rechavi, 2009). The tests comparing the
 243 nested models M1a/M2a, M7/M8 are usually recognized as having the best support (Yang,
 244 2007; Zhang, 2005) and their parameters are presented in **Table 1**.

245

246 **Table 1** | Neutral and positive selection site models based in codon substitution

Models	Possible ω classes	Site proportions	References
Neutral			
M1a	$\omega_0 < 1, \omega_1 = 1$	$p_0, (p_1 = 1 - p_0)$	(Wong, 2004)(Yang, 2005)
M7	Beta(p, q)	$p_0(p, q)$	(Yang et al., 2000)
Selection			
M2a	$\omega_0 < 1, \omega_1 = 1, \omega_2 > 1$	$p_0, p_1, (p_2 = 1 - p_0 - p_1)$	(Wong, 2004)(Yang, 2005)
M8	Beta, $\omega_s > 1$	$p_0, (p_1 = 1 - p_0)$	(Yang et al., 2000)

247 The value ω represents the proportion of non-synonymous (d_N) by the synonymous (d_S),
 248 calculated as $\omega = d_N / d_S$.

249

250 Each model assumes that the sites (codons) belong to different classes according to ω
 251 values and that each class has its proportion. The null model M1a assumes a proportion p_0 of
 252 sites under negative selection ($\omega_0 < 1$), and a proportion p_1 with sites in neutral evolution (ω_1
 253 = 1). The selection model M2a assumes an extra class of sites under positive selection (ω_2 ,
 254 where $\omega > 1$) with proportion p_2 . The null model M7 assumes a proportion p_0 of sites in a
 255 class with beta distribution $B(p, q)$, where $0 \leq \omega \leq 1$. The selection model M8 adds an extra
 256 proportion p_1 of sites with $\omega_s > 1$. As each model has advantages and disadvantages, the use of
 257 multiple models in real data analysis is suggested (Anisimova et al., 2002).

258 To compare nested models (M1a/M2a and M7/M8) with a LRT, twice the log likelihood
 259 difference ($2\Delta l$) between the two compared models is compared against a χ^2 distribution with
 260 two degrees of freedom. If the more general model that allows sites with $\omega < 1$ (M2a or M8)
 261 fits the data significantly better, the alignment can be assumed to be affected by positive
 262 selection (Yang, 2007). The Bayesian Empirical Bayes (BEB) method which is implemented
 263 in CodeML can be applied to calculate the posterior probability of each site to be under
 264 positive selection (Yang, 2005).

265

266 5.8. Positive selection tests using branch-site models

267 The lineage and site models (branch-site models) were developed to detect positive selection
 268 in pre-specified lineages that affects only a few sites in the protein. It is useful to detect
 269 episodic positive selection and generate biological hypotheses for mutation studies and
 270 functional analyses (Yang and Dos Reis, 2011).

271 In these models, the branches that examined for positive selection are called foreground
 272 branches, while all others are the background branches (Yang and Nielsen, 2002; Zhang,
 273 2005). The parameters used in branch-site models are in **Table 2**. In site class 0, the codons
 274 are under negative selection ($0 < \omega_0 < 1$) throughout the tree. In the site class 1, the codons are
 275 under neutral evolution throughout the tree ($\omega_1 = 1$). In site classes 2a and 2b, the codons in
 276 the foreground are under positive selection ($\omega_2 > 1$) while in background they are under
 277 negative selection (class 2a) or neutral evolution (class 2b). The selection model A allows the
 278 classes 2a and 2b (positive selection in foreground) and the null model fix ω_2 to 1 (Zhang,
 279 2005). The LRT between model A and the null model is calculated with one degree of
 280 freedom (Zhang, 2005). Again the Bayes Empirical Bayes (BEB) method can be used to
 281 identify sites under positive selection (Yang, 2005).

282

283 **Table 2** | Parameters of the branch-site model A

Site class	Proportion	Background	Foreground
0	p_0	$0 < \omega_0 < 1$	$0 < \omega_0 < 1$
1	p_1	$\omega_1 = 1$	$\omega_1 = 1$
2a	$(1 - p_0 - p_1) p_0 / (p_0 + p_1)$	$0 < \omega_0 < 1$	$\omega_2 > 1$
2b	$(1 - p_0 - p_1) p_0 / (p_0 + p_1)$	$\omega_1 = 1$	$\omega_2 > 1$

284

285 Branch-site tests are very conservative and robust, but lacks power under synonymous
286 substitution saturation and variation in GC content (Gharib and Robinson-Rechavi, 2013;
287 Yang and Dos Reis, 2011).

288

289 **6. Softwares and pipelines for positive selection analysis**

290 The software CodeML in PAML package is used to perform LRTs using site, branch and
291 branch-site models (Yang, 2007). Two improvements were developed for codeml: gcodeml
292 distributes the computational jobs to different machines (grid) (Moretti et al., 2012) and
293 FastCodeML has optimization techniques to speed up the analysis and to analyze substantially
294 larger datasets (Valle et al., 2014).

295 On a genome scale this analysis is computationally costly, but pipelines are being
296 developed to perform this task in a general, automatic, large-scale and reliable manner
297 (Hongo et al., 2015). A list of pipelines and characteristics is in **Table 3**. Within these ones,
298 POTION (Hongo et al., 2015) is a pipeline that uses site models and PosiGene uses branch-
299 site models (Sahm et al., 2017).

Table 3 | Features of softwares used for positive selection analysis in genome scale

Features	Datamonkey	Selecton	JCoDA	IDEA	PhyleasProg	PSP	POTION	PosiGene
Ortholog detection	-	-	-	-	+	+	+-	+
Coding sequence alignments	-	+	+	-	+	+	+	+
Phylogenetic tree reconstruction	+	+	+	+	+	+	+	+
Use of multiple CPU cores	-	-	-	+	+	+	+	+
Sequence and group filtering steps	+-	-	-	-	+-	+	+	+
User of any user-defined input data	+	+	+	+	-	-	+	+
Site tests	+	+	+	+	+	+	+	-
Branch-site tests	+	-	-	+	+	+	-	+
Visualization of positively selected sites within alignment	+	+	+	+	-	-	+	+

References of each software: Datamonkey (Delpont et al., 2010), Selecton (Stern et al., 2007), JCoDA (Steinway et al., 2010), IDEA (Egan et al., 2008), PhyleasProg (Busset et al., 2011), PSP (Su et al., 2013), POTION (Hongo et al., 2015), PosiGene (Sahm et al., 2017).

7. Positive selection analysis in bacteria

Bacteria can be free living or adapted to colonize animals and plants, as mutualists or pathogens (Toft and Andersson, 2010). To colonize a host, bacteria use mechanisms involved in adhesion, mobility, evasion of host immune system, intra and extracellular survival, nutrient acquisition and competition with other microorganisms (Hill, 2012; Webb and Kahler, 2008). Additionally, pathogenic species have virulence factors, defined as structures and strategies that i) allows the bacteria to invade and survive in normally noncolonized body sites or cellular compartments, ii) cause damage to the body, iii) cause dysregulation of the immune system to the extent of creating disease symptoms, or iv) cause a neurological response that again leads to disease symptoms mechanisms that cause damage to the host. Those virulence factors include internalins and invasins, toxins, superantigens and neurotoxins (Hill, 2012). Genome variations were associated with adaptations to different lifestyles, from free living to obligate intracellular pathogen, and positive selection plays a role in those adaptations (Toft and Andersson, 2010). A list of bacterial species tested for positive selection is presented in **Table 4**.

Table 4 | Species of bacteria analyzed for positive selection

Species	Reference
<i>Actinobacillus pleuropneumoniae</i>	(Xu et al., 2011)
<i>Bacillus cereus</i>	(Su et al., 2013)
<i>Bacillus subtilis</i>	(Kopac et al., 2014)
<i>Bartonella</i>	(Nystedt et al., 2008)
<i>Brucella abortus</i>	(Kim et al., 2011; Vishnu et al., 2015; Yang et al., 2016)
<i>B. canis</i>	(Kim et al., 2011; Vishnu et al., 2015; Yang et al., 2016)
<i>B. ceti</i>	(Vishnu et al., 2015; Yang et al., 2016)
<i>B. melitensis</i>	(Kim et al., 2011; Vishnu et al., 2015; Yang et al., 2016)
<i>B. microti</i>	(Vishnu et al., 2015; Yang et al., 2016)
<i>B. ovis</i>	(Kim et al., 2011; Vishnu et al., 2015; Yang et al., 2016)
<i>B. pinnipedialis</i>	(Vishnu et al., 2015; Yang et al., 2016)
<i>B. suis</i>	(Kim et al., 2011; Vishnu et al., 2015; Yang et al., 2016)
<i>Campylobacter</i> (17 species)	(Lefébure and Stanhope, 2009)
<i>Chlamydia trachomatis</i>	(Joseph et al., 2011)
<i>Escherichia coli</i>	(Chen et al., 2006; Petersen et al., 2007; Su et al., 2013)
<i>Leptospira interrogans</i>	(Xu et al., 2016)
<i>Mycobacterium abscessus</i>	(Su et al., 2013; Tan et al., 2017)
<i>M. tuberculosis</i>	(Hongo et al., 2015; Osório et al., 2013; Wang et al., 2017; Zhang et al., 2011, 2013)
<i>M. bovis</i>	(Osório et al., 2013)
<i>Salmonella enterica</i>	(Soyer et al., 2009)
<i>Staphylococcus aureus</i>	(Guinane et al., 2010)
<i>Streptococcus agalactiae</i>	(Anisimova et al., 2007)

<i>S. mutans</i>	(Anisimova et al., 2007)
<i>S. pneumoniae</i>	(Anisimova et al., 2007)
<i>S. pyogenes</i>	(Anisimova et al., 2007)
<i>S. thermophilus</i>	(Anisimova et al., 2007)

7.1. Host colonization

Positive selection in bacteria is usually detected in genes that are involved in interactions with hosts, phages or other bacteria (Petersen et al., 2007; Su et al., 2013). These genes are involved in regulation, modulation and modification of the host immune response, membrane lipid metabolism, cell wall processes, and receptor mediated binding (Hongo et al., 2015). Core and accessory genes can be under positive selection, as they participate in complex networks that comprise the molecular basis of virulence (Su et al., 2013). These genes also can have specific patterns of expression, as found in *Streptococcus* tissue specific invasion (Anisimova et al., 2007; Su et al., 2013). An adaptive mutation is not always related to the main function of the protein. An amino acid changing in a transporter protein can be positively selected due to avoiding phage or antibody binding (Petersen et al., 2007).

Outer membranes proteins are commonly found to undergo positive selection, as they are preferential targets for host immune-system-related selective pressure (Osório et al., 2013; Xu et al., 2011) and there is a requirement for antigen variation (Zhang et al., 2011). Amino acid substitutions in porins, such as Omp proteins, and ferrichrome receptor FhuA were suggested to be involved in escaping recognition by the host immune system and phage binding, but also to avoid the uptake of antibiotics (Petersen et al., 2007; Soyer et al., 2009). Selection target not only proteins, but also processes.

Positive selection can result in adaptations beyond the protein structure. Lipopolysaccharide core region interacts with the host immune system and the proteins involved in its synthesis were also detected in *E. coli* (Petersen et al., 2007). SigM, an extracytoplasmatic sigma factor, regulates positively or negatively the expression of surface and secreted molecules involved in host-pathogen interaction and was detected in *M. tuberculosis* (Zhang et al., 2011).

Proteins related to adhesion, invasion and biofilm formation also have been detected (Petersen et al., 2007). Diversifying selection in a Type 4 Secretion System of *Bartonella* was suggested to change its function from plasmid conjugation to adherence to a divergent set of erythrocyte surface structures, allowing the invasion of those cells (Nystedt et al., 2008).

Analysis of *Staphylococcus aureus* genomes identified mutations involved in host adaptation of clonal complex CC133 from humans to ruminants. Gene decay and diversifying selection was found in proteins associated with cell wall, adherence, toxin production, metabolism, replication and repair and gene regulation (Guinane et al., 2010). Positive selection was found to contribute to host preference in *Brucella* (Kim et al., 2011), *Lepstospira* (Xu et al., 2016), and host restricted *Salmonella* serotypes Typhi and Paratyphi (Soyer et al., 2009).

7.2. Drug resistance, targeting and development

Adaptive mutations can contribute to drug resistance by avoiding binding or uptake (Su et al., 2013). Genes directly involved in drug resistance has been identified, as the multidrug resistance cluster *mdtABCD* of *E. coli* (Petersen et al., 2007) and the known resistance related genes *katG*,

rpoB, and *embB* in *M. tuberculosis*, validating the site test methods (Osório et al., 2013). An analysis of 161 *M. tuberculosis* isolates with a range of drug resistance profiles found a near-complete set of drug resistance-associated genes (Zhang et al., 2013). The positively selected sites found in *rpoB* of *Mycobacterium* the analysis of its protein structure allowed the prediction of a new drug binding site (Wang et al., 2017).

7.3. Speciation

By contributing to with adaptation to new niches, positive selection is also involved in bacterial speciation (Lassalle et al., 2015). Genomes of *Bacillus subtilis* subsp. *spizizenii* isolated from warmer and sunnier and isolated from cooler and shadier microhabitats were tested for positive selection. Diversifying selection was identified in 14 genes, while 38 genes were identified as undergoing directional selection across the different branches. Most genes were involved in metabolism, but also cell wall synthesis, transport and defense. Each strain was under different regimen of positive selection, suggesting distinct speciation processes and no unique gene conferred a metabolic system or subsystem function that was not already present in all isolates. The speciation rate was predicted to be faster than can be resolved with multilocus sequencing (Kopac et al., 2014).

8. Conclusion

Genomic data is being used to understand the contribution of natural selection to molecular evolution. In this review, we described the main methods used for identifying adaptation at molecular level in bacteria and the contributions of these methods to biology research and control of these organisms. The identification of selective pressures on the sequences of shared genes can complement the analysis of gene distribution and expression across bacterial strains. This information can be used to understanding the mechanisms involved in adaptation for an ecological niche and have applications for health and biotechnology.

Author Contributions

MV, AS, ALS, RP, AN and RW wrote the manuscript. AS, HF and VA revised for its integrity and accuracy. VA approved the final version of this manuscript.

Funding

This study was supported by: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (www.capes.gov.br), Conselho Nacional de Desenvolvimento Científico e Tecnológico (cnpq.br), and Pró-Reitoria de Pesquisa e Extensão of Universidade Federal de Minas Gerais (www.ufmg.br/prpq). A.R. Wattam was supported in part by federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health (www.nih.gov).

Conflict of Interest

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

9. References

- Altenhoff, A. M., and Dessimoz, C. (2009). Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput. Biol.* 5. doi:10.1371/journal.pcbi.1000262.
- Anisimova, M., Bielawski, J., Dunn, K., and Yang, Z. (2007). Phylogenomic analysis of natural selection pressure in *Streptococcus* genomes. *BMC Evol. Biol.* 7, 1–13. doi:10.1186/1471-2148-7-154.
- Anisimova, M., Bielawski, J. P., and Yang, Z. (2002). Accuracy and Power of Bayes Prediction of Amino Acid Sites Under Positive Selection. *Mol. Biol. Evol.* 19, 950–958. doi:10.1093/oxfordjournals.molbev.a004152.
- Anisimova, M., Nielsen, R., and Yang, Z. (2003). Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* 164, 1229–1236. doi:10.1093/bioinformatics/btn086.
- Anisimova, M., and Yang, Z. (2007). Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. *Mol. Biol. Evol.* 24, 1219–1228. doi:10.1093/molbev/msm042.
- Arenas, M. (2015). Trends in substitution models of molecular evolution. *Front. Genet.* 6. doi:10.3389/fgene.2015.00319.
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57, 289–300. doi:10.2307/2346101.
- Biswas, S., and Akey, J. M. (2006). Genomic insights into positive selection. *Trends Genet.* 22, 437–446. doi:10.1016/j.tig.2006.06.005.
- Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C. H., Xie, D., et al. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput. Biol.* 10, 1–6. doi:10.1371/journal.pcbi.1003537.
- Bruen, T. C., Philippe, H. H., and Bryant, D. (2006). A simple and robust statistical test for detecting the presence of recombination. *Genetics* 172, 2665–81. doi:10.1534/genetics.105.048975.
- Busset, J., Cabau, C., Meslin, C., and Pascal, G. (2011). PhyleasProg: a user-oriented web server for wide evolutionary analyses. *Nucleic Acids Res.* 39, W479–W485. doi:10.1093/nar/gkr243.
- Casillas, S., and Barbadilla, A. (2017). Molecular population genetics. *Genetics* 205, 1003–1035. doi:10.1534/genetics.116.196493.
- Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Mol. Biol. Evol.* 17, 540–552.

doi:10.1093/oxfordjournals.molbev.a026334.

- Chen, F., Mackey, A. J., Vermunt, J. K., and Roos, D. S. (2007). Assessing Performance of Orthology Detection Strategies Applied to Eukaryotic Genomes. *PLoS One* 2, e383. doi:10.1371/journal.pone.0000383.
- Chen, S. L., Hung, C., Xu, J., Reigstad, C. S., Magrini, V., Sabo, A., et al. (2006). Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach. *Proc. Natl. Acad. Sci. U. S. A.* 103, 5977–82. doi:10.1073/pnas.0600938103.
- Delport, W., Poon, A. F. Y., Frost, S. D. W., and Kosakovsky Pond, S. L. (2010). Datamonkey 2010: A suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* 26, 2455–2457. doi:10.1093/bioinformatics/btq429.
- Do, C. B., Mahabhashyam, M. S. P., Brudno, M., and Batzoglou, S. (2005). ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.* 15, 330–340. doi:10.1101/gr.2821705.1994.
- Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5, 113. doi:10.1186/1471-2105-5-113.
- Egan, A., Mahurkar, A., Crabtree, J., Badger, J. H., Carlton, J. M., and Silva, J. C. (2008). IDEA: Interactive Display for Evolutionary Analyses. *BMC Bioinformatics* 9, 1–9. doi:10.1186/1471-2105-9-524.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* 17, 368–376. doi:10.1007/BF01734359.
- Fischer, S., Brunk, B. P., Chen, F., Gao, X., Harb, O. S., Iodice, J. B., et al. (2011). “Using OrthoMCL to Assign Proteins to OrthoMCL-DB Groups or to Cluster Proteomes Into New Ortholog Groups,” in *Current Protocols in Bioinformatics* (Hoboken, NJ, USA: John Wiley & Sons, Inc.), 1–23. doi:10.1002/0471250953.bi0612s35.
- Fletcher, W., and Yang, Z. (2010). The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol. Biol. Evol.* 27, 2257–2267. doi:10.1093/molbev/msq115.
- Gharib, W. H., and Robinson-Rechavi, M. (2013). The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC. *Mol. Biol. Evol.* 30, 1675–1686. doi:10.1093/molbev/mst062.
- Goldman, N., and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11, 725–36. doi:10.1186/s13059-014-0542-8.
- Guinane, C. M., Ben Zakour, N. L., Tormo-Mas, M. A., Weinert, L. A., Lowder, B. V., Cartwright, R. A., et al. (2010). Evolutionary genomics of *Staphylococcus aureus* reveals insights into the origin and molecular basis of ruminant host adaptation. *Genome Biol. Evol.* 2, 454–66. doi:10.1093/gbe/evq031.
- Hill, C. (2012). Virulence or Niche Factors: What’s in a Name? *J. Bacteriol.* 194, 5725–5727. doi:10.1128/JB.00980-12.

- Hongo, J. A., de Castro, G. M., Cintra, L. C., Zerlotini, A., and Lobo, F. P. (2015). POTION: an end-to-end pipeline for positive Darwinian selection detection in genome-scale data through phylogenetic comparison of protein-coding genes. *BMC Genomics* 16, 567. doi:10.1186/s12864-015-1765-0.
- Jakobsen, I. B., and Easteal, S. (1996). A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *Comput. Appl. Biosci.* 12, 291–295. doi:10.1093/bioinformatics/12.4.291.
- Jordan, G., and Goldman, N. (2012). The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol. Biol. Evol.* 29, 1125–1139. doi:10.1093/molbev/msr272.
- Joseph, S. J., Didelot, X., Gandhi, K., Dean, D., and Read, T. D. (2011). Interplay of recombination and selection in the genomes of *Chlamydia trachomatis*. *Biol. Direct* 6, 28. doi:10.1186/1745-6150-6-28.
- Katoh, K., Kuma, K. I., Toh, H., and Miyata, T. (2005). MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 33, 511–518. doi:10.1093/nar/gki198.
- Kim, K. W. M., Kim, K. W. M., Sung, S., and Kim, H. (2011). A genome-wide identification of genes potentially associated with host specificity of *Brucella* species. *J. Microbiol.* 49, 768–775. doi:10.1007/s12275-011-1084-3.
- Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* 217, 624–6. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/5637732>.
- Koonin, E. V. (2005). Orthologs, Paralogs, and Evolutionary Genomics 1. *Annu. Rev. Genet.* 39, 309–338. doi:10.1146/annurev.genet.39.073003.114725.
- Kopac, S., Wang, Z., Wiedenbeck, J., Sherry, J., Wu, M., and Cohan, F. M. (2014). Genomic Heterogeneity and Ecological Speciation within One Subspecies of *Bacillus subtilis*. *Appl. Environ. Microbiol.* 80, 4842–4853. doi:10.1128/AEM.00576-14.
- Kosakovsky Pond, S. L., and Frost, S. D. W. (2005). Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* 22, 1208–1222. doi:10.1093/molbev/msi105.
- Kristensen, D. M. M., Wolf, Y. I. I., Mushegian, A. R. R., and Koonin, E. V. V. (2011). Computational methods for Gene Orthology inference. *Brief. Bioinform.* 12, 379–391. doi:10.1093/bib/bbr030.
- Kryazhimskiy, S., and Plotkin, J. B. (2008). The population genetics of dN/dS. *PLoS Genet.* 4. doi:10.1371/journal.pgen.1000304.
- Lassalle, F., Muller, D., and Nesme, X. (2015). Ecological speciation in bacteria: reverse ecology approaches reveal the adaptive part of bacterial cladogenesis. *Res. Microbiol.* 166, 729–41. doi:10.1016/j.resmic.2015.06.008.
- Lefebvre, T., and Stanhope, M. J. (2007). Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome Biol.* 8,

R71. doi:10.1186/gb-2007-8-5-r71.

- Lefébure, T., and Stanhope, M. J. (2009). Pervasive, genome-wide positive selection leading to functional divergence in the bacterial genus *Campylobacter*. *Genome Res.* 19, 1224–1232. doi:10.1101/gr.089250.108.
- Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–89. doi:10.1101/gr.1224503.
- Loytynoja, A., and Goldman, N. (2008). Phylogeny-Aware Gap Placement Prevents Errors in Sequence Alignment and Evolutionary Analysis. *Science (80-)*. 320, 1632–1635. doi:10.1126/science.1158395.
- Mahajan, S., and Agashe, D. (2018). Translational Selection for Speed is Not Sufficient to Explain Variation in Bacterial Codon Usage Bias. *Genome Biol. Evol.* 10, 562–576. doi:10.1093/gbe/evy018.
- Mallick, S., Gnerre, S., Muller, P., and Reich, D. (2009). The difficulty of avoiding false positives in genome scans for natural selection. *Genome Res.* 19, 922–33. doi:10.1101/gr.086512.108.
- Markova-Raina, P., and Petrov, D. (2011). High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 *Drosophila* genomes. *Genome Res.* 21, 863–74. doi:10.1101/gr.115949.110.
- McDonald, J. H., and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351, 652–654. doi:10.1038/351652a0.
- Miyata, T., and Yasunaga, T. (1980). Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J. Mol. Evol.* 16, 23–36. doi:10.1007/BF01732067.
- Moretti, S., Murri, R., Maffioletti, S., Kuzniar, A., Castella, B., Salamin, N., et al. (2012). Gcodeml: A grid-enabled tool for detecting positive selection in biological evolution. *Stud. Health Technol. Inform.* 175, 59–68. doi:10.3233/978-1-61499-054-3-59.
- Nei, M., and Gojoborit, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3, 418–426. doi:10.1093/oxfordjournals.molbev.a040410.
- Nielsen, R. (2001). Statistical tests of selective neutrality in the age of genomics. *Heredity (Edinb)*. 86, 641–647. doi:10.1046/j.1365-2540.2001.00895.x.
- Nielsen, R., and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148, 929–936.
- Noble, W. S. (2009). How does multiple testing correction work? *Nat. Biotechnol.* 27, 1135–1137. doi:10.1038/nbt1209-1135.
- Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302, 205–217. doi:10.1006/jmbi.2000.4042.

- Nystedt, B., Frank, A. C., Thollessen, M., and Andersson, S. G. E. (2008). Diversifying selection and concerted evolution of a type IV secretion system in *Bartonella*. *Mol. Biol. Evol.* 25, 287–300. doi:10.1093/molbev/msm252.
- Orsi, R. H., Ripoll, D. R., Yeung, M., Nightingale, K. K., and Wiedmann, M. (2007). Recombination and positive selection contribute to evolution of *Listeria monocytogenes* inLA. *Microbiology* 153, 2666–78. doi:10.1099/mic.0.2007/007310-0.
- Orsi, R. H., Sun, Q., and Wiedmann, M. (2008). Genome-wide analyses reveal lineage specific contributions of positive selection and recombination to the evolution of *Listeria monocytogenes*. *BMC Evol. Biol.* 8, 233. doi:10.1186/1471-2148-8-233.
- Osório, N. S., Rodrigues, F., Gagneux, S., Pedrosa, J., Pinto-Carbó, M., Castro, A. G., et al. (2013). Evidence for diversifying selection in a set of *Mycobacterium tuberculosis* genes in response to antibiotic- and nonantibiotic-related pressure. *Mol. Biol. Evol.* 30, 1326–1336. doi:10.1093/molbev/mst038.
- Pavlidis, P., and Alachiotis, N. (2017). A survey of methods and tools to detect recent and strong positive selection. *J. Biol. Res.* 24, 7. doi:10.1186/s40709-017-0064-0.
- Penn, O., Privman, E., Landan, G., Graur, D., and Pupko, T. (2010). An alignment confidence score capturing robustness to guide tree uncertainty. *Mol. Biol. Evol.* 27, 1759–1767. doi:10.1093/molbev/msq066.
- Petersen, L., Bollback, J. P., Dimmic, M., Hubisz, M., and Nielsen, R. (2007). Genes under positive selection in *Escherichia coli*. *Genome Res.* 17, 1336–1343. doi:10.1101/gr.6254707.
- Posada, D., Crandall, K. A., and Holmes, E. C. (2002). Recombination in Evolutionary Genomics. *Annu. Rev. Genet.* 36, 75–97. doi:10.1146/annurev.genet.36.040202.111115.
- Privman, E., Penn, O., and Pupko, T. (2012). Improving the performance of positive selection inference by filtering unreliable alignment regions. *Mol. Biol. Evol.* 29, 1–5. doi:10.1093/molbev/msr177.
- Sahm, A., Bens, M., Platzer, M., and Szafranski, K. (2017). PosiGene: automated and easy-to-use pipeline for genome-wide detection of positively selected genes. *Nucleic Acids Res.* 45, 1–11. doi:10.1093/nar/gkx179.
- Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–25. doi:10.1093/oxfordjournals.molbev.a040454.
- Sawyer, S. (1989). Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* 6, 526–38. doi:10.1093/oxfordjournals.molbev.a040567.
- Schneider, A., Souvorov, A., Sabath, N., Landan, G., Gonnet, G. H., and Graur, D. (2009). Estimates of Positive Darwinian Selection Are Inflated by Errors in Sequencing, Annotation, and Alignment. *Genome Biol. Evol.* 1, 114–118. doi:10.1093/gbe/evp012.
- Schwartz, A. S., and Pachter, L. (2007). Multiple alignment by sequence annealing. *Bioinformatics* 23, 24–29. doi:10.1093/bioinformatics/btl311.

- Shriner, D., Nickle, D. C., Jensen, M. A., and Mullins, J. I. (2003). Potential impact of recombination on sitewise approaches for detecting positive natural selection. *Genet. Res.* 81, 115–121. doi:10.1017/S0016672303006128.
- Smith, J. M. (1992). Analyzing the mosaic structure of genes. *J. Mol. Evol.* 34, 126–129. doi:10.1007/BF00182389.
- Soyer, Y. Y., Orsi, R. H., Rodriguez-Rivera, L. D., Sun, Q., and Wiedmann, M. (2009). Genome wide evolutionary analyses reveal serotype specific patterns of positive selection in selected *Salmonella* serotypes. *BMC Evol. Biol.* 9, 264. doi:10.1186/1471-2148-9-264.
- Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi:10.1093/bioinformatics/btu033.
- Steinway, S. N., Dannenfelser, R., Laucius, C. D., Hayes, J. E., and Nayak, S. (2010). JCoDA: A tool for detecting evolutionary selection. *BMC Bioinformatics* 11, 1–9. doi:10.1186/1471-2105-11-284.
- Stephan, W. (2010). Detecting strong positive selection in the genome. *Mol. Ecol. Resour.* 10, 863–872. doi:10.1111/j.1755-0998.2010.02869.x.
- Stern, A., Doron-Faigenboim, A., Erez, E., Martz, E., Bacharach, E., and Pupko, T. (2007). Selecton 2007: Advanced models for detecting positive and purifying selection using a Bayesian inference approach. *Nucleic Acids Res.* 35, 506–511. doi:10.1093/nar/gkm382.
- Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* 100, 9440–5. doi:10.1073/pnas.1530509100.
- Studer, R. A., and Robinson-Rechavi, M. (2009). “Large-Scale Analyses of Positive Selection Using Codon Models,” in *Evolutionary Biology* (Berlin, Heidelberg: Springer Berlin Heidelberg), 217–235. doi:10.1007/978-3-642-00952-5_13.
- Su, F., Ou, H. Y., Tao, F., Tang, H., and Xu, P. (2013). PSP: Rapid identification of orthologous coding genes under positive selection across multiple closely related prokaryotic genomes. *BMC Genomics* 14. doi:10.1186/1471-2164-14-924.
- Tan, J. L., Ng, K. P., Ong, C. S., and Ngeow, Y. F. (2017). Genomic Comparisons Reveal Microevolutionary Differences in *Mycobacterium abscessus* Subspecies. *Front. Microbiol.* 8, 1–10. doi:10.3389/fmicb.2017.02042.
- Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–80. doi:10.1093/nar/22.22.4673.
- Thurman, T. J., and Barrett, R. D. H. (2016). The genetic consequences of selection in natural populations. *Mol. Ecol.* 25, 1429–1448. doi:10.1111/mec.13559.
- Toft, C., and Andersson, S. G. E. (2010). Evolutionary microbial genomics: Insights into bacterial host adaptation. *Nat. Rev. Genet.* 11, 465–475. doi:10.1038/nrg2798.
- Valle, M., Schabauer, H., Pacher, C., Stockinger, H., Stamatakis, A., Robinson-Rechavi, M., et

- al. (2014). Optimization strategies for fast detection of positive selection on phylogenetic trees. *Bioinformatics* 30, 1129–1137. doi:10.1093/bioinformatics/btt760.
- Villanueva-Cañas, J. L., Laurie, S., and Albà, M. M. (2013). Improving genome-wide scans of positive selection by using protein isoforms of similar length. *Genome Biol. Evol.* 5, 457–467. doi:10.1093/gbe/evt017.
- Vishnu, U. S., Sankarasubramanian, J., Sridhar, J., Gunasekaran, P., and Rajendhran, J. (2015). Identification of Recombination and Positively Selected Genes in *Brucella*. *Indian J. Microbiol.* 55, 384–391. doi:10.1007/s12088-015-0545-5.
- Wang, Q., Xu, Y., Gu, Z., Liu, N., Jin, K., Li, Y., et al. (2017). Identification of new antibacterial targets in RNA polymerase of *Mycobacterium tuberculosis* by detecting positive selection sites. *Comput. Biol. Chem.* doi:10.1016/j.compbiolchem.2017.11.002.
- Webb, S. A., and Kahler, C. M. (2008). Bench-to-bedside review: Bacterial virulence and subversion of host defences. *Crit. Care* 12, 234. doi:10.1186/cc7091.
- Wolf, Y. I., and Koonin, E. V. (2012). A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. *Genome Biol. Evol.* 4, 1286–1294. doi:10.1093/gbe/evs100.
- Wong, K. M., Suchard, M. A., and Huelsenbeck, J. P. (2008). Alignment Uncertainty and Genomic Analysis Alignment Uncertainty and Genomic Analysis : Supplemental Material. 473, 1–4. doi:10.1126/science.1151532.
- Wong, W. S. W. (2004). Accuracy and Power of Statistical Methods for Detecting Adaptive Evolution in Protein Coding Sequences and for Identifying Positively Selected Sites. *Genetics* 168, 1041–1051. doi:10.1534/genetics.104.031153.
- Xu, Y., Zhu, Y. Y., Wang, Y., Chang, Y.-F., Zhang, Y., Jiang, X., et al. (2016). Whole genome sequencing revealed host adaptation-focused genomic plasticity of pathogenic *Leptospira*. *Sci. Rep.* 6, 20020. doi:10.1038/srep20020.
- Xu, Z., Chen, H., and Zhou, R. (2011). Genome-wide evidence for positive selection and recombination in *Actinobacillus pleuropneumoniae*. *BMC Evol. Biol.* 11, 203. doi:10.1186/1471-2148-11-203.
- Yang, X., Li, Y., Zang, J., Li, Y., Bie, P., Lu, Y., et al. (2016). Analysis of pan-genome to identify the core genes and essential genes of *Brucella* spp. *Mol. Genet. Genomics* 291, 905–912. doi:10.1007/s00438-015-1154-z.
- Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* 15, 568–573. doi:10.1093/oxfordjournals.molbev.a025957.
- Yang, Z. (2005). Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection. *Mol. Biol. Evol.* 22, 1107–1118. doi:10.1093/molbev/msi097.
- Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi:10.1093/molbev/msm088.

- Yang, Z., and Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* 15, 496–503. doi:10.1016/S0169-5347(00)01994-7.
- Yang, Z., and Dos Reis, M. (2011). Statistical properties of the branch-site test of positive selection. *Mol. Biol. Evol.* 28, 1217–1228. doi:10.1093/molbev/msq303.
- Yang, Z., and Nielsen, R. (1998). Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.* 46, 409–18. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/9541535>.
- Yang, Z., and Nielsen, R. (2000). Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models. *Mol. Biol. Evol.* 17, 32–43.
- Yang, Z., and Nielsen, R. (2002). Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages. *Mol. Biol. Evol.* 19, 908–917. doi:10.1093/oxfordjournals.molbev.a004148.
- Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A. M. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155, 431–49. doi:10.1093/oxfordjournals.molbev.a003981.
- Zhang, H., Li, D., Zhao, L., Fleming, J., Lin, N., Wang, T., et al. (2013). Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from China identifies genes and intergenic regions associated with drug resistance. *Nat. Genet.* 45, 1255–1260. doi:10.1038/ng.2735.
- Zhang, J. (2005). Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. *Mol. Biol. Evol.* 22, 2472–2479. doi:10.1093/molbev/msi237.
- Zhang, Y., Zhang, H., Zhou, T., Zhong, Y., and Jin, Q. (2011). Genes under positive selection in *Mycobacterium tuberculosis*. *Comput. Biol. Chem.* 35, 319–322. doi:10.1016/j.compbiolchem.2011.08.001.

IV.3 Chapter III. Rapidly evolving changes and gene loss associated with host switching in *Corynebacterium pseudotuberculosis*

Marcus Vinicius Canário Viana, Arne Sahm, Aristóteles Góes Neto, Henrique Cesar Pereira Figueiredo, Alice Rebecca Wattam, Vasco Azevedo. **Submitted to PLOS Pathogens.**

To accomplish the goal of identifying genes under positive selection in the *Corynebacterium pseudotuberculosis* species, we performed an analysis of the selective pressures and variation on genes from 29 *C. pseudotuberculosis* genomes that were isolated from different hosts, including representatives of both the Ovis and Equi biovars. Using a genome scale positive selection pipeline with branch-site models, we looked specifically at genes under positive selection in specific branches of *C. pseudotuberculosis* phylogeny. A total of 28 genes were identified, involved in metabolism, cell division, resistance, transport, adhesion, exposed on the cell surface or hypothetical proteins with unknown functions. Some of the genes are suggested to be drug and vaccine targets for future exploration. This data was combined with the literature to explain both biovars as independent groups based on niche diversification.

PLOS ONE

Rapidly evolving changes and gene loss associated with host switching in *Corynebacterium pseudotuberculosis*

--Manuscript Draft--

Manuscript Number:	PONE-D-18-14082
Article Type:	Research Article
Full Title:	Rapidly evolving changes and gene loss associated with host switching in <i>Corynebacterium pseudotuberculosis</i>
Short Title:	Host switching and adaptations in <i>Corynebacterium pseudotuberculosis</i>
Corresponding Author:	Marcus Vinicius Canário Viana, Ph.D. Universidade Federal de Minas Gerais Belo Horizonte, Minas Gerais BRAZIL
Keywords:	positive selection; phylogenomics; <i>Corynebacterium</i>
Abstract:	Phylogenomics and genome scale positive selection analyses were performed on 29 <i>Corynebacterium pseudotuberculosis</i> strains that were isolated from different hosts. Representatives of both the Ovis and Equi biovars were included. Ovis is a derived, monophyletic, clonal and highly specialized group, while Equi is a primitive paraphyletic group with higher diversity and capable to infect exclusive hosts. A total of 27 genes were identified as undergoing adaptive changes, and some of them had homology to known virulence factors and drug targets. In addition, we describe the biovars as independent groups based on ecological diversification.
Order of Authors:	Marcus Vinicius Canário Viana, Msc Arne Sahn Aristóteles Góes Neto Henrique Cesar Pereira Figueiredo Alice Rebecca Wattam Vasco Azevedo
Opposed Reviewers:	
Additional Information:	
Question	Response
Financial Disclosure Please describe all sources of funding that have supported your work. This information is required for submission and will be published with your article, should it be accepted. A complete funding statement should do the following: Include grant numbers and the URLs of any funder's website. Use the full name, not acronyms, of funding institutions, and use initials to identify authors who received the funding. Describe the role of any sponsors or funders in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. If the funders had no role in any of the above, include this sentence at the end of	This work was supported by: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (www.capes.gov.br), Conselho Nacional de Desenvolvimento Científico e Tecnológico (cnpq.br), and Pró-Reitoria de Pesquisa e Extensão of Universidade Federal de Minas Gerais (www.ufmg.br/prpq). A.R. Wattam was supported in part by federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health (www.nih.gov). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

<p>your statement: "<i>The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.</i>"</p> <p>However, if the study was unfunded, please provide a statement that clearly indicates this, for example: "<i>The author(s) received no specific funding for this work.</i>"</p> <p>* typeset</p>	
<p>Competing Interests</p> <p>You are responsible for recognizing and disclosing on behalf of all authors any competing interest that could be perceived to bias their work, acknowledging all financial support and any other relevant financial or non-financial competing interests.</p> <p>Do any authors of this manuscript have competing interests (as described in the PLOS Policy on Declaration and Evaluation of Competing Interests)?</p> <p>If yes, please provide details about any and all competing interests in the box below. Your response should begin with this statement: <i>I have read the journal's policy and the authors of this manuscript have the following competing interests:</i></p> <p>If no authors have any competing interests to declare, please enter this statement in the box: "<i>The authors have declared that no competing interests exist.</i>"</p> <p>* typeset</p>	<p>The authors have declared that no competing interests exist.</p>
<p>Ethics Statement</p> <p>You must provide an ethics statement if your study involved human participants, specimens or tissue samples, or vertebrate animals, embryos or tissues. All information entered here should also be included in the Methods section of your manuscript. Please write "N/A" if your study does not require an ethics</p>	<p>NA</p>

statement.

Human Subject Research (involved human participants and/or tissue)

All research involving human participants must have been approved by the authors' Institutional Review Board (IRB) or an equivalent committee, and all clinical investigation must have been conducted according to the principles expressed in the [Declaration of Helsinki](#). Informed consent, written or oral, should also have been obtained from the participants. If no consent was given, the reason must be explained (e.g. the data were analyzed anonymously) and reported. The form of consent (written/oral), or reason for lack of consent, should be indicated in the Methods section of your manuscript.

Please enter the name of the IRB or Ethics Committee that approved this study in the space below. Include the approval number and/or a statement indicating approval of this research.

Animal Research (involved vertebrate animals, embryos or tissues)

All animal work must have been conducted according to relevant national and international guidelines. If your study involved non-human primates, you must provide details regarding animal welfare and steps taken to ameliorate suffering; this is in accordance with the recommendations of the Weatherall report, "[The use of non-human primates in research](#)." The relevant guidelines followed and the committee that approved the study should be identified in the ethics statement.

If anesthesia, euthanasia or any kind of animal sacrifice is part of the study, please include briefly in your statement which substances and/or methods were applied.

Please enter the name of your Institutional Animal Care and Use Committee (IACUC) or other relevant ethics board, and indicate whether they approved this

<p>research or granted a formal waiver of ethical approval. Also include an approval number if one was obtained.</p> <p>Field Permit</p> <p>Please indicate the name of the institution or the relevant body that granted permission.</p>	
<p>Data Availability</p> <p>PLOS journals require authors to make all data underlying the findings described in their manuscript fully available, without restriction and from the time of publication, with only rare exceptions to address legal and ethical concerns (see the PLOS Data Policy and FAQ for further details). When submitting a manuscript, authors must provide a Data Availability Statement that describes where the data underlying their manuscript can be found.</p> <p>Your answers to the following constitute your statement about data availability and will be included with the article in the event of publication. Please note that simply stating 'data available on request from the author' is not acceptable. If, however, your data are only available upon request from the author(s), you must answer "No" to the first question below, and explain your exceptional situation in the text box provided.</p> <p>Do the authors confirm that all data underlying the findings described in their manuscript are fully available without restriction?</p>	<p>Yes - all data are fully available without restriction</p>
<p>Please describe where your data may be found, writing in full sentences. Your answers should be entered into the box below and will be published in the form you provide them, if your manuscript is accepted. If you are copying our sample text below, please ensure you replace any instances of XXX with the appropriate details.</p> <p>If your data are all contained within the paper and/or Supporting Information files, please state this in your answer below. For example, "All relevant data are within the paper and its Supporting Information files." If your data are held or will be held in a public repository, include URLs,</p>	<p>All relevant data are within the paper and its Supporting Information files.</p>

accession numbers or DOIs. For example, "All XXX files are available from the XXX database (accession number(s) XXX, XXX)." If this information will only be available after acceptance, please indicate this by ticking the box below. If neither of these applies but you are able to provide details of access elsewhere, with or without limitations, please do so in the box below. For example:

"Data are available from the XXX Institutional Data Access / Ethics Committee for researchers who meet the criteria for access to confidential data."

"Data are from the XXX study whose authors may be contacted at XXX."

* typeset

Additional data availability information:

Dear editorial board of PLOS ONE,

Please find enclosed the manuscript: “**Rapidly evolving changes and gene loss associated with host switching in *Corynebacterium pseudotuberculosis***” which we are submitting for exclusive consideration of publication as a Research Article in PLOS ONE.

From the time of the first bacterial genome being sequenced, researchers have been trying to identify genes or changes that allow a bacterial pathogen to infect a specific type of host. By using a combination of phylogeny, gene loss, and the identification of genes under positive selection, we have identified genes that we feel allowed the *Corynebacterium pseudotuberculosis* to switch to a new range of hosts. This species has been identified in both large and small ruminants. The phylogeny reflects the host range, and we found patterns in gene loss and adaptive change specific to these branches.

This article is a positive selection analysis on *Corynebacterium pseudotuberculosis*, a Gram-positive bacterium that causes economic losses for sheep and goat breeding and is increasing in frequency and spread in horses in North America. We identified positive selection in genes from eight lineages representing isolates from different hosts. Some of the genes can be used to develop control methods for isolates from specific hosts. We also mapped the literature information on the species phylogeny, to identify variations that could be related to lifestyle and explain the species evolution. We hope that the editorial board and the reviewers will agree on the interest and potential impact of this study.

Thank you for your consideration,

Marcus V. C. Viana, Alice Rebecca Wattam and Vasco Azevedo on behalf of the authors.

Corresponding author:

Vasco Azevedo, DVM, M.Sc, Ph.D, Full Professor, Depto de Biologia Geral, ICB/UFMG.
Av. Antonio Carlos,6627. Pampulha, Belo Horizonte, Minas Gerais, Brazil.
CP 486 CEP 31270-901,

Tel/FAX: +55 31 3409 2610

Online Curriculum Vitae: <http://lattes.cnpq.br/1020477751003832>

skype: [vascoaristondecarvalhoazevedo](https://www.skype.com/profile/vascoaristondecarvalhoazevedo)



1 Rapidly evolving changes and gene loss associated with host switching in *Corynebacterium*
2 *pseudotuberculosis*

3
4 Marcus Vinicius Canário Viana¹, Arne Sahn², Aristóteles Góes Neto³, Henrique Cesar Pereira
5 Figueiredo⁴, Alice Rebecca Wattam^{5¶}, Vasco Azevedo^{1¶*}

6
7 ¹Departament of General Biology, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais,
8 Brazil.

9 ²Leibniz Institute on Aging, Fritz Lipmann Institute, 07745 Jena, Germany.

10 ³Department of Microbiology, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais,
11 Brazil.

12 ⁴AQUACEN, National Reference Laboratory for Aquatic Animal Diseases, Ministry of Fisheries and
13 Aquaculture, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil.

14 ⁵Biocomplexity Institute of Virginia Tech, Virginia Tech, Blacksburg, Virginia, United States of America

15

16 *Corresponding author

17 E-mail: vasco@icb.ufmg.br (VA)

18

19 ¶These authors contributed equally to this work.

20

21

22 **Abstract**

23 Phylogenomics and genome scale positive selection analyses were performed on 29
24 *Corynebacterium pseudotuberculosis* strains that were isolated from different hosts. Representatives of
25 both the Ovis and Equi biovars were included. Ovis is a derived, monophyletic, clonal and highly
26 specialized group, while Equi is a primitive paraphyletic group with higher diversity and capable to infect
27 exclusive hosts. A total of 27 genes were identified as undergoing adaptive changes, and some of them had
28 homology to known virulence factors and drug targets. In addition, we describe the biovars as independent
29 groups based on ecological diversification.

30

31 **Author summary**

32 This article is a positive selection analysis on *Corynebacterium pseudotuberculosis*, a
33 Gram-positive bacterium that causes economic losses for sheep and goat breeding and is increasing in
34 frequency and spread in horses in North America. By using a combination of phylogeny, gene loss, and the
35 identification of genes under positive selection, we have identified genes that we feel allowed the *C.*
36 *pseudotuberculosis* to switch to a new range of hosts. The phylogeny reflects the host range, and we found
37 patterns in gene loss and adaptive change specific to these branches. We identified positive selection in
38 genes from eight lineages representing isolates from different hosts. Some of the genes can be used to
39 develop control methods for isolates from specific hosts. We also mapped the literature information on the
40 species phylogeny, to identify variations that could be related to lifestyle and explain the species evolution.

41

42 **Introduction**

43 Population genetics and genomic approaches increase our understanding of natural selection and
44 molecular evolution. Alleles with adaptive mutations increase in frequency by positive selection, and these
45 mutations can be identified by sequence comparison [1–3]. The use of codon substitution models, which
46 compare non-synonymous (d_N) to the synonymous (d_S) substitution rate, as $\omega = d_N / d_S$, can determine if
47 the mutations that change the amino acid (d_N) in a specific position are adaptive ($\omega > 1$, positive selection),
48 deleterious ($\omega < 1$, negative selection) or neutral ($\omega = 1$, neutral evolution) [4]. Research has shifted from
49 looking at selective pressures on individual genes to pipelines that examine genome scale positive selection
50 [5–7]. The pipelines that are used to examine positive selection often involve orthologous group
51 identification, codon based alignments, phylogenetic tree reconstruction, and fitting codon evolutionary
52 models [3,8].

53 Positive selection has previously been detected in genes that have a role in the host-pathogen
54 dynamic, particularly in the interaction related to the immune and defense mechanisms deployed by the
55 host. Genes determined to be under positive selection are commonly involved in regulation, modulation
56 and modification of the host immune response, membrane lipid metabolism, certain cell wall processes,
57 and receptor mediated binding [6,9]. Several studies have examined selective pressures and the response in
58 many important pathogenic taxa, including *Escherichia coli* [9,10], *Salmonella* [10], *Staphylococcus*
59 *aureus* [11], *Mycobacterium tuberculosis* [12–14], *Shigella flexneri* [9], *Streptococcus* [15], *Campylobacter*
60 [16] and *Leptospira* [17,18].

61 *Corynebacterium pseudotuberculosis* is a Gram-positive, pleomorphic and facultative intracellular
62 bacterium of veterinary and medical relevance that has a global distribution [19]. Infection with *C.*
63 *pseudotuberculosis* strains cause economic losses in animal production. Control methods, such as
64 diagnosis, vaccines and antibiotics remain elusive [19].

65 *C. pseudotuberculosis* is separated into two biovars based on host preference and nitrate reduction.
66 Isolates belonging to the Ovis biovar are nitrate negative, and biovar Equi strains are nitrate positive
67 [20,21]. Each of the biovars can infect a variety of mammals, with a different disease manifestation in each
68 type of host. Biovar Ovis, which causes the chronic disease Caseous Lymphadenitis (CLA) in goats and
69 sheep [20,22] has also been found in cattle [23], camels [21], and humans [20,24,25], causing skin lesions
70 or lymphadenitis. Isolates from the Equi biovar, known for causing Oedematous Skin Disease in buffaloes
71 [26], has also been found in horses [27,28], cattle [23,29] and camels [30], with different manifestations.
72 Cattle and camels are the only hosts that have been found to be infected by both, with a different disease
73 phenotype depending upon the biovar of the infecting agent. Biovar Ovis has not yet been isolated from
74 horses or buffalo, and to date, no sheep or goats have been found to be infected with the Equi biovar.
75 However, an experimental infection of a buffalo isolate (Equi) caused CLA in sheep [31].

76 Previous work has identified changes specific to each of the *C. pseudotuberculosis* biovars [32,33],
77 but to date, no one has been able to identify any differences that might be associated with the host
78 preference. The single exception is the identity of a particular prophage that is only found in strains isolated
79 from buffalo [33]. In this work, we examined the genes under positive selection in both biovars, exploring
80 the evolution of this pathogen and the genes that are associated with host preference.

81

82 **Methods**

83 **Genomes and reannotation**

84 We retrieved from GenBank 29 genomes of *C. pseudotuberculosis* isolated from different hosts,
85 including representatives from the Ovis and Equi biovars (Table 1). The genomes were all consistently
86 reannotated using the RASTtk (Rapid Annotation Using Subsystem Technology) [34] annotation service
87 in the Pathosystem Resource Integration Center (PATRIC) [35].

88 **Table 1. *Corynebacterium pseudotuberculosis* genomes used in positive selection analysis**

Strain	Biovar	Host	Country	Access no
E56	Ovis	Sheep	Egypt	CP013699.1
PA01	Ovis	Sheep	Brazil	CP013327.1
C231	Ovis	Sheep	Australia	CP001829.1
MEX25	Ovis	Sheep	Mexico	CP013697.1
N1	Ovis	Sheep	Equatorial Guinea	CP013146.1
1002B	Ovis	Goat	Brazil	CP012837.1
VD57	Ovis	Goat	Brazil	CP009927.1
PO222/4-1	Ovis	Goat	Portugal	CP013698.1
MEX1	Ovis	Goat	Mexico	CP017711.1
MEX9	Ovis	Goat	Mexico	CP014543.1
P54B96	Ovis	Wildebeest	South Africa	CP003385.1
267	Ovis	Llama	USA	CP003407.1
48252	Ovis	Human	Norway	CP008922.1
FRC41	Ovis	Human	France	CP002097.1
I19	Ovis	Cow	Israel	CP002251.1
29156	Ovis	Cow	Israel	CP010795.1
262	Equi	Cow	Belgium	CP012022.1
I37	Equi	Cow	Israel	CP017384.1
162	Equi	Camel	UK	CP013260.1
258	Equi	Horse	Belgium	CP003540.2
MB14	Equi	Horse	USA	CP013261.1
E19	Equi	Horse	Chile	CP003540.2

MEX30	Equi	Horse	Mexico	CP017291.1
CIP52.97	Equi	Horse	Kenya	CP003061.2
31	Equi	Buffalo	Egypt	CP003421.3
32	Equi	Buffalo	Egypt	CP015183.1
33	Equi	Buffalo	Egypt	CP015184.1
36	Equi	Buffalo	Egypt	CP015186.1
48	Equi	Buffalo	Egypt	CP015191.1

89

90

91 **Phylogenomic and phylogenetic species trees**

92 The phylogeny of *C. pseudotuberculosis* strains were estimated using three different methods, with
 93 *Corynebacterium ulcerans* strain 210932 (CP009500.1) [36] used to root the tree. The estimated branches
 94 were used to determine the groups to be analyzed for positive selection. A phylogenomic tree of the 29
 95 genomes was generated using PEPR (<https://github.com/enordber/pepr.git>), which uses the core proteome
 96 of the selected genomes and includes a progressive refinement step to resolve poorly supported branches.
 97 This tree was visualized using MEGA 7 [37] (Fig 1). A tree based on the pipeline implemented in PosiGene
 98 (SAHM et al., 2017) was generated using the modules “create_catalog” and “alignments” and was also
 99 visualized using MEGA 7 (S1 Fig). A *rpoB* gene phylogeny, known to better differentiate between
 100 *Corynebacterium* species than 16S [38], was generated using MEGA 7 and Maximum Likelihood [39] (S2
 101 Fig).

102

103 **Figure 3. Phylogenomic tree of 29 *Corynebacterium pseudotuberculosis* genomes.**

104

105 **Genome scale positive selection analysis**

106 Positive selection analysis using branch-site models can identify genes and specific codons (sites)
107 under positive selection, specifically directional selection, in different lineages that appear as branches on
108 a tree. In these types of analyses, the lineage to be tested for positive selection is identified as the
109 “foreground”, and the genomes compared to that tested, foreground lineage are identified as being in the
110 “background”. This type of analysis can identify sites that are under positive selection ($\omega > 1$) only in the
111 tested lineage, which are identified as adaptive mutations responding to selective pressures [40,41]. The
112 resulting information can be examined in terms of function, and used for hypothesis generation [42].

113 We used the PosiGene pipeline [7] to perform genome-scale positive selection in this analysis using
114 branch-site models. Multifasta files containing the protein-coding sequences of each gene of the 29
115 genomes were used as input files. The sequence IDs annotated by RASTtk were changed to a suitable
116 format (RASTtk-based IDs) for PosiGene using a modified version of the script `extract_aa_nt_from_gb.pl`
117 (S1 File) [6]. The generated input files are in S2 File. Furthermore, eight different target groups were
118 defined as input for PosiGene based on the clades, as defined in the phylogenetic tree in Fig 1, to identify
119 adaptive mutations that occurred only in the last common ancestor. The target groups are listed in Table 2
120 and are represented in the phylogenomic trees of Figs 2 and 3, which were generated by PosiGene using
121 the 29 genomes (see tree method below). Each of these last common ancestors was examined separately as
122 a foreground branch, while all other branches in the respective run were defined as background.

123

124 **Figure 2. Target groups (foreground branches) 1 to 6 of a *Corynebacterium pseudotuberculosis***
125 **phylogeny.**

126

127 **Figure 3. Target groups (foreground branches) 7 and 8 of a *Corynebacterium pseudotuberculosis***
 128 **phylogeny excluding the Equi strains 262, I37 and 162.**

129

130 **Table 2. Groups of foreground and background lineages of *Corynebacterium pseudotuberculosis***
 131 **analyzed by branch-site models**

Group number	Group name	Target group (foreground)	Background	Reference/anchor genome
1	Ovis	All Ovis strains	All Equi strains	Cp1002B (Ovis)
2	OvisEqui262	All Ovis strains and Equi 262	All Equi strains except 262	Cp1002B (Ovis)
3	EquiExcept262	All Equi strains except 262	All Ovis strains and Equi 262	Cp31 (Equi)
4	EquiBuffaloHose	Equi strains from buffalo and horse only	All other Ovis and Equi strains	Cp31 (Equi)
5	EquiBuffalo	Equi strains from buffalo only	All other Equi and Ovis strains	Cp31 (Equi)
6	EquiHorse	Equi strains from horse only	All other Equi and Ovis strains	Cp31 (Equi)
7	Ovis2	All Ovis strains	Equi strains from buffalo and horse only	Cp1002B (Ovis)
8	StraightEqui	Equi strains from buffalo and horse only	All Ovis strains	Cp31 (Equi)

132 Ortholog groups were determined by a BLASTp best-bidirectional hit analysis [43,44]. The gene
133 annotation was given based on a reference genome. We used reference and anchor genomes from *Ovis* and
134 *Equi*, depending on the foreground branch, to sample genes that are shared within most genomes of one
135 biovar but shared with some genomes from the other. The same applies for the selection of anchor genomes
136 for subsequent filtering steps. From all the orthologs groups, only those that had a sequence from the anchor
137 genome were analyzed. For each gene sequence from the anchor genome, the orthologs from all the 29
138 genomes was assigned by progressive protein alignments using CLUSTALW [45,46]. The minimum
139 sequence identity was set to 50%. For the *Equi* foreground groups, the reference and the anchor genome
140 was strain 31 (buffalo). For the *Ovis* foreground groups, the reference and the anchor genome was strain
141 1002B (goat).

142 The PosiGene pipeline has its own tree building method. No outgroup was used, as we only
143 compared *C. pseudoduberculosis* genomes. A phylogenetic tree of each ortholog group was generated by
144 alignment filtering using GBLOCKS [47] and phylogeny reconstruction by the parsimony method and
145 jackknifing using DNAPARS from PHYLIP package [48]. A consensus tree was calculated using the
146 CONSENSE program, from the PHYLIP package. Due to the absence of an outgroup species, this tree was
147 manually rooted according to the tree previously generated by PEPR, using MEGA 7 [37]. For each
148 ortholog group that comprise at least three sequences, alignments at codon level was performed using
149 PRANK [49] and the species tree.

150 The codeml program of the PAML package [8] was used to identify sites under positive selection
151 by a branch-site test [40,41], using as input files each gene sequence alignment and its phylogenetic gene
152 tree. The likelihood ratio test (LRT) calculates and compares the likelihood of a null model, under which
153 all sites evolve under neutral ($\omega = 1$) or negative selection ($\omega < 1$) and an alternative model, under which
154 the sites are additionally allowed to evolve under positive selection ($\omega > 1$) on the foreground branch. The
155 p -value for the LRT is calculated via a χ^2 distribution, with one degree of freedom. For each site with a
156 significant p -value, the Bayes empirical Bayes (BEB) method was used to calculate the posterior probability

157 [50]. Besides the p -value, the PosiGene pipeline provides the significance value for the Bonferroni
158 correction and Benjamini–Hochberg false discovery rate (FDR) [51]. We considered positive selection
159 when $p < 0.05$ for FDR only, as Bonferroni is too conservative and can lead to many false negatives [52].
160 For each gene that was identified as being under positive selected, the sequence alignment was tested for
161 evidence of intragenic recombination. Recombination leads to alignment of non-homologous codons and
162 possible false positive results [53,54]. As no single method performs optimally under all scenarios, the best
163 strategy is to use a combination of them [55]. We used PhiPack [56] to test for evidence of recombination
164 using the methods Pairwise Homoplasy Index (PHI) [56], Neighbor Similarity Score (NSS) [57] and
165 Maximum Chi-Square [58]. We considered recombination when $q < 0.05$ for PHI and at least one another
166 test [6].

167

168 **Enrichment analysis, comparative genomics, pathogenicity island** 169 **analysis and circular map**

170 For each positively selected gene identified by the pipeline, the sequence from the anchor genome
171 was checked for the presence of functional domains using the InterProScan Database
172 (<https://www.ebi.ac.uk/interpro/search/sequence-search>). The participation of the genes in metabolic
173 pathways was verified in PATRIC’s Pathway Summary [35]. PATRIC’s Protein Family Sorter was used to
174 verify the distribution of specific genes across the genomes. GIPSy [59] was used to verify the location of
175 positively selected genes in relation to 16 pathogenicity islands that have been previously described [32],
176 using *C. glutamicum* ATCC1302 (NC_006958.1) as the non-pathogenic reference. The positions of the
177 positively selected genes were plotted in a circular map generated using BRIG [60].

178

179 **Results and discussion**

180 We used genome scale positive selection analyses to identify adaptive mutations in specific
181 lineages (branches) of *C. pseudotuberculosis* that could be related to each biovar and host preference.

182

183 **Positively selected genes**

184 The complete results for positive selection analysis for each foreground (tested branch) are
185 provided (S3 File) as are the RASTtk-based and GenBank locus tags for each gene (S1 Table). Seven of
186 the eight branches had genes that were under positive selection, with the sole exception being Branch 6
187 (EquiHorse). Among these seven branches, 27 genes were identified as having evolved under positive
188 selection. None of these 27 genes were significant for the recombination detection method (S2 Table). The
189 genes under positive selection were involved in metabolism, cell division, resistance, transport, adhesion,
190 are exposed on the surface, or are identified as hypothetical proteins with unknown functions. Many of
191 these have previously been suggested as drug or vaccine targets (Table 3). Seven of the positively selected
192 genes are in areas that have been previously identified as pathogenicity islands [32]. The specific genes,
193 and the islands they are located on, are provided (Fig 4, Table 3). The functional categories assigned to
194 these genes have previously been included in a list of niche/virulence factors involved in pathogenesis for
195 the *Corynebacterium* genus [61]. Proteins located at the interface between bacteria and the environment,
196 like surface-exposed or secreted proteins, are more likely to undergo positive selection, so it is not surprising
197 that many of the genes we detected have a role in the dynamics of the host-pathogen interaction. Some of
198 the processes that had genes identified as being under positive selection include nutrient uptake, modulation
199 of the host immune response, resistance and receptor-mediated binding [6,9]. In those proteins, positive
200 selection could act as a protective measure to avoid attachment by antibodies or phages, instead of a
201 response related to the protein function [9]. A detailed discussion about each detected gene is in S2 File.

202 **Table 3. List of positively selected genes in *Corynebacterium pseudotuberculosis* in different branches (FDR < 0.05).**

GenBank ID (Equi, Ovis)	Branch (foreground)	Product (Gene)	Function	PAI	Drug target or vaccine reference
Cp31_0206, Cp1002B_0207	2. OvisEqui262, 3. EquiExcept262	Sialidase 1	Metabolism	PiCp13	[62,63]
Cp31_0399, Cp1002B_0408	7. Ovis2, 8. StraightEqui	Sialidase 2 (<i>nanH</i>)	Metabolism	-	[62,63]
Cp31_1168, Cp1002B_1500	1. Ovis	Citrate lyase beta chain (<i>citE</i>)	Metabolism	-	[64]
Cp31_0985, Cp1002B_1689	2. OvisEqui262	Dethiobiotin (<i>bioD</i>)	Metabolism	-	[65,66]
Cp31_0638, Cp1002B_2037	2. OvisEqui262	Dihydrofolate reductase (<i>folA</i>)	Metabolism	-	[67]
Cp31_0945, Cp1002B_1731	2. OvisEqui262	Coenzyme PQQ biosynthesis protein E (<i>pqqE</i>)	Metabolism	-	
Cp31_1044, Cp1002B_1624	2. OvisEqui262	Pup deaminase (<i>dop</i>)	Metabolism	-	[68,69]
Cp31_0279, Cp1002B_0289	5. EquiBuffalo	Glutamyl-tRNA reductase (<i>hemA</i>)	Metabolism	-	[70]

Cp31_1028, Cp1002B_1640	5. EquiBuffalo	Cobaltochelate subunit CobN (<i>cobN</i>)	Metabolism	-	
Cp31_1309, Cp1002B_1363	2. OvisEqui262, 3. EquiExcept262, 7. Ovis2, 8. StraightEqui	Cobalt chelate subunit CobS (<i>cobS</i>)	Metabolism	-	
Cp31_1117, Cp1002B_1551	5. EquiBuffalo	Sporulation regulator WhiA-like (<i>whiA</i>)	Cell division	-	-
Cp31_0950, Cp1002B_1726	2. OvisEqui262	Metallo-beta-lactamase	Resistance	-	[71]
Cp31_0488, Cp1002B_0499	1. Ovis	Drug resistance transporter	Resistance	-	[72,73]
Cp31_0893, Cp1002B_1784	7. Ovis2, 8. StraightEqui	Lysine exporter protein (<i>lysE</i>)	Transport	-	[74]
Cp31_1468, Cp1002B_1186	1. Ovis	Cell-surface heme receptor (<i>hatF</i>)	Transport	PiCp5	-
Cp31_2279, -	3. EquiExcept262, 4. EquiBuffaloHorse	Adhesin 1 (membrane anchored)	Adhesion	-	-
Cp31_1094, Cp1002B_1575	4. EquiBuffaloHorse	Adhesin 2 (membrane anchored)	Adhesion	-	-

Cp31_0180, Cp1002B_0178	7. Ovis2, 8. StraightEqui	Adhesin 3 (thioester domain)	Adhesion	PiCp8	-
Cp31_0109, Cp1002B_0104	3. EquiExcept262	Alpha/beta hydrolase	Unknown	-	[75]
Cp31_1868, Cp1002B_0763	2. OvisEqui262	Membrane anchored protein 1	Unknown	PiCp13	-
Cp31_1977, Cp1002B_0655	4. EquiBuffaloHorse	Membrane anchored protein 2	Unknown	-	-
Cp31_2015, Cp1002B_2083	3. EquiExcept262	Transmembrane protein	Unknown	PiCp16	-
Cp31_0142, Cp1002B_0139	7. Ovis2, 8. StraightEqui	Secreted protein	Unknown	PiCp12	-
Cp31_2169, Cp1002B_0189	1. Ovis	Hypothetical protein 1 (no domains)	Unknown	PiCp3	-
Cp31_0366, Cp1002B_0381	4. EquiBuffaloHorse	Hypothetical protein 2 (no domains)	Unknown	-	-
Cp31_1724, Cp1002B_0908	2. OvisEqui262, 3. EquiExcept262, 7. Ovis2, 8. StraightEqui	Hypothetical protein 3 (no domains)	Unknown	-	-

Cp31_2281, Cp1002B_0835	7. Ovis2	Hypothetical protein 4 (no domains)	Unknown	-	
----------------------------	----------	-------------------------------------	---------	---	--

203

PAI – Pathogenicity island

204 **Figure 4. Circular map showing the position of pathogenicity islands and positively selected genes in**
205 **relation to *Corynebacterium pseudotuberculosis* strain 31 genome.** PAI – Pathogenicity Island, PS –
206 positively selected, CDS – coding sequences.

207

208 **Positive selection in each target group**

209 **Adaptations in Ovis biovar (Branch 1: Ovis)**

210 Using all the Ovis genomes as a target group, we identified adaptive mutations unique to
211 this biovar. The results suggest that since separation from Equi, Ovis acquired adaptations for the
212 use of carbon and iron sources (citE, Cp31_1168 and htaF, Cp31_1468) and competition with
213 other microorganisms (Drug transporter, Cp31_0488) (Table 3). Within the detected genes, the
214 suggested targets for inhibition are citrate lyase (citE) [64] and the Drug resistance transporter
215 [72]. The function of the Hypothetical protein 1 (Cp31_2169) in PiCp3 must be identified.

216 These newly identified adaptive mutations add to the known characteristics that differentiate the
217 two biovars, which include differences in nitrate reduction [76], changes in serotype and disease
218 manifestation in the guinea pig model host [77] and the content of pathogenicity islands [32]. The higher
219 genomic similarity within strains of this biovar across the world could be explained by decrease of genetic
220 diversity by periodic selection as predicted by the “stable ecotype” model of diversification [78] and
221 extensive exportation of sheep from Europe to South Africa, Australia and the Americas [20]. However, it
222 was shown that experimental infection of an *Ovis aries*, the sheep host, by an Equi strain caused the same
223 disease manifestation [31]. As the infection of “Ovis hosts” by Equi is possible, but not common, this could
224 suggest that i) Ovis lost the capacity to infect Equi exclusive hosts (horse and buffalo), and acquired unique
225 adaptations that allows it to live more efficiently in the sheep and goats, and/or ii) the Equi exclusive hosts
226 are less likely to transmit the disease to sheep and goats because they are not in close proximity in farms.

227 This data could explain how the Ovis Biovar emerged as a more clonal and highly specialized parasite for
228 hosts that Equi could experimentally infect. It appears that certain genetic changes enabled this transfer,
229 and these included the loss of a Sodium/alanine symporter (False positives for positive selection, Table 4),
230 and specific changes to a number of genes that appear to be under positive selection (Table 3).

231

232 **Adaptations shared by Ovis and Equi strain 262 (Branch 2: OvisEqui262)**

233 To identify probable adaptive mutations prior to the separation of the Ovis biovar from an Equi
234 ancestor, we considered the sister groups Ovis and Equi strain 262 as target strains. In this group, positive
235 selection was found in genes related to nutrition and evasion of the host immune response (Sialidase 1,
236 Cp31_0206); Krebs cycle, membrane biogenesis and maintenance (*bioD*, Cp31_0985); DNA biosynthesis
237 (*folA*, Cp31_0638); oxidative stress response (*pqqE*, Cp31_0945); starvation response (*dop*, Cp31_1044);
238 acetyl-CoA and DNA synthesis, fermentation (*cobS*, Cp31_1309); and drug resistance (Metallo-beta-
239 lactamase, Cp31_0950). As Sialidase 1 (Cp31_0206) is the unique gene in PiCp13 of Equi strains, this
240 protein may play an important role in infection. The functions of the membrane anchored protein in PiCp13
241 and the Hypothetical protein 3 are not yet known.

242 Several of the genes identified as under positive selection have previously been suggested
243 as possible drug targets. Included among these are sialidase [63], dethiobiotin synthetase (*bioD*)
244 [65], dihydrofolate reductase (*folA*) [67], dup deamidase (*dop*) [68,69] and metallo-beta-lactamase
245 [71].

246

247 **Adaptations in the monophyletic Equi clade (Branch 3: EquiExcept262)**

248 In this group, we searched for positive selection only in the monophyletic lineage of Equi composed
249 from buffalo, horse, camel and the strain I37 from cow. Genes related to nutrition and evasion of the host
250 immune response (Sialidase 1, Cp31_0206), acetyl-CoA and DNA synthesis, fermentation (*cobS*,

251 Cp31_1309) and adhesion (Adhesin 1, Cp31_2279) were found to be positively selected. Sialidase 1 has
252 been previously suggested as a possible drug target [63].

253 Adhesin 1 (Cp31_2279) is exclusive to Equi biovar and has 20 sites under positive selection in this
254 foreground. This suggests that this protein is a primitive and important niche factor of Equi and that adaptive
255 mutations occurred after divergence from the common ancestor with strain 262. Other genes found to be
256 under positive selection in this group include Alpha/beta hydrolase (Cp31_0109), transmembrane protein
257 (Cp31_2015), and a Hypothetical protein 3 (Cp31_1724). The role of that these genes play in the interaction
258 between these organisms and the hosts they infect have yet to be determined.

259

260 **Adaptations shared by strains isolated from buffalo and horse (Branch 4:** 261 **EquiBuffaloHorse)**

262 Strains on the fourth branch were isolated from buffalo and horses. We used the isolates from these
263 two hosts as a unique foreground and the other strains as background. Only surface exposed proteins and a
264 hypothetical protein were found. Positive selection was found in Adhesin 2 (Cp31_1094, 14 sites) and in
265 the Equi exclusive Adhesin 1 (Cp31_2279, 23 sites). Seeing the adhesin genes responding to selective
266 pressure in the Equi biovar indicates that these proteins, and the function encoded within them, play
267 important role in the particular niche these organisms inhabit. These differences could help the Equi isolates
268 adapt to the different hosts that they are able to utilize, which presumably includes adhesion to specific cell
269 receptors.

270 Additional genes were found to be under positive selection on this branch. The function of the
271 Membrane anchored protein 2 (Cp31_1977) and the Hypothetical protein 2 (Cp31_0366) are currently
272 unknown.

273

274 **Adaptations in strains isolated from buffalo (Branch 5: EquiBuffalo) and horse**
275 **(Branch 6: EquiHorse)**

276 We searched for positive selection in Equi lineages isolated from buffalo and horses separately.
277 Genes under positive selection were found only in isolates from buffalo. The genes *hemA* (Cp31_0279),
278 *cobN* (Cp31_1028), *cobS* (Cp31_1309) are related to biosynthesis of cofactors used in important biological
279 process, while *whiA* (Cp31_1117) is involved in cell division regulation. This suggests adaptations related
280 to a wide range of cellular processes as nutrition, respiration, biosynthesis, resistance and growth. Within
281 the those genes, *hemA* has been previously suggested as a drug target [70].

282 In a previous analysis, buffalo strains were shown to be clonal, with 94.7% shared genes in the core
283 genome. They compose a monophyletic cluster and to differ from the horse isolates mainly by an exclusive
284 *tox*⁺ prophage [33]. Isolates from buffalo were the only *C. pseudotuberculosis* strains shown to produce
285 diphtheria toxin [31,79–83]. This information supports the hypothesis in which the presence of the
286 prophage, specifically its diphtheria toxin (*tox*), as required for *C. pseudotuberculosis* to infect this host,
287 and this has been suggested as a potential vaccine target [33].

288 Strains isolated from horses were shown to be the most genetically diverse group. The genomes
289 isolated from these strains only share 42.5% of their genes, but no genes related to the different disease
290 phenotypes were found [84]. It is clear that one of the main differences between the horse and buffalo
291 branches are the presence of the prophage and the diphtheria toxin [33], which fits the “stable ecotype”
292 model where adaptive genes allowed expansion into a new niche (the buffalo host), and then the founder
293 mutant reproduce clonally [78].

294

295 **Adaptation in Ovis (Branch 7: Ovis2) and the monophyletic Equi clade (Branch**
296 **8: StraightEqui)**

297 In this analysis, we identified genes undergoing positive selection in the Ovis biovar, and in what
298 we consider to be “straight Equi” (Buffalo and Horse strains) by removing the Equi strains from cow (I37

299 and 262) and camel (162), due to their closer position to biovar Ovis in the phylogenetic tree (Fig 1). Apart
300 from this, the Ovis2 group is similar to the Ovis group, i.e. all Ovis genomes were chosen as target strains.
301 In StraightEqui Equi from buffalo and horse were used as target strains (Table 2).

302 On both analyzed lineages positive selection was identified in Sialidase 2 (*nanH*, Cp31_0399),
303 Cobaltochelatae subunit CobS (*cobS*, Cp31_1309), Lysine exporter protein (*lysE*, Cp31_0893), Adhesin 3
304 (Cp31_0180), Secreted protein (Cp31_0142), Hypothetical protein 3 (Cp31_1724). Only Ovis2 had
305 positive selection in Hypothetical protein 4 (Cp31_2281) (Table 3). Sialidase 2 (*nanH*) is also found in *C.*
306 *diphtheria* and *C. ulcerans* [85]. Different sialidases in a bacteria can have differences in their substrate
307 specificities and could play important roles in the interaction with other organisms or in the infection of a
308 specific tissue [63]. In *C. pseudotuberculosis*, we detected positive selection in 92 sites of sialidase *nanH*
309 and 31 sites in Adhesin 3, what suggests adaptation to different hosts of each biovar.

310

311 **Phylogeny and adaptive selection**

312 The phylogenomic tree separates biovar Ovis from Equi with 100 percent jackknife value as a
313 monophyletic group (clade) (Fig 1), which has also been seen in previous studies [32,33,86]. In addition,
314 the Equi from buffalo and horse formed a clade with two different clusters representing each host. Equi
315 strain 262 was found to be a sister group of Ovis, as was found in a previous phylogenetic tree using 44
316 genomes [33]. This tree topology suggests that Ovis originated from an Equi ancestor, and that it is a
317 paraphyletic group to that biovar [87]. The species tree generated with other methodologies supports Ovis
318 as monophyletic, and Equi as a paraphyletic group, with i) Equi strain 262 as a sister group to Ovis (S1 Fig)
319 or ii) external to Ovis and the other Equi members (S2 Fig). The *rpoB* gene is more efficient at
320 differentiating *Corynebacterium* species than 16S gene [38] and was also used to differentiate the *C.*
321 *pseudotuberculosis* strains. PEPR and RpoB trees both were able to distinguish differences in the Equi
322 strains but could not identify groups within Ovis due to their similarity (Fig 1 and S2 Fig).

323 In a previous study, *C. pseudotuberculosis* was suggested to be under anagenesis and that Ovis
324 would replace Equi [86]. However, Equi has horse and buffalo as exclusive hosts [19,31] and infections
325 were described in Egypt [26], Chile [88] and with increasing frequency in North America [28]. This implies
326 that Equi has exclusive hosts and that both biovars are independent and will probably continue to coexist
327 [89], as proposed for *Bacillus subtilis* [90].

328 *C. pseudotuberculosis* evolution probably fits the “stable ecotype” model of ecological
329 diversification, in which the acquisition of adaptive genes and mutations allows an exploration of a new
330 resource, in this case a new host, creating a new “ecotype”. This would result in unique selective pressure
331 during the initial expansion by the new clonal population, decrease genetic diversity within the new
332 population by periodic positive selection and genetic drift, and decrease the fitness for the ancestral niche.
333 Both populations coexist long enough to accumulate neutral sequence divergence at every locus, being
334 distinguished as multilocus sequence clusters [78]. Indeed, Ovis was shown to be derived from and more
335 clonal than Equi and the discrimination of biovars, hosts and countries was verified using ERIC-PCR
336 analysis on 101 strains [91]. Our results identified genes under different selection pressures across lineages
337 of *C. pseudotuberculosis* that are probably related to changes in ecological niches represented by new host
338 ranges.

339

340 **False positives for positive selection**

341 The codon models of positive selection analysis are sensitive to data quality. Errors in sequencing,
342 assembly, annotation, alignment and ortholog assignment can lead to false polymorphisms and alignments
343 of non-homologous sites resulting in a statistical signal that is misinterpreted as positive selection [7,92–
344 94]. In this work, five of the total results were identified as false positives (Table 4).

345

346 **Table 4. False positive for positive selection in *Corynebacterium pseudotuberculosis***

Product	Artifact	Branch	GenBank ID
Sodium/alanine symporter family protein	Frameshifts in Ovis and Equi strain 262 (cow)	1: Ovis	Cp1002B_0653
Zinc ABC transporter, permease protein (<i>znuB1</i>)	Frameshift in Equi strains I37 (cow) and 162 (camel)	1. Ovis	Cp1002B_0053
HNH endonuclease	Frameshift in Ovis and Equi MEX30	1. Ovis, 7. Ovis2	Cp1002B_1784
Lysine exporter protein	Two frameshifts in Ovis	1. Ovis	Cp1002B_1784

347
 348 Frameshifts causing alignment of non-homologous codons were identified in proteins mainly
 349 related to transport. The false positive found in a Sodium/alanine symporter due to different frameshifts in
 350 Ovis and Equi 262 suggest independent loss of function, but the benefits of these mutations are unknown.

351 In addition to the frameshifts in *znuB1* from Equi strains I37 and 262, the entire *znuB1CIA1* operon
 352 of zinc transporter is frameshifted in all the other Equi strains. This operon is in pathogenicity island PiCp2,
 353 but another zinc transport operon (*znuB2C2A2*) is found in *C. pseudotuberculosis*, but not located in a
 354 pathogenicity island. The loss of function in the zinc transport operon *znuB1CIA1* only in Equi suggests a
 355 different selective pressure in its niche [95]. The loss of gene functions in specific branches could be the
 356 related to the different selective pressures in their respective niches. In bacteria, there is a strong mutational
 357 bias toward deleting superfluous sequences by mutation, drift, and selective pressure to reduce the genome
 358 [89,95].

359

360 **Genome variation and the evolution of *C. pseudotuberculosis***

361 In the circular map (Fig 3), there is a gap between PiCp3 and PiCp8 of Cp1002B genome. We
362 examined this region and found an adhesin containing the “Fibrogen-binding domain 1” (RASTtk
363 Cp31_247, GenBank Cp31_2168). Both biovars have this adhesin, but the difference in nucleotide
364 sequence was high enough to be considered a non-homologous sequence by BRIG. The identity between
365 the sequences of the protein in Cp31 and Cp1002B (RASTtk Cp1002B_180, GenBank Cp1002B_184) is
366 39% with a coverage of 98%. This variation is probably related to adhesion to tissues from different hosts,
367 within the range of each biovar.

368 Various comparative genomics studies have been done in *C. pseudotuberculosis* [32,33,84,96–98].
369 Two additional characteristics that differentiate the biovars were recently identified, which include a type
370 III restriction-modification system found only in Ovis, and a CRISPR-Cas system found only in Equi
371 (Parise et al, accepted). We verified the position of this systems in relation to the pathogenicity islands and
372 whether the presence CRISPR-Cas system is primitive or derived in *C. pseudotuberculosis*, by checking its
373 presence across Equi strains. The Type III restriction-modification system is in PiCp15, suggesting that is
374 was acquired by horizontal gene transfer, while CRISPR-Cas genes are in PiCp1 and is probably primitive,
375 due to its presence in all Equi strains, including strain 262 and one reminiscent gene in Ovis. This suggests
376 that the CRISPR-Cas genes were lost from the Ovis biovar.

377 We mapped our data and data from literature to a phylogeny in Fig 5, to better understand the
378 evolution of *C. pseudotuberculosis*. In Ovis, previous analyses have documented the loss of nitrate
379 reduction related genes [33,76,99], change in serotype [77], acquisition of a Type III restriction-
380 modification system (Parise et al, accepted) and acquisition of a sigma factor in PiCp5 [32,98]. In Equi,
381 other work has described frameshifts in pilus genes [32,33] and acquisition of a *tox+* prophage in PiCp12
382 [33,100]. Prior to this study, variations in the pathogenicity islands were said to explain most of the

383 differences between biovars [32]. Here we can see that selective pressures have also played a role in
384 allowing *C. pseudotuberculosis* to fine-tune its adaptation as it expands into a new host environment.

385

386 **Figure 5. Genome variations in different branches of *Corynebacterium pseudotuberculosis*.** HGT –
387 horizontal gene transfer, PS – positive selection.

388

389 Performing genome scale positive selection analysis, we identified what appear to be adaptive
390 mutations in specific genes found in separate clades that comprise *C. pseudotuberculosis*. These differences
391 are associated with biovar differentiation and host preference. Many of the proteins under selection are
392 involved in important processes that have been shown to increase of survival, including metabolism, cell
393 division, resistance, transport, adhesion. Some of the genes that are under positive selection have previously
394 been identified as potential drug targets that could help the control of this species. We have correlated this
395 information with the presence or absence of specific genes in each clade, including information on gene
396 loss, including the ones triggered by frameshift. In addition, we used phylogenomic analysis and data from
397 literature to explain the genetic structure and evolution of the species based on ecological diversification as
398 its host range expands. The model could explain biovar Ovis as a clonal and highly specialized parasite,
399 while Equi maintain its ability to infect exclusive hosts.

400

401 **References**

- 402 1. Stephan W. Detecting strong positive selection in the genome. *Mol Ecol Resour.* 2010;10: 863–872.
403 doi:10.1111/j.1755-0998.2010.02869.x
- 404 2. Casillas S, Barbadilla A. Molecular population genetics. *Genetics.* 2017;205: 1003–1035.

- 405 doi:10.1534/genetics.116.196493
- 406 3. Hedge J, Wilson DJ. Practical Approaches for Detecting Selection in Microbial Genomes. PLoS
407 Comput Biol. 2016;12: 1–12. doi:10.1371/journal.pcbi.1004739
- 408 4. Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA
409 sequences. Mol Biol Evol. 1994;11: 725–36. doi:10.1186/s13059-014-0542-8
- 410 5. Moretti S, Murri R, Maffioletti S, Kuzniar A, Castella B, Salamin N, et al. Gcodeml: A grid-enabled
411 tool for detecting positive selection in biological evolution. Stud Health Technol Inform. 2012;175:
412 59–68. doi:10.3233/978-1-61499-054-3-59
- 413 6. Hongo JA, de Castro GM, Cintra LC, Zerlotini A, Lobo FP. POTION: an end-to-end pipeline for
414 positive Darwinian selection detection in genome-scale data through phylogenetic comparison of
415 protein-coding genes. BMC Genomics. 2015;16: 567. doi:10.1186/s12864-015-1765-0
- 416 7. Sahm A, Bens M, Platzer M, Szafranski K. PosiGene: automated and easy-to-use pipeline for
417 genome-wide detection of positively selected genes. Nucleic Acids Res. 2017;45: 1–11.
418 doi:10.1093/nar/gkx179
- 419 8. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol. 2007;24: 1586–
420 1591. doi:10.1093/molbev/msm088
- 421 9. Petersen L, Bollback JP, Dimmic M, Hubisz M, Nielsen R. Genes under positive selection in
422 *Escherichia coli*. Genome Res. 2007;17: 1336–1343. doi:10.1101/gr.6254707
- 423 10. Chattopadhyay S, Paul S, Kisiela DI, Linardopoulou E V., Sokurenko E V. Convergent molecular
424 evolution of genomic cores in *Salmonella enterica* and *Escherichia coli*. J Bacteriol. 2012;194:
425 5002–5011. doi:10.1128/JB.00552-12
- 426 11. Guinane CM, Ben Zakour NL, Tormo-Mas MA, Weinert LA, Lowder B V, Cartwright RA, et al.
427 Evolutionary genomics of *Staphylococcus aureus* reveals insights into the origin and molecular basis

- 428 of ruminant host adaptation. *Genome Biol Evol.* 2010;2: 454–66. doi:10.1093/gbe/evq031
- 429 12. Osório NS, Rodrigues F, Gagneux S, Pedrosa J, Pinto-Carbó M, Castro AG, et al. Evidence for
430 diversifying selection in a set of *Mycobacterium tuberculosis* genes in response to antibiotic- and
431 nonantibiotic-related pressure. *Mol Biol Evol.* 2013;30: 1326–1336. doi:10.1093/molbev/mst038
- 432 13. Wang Q, Xu Y, Gu Z, Liu N, Jin K, Li Y, et al. Identification of new antibacterial targets in RNA
433 polymerase of *Mycobacterium tuberculosis* by detecting positive selection sites. *Comput Biol*
434 *Chem.* 2017; doi:10.1016/j.compbiolchem.2017.11.002
- 435 14. Tan JL, Ng KP, Ong CS, Ngeow YF. Genomic Comparisons Reveal Microevolutionary Differences
436 in *Mycobacterium abscessus* Subspecies. *Front Microbiol.* 2017;8: 1–10.
437 doi:10.3389/fmicb.2017.02042
- 438 15. Lefébure T, Stanhope MJ. Evolution of the core and pan-genome of *Streptococcus*: positive
439 selection, recombination, and genome composition. *Genome Biol.* 2007;8: R71. doi:10.1186/gb-
440 2007-8-5-r71
- 441 16. Lefébure T, Stanhope MJ. Pervasive, genome-wide positive selection leading to functional
442 divergence in the bacterial genus *Campylobacter*. *Genome Res.* 2009;19: 1224–1232.
443 doi:10.1101/gr.089250.108
- 444 17. Lehmann JS, Corey VC, Ricaldi JN, Vinetz JM, Winzeler EA, Matthias MA. Whole Genome
445 Shotgun Sequencing Shows Selection on *Leptospira* Regulatory Proteins During in vitro Culture
446 Attenuation. *Am J Trop Med Hyg.* 2016;94: 302–313. doi:10.4269/ajtmh.15-0401
- 447 18. Xu Y, Zhu YY, Wang Y, Chang Y-F, Zhang Y, Jiang X, et al. Whole genome sequencing revealed
448 host adaptation-focused genomic plasticity of pathogenic *Leptospira*. *Sci Rep.* Nature Publishing
449 Group; 2016;6: 20020. doi:10.1038/srep20020
- 450 19. Dorella FA, Carvalho Pacheco L, Oliveira SC, Miyoshi A, Azevedo V. *Corynebacterium*

- 451 *pseudotuberculosis* : microbiology, biochemical properties, pathogenesis and molecular studies of
452 virulence. Vet Res. 2006;37: 201–218. doi:10.1051/vetres:2005056
- 453 20. Baird GJ, Fontaine MC. *Corynebacterium pseudotuberculosis* and its Role in Ovine Caseous
454 Lymphadenitis. J Comp Pathol. 2007;137: 179–210. doi:10.1016/j.jcpa.2007.07.002
- 455 21. Hawari AD. *Corynebacterium pseudotuberculosis* Infection (Caseous Lymphadenitis) in Camels
456 (*Camelus dromedarius*) in Jordan. Am J Anim Vet Sci. 2008;3: 68–72.
457 doi:10.3844/ajavsp.2008.68.72
- 458 22. Williamson LH. Caseous Lymphadenitis in Small Ruminants. Vet Clin North Am Food Anim Pract.
459 2001;17: 359–371. doi:10.1016/S0749-0720(15)30033-5
- 460 23. Yeruham I, Friedman S, Perl S, Elad D, Berkovich Y, Kalgard Y. A herd level analysis of a
461 *Corynebacterium pseudotuberculosis* outbreak in a dairy cattle herd. Vet Dermatol. 2004;15: 315–
462 320. doi:10.1111/j.1365-3164.2004.00388.x
- 463 24. Trost E, Ott L, Schneider J, Schröder J, Jaenicke S, Goesmann A, et al. The complete genome
464 sequence of *Corynebacterium pseudotuberculosis* FRC41 isolated from a 12-year-old girl with
465 necrotizing lymphadenitis reveals insights into gene-regulatory networks contributing to virulence.
466 BMC Genomics. BioMed Central Ltd; 2010;11: 728. doi:10.1186/1471-2164-11-728
- 467 25. Heggelund L, Gaustad P, Havelsrud OE, Blom J, Borgen L, Sundset A, et al. *Corynebacterium*
468 *pseudotuberculosis* Pneumonia in a Veterinary Student Infected During Laboratory Work. Open
469 Forum Infect Dis. 2015;2: ofv053-ofv053. doi:10.1093/ofid/ofv053
- 470 26. Selim SA. Oedematous Skin Disease of Buffalo in Egypt. J Vet Med Ser B. 2001;48: 241–258.
471 doi:10.1046/j.1439-0450.2001.00451.x
- 472 27. Foley JE, Spier SJ, Mihalyi J, Drazenovich N, Leutenegger CM. Molecular epidemiologic features
473 of *Corynebacterium pseudotuberculosis* isolated from horses. Am J Vet Res. 2004;65: 1734–1737.

474 doi:10.2460/ajvr.2004.65.1734

475 28. Spier SJ, Azevedo V. *Corynebacterium pseudotuberculosis* infection in horses: Increasing
476 frequency and spread to new regions of North America. Equine Vet Educ. 2016;
477 doi:10.1111/eve.12589

478 29. Yeruham I, Braverman Y, Shpigel NY, Chizov- Ginzburg A, Saran A, Winkler M. Mastitis in Dairy
479 Cattle Caused by *Corynebacterium pseudotuberculosis* and the Feasibility Of Transmission by
480 Houseflies I. Vet Q. 1996;18: 87–89. doi:10.1080/01652176.1996.9694623

481 30. Tejedor-Junco MT, Lupiola P, Schulz U, Gutierrez C. Isolation of nitrate-reductase positive
482 *Corynebacterium pseudotuberculosis* from dromedary camels. Trop Anim Health Prod. 2008;40:
483 165–167. doi:10.1007/s11250-007-9077-2

484 31. Moussa IM, Ali MS, Hessain AM, Kabli SA, Hemeg HA, Selim SA. Vaccination against
485 *Corynebacterium pseudotuberculosis* infections controlling caseous lymphadenitis (CLA) and
486 oedematousskin disease. Saudi J Biol Sci. 2016;23: 718–723. doi:10.1016/j.sjbs.2016.06.005

487 32. Soares SC, Silva A, Trost E, Blom J, Ramos R, Carneiro A, et al. The Pan-Genome of the Animal
488 Pathogen *Corynebacterium pseudotuberculosis* Reveals Differences in Genome Plasticity between
489 the Biovar ovis and equi Strains. PLoS One. 2013;8. doi:10.1371/journal.pone.0053818

490 33. Viana MVC, Figueiredo H, Ramos R, Guimarães LC, Pereira FL, Dorella FA, et al. Comparative
491 genomic analysis between *Corynebacterium pseudotuberculosis* strains isolated from buffalo. Lin
492 B, editor. PLoS One. 2017;12: e0176347. doi:10.1371/journal.pone.0176347

493 34. Brettin T, Davis JJ, Disz T, Edwards R a, Gerdes S, Olsen GJ, et al. RASTtk: A modular and
494 extensible implementation of the RAST algorithm for building custom annotation pipelines and
495 annotating batches of genomes. Sci Rep. 2015;5: 8365. doi:10.1038/srep08365

496 35. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, et al. Improvements to PATRIC, the

- 497 all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Res.* 2017;45:
498 D535–D542. doi:10.1093/nar/gkw1017
- 499 36. Viana MVC, Benevides LJ, Mariano DCB, Rocha FS, Vilas Boas PCB, Folador EL, et al. Genome
500 Sequence of *Corynebacterium ulcerans* Strain 210932. *Genome Announc.* 2014;2: 1–2.
- 501 37. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0
502 for Bigger Datasets. *Mol Biol Evol.* 2016;33: 1870–1874. doi:10.1093/molbev/msw054
- 503 38. Khamis A, Raoult D, La Scola B. Comparison between rpoB and 16S rRNA gene sequencing for
504 molecular identification of 168 clinical isolates of *Corynebacterium*. *J Clin Microbiol.* 2005;43:
505 1934–1936. doi:10.1128/JCM.43.4.1934-1936.2005
- 506 39. Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of
507 mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol.* 1993;10: 512–26.
508 doi:10.1093/oxfordjournals.molbev.a040023
- 509 40. Yang Z, Nielsen R. Codon-Substitution Models for Detecting Molecular Adaptation at Individual
510 Sites Along Specific Lineages. *Mol Biol Evol.* 2002;19: 908–917.
511 doi:10.1093/oxfordjournals.molbev.a004148
- 512 41. Zhang J. Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive
513 Selection at the Molecular Level. *Mol Biol Evol.* 2005;22: 2472–2479. doi:10.1093/molbev/msi237
- 514 42. Yang Z, Dos Reis M. Statistical properties of the branch-site test of positive selection. *Mol Biol*
515 *Evol.* 2011;28: 1217–1228. doi:10.1093/molbev/msq303
- 516 43. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture
517 and applications. *BMC Bioinformatics.* 2009;10: 421. doi:10.1186/1471-2105-10-421
- 518 44. Altenhoff AM, Dessimoz C. Phylogenetic and functional assessment of orthologs inference projects
519 and methods. *PLoS Comput Biol.* 2009;5. doi:10.1371/journal.pcbi.1000262

- 520 45. Larkin MA, Blackshields G, Brown NP, Chenna R, Mcgettigan PA, McWilliam H, et al. Clustal W
521 and Clustal X version 2.0. *Bioinformatics*. 2007;23: 2947–2948.
522 doi:10.1093/bioinformatics/btm404
- 523 46. Liu K, Linder CR, Warnow T. Multiple sequence alignment: a major challenge to large-scale
524 phylogenetics. *PLoS Curr*. 2011;2: RRN1198. doi:10.1371/currents.RRN1198
- 525 47. Castresana J. Selection of Conserved Blocks from Multiple Alignments for Their Use in
526 Phylogenetic Analysis. *Mol Biol Evol*. 2000;17: 540–552.
527 doi:10.1093/oxfordjournals.molbev.a026334
- 528 48. Felsenstein J. PHYLIP (Phylogeny Inference Package) version 3.6. Department of Genome
529 Sciences, University of Washington, Seattle; 2005.
- 530 49. Loytynoja A, Goldman N. Phylogeny-Aware Gap Placement Prevents Errors in Sequence
531 Alignment and Evolutionary Analysis. *Science*. 2008;320: 1632–1635.
532 doi:10.1126/science.1158395
- 533 50. Yang Z. Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection. *Mol Biol*
534 *Evol*. 2005;22: 1107–1118. doi:10.1093/molbev/msi097
- 535 51. Noble WS. How does multiple testing correction work? *Nat Biotechnol*. 2009;27: 1135–1137.
536 doi:10.1038/nbt1209-1135
- 537 52. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*.
538 2003;100: 9440–5. doi:10.1073/pnas.1530509100
- 539 53. Shriner D, Nickle DC, Jensen MA, Mullins JI. Potential impact of recombination on sitewise
540 approaches for detecting positive natural selection. *Genet Res*. 2003;81: 115–121.
541 doi:10.1017/S0016672303006128
- 542 54. Anisimova M, Nielsen R, Yang Z. Effect of recombination on the accuracy of the likelihood method

- 543 for detecting positive selection at amino acid sites. *Genetics*. 2003;164: 1229–1236.
544 doi:10.1093/bioinformatics/btn086
- 545 55. Posada D, Crandall KA, Holmes EC. Recombination in Evolutionary Genomics. *Annu Rev Genet*.
546 2002;36: 75–97. doi:10.1146/annurev.genet.36.040202.111115
- 547 56. Bruen TC, Philippe HH, Bryant D. A simple and robust statistical test for detecting the presence of
548 recombination. *Genetics*. 2006;172: 2665–81. doi:10.1534/genetics.105.048975
- 549 57. Jakobsen IB, Easteal S. A program for calculating and displaying compatibility matrices as an aid
550 in determining reticulate evolution in molecular sequences. *Comput Appl Biosci*. 1996;12: 291–
551 295. doi:10.1093/bioinformatics/12.4.291
- 552 58. Smith JM. Analyzing the mosaic structure of genes. *J Mol Evol*. 1992;34: 126–129.
553 doi:10.1007/BF00182389
- 554 59. Soares SC, Geyik H, Ramos RTJ, de Sá PHCG, Barbosa EGV, Baumbach J, et al. GIPSY: Genomic
555 island prediction software. *J Biotechnol*. 2016;232: 2–11. doi:10.1016/j.jbiotec.2015.09.008
- 556 60. Alikhan N-F, Petty NK, Ben Zakour NL, Beatson SA. BLAST Ring Image Generator (BRIG):
557 simple prokaryote genome comparisons. *BMC Genomics*. 2011;12: 402. doi:10.1186/1471-2164-
558 12-402
- 559 61. Tauch A, Burkovski A. Molecular armory or niche factors: virulence determinants of
560 *Corynebacterium* species. *FEMS Microbiol Lett*. 2015;67: fnv185. doi:10.1093/femsle/fnv185
- 561 62. Kim S, Oh D-B, Kang HA, Kwon O. Features and applications of bacterial sialidases. *Appl*
562 *Microbiol Biotechnol*. 2011;91: 1–15. doi:10.1007/s00253-011-3307-2
- 563 63. Kim S, Oh D-B, Kwon O. Sialidases of *Corynebacteria* and their Biotechnological Applications.
564 *Corynebacterium diphtheriae* and Related Toxigenic Species. Dordrecht: Springer Netherlands;
565 2014. pp. 247–262. doi:10.1007/978-94-007-7624-1_13

- 566 64. Goulding CW, Bowers PM, Segelke B, Lakin T, Kim CY, Terwilliger TC, et al. The Structure and
567 Computational Analysis of *Mycobacterium tuberculosis* Protein CitE Suggest a Novel Enzymatic
568 Function. *J Mol Biol.* 2007;365: 275–283. doi:10.1016/j.jmb.2006.09.086
- 569 65. Salaemae W, Azhar A, Booker GW, Polyak SW. Biotin biosynthesis in *Mycobacterium*
570 *tuberculosis*: physiology, biochemistry and molecular intervention. *Protein Cell.* 2011;2: 691–695.
571 doi:10.1007/s13238-011-1100-8
- 572 66. Salaemae W, Yap MY, Wegener KL, Booker GW, Wilce MCJ, Polyak SW. Nucleotide triphosphate
573 promiscuity in *Mycobacterium tuberculosis* dethiobiotin synthetase. *Tuberculosis.* Elsevier Ltd;
574 2015;95: 259–266. doi:10.1016/j.tube.2015.02.046
- 575 67. Santa Maria JP, Park Y, Yang L, Murgolo N, Altman MD, Zuck P, et al. Linking High-Throughput
576 Screens to Identify MoAs and Novel Inhibitors of *Mycobacterium tuberculosis* Dihydrofolate
577 Reductase. *ACS Chem Biol.* 2017;12: 2448–2456. doi:10.1021/acscchembio.7b00468
- 578 68. Zhang S, Burns-Huang KE, Janssen G V, Li H, Ovaa H, Hedstrom L, et al. *Mycobacterium*
579 *tuberculosis* Proteasome Accessory Factor A (PafA) Can Transfer Prokaryotic Ubiquitin-Like
580 Protein (Pup) between Substrates. Rubin EJ, editor. *MBio.* 2017;8: e00122-17.
581 doi:10.1128/mBio.00122-17
- 582 69. Delley CL, Müller AU, Ziemski M, Weber-Ban E. Prokaryotic Ubiquitin-Like Protein and Its
583 Ligase/Delignase Enzymes. *J Mol Biol.* Elsevier Ltd; 2017;429: 3486–3499.
584 doi:10.1016/j.jmb.2017.04.020
- 585 70. Ravichandran M, Ali SA, Rashid NHA, Kurunathan S, Yean CY, Ting LC, et al. Construction and
586 evaluation of a O139 *Vibrio cholerae* vaccine candidate based on a hemA gene mutation. *Vaccine.*
587 2006;24: 3750–3761. doi:10.1016/j.vaccine.2005.07.016
- 588 71. Palzkill T. Metallo- β -lactamase structure and function. *Ann N Y Acad Sci.* 2013;1277: 91–104.

- 589 doi:10.1111/j.1749-6632.2012.06796.x
- 590 72. Li X, Nikaido H. Efflux-mediated drug resistance in bacteria. *Drugs*. 2009;69: 1555–1623.
591 doi:10.2165/11317030-000000000-00000.Efflux-Mediated
- 592 73. Schroeder M, Brooks B, Brooks A. The Complex Relationship between Virulence and Antibiotic
593 Resistance. *Genes (Basel)*. 2017;8: 39. doi:10.3390/genes8010039
- 594 74. Gideon HP, Wilkinson KA, Rustad TR, Oni T, Guio H, Kozak RA, et al. Hypoxia Induces an
595 Immunodominant Target of Tuberculosis Specific T Cells Absent from Common BCG Vaccines.
596 Deretic V, editor. *PLoS Pathog*. 2010;6: e1001237. doi:10.1371/journal.ppat.1001237
- 597 75. Carr PD, Ollis DL. Alpha/beta hydrolase fold: an update. *Protein Pept Lett*. 2009;16: 1137–1148.
598 doi:10.2174/092986609789071298
- 599 76. Biberstein EL, Knight HD, Jang S. Two biotypes of *Corynebacterium pseudotuberculosis*. *Vet Rec*.
600 1971;89: 691–692. doi:10.1136/vr.89.26.691
- 601 77. Barakat A., Selim SA, Atef A, Saber MS, Nafie EK, El-Ebeedy AA. Two serotypes of
602 *Corynebacterium pseudotuberculosis* isolated from different animal species. *Rev Sci Tech Off Int*
603 *Epiz*. 1984;3: 151–163.
- 604 78. Cohan FM, Koeppl AF. The origins of ecological diversity in prokaryotes. *Curr Biol*. Elsevier Ltd;
605 2008;18: R1024-34. doi:10.1016/j.cub.2008.09.014
- 606 79. Maximescu P, Oprișan A, Pop A, Potorac E. Further studies on *Corynebacterium* species capable
607 of producing diphtheria toxin (*C. diphtheriae*, *C. ulcerans*, *C. ovis*). *J Gen Microbiol*. 1974;82: 49–
608 56. doi:10.1099/00221287-82-1-49
- 609 80. Wong TP, Groman N. Production of diphtheria toxin by selected isolates of *Corynebacterium*
610 *ulcerans* and *Corynebacterium pseudotuberculosis*. *Infect Immun*. 1984;43: 1114–6. Available:
611 <http://www.ncbi.nlm.nih.gov/pubmed/6321350>

- 612 81. Groman N, Schiller J, Russell J. *Corynebacterium ulcerans* and *Corynebacterium*
613 *pseudotuberculosis* responses to DNA probes derived from corynephage beta and *Corynebacterium*
614 *diphtheriae*. Infect Immun. 1984;45: 511–7. doi:10.1159/000114687
- 615 82. Syame SM, Hakim AS, Hedia RH, Marie HSH, Selim SA. Characterization of Virulence Genes
616 Present in *Corynebacterium pseudotuberculosis* Strains Isolated From Buffaloes. 2013;10: 585–
617 591. doi:10.5829/idosi.gv.2013.10.5.7388
- 618 83. Selim SA, Mohamed FH, Hessain AM, Moussa IM. Immunological characterization of diphtheria
619 toxin recovered from *Corynebacterium pseudotuberculosis*. Saudi J Biol Sci. King Saud University;
620 2015; 0–5. doi:10.1016/j.sjbs.2015.11.004
- 621 84. Baraúna RA, Ramos RTJ, Veras AAO, Pinheiro KC, Benevides LJ, Viana MVC, et al. Assessing
622 the Genotypic Differences between Strains of *Corynebacterium pseudotuberculosis* biovar equi
623 through Comparative Genomics. Munderloh UG, editor. PLoS One. 2017;12: e0170676.
624 doi:10.1371/journal.pone.0170676
- 625 85. Ott L, Burkovski A. Toxigenic Corynebacteria: Adhesion, Invasion and Host Response. In:
626 Burkovski A, editor. *Corynebacterium diphtheriae* and Related Toxigenic Species. 1st ed.
627 Dordrecht: Springer Netherlands; 2014. pp. 143–170. doi:10.1007/978-94-007-7624-1_8
- 628 86. Oliveira A, Teixeira P, Azevedo M, Jamal SB, Tiwari S, Almeida S, et al. *Corynebacterium*
629 *pseudotuberculosis* may be under anagenesis and biovar Equi forms biovar Ovis: a phylogenic
630 inference from sequence and structural analysis. BMC Microbiol. 2016;16: 100.
631 doi:10.1186/s12866-016-0717-4
- 632 87. Vandamme A-M. Basic concepts of molecular evolution. In: Lemey P, Salemi M, Vandamme A-
633 M, editors. The Phylogenetic Handbook - A Practical Approach to Phylogenetic Analysis and
634 Hypothesis Testing. 2nd ed. Cambridge: Cambridge University Press; 2009. pp. 3–29.

- 635 88. Cavalcante ALQ, Dias LM, Alves JTC, Veras AAO, Guimarães LC, Rocha FS, et al. Complete
636 Genome Sequence of *Corynebacterium pseudotuberculosis* Strain E19, Isolated from a Horse in
637 Chile. *Genome Announc.* 2015;3: e01385-15. doi:10.1128/genomeA.01385-15
- 638 89. Lassalle F, Muller D, Nesme X. Ecological speciation in bacteria: reverse ecology approaches reveal
639 the adaptive part of bacterial cladogenesis. *Res Microbiol.* 2015;166: 729–41.
640 doi:10.1016/j.resmic.2015.06.008
- 641 90. Kopac S, Wang Z, Wiedenbeck J, Sherry J, Wu M, Cohan FM. Genomic Heterogeneity and
642 Ecological Speciation within One Subspecies of *Bacillus subtilis*. *Appl Environ Microbiol.* 2014;80:
643 4842–4853. doi:10.1128/AEM.00576-14
- 644 91. Dorneles EMS, Santana JA, Ribeiro D, Dorella FA, Guimarães AS, Moawad MS, et al. Evaluation
645 of ERIC-PCR as Genotyping Method for *Corynebacterium pseudotuberculosis* Isolates. Hozbor DF,
646 editor. *PLoS One.* 2014;9: e98758. doi:10.1371/journal.pone.0098758
- 647 92. Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH, Graur D. Estimates of Positive
648 Darwinian Selection Are Inflated by Errors in Sequencing, Annotation, and Alignment. *Genome*
649 *Biol Evol.* 2009;1: 114–118. doi:10.1093/gbe/evp012
- 650 93. Mallick S, Gnerre S, Muller P, Reich D. The difficulty of avoiding false positives in genome scans
651 for natural selection. *Genome Res.* 2009;19: 922–33. doi:10.1101/gr.086512.108
- 652 94. Markova-Raina P, Petrov D. High sensitivity to aligner and high rate of false positives in the
653 estimates of positive selection in the 12 *Drosophila* genomes. *Genome Res.* 2011;21: 863–74.
654 doi:10.1101/gr.115949.110
- 655 95. Bobay L-M, Ochman H. The Evolution of Bacterial Genome Architecture. *Front Genet.* 2017;8: 1–
656 6. doi:10.3389/fgene.2017.00072
- 657 96. Soares SC, Abreu VAC, Ramos RTJ, Cerdeira L, Silva A, Baumbach J, et al. PIPS: Pathogenicity

- 658 Island Prediction Software. Mokrousov I, editor. PLoS One. 2012;7: e30848.
659 doi:10.1371/journal.pone.0030848
- 660 97. Soares SC, Trost E, Ramos RTJ, Carneiro AR, Santos AR, Pinto AC, et al. Genome sequence of
661 *Corynebacterium pseudotuberculosis* biovar equi strain 258 and prediction of antigenic targets to
662 improve biotechnological vaccine production. J Biotechnol. 2013;167: 135–141.
663 doi:10.1016/j.jbiotec.2012.11.003
- 664 98. Ruiz JC, D’Afonseca V, Silva A, Ali A, Pinto AC, Santos AR, et al. Evidence for reductive genome
665 evolution and lateral acquisition of virulence functions in two *Corynebacterium pseudotuberculosis*
666 strains. PLoS One. 2011;6. doi:10.1371/journal.pone.0018551
- 667 99. Almeida S, Sousa C, Abreu V, Diniz C, Dorneles EMS, Lage AP, et al. Exploration of Nitrate
668 Reductase Metabolic Pathway in *Corynebacterium pseudotuberculosis*. Int J Genomics. 2017;2017:
669 1–12. doi:10.1155/2017/9481756
- 670 100. Ramos RTJ, Carneiro AR, de Castro Soares S, Barbosa S, Varuzza L, Orabona G, et al. High
671 efficiency application of a mate-paired library from next-generation sequencing to postlight
672 sequencing: *Corynebacterium pseudotuberculosis* as a case study for microbial de novo genome
673 assembly. J Microbiol Methods. 2013;95: 441–447. doi:10.1016/j.mimet.2013.06.006

674

675 **Supporting information**

676 **S1 Fig. Phylogenomic tree of *Corynebacterium pseudotuberculosis* generated with the PosiGene**
677 **pipeline.**

678 **S2 Fig. Phylogenetic tree of *Corynebacterium pseudotuberculosis* species tree based on the *rpoB* gene.**

679 **S1 Table. Mapping of RASTtk-based and GenBank IDs of each positively selected gene of strain 31**
680 **(Equi) and strain 1002B (Ovis).**

681 **S2 Table. List of positively selected genes and probability of recombination in *Corynebacterium***
682 ***pseudotuberculosis* ($q < 0.05$ for PHI and at least one other test).**

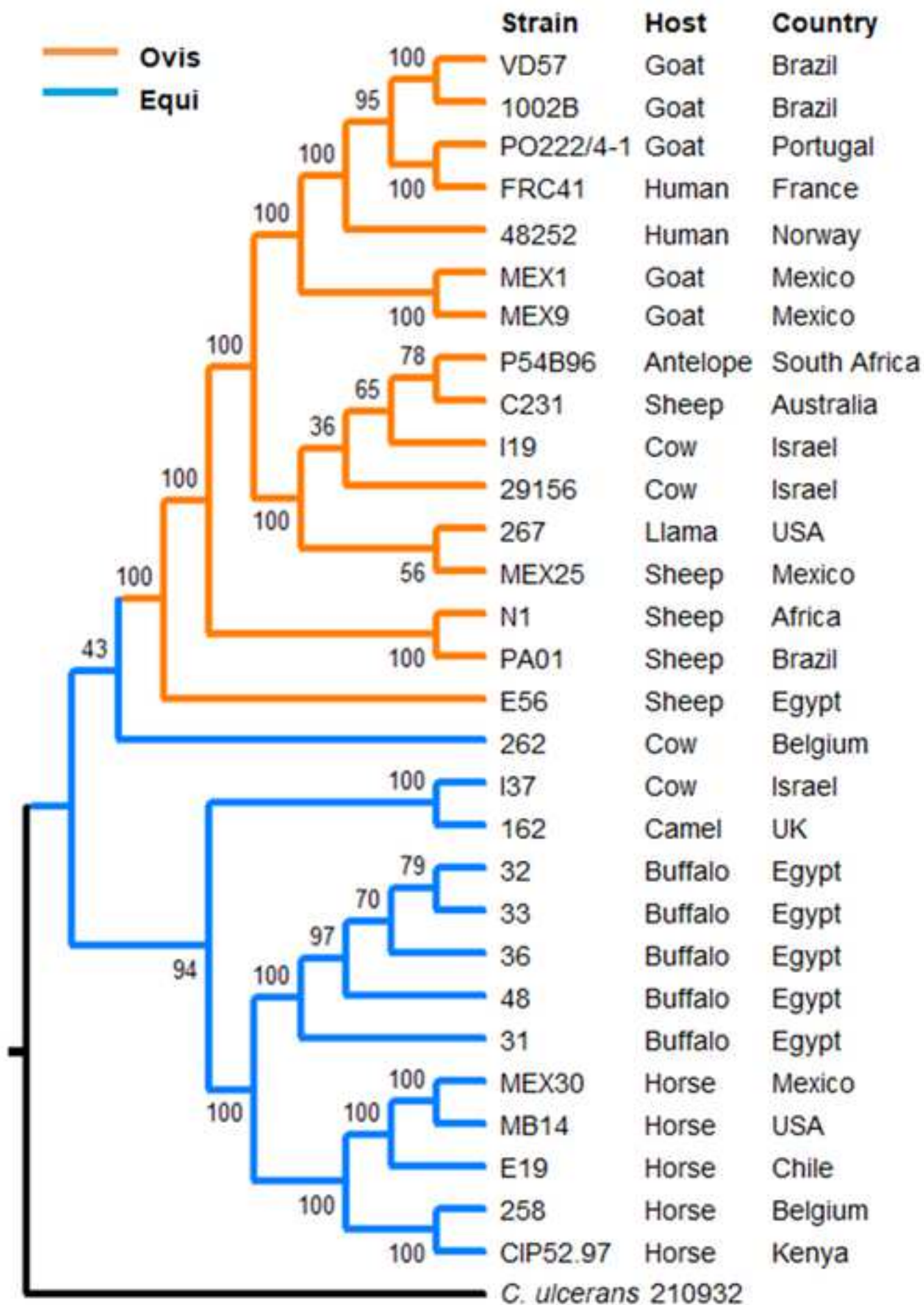
683 **S1 File. Script used to extract multifasta amino acid files with suitable identifiers from genbank files**

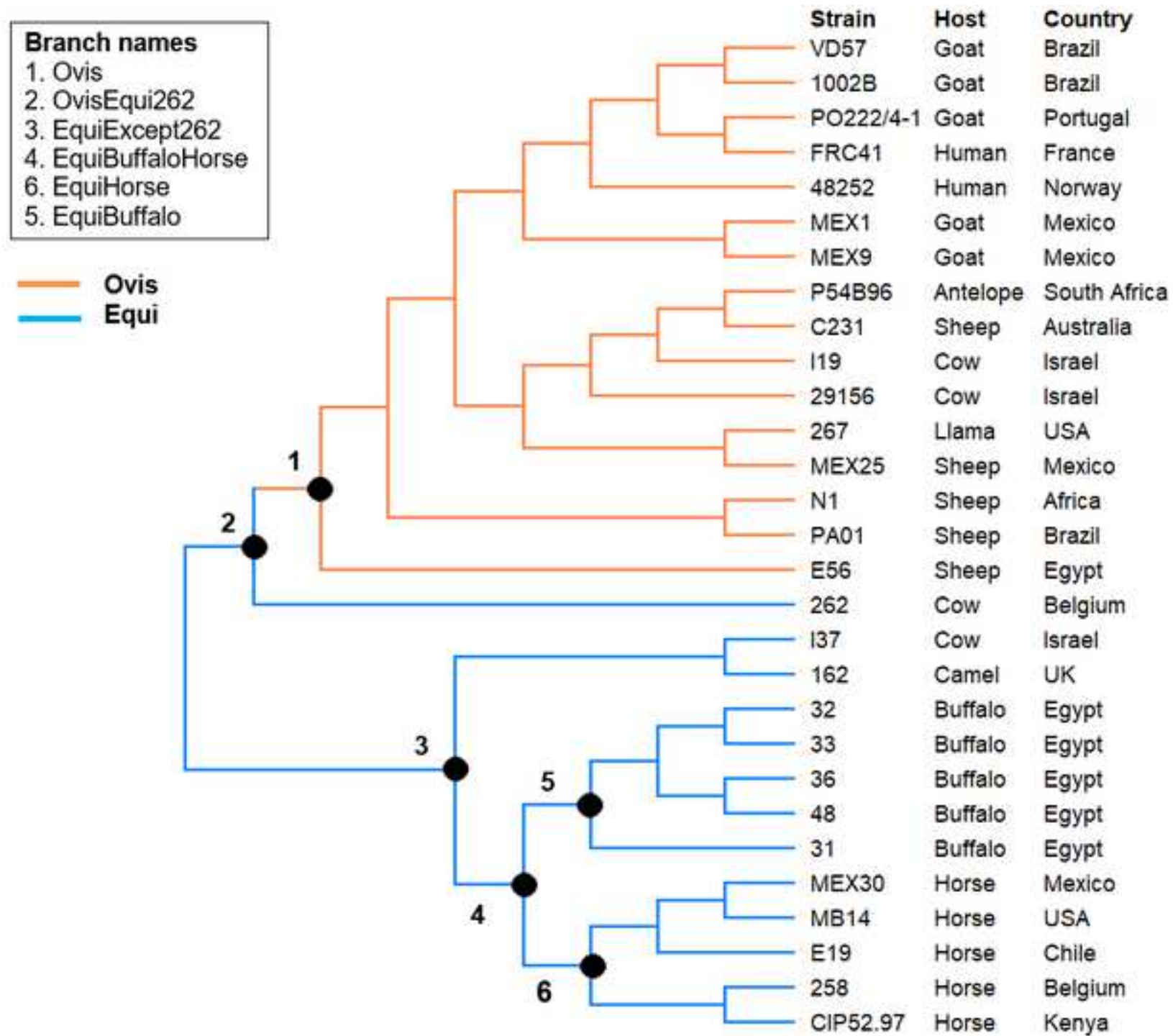
684 **S2 File. Input files used in PosiGene pipeline.**

685 **S3 File. Individual results of each branch-site analysis.**

686 **S4 File. Detailed discussion of each positively selected gene.**

687





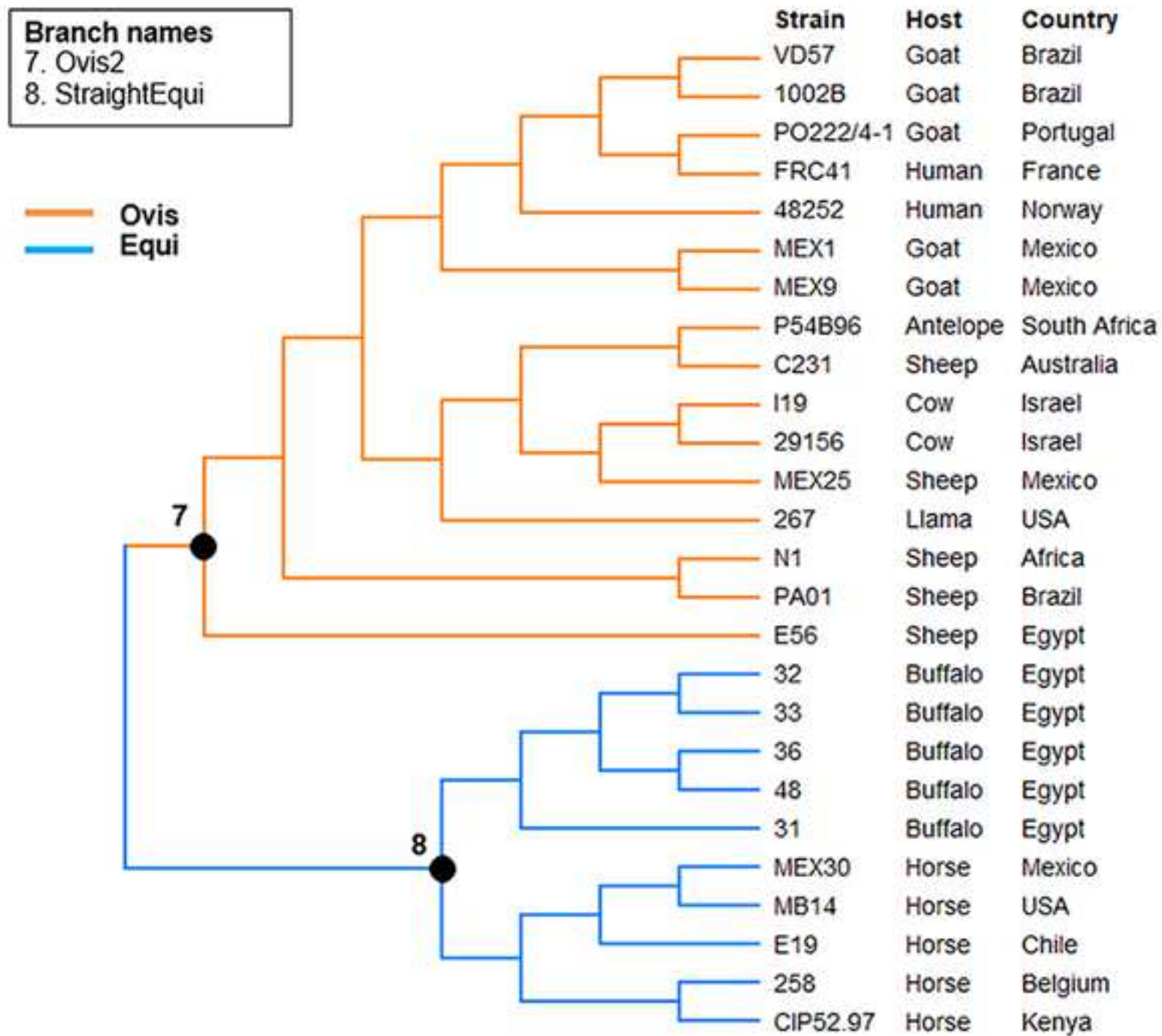
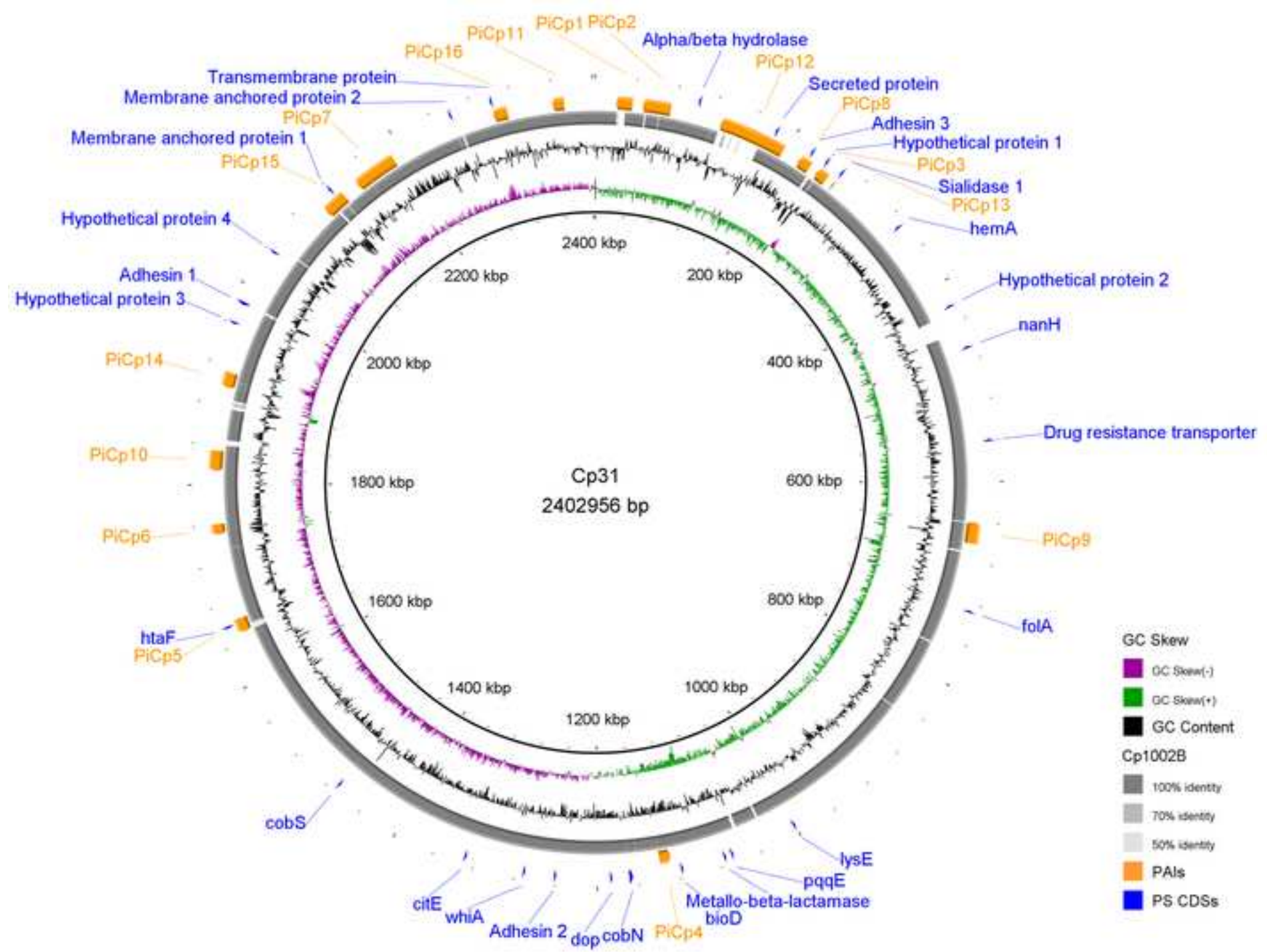
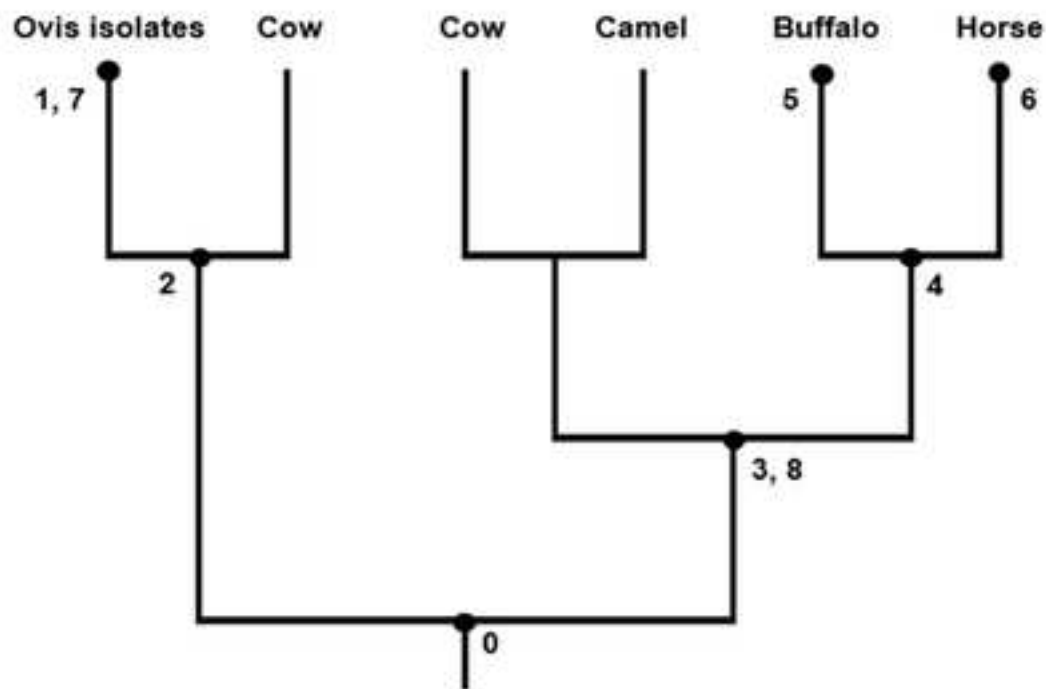


Figure 4



**1. Ovis**

Change in serotype
 (HGT) Type III restriction-modification system
 (HGT) Sigma factor
 (Loss) CRISPR-Cas system
 (Loss) Nitrate reduction
 (Loss) Adhesin 1
 (PS) Citrate lyase (*citE*)
 (PS) Drug resistance transporter
 (PS) Hemin receptor (*htaF*)

2. OvisEqui262

(Frameshift) Sodium/alanine symporter
 (PS) Sialidase 1
 (PS) Dethiobiotin synthetase (*bioD*)
 (PS) Dihydrofolate reductase (*folA*)
 (PS) Coenzyme PQQ biosynthesis protein E (*pqqE*)
 (PS) Pup deaminase (*dop*)
 (PS) Cobalt chelatase subunit CobS
 (PS) Metallo-beta-lactamase
 (PS) Membrane anchored protein 1

3. EquiExcept262

Changes in pilus genes
 (Frameshift) Zinc transporter *znuA1B1C1*
 (PS) Sialidase 1
 (PS) Cobalt chelatase subunit CobS
 (PS) Alpha/beta hydrolase
 (PS) Adhesin 1
 (PS) Transmembrane protein

4. EquiBuffaloHorse

(PS) Adhesin 1
 (PS) Adhesin 2
 (PS) Membrane anchored protein 2
 (PS) Hypothetical protein 2

5. EquiBuffalo

(HGT) *tox* + prophage
 (PS) Glutamyl-tRNA reductase (*hemA*)
 (PS) Cobaltochelate subunit CobN
 (PS) Cobaltochelate subunit CobS
 (PS) Sporulation regulator WhiA

6. EquiHorse

No results

7. Ovis2

(PS) Sialidase 2 (*nanH*)
 (PS) Lysine exporter protein (*lysE*)

8. StraightEqui

(PS) Sialidase 2 (*nanH*)
 (PS) Lysine exporter protein (*lysE*)
 (PS) Cobalt chelatase subunit CobS

0. Common ancestor

Nitrate reduction
 CRISPR-Cas system
 Adhesin 1
 Sodium/alanine symporter
 Zinc transporter *znuA1B1C1*







Click here to access/download
Supporting Information
S1_Table.xlsx



Click here to access/download
Supporting Information
S2_Table.xlsx




Click here to access/download
Supporting Information
S4_File.docx



Click here to access/download


Supporting Information - Compressed/ZIP File Archive
S1_File.zip





Click here to access/download

Supporting Information - Compressed/ZIP File Archive
S2_File.zip





Click here to access/download

Supporting Information - Compressed/ZIP File Archive
S3_File.zip



V. Discussion

In the introduction of this thesis, we reviewed the current knowledge about *C. pseudotuberculosis* biology, difficulties for control, structural genomics and introduced positive selection analysis as a tool for studying bacterial biology.

In Chapter I, we performed a comparative genomics analysis of *C. pseudotuberculosis* with focus on 11 strains isolated from buffalo with the goal of better understand the pathogenic mechanisms of Oedematous Skin Disease. The buffalo isolates had all the 16 previously described pathogenicity islands, formed a monophyletic clade and were highly similar to each other in gene content, sharing 2,058 of 2,172 genes (94.75%). We saw that strains isolated from buffalo had fusions in some of their pilin genes when compared to strains isolated from other hosts, suggesting that this is an adaptation for adhesion specifically to buffalo tissues. The 11 genomes from buffalo showed variation in synteny in the pathogenicity island PiCp12, where a 36.6 Kb sequence unique to buffalo isolates is inserted in one of two tRNA-Arg genes or is missing. A complete prophage was predicted in this region, harboring the diphtheria toxin gene (DT). In *C. pseudotuberculosis*, DT is restricted to in buffalo isolates and all the tested strains were positive for the toxin. Two of the sequenced genomes were positive for DT but the phage was not found in the assembled genome. Due to the presence only in buffalo isolates, differences in positions across the genomes and the loss after isolation, we suggested that the prophage is a volatile region that is maintained in the chromosome as a requirement to infect buffalo. We also suggested that a vaccine could be developed for this specific host based on the diphtheria toxin.

In Chapter II, we did a literature review on the methodology for detecting positive selection at the molecular level and its contributions to bacterial biology. Methods for sequence sampling, alignment and phylogeny, recombination detection and correction for multiple tests were discussed. We introduced the positive selection analysis using codon models, in which the proportion ω of non-synonymous (d_N) to synonymous (d_S) mutations ($\omega = d_N / d_S$) identifies genes as being under positive or adaptive selection ($\omega > 1$), negative or purifying selection ($\omega < 1$) or neutral evolution ($\omega = 1$). We described the most used Maximum Likelihood models for codon evolution, the Likelihood Ratio Tests used to estimate ω , and their purpose in detecting positive selection as diversifying selection (site models) and directional selection (branch-site models). Then, we did a literature review on the findings of these analyses in bacteria. Usually, the identified genes code membrane exposed proteins involved in interactions with the environment, including other bacteria, phages and host immune system. Analysis of free living bacteria shows genes related to adaptation to environmental changes and speciation, while pathogens shows genes related to host interaction and proteins that can be used as drug and vaccine targets. Among the pathogens, we showed the results for *Mycobacterium tuberculosis*, *M. abscessus*, *Staphylococcus aureus*, *Streptococcus*, *Campylobacter*, *Chlamydia*,

Leptospira, *Salmonella*. This review was important to direct the positive selection study for *C. pseudotuberculosis*.

In Chapter III, we performed a positive selection analysis on 29 genomes of *C. pseudotuberculosis* to identify adaptive mutations related to the lifestyle of 8 different phylogenetic lineages (branches) representing the different biovars and hosts that the genomes were isolated from. A total of 27 genes were identified. These genes were found to be involved in metabolism, cell division, resistance, transport, and adhesion. They also included hypothetical proteins with unknown functions. Some of these genes have been previously suggested to be drug or vaccine targets. By comparing the genes under positive selection with data from the literature, and examining the phylogeny of these strains, we concluded that: i) Ovis is a monophyletic group derived from Equi, ii) Ovis lost the ability to infect Equi hosts after adapting to a new niche, and these changes included a different serotype, losing nitrate reduction, and positive selection of the hemin receptor; iii) the Ovis and Equi biovars are separate clades, each with adaptations for their different niches. We suggest that these specific genes could be used to develop control methods targeting isolates from a specific biovar and/or host.

VI. Conclusion and perspectives

In our genomics analysis of *Corynebacterium pseudotuberculosis*, we increased the basic knowledge related to the host preference and biovar evolution. We have:

- a. compared different genomes of *C. pseudotuberculosis* isolated from buffalo to better understand the pathogenicity mechanisms required to infection of that host. The results showed overall synteny, high genome identity, and a phage containing the diphtheria toxin that was inserted in a hotspot in the pathogenicity island PiCp12. The presence of the phage in all strains isolated from buffalo and the loss of the phage in vitro by recombination in the hotspot suggest that the diphtheria toxin is required to infect the host;
- b. performed pan-genomics, phylogenomics, and plasticity analysis on 44 genomes of *C. pseudotuberculosis*. We identified the prophage harboring diphtheria toxin only in buffalo isolates and genes related to nitrate reduction. We also identified assembly errors in Equi strains and fixed the buffalo isolated strain 31 assembly;
- c. performed a genome scale positive selection analysis in 29 genomes of *C. pseudotuberculosis*, detecting involved in metabolism, cell division, resistance, transport, adhesion, exposed on the cell surface or hypothetical proteins with unknown function.

As perspectives to future works, we intend to:

- a. validate the identified fusions of pilin genes from buffalo isolated strains by PCR;
- b. perform a network analysis of the positively selected genes;
- c. predict the existence of different ecotypes within *C. pseudotuberculosis* genomes using the softwares AdaptML and Ecotype Simulator;
- d. analyze the polymorphism of promoter regions and correlate it to probable differentiation of gene expression within ecotypes;
- e. perform a genome scale positive selection analysis using genomes from *C. pseudotuberculosis*, *C. ulcerans* and *C. diphtheriae* to identify selective pressures involved in their speciation.

VII. Bibliography

- AHMED, I. M.; EL-TAHAWY, A. S. Prevalence of So-called Oedematous Skin Disease in Egyptian buffaloes with particular study on its economic influence. **Alexandria Journal of Veterinary Sciences**, v. 37, n. October, p. 129–133, 2012.
- ALMEIDA, S. et al. Exploration of Nitrate Reductase Metabolic Pathway in *Corynebacterium pseudotuberculosis*. **International Journal of Genomics**, v. 2017, p. 1–12, 2017.
- ALTENHOFF, A. M.; DESSIMOZ, C. Phylogenetic and functional assessment of orthologs inference projects and methods. **PLoS Computational Biology**, v. 5, n. 1, 2009.
- ANDREWS, S. C.; ROBINSON, A. K.; RODRÍGUEZ-QUIÑONES, F. Bacterial iron homeostasis. **FEMS Microbiology Reviews**, v. 27, n. 2–3, p. 215–237, jun. 2003.
- AUGUSTINE, J. L.; RENSCHAW, H. W. Survival of *Corynebacterium pseudotuberculosis* in axenic purulent exudate on common barnyard fomites. **American Journal of Veterinary Research**, v. 47, n. 4, p. 713–715, 1986.
- BAIRD, G. J.; FONTAINE, M. C. *Corynebacterium pseudotuberculosis* and its Role in Ovine Caseous Lymphadenitis. **Journal of Comparative Pathology**, v. 137, n. 4, p. 179–210, nov. 2007.
- BARAKAT, A. . et al. Two serotypes of *Corynebacterium pseudotuberculosis* isolated from different animal species. **Rev Sci Tech Off Int Epiz**, v. 3, n. 1, p. 151–163, 1984.
- BARAÚNA, R. A. et al. Assessing the Genotypic Differences between Strains of *Corynebacterium pseudotuberculosis* biovar equi through Comparative Genomics. **PLOS ONE**, v. 12, n. 1, p. e0170676, 26 jan. 2017.
- BARBA, M. et al. Experimental Transmission of *Corynebacterium pseudotuberculosis* Biovar equi in Horses by House Flies. **Journal of Veterinary Internal Medicine**, v. 29, n. 2, p. 636–643, mar. 2015.
- BERNARD, A. L.; FUNKE, G. *Corynebacterium*. In: **Bergey's Manual of Systematic of Archaea and Bacteria**. [s.l: s.n.]. p. 1–70.
- BIBERSTEIN, E. L.; KNIGHT, H. D.; JANG, S. Two biotypes of *Corynebacterium pseudotuberculosis*. **The Veterinary Record**, v. 89, n. 26, p. 691–692, 1971.
- BILLINGTON, S. J. et al. Identification and role in virulence of putative iron acquisition genes from *Corynebacterium pseudotuberculosis*. **FEMS Microbiology Letters**, v. 208, n. 1, p. 41–45, fev. 2002.
- BOBAY, L.-M.; OCHMAN, H. The Evolution of Bacterial Genome Architecture. **Frontiers in Genetics**, v. 8, n. May, p. 1–6, 2017.

- BOLOTIN, E.; HERSHBERG, R. Bacterial intra-species gene loss occurs in a largely clocklike manner mostly within a pool of less conserved and constrained genes. **Scientific Reports**, v. 6, n. October, p. 1–9, 2016.
- BRAVERMAN, Y. et al. The role of houseflies (*Musca domestica*) in harbouring *Corynebacterium pseudotuberculosis* in dairy herds in Israel. **Revue Scientifique et Technique (International Office of Epizootics)**, v. 18, n. 3, p. 681–90, dez. 1999.
- BRETTIN, T. et al. RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. **Scientific Reports**, v. 5, p. 8365, 10 fev. 2015.
- CARNE, H. R.; WICKHAM, N.; KATER, J. C. A Toxic Lipid from the Surface of *Corynebacterium ovis*. **Nature**, v. 178, n. 4535, p. 701–702, 29 set. 1956.
- CERDEÑO-TÁRRAGA, A. M. et al. The complete genome sequence and analysis of *Corynebacterium diphtheriae* NCTC13129. **Nucleic acids research**, v. 31, n. 22, p. 6516–23, 15 nov. 2003.
- CHAN, C. X.; RAGAN, M. A. Next-generation phylogenomics. p. 1–6, 2013.
- CHEN, F. et al. Assessing Performance of Orthology Detection Strategies Applied to Eukaryotic Genomes. **PLoS ONE**, v. 2, n. 4, p. e383, 18 abr. 2007.
- COLLETT, M.; BATH, G.; CAMERON, C. *Corynebacterium pseudotuberculosis* infections. In: **Infections diseases of livestock with special reference to Southern Africa**. [s.l.] Oxford University Press, 1994. p. 1387–1395.
- COSTA, L. R.; SPIER, S. J.; HIRSH, D. C. Comparative molecular characterization of *Corynebacterium pseudotuberculosis* of different origin. **Veterinary microbiology**, v. 62, n. 2, p. 135–143, 1998.
- D'AFONSECA, V. et al. A description of genes of *Corynebacterium pseudotuberculosis* useful in diagnostics and vaccine applications. **Genetics and Molecular Research**, v. 7, n. 1, p. 252–260, 2008.
- DARLING, A. C. E. Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. **Genome Research**, v. 14, n. 7, p. 1394–1403, 14 jun. 2004.
- DARLING, A. E.; MAU, B.; PERNA, N. T. Progressivemaue: Multiple genome alignment with gene gain, loss and rearrangement. **PLoS ONE**, v. 5, n. 6, 2010.
- DAVIS, J. J. et al. PATtyFams: Protein Families for the Microbial Genomes in the PATRIC Database. **Frontiers in Microbiology**, v. 7, n. February, p. 1–12, 8 fev. 2016.

- DORELLA, F. A. et al. *Corynebacterium pseudotuberculosis* : microbiology, biochemical properties, pathogenesis and molecular studies of virulence. **Veterinary Research**, v. 37, n. 2, p. 201–218, mar. 2006.
- DORELLA, F. A. et al. Antigens of *Corynebacterium pseudotuberculosis* and prospects for vaccine development. **Expert Review of Vaccines**, v. 8, n. 2, p. 205–213, 9 fev. 2009.
- FOLEY, J. E. et al. Molecular epidemiologic features of *Corynebacterium pseudotuberculosis* isolated from horses. **American Journal of Veterinary Research**, v. 65, n. 12, p. 1734–1737, 2004.
- GHONEIM, M. A. et al. Role of *Hippobosca equina* as a transmitter of *C. pseudotuberculosis* among buffaloes as revealed by PCR and dot blot hybridization. **Journal of Egyptian Veterinary Medicine Association**, v. 61, p. 165–176, 2001.
- GILLESPIE, J. J. et al. Patric: The comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. **Infection and Immunity**, v. 79, n. 11, p. 4286–4298, 2011.
- HAFEZ, M.; HILALI, M. Biology of *Hippobosca longipennis* (Fabricius, 1805) in Egypt [Diptera. Hippoboscidae]. **Vet Parasitol**, v. 4, p. 275±288, 1978.
- HAGGERTY, L. S. et al. **A pluralistic account of homology: Adapting the models to the data** *Molecular Biology and Evolution*, 2014.
- HALL, J. P. J.; BROCKHURST, M. A.; HARRISON, E. Sampling the mobile gene pool: innovation via horizontal gene transfer in bacteria. **Philosophical transactions of the Royal Society of London. Series B, Biological sciences**, v. 372, n. 1735, p. 20160424, 2017.
- HARD, G. C. Comparative toxic effect of the surface lipid of *Corynebacterium ovis* on peritoneal macrophages. **Infection and Immunity**, v. 12, n. 6, p. 1439–1449, 1975.
- HAWARI, A. D. *Corynebacterium pseudotuberculosis* Infection (Caseous Lymphadenitis) in Camels (*Camelus dromedarius*) in Jordan. **American Journal of Animal and Veterinary Sciences**, v. 3, n. 3, p. 68–72, 1 mar. 2008.
- HEGGELUND, L. et al. *Corynebacterium pseudotuberculosis* Pneumonia in a Veterinary Student Infected During Laboratory Work. **Open Forum Infectious Diseases**, v. 2, n. 2, p. ofv053-ofv053, 14 maio 2015.
- HILL, C. Virulence or Niche Factors: What's in a Name? **Journal of Bacteriology**, v. 194, n. 21, p. 5725–5727, 1 nov. 2012.
- HOLMES, R. K. Biology and molecular epidemiology of diphtheria toxin and the *tox* gene.

The Journal of Infectious Diseases, v. 181, n. s1, p. S156–S167, 2000.

HOOD, M. I.; SKAAR, E. P. Nutritional immunity: Transition metals at the pathogen-host interface. **Nature Reviews Microbiology**, v. 10, n. 8, p. 525–537, 2012.

HUSSEIN, K. H. An unusual case of a huge abscess in a buffalo bull (*Bubalus bubalis*). **Buffalo Bulletin**, v. 31, n. 4, p. 183–185, 2012.

IKEDA, M.; NAKAGAWA, S. **The Corynebacterium glutamicum genome: Features and impacts on biotechnological processes** **Applied Microbiology and Biotechnology**, 2003.

JONES, T. C.; HUNT, R. D.; KING, N. W. **Veterinary Pathology**. 6. ed. [s.l.] Wiley-Blackwell, 2000.

KARAOLIS, D. K. et al. A *Vibrio cholerae* pathogenicity island associated with epidemic and pandemic strains. **Proceedings of the National Academy of Sciences of the United States of America**, v. 95, n. 6, p. 3134–9, 17 mar. 1998.

KHAMIS, A.; RAOULT, D.; LA SCOLA, B. Comparison between *rpoB* and 16S rRNA gene sequencing for molecular identification of 168 clinical isolates of *Corynebacterium*. **Journal of Clinical Microbiology**, v. 43, n. 4, p. 1934–1936, abr. 2005.

KOONIN, E. V. Orthologs, Paralogs, and Evolutionary Genomics 1. **Annual review of genetics**, v. 39, n. 1, p. 309–338, dez. 2005.

LAND, M. et al. Insights from 20 years of bacterial genome sequencing. **Functional & Integrative Genomics**, v. 15, n. 2, p. 141–161, 2015.

LASSALLE, F.; MULLER, D.; NESME, X. Ecological speciation in bacteria: reverse ecology approaches reveal the adaptive part of bacterial cladogenesis. **Research in microbiology**, v. 166, n. 10, p. 729–41, dez. 2015.

LOMAN, N. J.; PALLEN, M. J. Twenty years of bacterial genome sequencing. **Nature Reviews Microbiology**, v. 13, n. 12, p. 787–794, 2015.

MANDLIK, A. et al. Pili in Gram-positive bacteria: assembly, involvement in colonization and biofilm development. **Trends in microbiology**, v. 16, n. 1, p. 33–40, 2008.

MARIANO, D. C. B. et al. Whole-genome optical mapping reveals a mis-assembly between two rRNA operons of *Corynebacterium pseudotuberculosis* strain 1002. **BMC genomics**, v. 17, n. 1, p. 315, 2016.

MCKEAN, S. C.; DAVIES, J. K.; MOORE, R. J. Expression of phospholipase D, the major virulence factor of *Corynebacterium pseudotuberculosis* is regulated by multiple

environmental factors and plays a role in macrophage death. **Microbiology**, v. 153, n. 7, p. 2203–2211, 2007.

MEDINI, D. et al. The microbial pan-genome. **Current Opinion in Genetics & Development**, v. 15, n. 6, p. 589–594, dez. 2005.

MEYER, F.; OVERBEEK, R.; RODRIGUEZ, A. FIGfams: yet another set of protein families. **Nucleic Acids Research**, v. 37, n. 20, p. 6643–6654, 1 nov. 2009.

MOHAMED MOUSSA, I. et al. Single-point mutation as a molecular tool for preparation of recombinant vaccine against Caseous Lymphadenitis. **Journal of Food, Agriculture and Environment**, v. 12, n. 2, p. 626–629, 2014.

MOUSSA, I. M. et al. Vaccination against *Corynebacterium pseudotuberculosis* infections controlling caseous lymphadenitis (CLA) and oedematous skin disease. **Saudi Journal of Biological Sciences**, v. 23, n. 6, p. 718–723, nov. 2016.

MURPHY, J. R. Mechanism of Diphtheria Toxin Catalytic Domain Delivery to the Eukaryotic Cell Cytosol and the Cellular Factors that Directly Participate in the Process. **Toxins**, v. 3, n. 12, p. 294–308, 21 mar. 2011.

MUZZI, A.; MASIGNANI, V.; RAPPUOLI, R. The pan-genome: towards a knowledge-based discovery of novel targets for vaccines and antibacterials. **Drug Discovery Today**, v. 12, n. 11–12, p. 429–439, 2007.

NAKAMURA, Y. et al. The genome stability in *Corynebacterium* species due to lack of the recombinational repair system. **Gene**, v. 317, p. 149–155, out. 2003.

OVERBEEK, R. The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes. **Nucleic Acids Research**, v. 33, n. 17, p. 5691–5702, 25 set. 2005.

OVERBEEK, R. et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). **Nucleic acids research**, v. 42, n. Database issue, p. D206-14, jan. 2014.

PEARSON, W. R. An Introduction to Sequence Similarity (“Homology”) Searching. In: **Current Protocols in Bioinformatics**. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2013. v. 1p. 1286–1292.

PEEL, M. M. et al. Human lymphadenitis due to *Corynebacterium pseudotuberculosis*: report of ten cases from Australia and review. **Clinical infectious diseases : an official publication of the Infectious Diseases Society of America**, v. 24, n. 2, p. 185–91, fev.

1997.

PERIWAL, V.; SCARIA, V. Insights into structural variations and genome rearrangements in prokaryotic genomes. **Bioinformatics**, v. 31, n. 1, p. 1–9, 2015.

PINTO, A. C. et al. Differential transcriptional profile of *Corynebacterium pseudotuberculosis* in response to abiotic stresses. **BMC genomics**, v. 15, p. 14, 2014.

RAYNAL, J. T. et al. Identification of membrane-associated proteins with pathogenic potential expressed by *Corynebacterium pseudotuberculosis* grown in animal serum. **BMC Research Notes**, v. 11, n. 1, p. 1–6, 2018.

ROULI, L. et al. The bacterial pangenome as a new tool for analysing pathogenic bacteria. **New Microbes and New Infections**, v. 7, p. 72–85, set. 2015.

RUIZ, J. C. et al. Evidence for reductive genome evolution and lateral acquisition of virulence functions in two *Corynebacterium pseudotuberculosis* strains. **PLoS ONE**, v. 6, n. 4, 2011.

SELIM, S. A. Oedematous Skin Disease of Buffalo in Egypt. **Journal of Veterinary Medicine Series B**, v. 48, n. 4, p. 241–258, 24 maio 2001.

SELIM, S. A. et al. Immunological characterization of diphtheria toxin recovered from *Corynebacterium pseudotuberculosis*. **Saudi Journal of Biological Sciences**, p. 0–5, 2015.

SING, A. et al. *Corynebacterium diphtheriae* in a free-roaming red fox: case report and historical review on diphtheria in animals. **Infection**, 2015.

SNYDER, E. E. et al. PATRIC: The VBI PathoSystems Resource Integration Center. **Nucleic Acids Research**, v. 35, n. Database, p. D401–D406, 3 jan. 2007.

SOARES, S. C. et al. 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, n. 1, 2013a.

SOARES, S. C. et al. Genome sequence of *Corynebacterium pseudotuberculosis* biovar *equi* strain 258 and prediction of antigenic targets to improve biotechnological vaccine production. **Journal of Biotechnology**, v. 167, n. 2, p. 135–141, ago. 2013b.

SOARES, S. C. et al. GIPSY: Genomic island prediction software. **Journal of Biotechnology**, v. 232, p. 2–11, ago. 2016.

SONGER, J. G. et al. Biochemical and genetic characterization of *Corynebacterium pseudotuberculosis*. **American Journal of Veterinary Research**, v. 49, n. 2, p. 223–226, 1988.

SPIER, S. J.; AZEVEDO, V. *Corynebacterium pseudotuberculosis* infection in horses: Increasing frequency and spread to new regions of North America. **Equine Veterinary Education**, maio 2016.

SUTHERLAND, S. S.; HART, R. A.; BULLER, N. B. Genetic differences between nitrate-negative and nitrate-positive *C. pseudotuberculosis* strains using restriction fragment length polymorphisms. **Veterinary Microbiology**, v. 49, n. 1–2, p. 1–9, 1996.

SYAME, S. M.; EL-HEWAIRY, H. M.; SELIM, S. A. Protection of Buffaloes Against Oedematous Skin Disease by Recombinant-bacterin and Toxoid-bacterin Vaccines. **Global Veterinaria**, v. 2, n. 4, p. 151–156, 2008.

TAKAHASHI, T. et al. Phylogenetic positions and assignment of swine and ovine corynebacterial isolates based on the 16S rDNA sequence. **Microbiology and immunology**, v. 41, n. 9, p. 649–655, 1997.

TAUCH, A. et al. Complete genome sequence and analysis of the multiresistant nosocomial pathogen *Corynebacterium jeikeium* K411, a lipid-requiring bacterium of the human skin flora. **Journal of bacteriology**, v. 187, n. 13, p. 4671–82, 1 jul. 2005.

TAUCH, A.; BURKOVSKI, A. Molecular armory or niche factors: virulence determinants of *Corynebacterium* species. **FEMS Microbiology Letters**, v. 67, n. 2, p. fnv185, 7 out. 2015.

TEJEDOR-JUNCO, M. T. et al. Isolation of nitrate-reductase positive *Corynebacterium pseudotuberculosis* from dromedary camels. **Tropical Animal Health and Production**, v. 40, n. 3, p. 165–167, 2008.

TETTELIN, H. et al. Comparative genomics: the bacterial pan-genome. **Current Opinion in Microbiology**, v. 11, n. 5, p. 472–477, out. 2008.

TROST, E. et al. The complete genome sequence of *Corynebacterium pseudotuberculosis* FRC41 isolated from a 12-year-old girl with necrotizing lymphadenitis reveals insights into gene-regulatory networks contributing to virulence. **BMC genomics**, v. 11, n. 1, p. 728, 2010.

VANEECHOUTTE, M. et al. Evaluation of the applicability of amplified rDNA-restriction analysis (ARDRA) to identification of species of the genus *Corynebacterium*. **Research in microbiology**, v. 146, n. 8, p. 633–641, 1995.

VIANA, M. V. C. et al. Comparative genomic analysis between *Corynebacterium pseudotuberculosis* strains isolated from buffalo. **PLOS ONE**, v. 12, n. 4, p. e0176347, 26 abr. 2017.

- VOS, M. A species concept for bacteria based on adaptive divergence. **Trends in microbiology**, v. 19, n. 1, p. 1–7, jan. 2011.
- WATTAM, A. R. et al. PATRIC, the bacterial bioinformatics database and analysis resource. **Nucleic acids research**, v. 42, n. Database issue, p. D581-91, jan. 2014a.
- WATTAM, A. R. et al. Comparative phylogenomics and evolution of the Brucellae reveal a path to virulence. **Journal of bacteriology**, v. 196, n. 5, p. 920–30, mar. 2014b.
- WEBB, S. A.; KAHLER, C. M. Bench-to-bedside review: Bacterial virulence and subversion of host defences. **Critical Care**, v. 12, n. 6, p. 234, 2008.
- WILLIAMSON, L. H. Caseous Lymphadenitis in Small Ruminants. **Veterinary Clinics of North America: Food Animal Practice**, v. 17, n. 2, p. 359–371, jul. 2001.
- WOLF, Y. I.; KOONIN, E. V. A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. **Genome Biology and Evolution**, v. 4, n. 12, p. 1286–1294, 2012.
- YANAGAWA, R.; HONDA, E. Presence of pili in species of human and animal parasites and pathogens of the genus *Corynebacterium*. **Infection and Immunity**, v. 13, n. 4, p. 1293–1295, 1976.
- YERUHAM, I. et al. Mastitis in Dairy Cattle Caused by *Corynebacterium pseudotuberculosis* and the Feasibility Of Transmission by Houseflies I. **Veterinary Quarterly**, v. 18, n. 3, p. 87–89, set. 1996.
- YERUHAM, I. et al. A herd level analysis of a *Corynebacterium pseudotuberculosis* outbreak in a dairy cattle herd. **Veterinary Dermatology**, v. 15, n. 5, p. 315–320, 2004.

VIII. Appendix

A. Published, accepted and submitted research articles

Genome Announcements

(My contributions to these papers)

These papers have work from national and international collaborations made by our groups, with the purpose of increasing the available genomic data. With this genomics projects, I got familiar with Next Generation Sequencing, genome assembly, and automatic and manual genome annotation.

My contribution to these papers were genome assembly, annotation, and manuscript preparation.

Research Article 1

Marcus Vinicius Canário Viana, Alice Rebecca Wattam, Dhvani Govil Batra, Sébastien Boisvert, Thomas Scott Brettin, Michael Frace, Fangfang Xia, Vasco Azevedo, Rebekah Tiller, Alex R. Hoffmaster. Genome sequences of three *Brucella canis* strains isolated from humans and a dog. **Genome Announcements**, v. 5, p. e01688-16- e01688-16, 2017. doi: **10.1128/genomeA.01688-16**.

Abstract

Brucella canis is a facultative intracellular pathogen that has members of Canidae family as its reservoir hosts. We report the genome sequencing of two *Brucella canis* strains isolated from a human and one isolated from a dog host.



Genome Sequences of Three *Brucella canis* Strains Isolated from Humans and a Dog

Marcus Vinicius Canário Viana,^a Alice Rebecca Wattam,^b Dhvani Govil Batra,^c Sébastien Boisvert,^d Thomas Scott Brettn,^{e,f} Michael Frace,^g Fangfang Xia,^h Vasco Azevedo,^a Rebekah Tiller,^c Alex R. Hoffmasterⁱ

Laboratório de Genética Celular e Molecular, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil^a; Biocomplexity Institute of Virginia Tech, Virginia Tech University, Blacksburg, Virginia, USA^b; Centers for Disease Control and Prevention, Atlanta, Georgia, USA^c; Gydle, Inc., Québec, Canada^d; Computing, Environment and Life Sciences, Argonne National Laboratory, Argonne, Illinois, USA^e; Computation Institute, University of Chicago, Chicago, Illinois, USA^f; Biotechnology Core Facility Branch, National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, USA^g; Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois, USA^h; Bacterial Special Pathogens Branch, Division of High-Consequence Pathogens and Pathology, Centers for Disease Control and Prevention, Atlanta, Georgia, USAⁱ

ABSTRACT *Brucella canis* is a facultative intracellular pathogen that preferentially infects members of the Canidae family. Here, we report the genome sequencing of two *Brucella canis* strains isolated from humans and one isolated from a dog host.

Brucellosis is a zoonotic disease caused by *Brucella* species that infect a diverse array of land and aquatic mammals, with humans as an accidental host (1). It is the most common zoonosis worldwide and induces an often chronic and incapacitating disease in humans, with low mortality (2). *Brucella* are stealth pathogens that have adapted to cause long-term chronic infections in their hosts (3). They are Gram-negative coccobacilli that live intracellularly within vertebrate hosts. *Brucella canis* is named due to its preference for hosts that belong to species of Canidae family (4), which includes dogs, coyotes, wolves, and foxes. Clinical signs of canine brucellosis range from asymptomatic infections to abortion and testicular atrophy. It is sexually transmitted, or by contact with aborted fetuses, food, or from an environment that has been contaminated by aborted material or excreta (5). *B. canis* infrequently infects humans, and the reported cases from such infection are usually described as mild (5).

Herein, we report the sequencing of three *B. canis* genomes. Strain 2009004498 was isolated from a human in Louisiana, USA, in January 2009. Strain 2009013648 was isolated from a human host in Arizona, USA, in April 2009. Strain 2010009751 was isolated from a dog (*Canis lupus familiaris*) in Massachusetts, USA, in January 2010.

The isolates were typed as *B. canis* sequence type 20 (ST20) by multilocus sequence typing (MLST) (6) and clustered with *B. canis* strains by multiple-locus variable-number tandem-repeat analysis (MLVA) (7). The genomes were sequenced using Roche 454 platform. The reads were checked for quality by FastQC 0.11.4 (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) and assembled by Newbler version 2.9 using a *de novo* strategy. Scaffolding was performed by CONTIGuator 2.7 (8), using *B. canis* ATCC 23365 (accession numbers CP000872.1 and CP000873.1) as reference. The gaps were closed by mapping the sequencing reads of each genome to *B. canis* ATCC 23365 and by extracting the consensus sequences (9) using CLC Genomics Workbench 6.5 (Qiagen, USA). A consistent automatic annotation by was generated by RASTtk (10) at the PATRIC bioinformatics resource center (11, 12).

Received 13 December 2016 Accepted 16 December 2016 Published 23 February 2017

Citation Viana MVC, Wattam AR, Govil Batra D, Boisvert S, Brettn TS, Frace M, Xia F, Azevedo V, Tiller R, Hoffmaster AR. 2017. Genome sequences of three *Brucella canis* strains isolated from humans and a dog. Genome Announc 5:e01688-16. <https://doi.org/10.1128/genomeA.01688-16>.

Copyright © 2017 Viana et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Address correspondence to Alice Rebecca Wattam, rwattam@vbl.vt.edu.

Research Article 2

Marcus Vinicius Canário Viana, Alice Rebecca Wattam, Dhvani Govil Batra, Sébastien Boisvert, Thomas Scott Brettin, Michael Frace, Fangfang Xia, Vasco Azevedo, Rebekah Tiller, Alex R. Hoffmaster. Genome sequences of two *Brucella suis* strains isolated from the same patient, 8 years apart. **Genome Announcements**, v. 5, p. e01687-16- e01687-16, 2017. doi: **10.1128/genomeA.01687-16**.

Abstract

Brucella suis is a Gram-negative, facultative intracellular pathogen that has pigs as its preferred host but can infect humans also. Herein, we report the draft genome sequences of two *B. suis* strains, isolated from the same patient, 8 years apart.



Genome Sequences of Two *Brucella suis* Strains Isolated from the Same Patient, 8 Years Apart

Marcus Vinicius Canário Viana,^a Alice Rebecca Wattam,^b Dhvani Govil Batra,^c Sébastien Boisvert,^d Thomas Scott Brettn,^{e,f} Michael Frace,^g Fangfang Xia,^h Vasco Azevedo,^a Rebekah Tiller,^c Alex R. Hoffmasterⁱ

Laboratório de Genética Celular e Molecular, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil^a; Biocomplexity Institute of Virginia Tech, Virginia Tech University, Blacksburg, Virginia, USA^b; Centers for Disease Control and Prevention, Atlanta, Georgia, USA^c; Life Sciences, Argonne National Laboratory, Argonne, Illinois, USA^d; Computation Institute, University of Chicago, Chicago, Illinois, USA^e; Biotechnology Core Facility Branch, National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, USA^f; Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois, USA^g; Bacterial Special Pathogens Branch, Division of High-Consequence Pathogens and Pathology, Centers for Disease Control and Prevention, Atlanta, Georgia, USA^h

ABSTRACT *Brucella suis* is a Gram-negative, facultative intracellular pathogen that has pigs as its preferred host, but it can also infect humans. Here, we report the draft genome sequences of two *B. suis* strains that were isolated from the same patient, 8 years apart.

Brucellosis is the most common zoonosis worldwide caused by bacteria of the *Brucella* genus, transmitted by direct or indirect contact with infected animals or their products. Its chronic infection causes infertility and abortion in animals, which is not only a burden for the animals but also has an economic impact. In humans, infection with *Brucella* causes an acute or subacute febrile illness, along with varied and nonspecific clinical manifestations that include fever, sweats, fatigue, malaise, anorexia, weight loss, headache, arthralgia, and back pain (1). *Brucella* is part of the family *Brucellaceae*, order *Rhizobiales*, and class *Alphaproteobacteria* (2). All members of this genus are Gram-negative, facultative intracellular pathogens, and 10 species have been described that are differentiated by the mammalian host that they prefer to infect, as well as a set of antigenic and metabolic phenotypes (3). *Brucella* spp. cause long-term chronic infections as intracellular pathogens, and they are noted as having reduced genomes, a type IV secretion system, a perosamine-based O antigen, and reduced virulence (4).

Brucella suis generally is found to infect pigs, but it can also infect other animals, including humans. Although *B. melitensis* is the most common species causing human disease, *B. suis* also commonly affects them and is the most common *Brucella* sp. to infect people in Argentina (5, 6).

B. suis strains 2004000577 and 2011017258 were isolated from a recrudescence case from a single human patient but were collected 8 years apart (2003 and 2011, respectively) in Massachusetts, USA. Both genomes were collected from the same leg wound, the first collection in October 2003, and the second in April 2011. The genomes were sequenced using the Roche 454 platform. For strain 2004000577, a *de novo* assembly was performed with Newbler version 2.9. Scaffolding was done with CONTIGuator version 2.7, using *B. suis* 1330 (AE014291.4, AE014292.2) as the reference genome. Gaps closure was performed by reference assembly (7), using CLC Genomics Workbench version 6.5 (Qiagen, USA) and *B. suis* 1330 as the reference genome. For strain 2011017258, a reference assembly was done by mapping the sequencing reads

Received 13 December 2016 Accepted 23 December 2016 Published 2 March 2017

Citation Viana MVC, Wattam AR, Govil Batra D, Boisvert S, Brettn TS, Frace M, Xia F, Azevedo V, Tiller R, Hoffmaster AR. 2017. Genome sequences of two *Brucella suis* strains isolated from the same patient, 8 years apart. *Genome Announc* 5:e01687-16. <https://doi.org/10.1128/genomeA.01687-16>.

Copyright © 2017 Viana et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Alice Rebecca Wattam, rwattam@vbl.vt.edu.

Research Article 3

Raquel Enma Hurtado, Flavia Aburjaile, Diego Mariano, **Marcus Vinicius Canário**, Leandro Benevides, Daniel Antonio Fernandez, Nataly Olivia Allasi, Rocio Rimac, Julio Eduardo Juscamayta, Jorge Enrique Maximiliano, Raul Hector Rosadio, Vasco Azevedo, Lenin Maturrano. Draft Genome Sequence of a Virulent Strain of *Pasteurella multocida* Isolated From Alpaca. **Journal of Genomics**, v. 5, p. 68-70, 2017. doi:10.7150/jgen.19297.

Abstract

Pasteurella multocida is one of the most frequently isolated bacteria in acute pneumonia cases, being responsible for high mortality rates in Peruvian young alpacas, with consequent social and economic costs. Here we report the genome sequence of *P. multocida* strain UNMSM, isolated from the lung of an alpaca diagnosed with pneumonia, in Peru. The genome consists of 2,439,814 base pairs assembled into 82 contigs and 2,252 protein encoding genes, revealing the presence of known virulence-associated genes (*ompH*, *ompA*, *tonB*, *tbpA*, *nanA*, *nanB*, *nanH*, *sodA*, *sodC*, *plpB* and *toxA*). Further analysis could provide insights about bacterial pathogenesis and control strategies of this disease in Peruvian alpacas.



Short Research Paper

Draft Genome Sequence of a Virulent Strain of *Pasteurella Multocida* Isolated From Alpaca

Raquel Erma Hurtado¹, Flavia Aburjaile², Diego Mariano², Marcus Vinicius Canário², Leandro Benevides², Daniel Antonio Fernandez¹, Nataly Olivia Allasi¹, Rocio Rimac¹, Julio Eduardo Juscamayta¹, Jorge Enrique Maximiliano¹, Raul Hector Rosadio¹, Vasco Azevedo² and Lenin Maturrano¹✉

1. Laboratory of Molecular Biology and Genetics, Veterinary Medicine Faculty, San Marcos University, Lima, Peru;
2. Laboratory of Cellular and Molecular Genetics, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil.

✉ Corresponding author: Tel.: +551 956533581, E-mail: amaturrano@unmsm.edu.pe (L. Maturrano)

© Ivyspring International Publisher. This is an open access article distributed under the terms of the Creative Commons Attribution (CC BY-NC) license (<https://creativecommons.org/licenses/by-nc/4.0/>). See <http://ivyspring.com/terms> for full terms and conditions.

Received: 2017.01.22; Accepted: 2017.05.14; Published: 2017.06.28

Abstract

Pasteurella multocida is one of the most frequently isolated bacteria in acute pneumonia cases, being responsible for high mortality rates in Peruvian young alpacas, with consequent social and economic costs. Here we report the genome sequence of *P. multocida* strain UNMSM, isolated from the lung of an alpaca diagnosed with pneumonia, in Peru. The genome consists of 2,439,814 base pairs assembled into 82 contigs and 2,252 protein encoding genes, revealing the presence of known virulence-associated genes (*ompH*, *ompA*, *tonB*, *tbpA*, *nanA*, *nanB*, *nanH*, *sodA*, *sodC*, *plpB* and *toxA*). Further analysis could provide insights about bacterial pathogenesis and control strategies of this disease in Peruvian alpacas.

Key words: Alpaca, genome, pasteurellosis, pneumonia.

Introduction

Pasteurella multocida is a commensal bacteria from the upper respiratory tract [1], which affects a wide range of hosts [2, 3]. This bacteria is the primary agent of many infections such as; avian cholera, hemorrhagic septicemia in ungulates, atrophic rhinitis in pigs and snuffles in rabbits [3], and acts as a secondary agent in infectious pneumonia, including cases of acute or chronic pneumonia in different hosts such as swine, calves, sheep, bovine and alpaca [3-5].

In Peru, alpaca raising represents an important economic activity for the High Andean population. However, acute pneumonia causes high mortality rates in young alpacas, in which *P. multocida* has been principally isolated [6]. *P. multocida* has a large number of virulence factors that play a role in pathogenesis, including capsule, lipopolysaccharide, fimbriae, adhesins, toxins, outer membrane proteins, iron regulated and iron acquisition proteins, acquisition proteins, hyaluronidase and sialidase [6].

In this study, we announce the draft genome of *P. multocida* strain UNMSM isolated from an alpaca lung affected with pneumonia.

Pasteurella multocida strain UNMSM is a gram negative, short rod shaped bacteria, oxidase and catalase positive and nonhemolytic, with approximate measures of 0.3 - 0.6 µm in width and 0.8 - 2.0 µm in length (Figure 1). Genome sequencing was performed using Illumina HiSeq sequencing platform. The paired-end library contained inserts of an average size of 100 bp. *De novo* assembly was performed using Edena v3.131028 and SIMBA v1.4 software [7], which produced 82 contigs, with a N50 value of 70,838, 2.4 Mb of size and mean depth coverage ~400-fold. The genome was annotated using the Rapid Annotations using Subsystems Technology (RAST) [8], following by manual curation of the predicted CDSs (Coding Sequences). The genome presents GC content around 40.2%. A total of 2,434

Research Article 4

Luis C. Guimarães, **Marcus V. C. Viana**, Leandro J. Benevides, Diego C. B. Mariano, Adooney A. O. Veras, Pablo H. C. Sá, Flávia S. Rocha, Priscilla C. B. Vilas Boas, Siomar C. Soares, Maria S. Barbosa, Nicole Guiso, Edgar Badell, Adriana R. Carneiro, Vasco Azevedo, Rommel T. J. Ramos, Artur Silva. Draft Genome Sequence of Toxigenic *Corynebacterium ulcerans* Strain 04-7514, Isolated from a Dog in France. **Genome Announcements**, v. 4, p. e00172-16, 2016. doi: 10.1128/genomeA.00172-16.

Abstract

Here, we present the draft genome of toxigenic *Corynebacterium ulcerans* strain 04-7514. The draft genome has 2,497,845 bp, 2,059 coding sequences, 12 rRNA genes, 46 tRNA genes, 150 pseudogenes, 1 clustered regularly interspaced short palindromic repeat (CRISPR) array, and a G+C content of 53.50%.



Draft Genome Sequence of Toxigenic *Corynebacterium ulcerans* Strain 04-7514, Isolated from a Dog in France

Luís C. Guimarães,^a Marcus V. C. Viana,^b Leandro J. Benevides,^b Diego C. B. Mariano,^b Adooney A. O. Veras,^a Pablo H. C. Sá,^a Flávia S. Rocha,^b Priscilla C. B. Vilas Boas,^b Siomar C. Soares,^c Maria S. Barbosa,^a Nicole Guiso,^d Edgar Badell,^d Adriana R. Carneiro,^a Vasco Azevedo,^b Rommel T. J. Ramos,^a Artur Silva^a

Institute of Biological Sciences, Federal University of Pará (UFPA), Belém, Pará, Brazil^a; Institute of Biological Sciences, Federal University of Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brazil^b; Department of Immunology, Microbiology and Parasitology, Institute of Biological Sciences and Natural Sciences, Federal University of Triângulo Mineiro (UFTRM), Uberaba, Belo Horizonte, Brazil^c; Institut Pasteur, Unité de Prévention et Thérapies Moléculaires des Maladies Humaines, National Centre of Reference of Toxigenic *Corynebacterium*, Paris, France^d

Here, we present the draft genome of toxigenic *Corynebacterium ulcerans* strain 04-7514. The draft genome has 2,497,845 bp, 2,059 coding sequences, 12 rRNA genes, 46 tRNA genes, 150 pseudogenes, 1 clustered regularly interspaced short palindromic repeat (CRISPR) array, and a G+C content of 53.50%.

Received 4 February 2016 Accepted 15 February 2016 Published 31 March 2016

Citation Guimarães LC, Viana MVC, Benevides LJ, Mariano DCB, Veras AAO, Sá PHC, Rocha FS, Vilas Boas PCB, Soares SC, Barbosa MS, Guiso N, Badell E, Carneiro AR, Azevedo V, Ramos RTJ, Silva A. 2016. Draft genome sequence of toxigenic *Corynebacterium ulcerans* strain 04-7514, isolated from a dog in France. *Genome Announc* 4(2):e00172-16. doi:10.1128/genomeA.00172-16.

Copyright © 2016 Guimarães et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Artur Silva, asilva@ufpa.br.

Corynebacterium ulcerans is an important zoonotic bacterium, and case report infections have been increasing worldwide during the last decade (1–3). This bacterium is Gram-positive, nonmotile, pleomorphic, arranged in palisades or V-shaped forms, and non-spore forming. It is facultatively anaerobic, catalase positive, nitrate negative, oxidase negative, and differs from other species of the genus by the fermentation of glycogen and starch (4).

C. ulcerans exhibits levels of genomic DNA relatedness with *Corynebacterium diphtheriae* and *Corynebacterium pseudotuberculosis*. Furthermore, taxonomic analyzes of 16S rRNA gene sequences highlight the close phylogenetic relationship between these three species, putting them in a distinct cluster of the genus *Corynebacterium* (1, 4).

Additionally, in 1926, a strain of *C. ulcerans* coding for diphtheria toxin was isolated from the human throat (5). This diphtheria toxin has 95% similarity compared to the diphtheria toxin present in *C. diphtheriae* (2). Nevertheless, nontoxigenic *C. ulcerans* strains have been reported to code for a powerful and severe dermonecrotic toxin similar to phospholipase D from *C. pseudotuberculosis* (6). This repertoire of potent toxins shared for these three species corroborates the apparent relationship between them.

In this study, we present the draft genome sequence of toxigenic *C. ulcerans* strain 04-7514. This strain was isolated from a dog in Bourges, France. The strain is part of the Collection of Institut Pasteur (CIP) (<https://www.pasteur.fr/en>). These were kindly given to the Laboratory of Genomics and System Biology located at the Federal University of Pará, Belém, Pará, Brazil, and the Laboratory of Cellular and Molecular Genetics located at the Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil.

The genome sequencing was performed using the next-

generation sequencing SOLiD platform, using a fragment library. The predicted genome coverage was approximately 6,000×, based on *C. ulcerans* genomes available in GenBank (<http://www.ncbi.nlm.nih.gov/genbank/>). The *de novo* assembly strategy was performed using Velvet software version 1.2.10 (7), generating 28 contigs with 2,497,845 bp. The contigs were submitted to GenBank for automatic annotation. The genome has 2,059 coding sequences, 12 rRNA genes, 46 tRNA genes, 150 pseudogenes, 1 clustered regularly interspaced short palindromic repeat (CRISPR) array, and a G+C content of 53.50%. This genome is part of further studies of comparative genomics, pathogenicity, and vaccine and drug targets of the species.

Nucleotide sequence accession numbers. The *C. ulcerans* whole-genome shotgun (WGS) project has the project accession no. LJVH00000000. The version described in this paper is version LJVH01000000 and consists of sequences LJVH01000001 to LJVH01000028.

ACKNOWLEDGMENTS

This work was supported by the Coordination for the Improvement of Higher Education Personnel (CAPES) and National Council for Scientific and Technological Development (CNPq) and the Genome and Proteome Network of the State of Pará (RPGP).

REFERENCES

- Trost E, Al-Dilaimi A, Papavasiliou P, Schneider J, Viehoveer 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. 2011. Comparative analysis of two complete *Corynebacterium ulcerans* genomes and detection of candidate virulence factors. *BMC Genomics* 12:383. <http://dx.doi.org/10.1186/1471-2164-12-383>.
- Sing A, Bierschenk S, Heesemann J. 2004. Classical diphtheria caused by *Corynebacterium ulcerans* in Germany: amino acid sequence differences

Research Article 5

Luis C. Guimarães, **Marcus V. C. Viana**, Leandro J. Benevides, Diego C. B. Mariano, Adonney A. O. Veras, Pablo H. C. Sá, Flávia S. Rocha, Priscilla C. B. Vilas Boas, Siomar C. Soares, Maria S. Barbosa, Nicole Guiso, Edgar Badell, Vasco Azevedo, Rommel T. J. Ramos, Artur Silva. Draft Genome Sequence of *Corynebacterium ulcerans* Strain 04-3911, Isolated from Humans. **Genome Announcements**, v. 4, p. e00171-16, 2016. doi: **10.1128/genomeA.00171-16**.

Abstract

Corynebacterium ulcerans is an emergent pathogen infecting wild and domesticated animals worldwide that may serve as reservoirs for zoonotic infections. In this study, we present the draft genome of *C. ulcerans* strain 03-8664. The draft genome has 2,428,683 bp, 2,262 coding sequences, and 12 rRNA genes.



Draft Genome Sequence of *Corynebacterium ulcerans* Strain 04-3911, Isolated from Humans

Luis C. Guimarães,^a Marcus V. C. Viana,^b Leandro J. Benevides,^b Diego C. B. Mariano,^b Adooney A. O. Veras,^a Pablo H. C. Sá,^a Flávia S. Rocha,^b Priscilla C. B. Vilas Boas,^b Siomar C. Soares,^c Maria S. Barbosa,^a Nicole Guiso,^d Edgar Badell,^d Adriana R. Carneiro,^a Vasco Azevedo,^a Rommel T. J. Ramos,^a Artur Silva^a

Institute of Biological Sciences, Federal University of Pará (UFPA), Belém, Pará, Brazil^a; Institute of Biological Sciences, Federal University of Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brazil^b; Department of Immunology, Microbiology and Parasitology, Institute of Biological Sciences and Natural Sciences, Federal University of Triângulo Mineiro (UFTM), Uberaba, Minas Gerais, Brazil^c; Institut Pasteur, Unité de Prévention et Thérapies Moléculaires des Maladies Humaines, National Centre of Reference of Toxicogenic *Corynebacteria*, Paris, France^d

***Corynebacterium ulcerans* is a pathogenic bacterium infecting wild and domesticated animals; some infection cases in humans have increased throughout the world. The current study describes the draft genome of strain 04-3911, isolated from humans. The draft genome has 2,492,680 bp, 2,143 coding sequences, 12 rRNA genes, and 50 tRNA genes.**

Received 4 February 2016 Accepted 15 February 2016 Published 31 March 2016

Citation Guimarães LC, Viana MVC, Benevides LJ, Mariano DCB, Veras AAO, Sá PHC, Rocha FS, Vilas Boas PCB, Soares SC, Barbosa MS, Guiso N, Badell E, Carneiro AR, Azevedo V, Ramos RTJ, Silva A. 2016. Draft genome sequence of *Corynebacterium ulcerans* strain 04-3911, isolated from humans. *Genome Announc* 4(2):e00171-16. doi:10.1128/genomeA.00171-16.

Copyright © 2016 Guimarães et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Artur Silva, asilva@ufpa.br.

Corynebacterium ulcerans belongs to a suprageneric group of actinomycetes that also includes the genera *Mycobacterium*, *Nocardia*, and *Rhodococcus*, termed the CMNR group (*Corynebacterium*, *Mycobacterium*, *Nocardia*, and *Rhodococcus*). This bacterium is facultative anaerobic, nonsporulating, nonmotile, catalase positive, and nitrate and oxidase negative (1, 2). Analyses of 16S rRNA gene sequences showed that *C. ulcerans*, *Corynebacterium pseudotuberculosis*, and *Corynebacterium diphtheriae* are closely related (2). Otherwise, there are *C. ulcerans* strains coding for a diphtheria toxin similar to that encoded by toxigenic strains of *Corynebacterium diphtheriae* (3), as well as *C. ulcerans* strains coding for a dermonecrotic toxin with similarity to toxic phospholipase D (PLD) from *C. pseudotuberculosis* (4).

C. ulcerans has been detected in a variety of wild and domesticated animals, suggesting that both groups may attend as reservoirs for zoonotic transmission (5). This bacterium is an emergent pathogen due its frequency and severity of human infections reported during the last two decades in various countries (6). In humans, it causes diphtheria-like disease, pharyngitis, sinusitis, tonsillitis, pulmonary nodules, and skin ulcers (7). Here, we present the draft genome sequence of *C. ulcerans* 04-3911, isolated from humans. The genome sequencing was performed by the SOLiD platform, using a fragment library. The predicted genome coverage was approximately 6,000×, based on *C. ulcerans* genomes available in GenBank (<http://www.ncbi.nlm.nih.gov/genbank/>). The software Velvet version 1.2.10 (8) was used with a *de novo*-assembled strategy. The assembly generated 40 contigs, with 2,492,680 bp, which were submitted to GenBank to automatic annotation. The genome has 2,143 coding sequences, 12 rRNA genes, 50 tRNA genes, 55 pseudogenes, 2 clustered regularly interspaced short palindromic repeat (CRISPR) arrays, and

a G+C content of 53.30%. This genome is part of further studies of comparative genomics, pathogenicity, and vaccine and drug targets of the species.

Nucleotide sequence accession numbers. The *C. ulcerans* whole-genome shotgun (WGS) project has the project accession no. LGSX000000000. The version described in this paper is version LGSX01000000 and consists of sequences LGSX01000001 to LGSX01000040.

ACKNOWLEDGMENTS

This work was supported by the Coordination for the Improvement of Higher Education Personnel (CAPES) and National Council for Scientific and Technological Development (CNPq) and the Genome and Proteome Network of the State of Pará (RPGP).

REFERENCES

- Dorella FA, Pacheco LGC, Oliveira SC, Miyoshi A, Azevedo V. 2006. *Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. *Vet Res* 37:201–218. <http://dx.doi.org/10.1051/vetres:2005056>.
- Riegel P, Ruimy R, De Briel B, Prevost G, Jehl F, Christen R, Monteil H. 1995. Taxonomy of *Corynebacterium diphtheriae* and related taxa, with recognition of *Corynebacterium ulcerans* sp. nov. *nom. rev.* *FEMS Microbiol Lett* 126:271–276. <http://dx.doi.org/10.1111/j.1574-6968.1995.tb07429.x>.
- Sing A, Hogardt M, Bierschenk S, Heesemann J. 2003. Detection of differences in the nucleotide and amino acid sequences of diphtheria toxin from *Corynebacterium diphtheriae* and *Corynebacterium ulcerans* causing extrapharyngeal infections. *J Clin Microbiol* 41:4848–4851. <http://dx.doi.org/10.1128/JCM.41.10.4848-4851.2003>.
- McNamara PJ, Cuevas WA, Songer JG. 1995. Toxic phospholipases D of *Corynebacterium pseudotuberculosis*, *C. ulcerans* and *Arcanobacterium haemolyticum*: cloning and sequence homology. *Gene* 156:113–118. [http://dx.doi.org/10.1016/0378-1119\(95\)00002-N](http://dx.doi.org/10.1016/0378-1119(95)00002-N).

Research Article 6

Luis C. Guimarães, **Marcus V. C. Viana**, Leandro J. Benevides, Diego C. B. Mariano, Adonney A. O. Veras, Pablo H. C. Sá, Flávia S. Rocha, Priscilla C. B. Vilas Boas, Siomar C. Soares, Maria S. Barbosa, Nicole Guiso, Edgar Badell, Vasco Azevedo, Rommel T. J. Ramos, Artur Silva. Draft Genome Sequence of Toxigenic *Corynebacterium ulcerans* Strain 03-8664 Isolated from a Human Throat. **Genome Announcements**, v. 4, p. e00719-16, 2016. doi: **10.1128/genomeA.00719-16**.

Abstract

Corynebacterium ulcerans is an emergent pathogen infecting wild and domesticated animals worldwide that may serve as reservoirs for zoonotic infections. In this study, we present the draft genome of *C. ulcerans* strain 03-8664. The draft genome has 2,428,683 bp, 2,262 coding sequences, and 12 rRNA genes.



Draft Genome Sequence of Toxigenic *Corynebacterium ulcerans* Strain 03-8664 Isolated from a Human Throat

Luis C. Guimarães,^a Marcus V. C. Viana,^b Leandro J. Benevides,^b Diego C. B. Mariano,^b Adonney A. O. Veras,^a Pablo H. C. Sá,^a Flávia S. Rocha,^b Priscilla C. B. Vilas Boas,^b Siomar C. Soares,^c Maria S. Barbosa,^a Nicole Guiso,^d Edgar Badell,^d Vasco Azevedo,^b Rommel T. J. Ramos,^a Artur Silva^a

Institute of Biological Sciences, Federal University of Pará (UFPA), Belém, Pará, Brazil^a; Institute of Biological Sciences, Federal University of Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brazil^b; Department of Immunology, Microbiology and Parasitology, Institute of Biological Sciences and Natural Sciences, Federal University of Triângulo Mineiro (UFTM), Uberaba, Minas Gerais, Brazil^c; Institut Pasteur, Unité de Prévention et Thérapies Moléculaires des Maladies Humaines, National Centre of Reference of Toxigenic *Corynebacteria*, Paris, France^d

***Corynebacterium ulcerans* is an emergent pathogen infecting wild and domesticated animals worldwide that may serve as reservoirs for zoonotic infections. In this study, we present the draft genome of *C. ulcerans* strain 03-8664. The draft genome has 2,428,683 bp, 2,262 coding sequences, and 12 rRNA genes.**

Received 8 June 2016 Accepted 10 June 2016 Published 28 July 2016

Citation Guimarães LC, Viana MVC, Benevides LJ, Mariano DCB, Veras AAO, Sá PHC, Rocha FS, Vilas Boas PCB, Soares SC, Barbosa MS, Guiso N, Badell E, Azevedo V, Ramos RTJ, Silva A. 2016. Draft genome sequence of toxigenic *Corynebacterium ulcerans* strain 03-8664 isolated from a human throat. *Genome Announc* 4(4):e00719-16. doi:10.1128/genomeA.00719-16.

Copyright © 2016 Guimarães et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Artur Silva, asilva@ufpa.br.

The *Corynebacterium ulcerans* *tox* gene was first described in 1926, isolated from a human throat (1). The *tox* gene present in *C. ulcerans* has 95% similarity compared to the *tox* gene present in *Corynebacterium diphtheriae* (2). This gene encodes diphtheria toxin present in lysogenic β -corynephages described in *C. ulcerans* 0102 but not in *C. ulcerans* 809 (both strains available at GenBank) (3).

However, the virulence of *C. ulcerans* does not necessarily depend on the production of diphtheria toxin; some strains have been reported to produce a powerful and severe dermonecrotic toxin similar to phospholipase D from *Corynebacterium pseudotuberculosis* (4). This repertoire of potent toxins shared by *C. ulcerans*, *C. diphtheriae*, and *C. pseudotuberculosis*, along with levels of genomic DNA relatedness and taxonomic analyzes of 16S rRNA gene sequences, particularly highlights the close phylogenetic relationship, putting these three species in a distinct cluster of the genus *Corynebacterium* (5).

C. ulcerans is an emergent pathogen infecting wild and domesticated animals that may serve as reservoirs for zoonotic infections. The frequency and severity of human infections reported worldwide during the past two decades have increased its medical importance (5, 6). Nevertheless, little knowledge about the lifestyle and associated virulence factors of *C. ulcerans* was available until recently. In humans, it may cause diphtheria-like disease, pharyngitis, sinusitis, tonsillitis, pulmonary nodules, and skin ulcers (7).

In this study, we present the draft genome sequence of toxigenic *Corynebacterium ulcerans* strain 03-8664. This strain was isolated from a human throat in France. The strain is part of the Collection of Institut Pasteur (CIP) (<https://www.pasteur.fr/en>) and was kindly given to the Laboratory of Genomics and System Biology located at Federal University of Pará, Belém, Pará, Brazil,

and the Laboratory of Cellular and Molecular Genetics Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil.

The SOLiD platform was used to perform the genome sequencing, using a fragment library. The predicted genome coverage was approximately 1,900 \times based on *C. ulcerans* genomes available in GenBank (<http://www.ncbi.nlm.nih.gov/genbank/>). The *de novo* assembly strategy was performed through SOAPdenovo version 2.04 (8) and Velvet version 1.0.13 (9). The assembly generated 258 contigs with 2,428,683 bp. The contigs were submitted to GenBank for automatic annotation. The genome has 2,270 coding sequences, 12 rRNA genes, 49 tRNA genes, 272 pseudogenes, and a G+C content of 53.60%. This genome is part of further studies of comparative genomics, pathogenicity, and vaccine and drug targets of the species.

Nucleotide sequence accession numbers. The *C. ulcerans* whole-genome shotgun (WGS) project has the project accession number LGSY00000000. This version of the project has the accession number LGSY02000000, and consists of sequences LGSY02000001 to LGSY02000258.

ACKNOWLEDGMENTS

This work was supported by the Coordination for the Improvement of Higher Education Personnel (CAPES), National Council for Scientific and Technological Development (CNPq), and the Genome and Proteome Network of the State of Pará (RPGP).

REFERENCES

- Gulbert R, Stewart FC. 1926. *Corynebacterium ulcerans*: a pathogenic microorganism resembling *C. diphtheriae*. *J Lab Clin Med* 12:756–761.
- Sing A, Bierschenk S, Heesemann J. 2004. Classical diphtheria caused by *Corynebacterium ulcerans* in Germany: amino acid sequence differences between diphtheria toxins from *Corynebacterium diphtheriae* and *C. ulcerans*. *Clin Infect Dis* 40:325–326.
- Sekizuka T, Yamamoto A, Komiya T, Kenri T, Takeuchi F, Shibayama K,

Research Article 7

Edgar Lacerda de Aguiar, Diego César Batista Mariano, **Marcus Vinícius Canário Viana**, Leandro de Jesus Benevides, Flávia de Souza Rocha, Letícia de Castro Oliveira, Felipe Luiz Pereira, Fernanda Alves Dorella, Carlos Augusto Gomes Leal, Alex Fiorini de Carvalho, Gabriela Silva Santos, Ana Luiza Mattos-Guaraldi, Prescilla Emy Nagao, Siomar de Castro Soares, Syed Shah Hassan, Anne Cybele Pinto, Henrique César Pereira Figueiredo and Vasco Azevedo. Complete genome sequence of *Streptococcus agalactiae* strain GBS85147 serotype of type Ia isolated from human oropharynx. **Standards in Genomic Sciences**, v. 11, p. 1-8, 2016. doi: 10.1186/s40793-016-0158-6.

Abstract

Streptococcus agalactiae, also referred to as Group B *Streptococcus*, is a frequent resident of the rectovaginal tract in humans, and a major cause of neonatal infection. The pathogen can also infect adults with underlying disease, particularly the elderly and immunocompromised ones. In addition, *S. agalactiae* is a known fish pathogen, which compromises food safety and represents a zoonotic hazard. This study provides valuable structural, functional and evolutionary genomic information of a human *S. agalactiae* serotype Ia (ST-103) GBS85147 strain isolated from the oropharynx of an adult patient from Rio de Janeiro, thereby representing the first human isolate in Brazil. We used the Ion Torrent PGM platform with the 200 bp fragment library sequencing kit. The sequencing generated 578,082,183 bp, distributed among 2,973,022 reads, resulting in an approximately 246-fold mean coverage depth and was assembled using the Mira Assembler v3.9.18. The *S. agalactiae* strain GBS85147 comprises of a circular chromosome with a final genome length of 1,996,151 bp containing 1,915 protein-coding genes, 18 rRNA, 63 tRNA, 2 pseudogenes and a G + C content of 35.48 %

SHORT GENOME REPORT

Open Access



Complete genome sequence of *Streptococcus agalactiae* strain GBS85147 serotype of type Ia isolated from human oropharynx

Edgar Lacerda de Aguiar¹, Diego César Batista Mariano¹, Marcus Vinícius Canário Viana¹, Leandro de Jesus Benevides¹, Flávia de Souza Rocha¹, Letícia de Castro Oliveira¹, Felipe Luiz Pereira², Fernanda Alves Dorella², Carlos Augusto Gomes Leal², Alex Fiorini de Carvalho², Gabriela Silva Santos³, Ana Luiza Mattos-Guaraldi⁴, Prescilla Emy Nagao³, Siomar de Castro Soares⁵, Syed Shah Hassan¹, Anne Cybele Pinto¹, Henrique César Pereira Figueiredo² and Vasco Azevedo^{1*}

Abstract

Streptococcus agalactiae, also referred to as Group B *Streptococcus*, is a frequent resident of the rectovaginal tract in humans, and a major cause of neonatal infection. The pathogen can also infect adults with underlying disease, particularly the elderly and immunocompromised ones. In addition, *S. agalactiae* is a known fish pathogen, which compromises food safety and represents a zoonotic hazard. This study provides valuable structural, functional and evolutionary genomic information of a human *S. agalactiae* serotype Ia (ST-103) GBS85147 strain isolated from the oropharynx of an adult patient from Rio de Janeiro, thereby representing the first human isolate in Brazil. We used the Ion Torrent PGM platform with the 200 bp fragment library sequencing kit. The sequencing generated 578,082,183 bp, distributed among 2,973,022 reads, resulting in an approximately 246-fold mean coverage depth and was assembled using the Mira Assembler v3.9.18. The *S. agalactiae* strain GBS85147 comprises of a circular chromosome with a final genome length of 1,996,151 bp containing 1,915 protein-coding genes, 18 rRNA, 63 tRNA, 2 pseudogenes and a G + C content of 35.48 %.

Keywords: *Streptococcus agalactiae*, Human pathogenic bacteria, Oropharynx, Complete genome sequence, Ion torrent

Abbreviations: CPS, Capsular polysaccharides; GBS, Group B Streptococcus; NT, Not type; PGM, Personal genome machine.

Introduction

Streptococcus agalactiae is a bacterial pathogen, distributed worldwide, that causes diseases in humans and animals [1]. In humans, it is frequently associated with meningitis, neonatal sepsis and may also affect immunocompromised adults and the elderly [2]. *S. agalactiae* is responsible for the most fatal bacterial infections in human newborns [3]. In fish, the pathogen causes meningoencephalitis and septicemia worldwide, in both freshwater and salt-water species [4, 5]. Consumption of fish

has been associated with an increased risk of colonization by *S. agalactiae* serotypes Ia and Ib in people [6]. *S. agalactiae* continues to be a major cause of subclinical mastitis in dairy cattle, which is the dominant health disorder affecting milk production in the dairy industry, and is responsible for substantial financial losses in that industry worldwide [7].

S. agalactiae is of great medical and veterinary importance due to a high social and economic impact [8], together with the incidence of disease in different hosts [9]. The incidence of invasive infections unrelated to pregnancy in human adults and animals is increasing worldwide [10]. Therefore, further studies in the area remains necessary. Since the 1990s, serotype V emerged

* Correspondence: vasco@icb.ufmg.br

¹Laboratory of Cellular and Molecular Genetics (LGCM), Federal University of Minas Gerais, Belo Horizonte, Brazil

Full list of author information is available at the end of the article



© 2016 The Author(s). Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

Research Article 8

Leandro de Jesus Benevides, **Marcus Vinicius Canário Viana**, Diego César Batista Mariano, Flávia de Souza Rocha, Priscilla Carolinne Bagano, Edson Luiz Folador, Felipe Luiz Pereira, Fernanda Alves Dorella, Carlos Augusto Gomes Leal, Alex Fiorini Carvalho, Siomar de Castro Soares, Adriana Carneiro, Rommel Ramos, Edgar Badell-Ocando, Nicole Guiso, Artur Silva, Henrique Figueiredo, Vasco Azevedo, Luis Carlos Guimarães. Genome Sequence of *Corynebacterium ulcerans* Strain FRC11. **Genome Announcements**, v. 3, p. e00112-15, 2015. doi: [10.1128/genomeA.00112-15](https://doi.org/10.1128/genomeA.00112-15).

Abstract

Here, we present the genome sequence of *Corynebacterium ulcerans* strain FRC11. The genome includes one circular chromosome of 2,442,826 bp (53.35% G+C content), and 2,210 genes were predicted, 2,146 of which are putative protein-coding genes, with 12 rRNAs and 51 tRNAs; 1 pseudogene was also identified.

Genome Sequence of *Corynebacterium ulcerans* Strain FRC11

Leandro de Jesus Benevides,^a Marcus Vinicius Canário Viana,^a Diego César Batista Mariano,^a Flávia de Souza Rocha,^a Priscilla Carolinne Bagano,^a Edson Luiz Folador,^a Felipe Luiz Pereira,^b Fernanda Alves Dorella,^b Carlos Augusto Gomes Leal,^b Alex Fiorini Carvalho,^b Siomar de Castro Soares,^b Adriana Carneiro,^c Rommel Ramos,^c Edgar Badell-Ocando,^d Nicole Guiso,^d Artur Silva,^c Henrique Figueiredo,^b Vasco Azevedo,^a Luis Carlos Guimarães^{a*}

Laboratory of Cellular and Molecular Genetics (LGCM), Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil^a; National Reference Laboratory for Aquatic Animal Diseases (AQUACEN), Ministry of Fisheries and Aquaculture, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil^b; Laboratory of Polymorphic DNA (LPDNA), Federal University of Pará, Belém, Brazil^c; Institut Pasteur, Unité de Prévention et Thérapies Moléculaires des Maladies Humaines, National Centre of Reference of Toxigenic *Corynebacteria*, Paris, France^d

* Present address: Luis Carlos Guimarães, Departamento de Biologia Geral, ICB/UFMG, Pampulha, Belo Horizonte, Minas Gerais, Brazil.

Here, we present the genome sequence of *Corynebacterium ulcerans* strain FRC11. The genome includes one circular chromosome of 2,442,826 bp (53.35% G+C content), and 2,210 genes were predicted, 2,146 of which are putative protein-coding genes, with 12 rRNAs and 51 tRNAs; 1 pseudogene was also identified.

Received 30 January 2015 Accepted 4 February 2015 Published 12 March 2015

Citation Benevides LDJ, Viana MVC, Mariano DCB, Rocha FDS, Bagano PC, Folador EL, Pereira FL, Dorella FA, Leal CAG, Carvalho AF, Soares SDC, Carneiro A, Ramos R, Badell-Ocando E, Guiso N, Silva A, Figueiredo H, Azevedo V, Guimarães LC. 2015. Genome sequence of *Corynebacterium ulcerans* strain FRC11. *Genome Announc* 3(2):e00112-15. doi:10.1128/genomeA.00112-15.

Copyright © 2015 Benevides et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 3.0 Unported license](http://creativecommons.org/licenses/by/3.0/).

Address correspondence to Vasco Azevedo, vasco@icb.ufmg.br.

Corynebacterium ulcerans is a bacterium that presents catalase-positive, nitrate-negative, and urease-positive biochemical properties (1). This bacterium belongs to the *Actinobacteria* class, which includes the genera *Corynebacterium*, *Mycobacterium*, *Nocardia*, and *Rhodococcus*, collectively termed the CMNR group. This is a very heterogeneous group; however, most of the species share particular characteristics, such as (i) a specific organization of the cell wall, which is mainly composed of peptidoglycans, arabinogalactans, and mycolic acids, and (ii) high G+C content (2–4).

Although *C. ulcerans* has increasing medical and veterinary importance, little is known about its lifestyle and associated virulence factors (5). The sequencing of more *C. ulcerans* genomes of both toxigenic and nontoxigenic strains will help in the identification of distinctive features of strains from human and animal sources (6). In addition, the data generated by newly sequenced genomes are helpful for identifying antibiotic and vaccine targets by way of a comparative analysis (7).

Nowadays, only seven complete genomes and two drafts are available in the National Center for Biotechnology Information (NCBI) database (<http://www.ncbi.nlm.nih.gov/genome/>). This scenario shows that more genomic knowledge is required in order to better characterize the virulence mechanisms of this emergent pathogen.

In the current study, we present the genome sequence of *C. ulcerans* strain FRC11, isolated from a 74-year-old human with leg ulcerans infection in Toulouse, France. This strain was first identified as *Corynebacterium pseudotuberculosis* (8), but recent analysis shows that it belongs to *C. ulcerans*.

The sequencing, assembly, and annotation of this strain were performed by the teams from the Laboratory of Cellular and Molecular Genetics (LGCM) and the National Reference Laboratory for Aquatic Animal Diseases (AQUACEN), both located at the

Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, and the Laboratory of Polymorphic DNA (LPDNA) at the Federal University of Pará, Belém, Pará, Brazil.

The platform used for sequencing was the Ion Torrent Personal Genome Machine (PGM) system (Life Technologies), using a fragment library. The quality of the raw data was analyzed using the Web tool FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). The assembly was done using the Simple Manager for Bacterial Assemblies (SIMBA) interface (<http://ufmg-simba.sourceforge.net>). The reads with good quality were assembled using a *de novo* strategy with the software MIRA 4.0 (9).

The assembly produced a total of 30 contigs, with a coverage of 179.14× and an N_{50} contig length of 236,335. Additionally, a scaffold was created using the CONTIGuator 2 software (10), using the genome sequence of *C. ulcerans* strain 0102 (accession no. NC_018101.1) (11) as a reference. The gap closure was performed automatically using SIMBA and manually using the CLC Genomics Workbench 7 software.

The genome was automatically annotated using Rapid Annotations using Subsystems Technology (RAST) (12). The manual curation of the annotation was performed using the Artemis software (13) and the UniProt database (<http://www.uniprot.org>). The CLC Genomics Workbench 7 software was used to correct indel errors in the regions of homopolymers.

The genome includes one circular chromosome of 2,442,826 bp (53.35% G+C content), and 2,210 genes were predicted, 2,146 of which are putative protein-coding genes, with 12 rRNAs and 51 tRNAs; 1 pseudogene was also identified.

Nucleotide sequence accession number. This genome has been deposited in GenBank under the accession no. CP009622.

Research Article 9

Thiago Jesus Sousa, Diego Mariano, Douglas Parise, Mariana Parise, **Marcus Vinicius Canário Viana**, Luis Carlos Guimarães, Leandro Jesus Benevides, Flávia Rocha, Priscilla Bagano, Rommel Ramos, Artur Silva, Henrique Figueiredo, Sintia Almeida, Vasco Azevedo. Complete Genome Sequence of *Corynebacterium pseudotuberculosis* Strain 12C. **Genome Announcements**, v. 3, p. e00759-15, 2015. doi: [10.1128/genomeA.00759-15](https://doi.org/10.1128/genomeA.00759-15).

Abstract

We present here the complete genome sequence of *Corynebacterium pseudotuberculosis* strain 12C, isolated from a sheep abscess in the Brazil. The sequencing was performed with the Ion Torrent Personal Genome Machine (PGM) system, a fragment library, and a coverage of ~48-fold. The genome presented is a circular chromosome with 2,337,451 bp in length, 2,119 coding sequences, 12 rRNAs, 49 tRNAs, and a G+C content of 52.83%.

Complete Genome Sequence of *Corynebacterium pseudotuberculosis* Strain 12C

Thiago Jesus Sousa,^a Diego Mariano,^a Douglas Parise,^a Mariana Parise,^a Marcus Vinicius Canário Viana,^a Luis Carlos Guimarães,^b Leandro Jesus Benevides,^a Flávia Rocha,^a Priscilla Bagano,^a Rommel Ramos,^b Artur Silva,^b Henrique Figueiredo,^c Sintia Almeida,^a Vasco Azevedo^a

Laboratory of Cellular and Molecular Genetics, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil^a; Institute of Biologic Sciences, Federal University of Para, Belém, Para, Brazil^b; Aquacen, National Reference Laboratory for Aquatic Animal Diseases, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil^c

We present here the complete genome sequence of *Corynebacterium pseudotuberculosis* strain 12C, isolated from a sheep abscess in the Brazil. The sequencing was performed with the Ion Torrent Personal Genome Machine (PGM) system, a fragment library, and a coverage of ~48-fold. The genome presented is a circular chromosome with 2,337,451 bp in length, 2,119 coding sequences, 12 rRNAs, 49 tRNAs, and a G+C content of 52.83%.

Received 8 June 2015 Accepted 12 June 2015 Published 16 July 2015

Citation Sousa TJ, Mariano D, Parise D, Parise M, Viana MVC, Guimarães LC, Benevides LJ, Rocha F, Bagano P, Ramos R, Silva A, Figueiredo H, Almeida S, Azevedo V. 2015.

Complete genome sequence of *Corynebacterium pseudotuberculosis* strain 12C. *Genome Announc* 3(4):e00759-15. doi:10.1128/genomeA.00759-15.

Copyright © 2015 Sousa et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported license.

Address correspondence to Vasco Azevedo, vasco@icb.ufmg.br.

Corynebacterium pseudotuberculosis is a Gram-positive bacterium that belongs to the CMNR group, which includes species of the genera *Corynebacterium*, *Mycobacterium*, *Nocardia*, and *Rhodococcus* (1). This species is responsible for diseases that cause great economic losses in the whole world: caseous lymphadenitis (CLA) in sheep and goats (*C. pseudotuberculosis* bv. *ovis*) and ulcerative lymphangitis in horses (*C. pseudotuberculosis* bv. *equi*) (2). With the advent of the next-generation sequencing platforms, new strains have been sequenced, and 22 strains are currently deposited in the database on the National Center for Biotechnology Information (NCBI). The increase in deposited strains can help close the pangenome this species and help the production of new vaccines and treatment methods for diseases.

Here, we present the complete genome sequence of *C. pseudotuberculosis* strain 12C. 12C was isolated from abscess of a sheep diagnosed with CLA in Pernambuco, Brazil, in 2009. This strain is suggested to belong to *C. pseudotuberculosis* bv. *ovis*, as the biochemical test for nitrate reductase showed a negative result. The genome was sequenced using the Ion Torrent Personal Genome Machine (PGM) system, 200-bp fragment library kit, and coverage of ~48-fold. The quality of the reads was analyzed using the FastQC software (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>), and *de novo* assembly was performed using Newbler 2.9 (Roche, USA). The assembly produced 17 contigs and an N_{50} value of 255,829 bp. Scaffolding was performed with CONTIGuator 2.7 (3), using as a reference the genome of *C. pseudotuberculosis* PAT10. The gap-filling process was done using the software SIMBA (<http://ufmg-simba.sourceforge.net>), CLC Genomics Workbench 7.0 (Qiagen, USA), and in-house scripts. Automatic annotation was performed by transferring information from a curated database using the Dinnotator software (<http://lgcm.icb.ufmg.br/dinnotator>). The prediction of tRNAs, rRNAs, and some coding sequences (CDSs) that were absent in the transference by Dinnotator were determined using RAST (4). All

CDSs were manually curated using the Artemis software (5) and the UniProt database (<http://www.uniprot.org>).

The complete genome of *C. pseudotuberculosis* 12C showed a length of 2,337,451 bp in a circular chromosome, a G+C content of 52.83%, and a total of 2,119 CDSs, 12 rRNAs (5S, 16S, and 23S), 49 tRNAs, and 40 pseudogenes.

Nucleotide sequence accession number. This complete genome has been deposited in GenBank under the accession no. CP011474.

ACKNOWLEDGMENTS

This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

REFERENCES

- Dorella FA, Pacheco LG, Oliveira SC, Miyoshi A, Azevedo V. 2006. *Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. *Vet Res* 37:201–218. <http://dx.doi.org/10.1051/vetres:2005056>.
- Pacheco LG, Pena RR, Castro TL, Dorella FA, Bahia RC, Carminati R, Frota MN, Oliveira SC, Meyer R, Alves FS, Miyoshi A, Azevedo V. 2007. Multiplex PCR assay for identification of *Corynebacterium pseudotuberculosis* from pure cultures and for rapid detection of this pathogen in clinical samples. *J Med Microbiol* 56:480–486. <http://dx.doi.org/10.1099/jmm.0.46997-0>.
- Galardini M, Biondi EG, Bazzicalupo M, Mengoni A. 2011. CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. *Source Code Biol Med* 6:11. <http://dx.doi.org/10.1186/1751-0473-6-11>.
- Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. 2008. The RAST server: Rapid Annotations using Subsystems Technology. *BMC Genomics* 9:75. <http://dx.doi.org/10.1186/1471-2164-9-75>.
- Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B. 2000. Artemis: sequence visualization and annotation. *Bioinformatics* 16:944–945. <http://dx.doi.org/10.1093/bioinformatics/16.10.944>.

Research Article 10

Rafael A. Baraúna, Luís C. Guimarães, Adonney A. O. Veras, Pablo H. C. G. de Sá, Diego A. Graças, Kenny C. Pinheiro, Andreia S. S. Silva, Edson L. Folador, Leandro J. Benevides, **Marcus V. C. Viana**, Adriana R. Carneiro, Maria P. C. Schneider, Sharon J. Spier, Judy M. Edman, Rommel T. J. Ramos, Vasco Azevedo, Artur Silva. Genome Sequence of *Corynebacterium pseudotuberculosis* MB20 bv. equi Isolated from a Pectoral Abscess of an Oldenburg Horse in California. **Genome Announcements**, v. 2, p. e00977-14-e00977-14, 2014. doi: [10.1128/genomeA.00977-14](https://doi.org/10.1128/genomeA.00977-14).

Abstract

The genome of *Corynebacterium pseudotuberculosis* MB20 bv. equi was sequenced using the Ion Personal Genome Machine (PGM) platform, and showed a size of 2,363,089 bp, with 2,365 coding sequences and a GC content of 52.1%. These results will serve as a basis for further studies on the pathogenicity of *C. pseudotuberculosis* bv. equi.

Genome Sequence of *Corynebacterium pseudotuberculosis* MB20 bv. equi Isolated from a Pectoral Abscess of an Oldenburg Horse in California

Rafael A. Baraúna,^a Luís C. Guimarães,^b Adonney A. O. Veras,^a Pablo H. C. G. de Sá,^a Diego A. Graças,^a Kenny C. Pinheiro,^a Andrea S. S. Silva,^a Edson L. Folador,^b Leandro J. Benevides,^b Marcus V. C. Viana,^b Adriana R. Carneiro,^a Maria P. C. Schneider,^a Sharon J. Spier,^c Judy M. Edman,^c Rommel T. J. Ramos,^a Vasco Azevedo,^b Artur Silva^a

Institute of Biological Sciences, Federal University of Pará, Belém, PA, Brazil^a; Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil^b; Department of Medicine and Epidemiology, School of Veterinary Medicine, University of California Davis, California, USA^c

The genome of *Corynebacterium pseudotuberculosis* MB20 bv. equi was sequenced using the Ion Personal Genome Machine (PGM) platform, and showed a size of 2,363,089 bp, with 2,365 coding sequences and a GC content of 52.1%. These results will serve as a basis for further studies on the pathogenicity of *C. pseudotuberculosis* bv. equi.

Received 28 August 2014 Accepted 8 October 2014 Published 13 November 2014

Citation Baraúna RA, Guimarães LC, Veras AAO, de Sá PHCG, Graças DA, Pinheiro KC, Silva ASS, Folador EL, Benevides LJ, Viana MVC, Carneiro AR, Schneider MPC, Spier SJ, Edman JM, Ramos RTJ, Azevedo V, Silva A. 2014. Genome sequence of *Corynebacterium pseudotuberculosis* MB20 bv. equi isolated from a pectoral abscess of an Oldenburg horse in California. *Genome Announc.* 2(6):e00977-14. doi:10.1128/genomeA.00977-14.

Copyright © 2014 Baraúna et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 3.0 Unported license](http://creativecommons.org/licenses/by/3.0/).

Address correspondence to Rafael A. Baraúna, rabarauna@ufpa.br.

Recent advances have been made in the genomic analysis of *Corynebacterium pseudotuberculosis*, a species of veterinary and biotechnological interest. Molecular diagnosis of several diseases caused by this species was achieved using multiplex PCR (PCR) (1), and a description of its pangenome was published using 15 strains of *C. pseudotuberculosis* bv. ovis and equi (2). *C. pseudotuberculosis* bv. equi comprises strains that infect horses and cattle and cause infection of cutaneous lymphatic vessels, termed ulcerative lymphangitis, which is characterized by the development of multiple waxy ulcerative lesions. The incidence of this disease in horses has been reported in the literature since the 1910s (3) and remains prevalent in animals worldwide (4, 5); moreover, this infection is likely underreported and has been characterized as a neglected zoonosis (6).

The host-pathogen interaction in this disease has been studied using omics approaches (7). For instance, the differential gene expression of a *C. pseudotuberculosis* bv. ovis strain was analyzed using RNA-seq, and genes involved in the molecular responses of the bacterium to different stresses during infection were identified (8). A study of reverse vaccinology reported by Soares et al. (9) identified 49 possible antigens from the genome of the *C. pseudotuberculosis* bv. equi strain 258 that may serve as targets for the development of effective vaccines. In addition, a new *C. pseudotuberculosis* bv. equi strain was isolated and sequenced, which will aid in future broader studies. These new data combined with those already reported will serve as a basis for the development of studies aimed at a better understanding of the pathogenic potential of *C. pseudotuberculosis* bv. equi.

The MB20 strain was isolated from a pectoral abscess of a 4-year-old horse of the breed Oldenburg, raised in the city of Vacaville, CA, USA. Genomic DNA was sequenced from a fragment library on a 318 chip of the Ion Torrent Personal Genome Machine (PGM) platform (Life Technologies). A total of 2,331,864 reads were

generated with an average length of 420 bp, which were used for genome assembly using the software Mira (10). The contigs generated with Mira were analyzed using the SeqMan Pro tool of the software Lasergene 11 Core Suite (DNASTAR) to remove redundant sequences. This approach resulted in 3 contigs, which were sorted with the Artemis Comparison tool (11) using the genome of *C. pseudotuberculosis* 316 as a reference. The scaffold produced at the end of the assembly was 2,363,089 bp in size and underwent automatic annotation using Rapid Annotation using Subsystem Technology (RAST) (12). As a result, 2,365 coding sequences (CDSs), 11 rRNA genes, 51 tRNA genes and a 52.1% GC content were identified. Of the 2,365 CDSs, 790 (33.4%) were classified as hypothetical proteins.

Nucleotide sequence accession numbers. The genomic sequence obtained in this study was deposited in the DDBJ/EMBL/GenBank under accession number [JPUV00000000](https://www.ncbi.nlm.nih.gov/nuccore/JPUV00000000). The version described in this paper is version [JPUV01000000](https://www.ncbi.nlm.nih.gov/nuccore/JPUV01000000).

ACKNOWLEDGMENTS

This study was conducted by the Rede Paraense de Genômica e Proteômica, with support from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and the Fundação de Amparo a Pesquisa do Estado do Pará (FAPESPA).

REFERENCES

- Pacheco LGC, Pena RR, Castro TLP, Dorella FA, Bahia RC, Carminati R, Frota MNL, Oliveira SC, Meyer R, Alves FSF, Miyoshi A, Azevedo V. 2007. Multiplex PCR assay for the identification of *Corynebacterium pseudotuberculosis* from pure cultures and for rapid detection of this pathogen in clinical samples. *J. Med. Microbiol.* 56:480–486. <http://dx.doi.org/10.1099/jmm.0.46997-0>.
- Soares SC, Silva A, Trost E, Blom J, Ramos R, Carneiro A, Ali A, Santos AR, Pinto AC, Diniz C, Barbosa EG, Dorella FA, Aburjaile F, Rocha FS, Nascimento KK, Guimarães LC, Almeida S, Hassan SS, Bakhtiar SM,

Research Article 11

Marcus Vinicius Canário Viana, Leandro de Jesus Benevides, Diego Cesar Batista Mariano, Flávia de Souza Rocha, Priscilla Carolinne Bagano Vilas Boas, Edson Luiz Folador, Felipe Luiz Pereira, Fernanda Alves Dorella, Carlos Augusto Gomes Leal, Alex Fiorini de Carvalho, Artur Silva, Siomar de Castro Soares, Henrique Cesar Pereira Figueiredo, Vasco Azevedo, and Luis Carlos Guimarães. Genome Sequence of *Corynebacterium ulcerans* Strain 210932. **Genome Announcements**, v. 2, p. e01233-14-e01233-14, 2014. doi: **10.1128/genomeA.01233-14**.

Abstract

In this work, we present the complete genome sequence of *Corynebacterium ulcerans* strain 210932, isolated from a human. The species is an emergent pathogen that infects a variety of wild and domesticated animals and humans. It is associated with a growing number of cases of a diphtheria-like disease around the world.

Genome Sequence of *Corynebacterium ulcerans* Strain 210932

Marcus Vinicius Canário Viana,^a Leandro de Jesus Benevides,^a Diego Cesar Batista Mariano,^a Flávia de Souza Rocha,^a Priscilla Carolinne Bagano Vilas Boas,^a Edson Luiz Folador,^a Felipe Luiz Pereira,^b Fernanda Alves Dorella,^b Carlos Augusto Gomes Leal,^b Alex Fiorini de Carvalho,^b Artur Silva,^c Siomar de Castro Soares,^b Henrique Cesar Pereira Figueiredo,^b Vasco Azevedo,^a Luis Carlos Guimarães^a

Department of General Biology, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil^a; AQUACEN, National Reference Laboratory for Aquatic Animal Diseases, Ministry of Fisheries and Aquaculture, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil^b; Department of Genetics, Federal University of Pará, Belém, Pará, Brazil^c

In this work, we present the complete genome sequence of *Corynebacterium ulcerans* strain 210932, isolated from a human. The species is an emergent pathogen that infects a variety of wild and domesticated animals and humans. It is associated with a growing number of cases of a diphtheria-like disease around the world.

Received 15 October 2014 Accepted 21 October 2014 Published 26 November 2014

Citation Viana MVC, de Jesus Benevides L, Batista Mariano DC, de Souza Rocha F, Bagano Vilas Boas PC, Folador EL, Pereira FL, Alves Dorella F, Gomes Leal CA, Fiorini de Carvalho A, Silva A, de Castro Soares S, Pereira Figueiredo HC, Azevedo V, Guimarães LC. 2014. Genome sequence of *Corynebacterium ulcerans* strain 210932. *Genome Announc.* 2(6):e01233-14. doi:10.1128/genomeA.01233-14.

Copyright © 2014 Viana et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 3.0 Unported license](http://creativecommons.org/licenses/by/4.0/).

Address correspondence to Vasco Azevedo, vasco@icb.ufmg.br.

Corynebacterium ulcerans is a toxigenic zoonotic agent and Gram-positive bacterium that belongs to the *Actinobacteria* class, which includes the genera *Corynebacterium*, *Mycobacterium*, *Nocardia*, and *Rhodococcus* and is referred to as a CMNR group. Studies using the 16S rRNA gene showed that *Corynebacterium pseudotuberculosis* and *Corynebacterium diphtheriae* are closely related to *C. ulcerans*. The species is facultative anaerobic, non-spore forming, nonmotile, catalase positive, and nitrate and oxidase negative. It differs from other species of the genus by fermentation of glycogen and starch (1).

The species can infect a variety of wild and domesticated animals and humans (2). It causes bovine mastitis and other infections in cats, dogs, monkeys, squirrels, otters, orcas, camels, lions, pigs, and goats. In humans, it causes diphtheria-like disease, pharyngitis, sinusitis, tonsillitis, pulmonary nodules, and skin ulcers (3). Contaminations in humans have been associated with raw milk and derivatives and contact with cattle and infected domestic pets (4). *C. ulcerans* is considered an emergent pathogen because the number of cases of infection in humans has been constantly increasing in the last two decades in the United States, Brazil, Western Europe, and Japan (5).

This species has a varied set of virulence factors, including *diphtheriae*-like toxin, phospholipase D, neuraminidase H, endoglycosidase EndoE, and a novel type of ribosome-binding protein with structural similarity to Shiga-like toxins. The sequencing of more *C. ulcerans* genomes, both toxigenic and non-toxigenic, will help in the identification of distinctive features of strains from human and animal sources, as well as in describing the zoonotic transmission in more detail (6). In addition, the data generated by newly sequenced genomes is helpful in identifying antibiotic and vaccine targets by comparative analysis (7). To date, only three complete genomes of *C. ulcerans* and two drafts have been deposited in the NCBI database.

Herein, we present the complete genome sequence of *Corynebacterium*

ulcerans strain 210932, isolated from a human. Its genome sequencing was performed by the Ion Personal Genome Machine (PGM) System, using a fragment library. A total of 1,606,464 genomic reads were filtered by quality using the software FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) and *de novo* assembling was done using Mira software version 3.9.18. The assembling step generated 12 contigs with a mean coverage of 129.23× and an N_{50} of 487,508. The contigs were scaffolded using the *C. ulcerans* strain 0102 as reference. The gaps were closed using CONTIGuator software (<http://contiguator.sourceforge.net/>) via the web tool SIMBA (SIMple Manager for Bacterial Assemblies) (<http://lgcm.icb.ufmg.br/simba/>). CLC Workbench version 7 was used for manual curation of homopolymers, generating a final assembled genome with 2,484,335 bp.

An automatic annotation was done by RAST (<http://rast.nmpdr.org/>), followed by manual curation using Artemis software (<http://www.sanger.ac.uk/resources/software/artemis/>) and the Uniprot database (<http://www.uniprot.org/>). The genome has 2,282 coding sequences (from which 654, or 28.65%, were annotated as “hypothetical proteins”), 12 rRNAs, 51 tRNAs, and a G+C content of 53.32%.

Nucleotide sequence accession number. This whole-genome shotgun project has been deposited in GenBank under the accession number CP009500.

ACKNOWLEDGMENTS

This work was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and Ministério da Pesca e Agricultura.

REFERENCES

- Riegel P, Ruimy R, de Briel D, Prévost G, Jehl F, Christen R, Monteil H. 1995. Taxonomy of *Corynebacterium diphtheriae* and related taxa, with rec-

Comparative Genomics and Metagenomics

(My contributions to this paper)

One of the main research line of our group is bacterial comparative genomics. Another research line is metagenomics. Besides manuscript preparation, my contributions to the following papers were positive selection analysis (research article 12), retrieving genome sequences and running a modelome pipeline (research article 13), taxonomic assignments for metagenome (research article 14), and genome assembly and annotation (research article 15). With this projects, I had an initial experience with modelomics and metagenomics.

Research Article 12

Raquel Hurtado, Dennis Carhuaricra, Siomar Soares, **Marcus Canário Viana**, Vasco Azevedo, Lenin Maturrano, Flavia Aburjaile. Comparative analysis of *Pasteurella multocida* reveals pathogenic specialization. **Submitted to GENE.**

Abstract

Pasteurella multocida is a gram-negative, non-motile bacterial pathogen, which is associated with chronic and acute infections as snuffles, pneumonia, atrophic rhinitis and hemorrhagic septicemia. These diseases affect a wide range of domestic animals, leading to significant morbidity and mortality and causing significant economic losses worldwide. Due to the interest in deciphering the genetic diversity and process adaptive between *P. multocida* strains, this work aimed was to perform a pan-genome analyses to evidence horizontal gene transfer and positive selection among 23 *P. multocida* strains isolated from distinct diseases and hosts. The results revealed an open pan-genome containing 3,585 genes and an accessory genome presenting 1,200 genes. The phylogenomic analysis based on the presence /absence of genes and islands exhibit high levels of plasticity, which reflects a high intraspecific diversity and a possible adaptive mechanism responsible for the specific disease manifestation between the established groups (pneumonia, fowl cholera, hemorrhagic septicemia and snuffles). Additionally, we identified differences in accessory genes among groups, which are involved in sugar metabolism and transport systems, virulence-related genes and a high concentration of hypothetical proteins. However, there was no specific indispensable functional mechanism to decisively correlate the presence of genes and their adaptation to a specific host/disease. Also, positive selection was found only for two genes from sub-group hemorrhagic septicemia, serotype B. This comprehensive comparative genome analysis will provide new insights of horizontal gene transfers that play an essential role in the diversification and adaptation mechanism into *P. multocida* species strains to a specific host/ disease.

Research Article 13

Syed Babar Jamal, Syed Shah Hassan, Sandeep Tiwari, **Marcus V. Viana**, Leandro de Jesus Benevides, Asad Ullah, Adrián G. Turjanski, Debmalya Barh, Preetam Ghosh, Daniela Arruda Costa, Artur Silva, Richard Röttger, Jan Baumbach, Vasco A. C. Azevedo. An Integrative in-silico Approach for Therapeutic Target Identification in the Human Pathogen *Corynebacterium diphtheriae*. **PLoS One**, v. 12, p. e0186401, 2017. doi: 10.1371/journal.pone.0186401.

Abstract

Corynebacterium diphtheriae (Cd) is a Gram-positive human pathogen responsible for diphtheria infection and once regarded for high mortalities worldwide. The fatality gradually decreased with improved living standards and further alleviated when many immunization programs were introduced. However, numerous drug-resistant strains emerged recently that consequently decreased the efficacy of current therapeutics and vaccines, thereby obliging the scientific community to start investigating new therapeutic targets in pathogenic microorganisms. In this study, our contributions include the prediction of modelome of 13 *C. diphtheriae* strains, using the MHOLline workflow. A set of 463 conserved proteins were identified by combining the results of pangenomics based core-genome and core-modelome analyses. Further, using subtractive proteomics and modelomics approaches for target identification, a set of 23 proteins was selected as essential for the bacteria. Considering human as a host, eight of these proteins (*glpX*, *nusB*, *rpsH*, *hisE*, *smpB*, *bioB*, DIP1084, and DIP0983) were considered as essential and non-host homologs, and have been subjected to virtual screening using four different compound libraries (extracted from the ZINC database, plant-derived natural compounds and Di-terpenoid Iso-steviol derivatives). The proposed ligand molecules showed favorable interactions, lowered energy values and high complementarity with the predicted targets. Our proposed approach expedites the selection of *C. diphtheriae* putative proteins for broad-spectrum development of novel drugs and vaccines, owing to the fact that some of these targets have already been identified and validated in other organisms.

RESEARCH ARTICLE

An integrative *in-silico* approach for therapeutic target identification in the human pathogen *Corynebacterium diphtheriae*

Syed Babar Jamal^{1*}, Syed Shah Hassan^{1,2*}, Sandeep Tiwari¹, Marcus V. Viana¹, Leandro de Jesus Benevides¹, Asad Ullah², Adrián G. Turjanski³, Debmalya Barh⁴, Preetam Ghosh⁵, Daniela Arruda Costa¹, Artur Silva⁶, Richard Röttger⁷, Jan Baumbach⁷, Vasco A. C. Azevedo^{1,8*}

1 PG program in Bioinformatics (LGCM), Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil, **2** Department of Chemistry, Islamia College University Peshawar, KPK, Pakistan, **3** Departamento de Química Biológica, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Pabellón II, Buenos Aires, Argentina, **4** Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology, Nonakuri, Purba Medinipur, West Bengal, India, **5** Department of Computer Science, Virginia Commonwealth University, Richmond, VA, United States of America, **6** Institute of Biologic Sciences, Federal University of Para, Belém, PA, Brazil, **7** Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark, **8** Department of General Biology (LGCM), Institute of Biologic Sciences, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil

* These authors contributed equally to this work.

* vascoariston@gmail.com



OPEN ACCESS

Citation: Jamal SB, Hassan SS, Tiwari S, Viana MV, Benevides LdJ, Ullah A, et al. (2017) An integrative *in-silico* approach for therapeutic target identification in the human pathogen *Corynebacterium diphtheriae*. PLoS ONE 12(10): e0186401. <https://doi.org/10.1371/journal.pone.0186401>

Editor: Alexandre G. de Brevem, UMR-S1134, INSERM, Université Paris Diderot, INTS, FRANCE

Received: December 6, 2016

Accepted: September 29, 2017

Published: October 19, 2017

Copyright: This is an open access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the [Creative Commons CC0](https://creativecommons.org/licenses/by/4.0/) public domain dedication.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The study was supported by grant from the TWAS-CNPq Postgraduate Fellowship Programme (<https://twas.org/opportunity/twas-cnpq-postgraduate-fellowship-programme>) for granting a fellowship for doctoral studies and CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasil: <http://www.capes.gov.br/>).

Abstract

Corynebacterium diphtheriae (Cd) is a Gram-positive human pathogen responsible for diphtheria infection and once regarded for high mortalities worldwide. The fatality gradually decreased with improved living standards and further alleviated when many immunization programs were introduced. However, numerous drug-resistant strains emerged recently that consequently decreased the efficacy of current therapeutics and vaccines, thereby obliging the scientific community to start investigating new therapeutic targets in pathogenic microorganisms. In this study, our contributions include the prediction of modelome of 13 *C. diphtheriae* strains, using the MHOLine workflow. A set of 463 conserved proteins were identified by combining the results of pangenomics based core-genome and core-modelome analyses. Further, using subtractive proteomics and modelomics approaches for target identification, a set of 23 proteins was selected as essential for the bacteria. Considering human as a host, eight of these proteins (glpX, nusB, rpsH, hisE, smpB, bioB, DIP1084, and DIP0983) were considered as essential and non-host homologs, and have been subjected to virtual screening using four different compound libraries (extracted from the ZINC database, plant-derived natural compounds and Di-terpenoid Iso-steviol derivatives). The proposed ligand molecules showed favorable interactions, lowered energy values and high complementarity with the predicted targets. Our proposed approach expedites the selection of *C. diphtheriae* putative proteins for broad-spectrum development of novel drugs and vaccines, owing to the fact that some of these targets have already been identified and validated in other organisms.

Research Article 14

Madangchanok Imchen, Ranjith Kumavath, Debmalya Barh, Vasco Azevedo, Preetam Ghosh, **Marcus Viana**, Alice Rebecca Wattam. Searching for signatures across microbial communities: Metagenomic analysis of soil samples from mangrove and other ecosystems. **Scientific Reports**, v. 7, p. 8859, 2017. doi: [10.1038/s41598-017-09254-6](https://doi.org/10.1038/s41598-017-09254-6).

Abstract

In this study, we categorize the microbial community in mangrove sediment samples from four different locations within a vast mangrove system in Kerala, India. We compared this data to other samples taken from the other known mangrove data, a tropical rainforest, and ocean sediment. An examination of the microbial communities from a large mangrove forest that stretches across southwestern India showed strong similarities across the higher taxonomic levels. When ocean sediment and a single isolate from a tropical rain forest were included in the analysis, a strong pattern emerged with Bacteria from the phylum Proteobacteria being the prominent taxon among the forest samples. The ocean samples were predominantly Archaea, with Euryarchaeota as the dominant phylum. Principal component and functional analyses grouped the samples isolated from forests, including those from disparate mangrove forests and the tropical rain forest, from the ocean. Our findings show similar patterns in samples were isolated from forests, and these were distinct from the ocean sediment isolates. The taxonomic structure was maintained to the level of class, and functional analysis of the genes present also displayed these similarities. Our report for the first time shows the richness of microbial diversity in the Kerala coast and its differences with tropical rain forest and ocean microbiome.

OPEN

Searching for signatures across microbial communities: Metagenomic analysis of soil samples from mangrove and other ecosystems

Received: 31 January 2017
Accepted: 26 July 2017
Published online: 18 August 2017

Madangchanok Imchen¹, Ranjith Kumavath¹, Debmalya Barh^{2,3,4}, Vasco Avezedo⁴, Preetam Ghosh⁵, Marcus Viana⁴ & Alice R. Wattam⁶

In this study, we categorize the microbial community in mangrove sediment samples from four different locations within a vast mangrove system in Kerala, India. We compared this data to other samples taken from the other known mangrove data, a tropical rainforest, and ocean sediment. An examination of the microbial communities from a large mangrove forest that stretches across southwestern India showed strong similarities across the higher taxonomic levels. When ocean sediment and a single isolate from a tropical rain forest were included in the analysis, a strong pattern emerged with Bacteria from the phylum *Proteobacteria* being the prominent taxon among the forest samples. The ocean samples were predominantly Archaea, with *Euryarchaeota* as the dominant phylum. Principal component and functional analyses grouped the samples isolated from forests, including those from disparate mangrove forests and the tropical rain forest, from the ocean. Our findings show similar patterns in samples were isolated from forests, and these were distinct from the ocean sediment isolates. The taxonomic structure was maintained to the level of class, and functional analysis of the genes present also displayed these similarities. Our report for the first time shows the richness of microbial diversity in the Kerala coast and its differences with tropical rain forest and ocean microbiome.

The mangrove ecosystem plays a crucial role by acting as a buffer zone between land and sea, maintaining the sea level and protecting the coast¹. Mangroves are a crucial component of the food chain in the saline coastal biome of the tropics and subtropics. Mangrove trees convert solar energy into organic matter via photosynthesis, with their leaves and branches serving as a source of energy and providing a habitat for a variety of aquatic organisms, which in turn, support a higher level in the food chain. This ecosystem is an enormous food web, supplying a myriad of microorganisms with both protection and nutrients^{2,3}. It is considered to be one of the most critical in tropical regions, and also one of the most vulnerable to global climate change⁴.

The complexity of the mangrove microbial communities has generated deep interest among microbial ecologists. The dynamic environment of the mangrove ecosystem, brought about by the regular tidal variations, pH, temperature, salinity, light, rainfall and nutrient availability provides an excellent environment for a wide range of organisms with diversified functional roles⁵. Studies have shown that microbial communities play a vital role in this ecosystem, being essential for biogeochemical cycles and biocycling of most nutrients, including nitrogen^{6,7}.

¹Department of Genomic Science, School of Biological Sciences, Central University of Kerala, Periyar, Padanakkad P.O., Kasaragod, Kerala, 671314, India. ²Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology, Nonakuri, Purba Medinipur, West Bengal, 721172, India. ³Xcode Life Sciences, 3D Eldorado, 112 Nungambakkam High Road, Nungambakkam, Chennai, Tamil Nadu, 600034, India. ⁴Laboratório de Genética Celular e Molecular, Departamento de Biologia Geral, Instituto de Ciências Biológicas (ICB), Universidade Federal de Minas Gerais, Pampulha, Belo Horizonte, Minas Gerais, Brazil. ⁵Department of Computer Science, Virginia Commonwealth University, Richmond, Virginia, 23284, USA. ⁶Biocomplexity Institute, Virginia Tech University, Blacksburg, Virginia, 24061, USA. Correspondence and requests for materials should be addressed to R.K. (email: rnkumavath@gmail.com) or A.R.W. (email: rwattam@vt.edu)

Research Article 15

Rafael A. Baraúna; Rommel T. J. Ramos; Adonney A. O. Veras; Kenny C. Pinheiro; Leandro J. Benevides; **Marcus V. C. Viana**; Luís C. Guimarães; Judy M. Edman; Sharon J. Spier; Vasco Azevedo; Artur Silva. Assessing the Genotypic Differences between Strains of *Corynebacterium pseudotuberculosis* biovar equi through Comparative Genomics. **PLoS ONE**, v. 12, n.1, p. e0170676, 2017. doi: 10.1371/journal.pone.0170676.

Abstract

Seven genomes of *Corynebacterium pseudotuberculosis* biovar equi were sequenced on the Ion Torrent PGM platform, generating high-quality scaffolds over 2.35 Mbp. This bacterium is the causative agent of disease known as “pigeon fever” which commonly affects horses worldwide. The pangenome of biovar equi was calculated and two phylogenomic approaches were used to identify clustering patterns within *Corynebacterium* genus. Furthermore, other comparative analyses were performed including the prediction of genomic islands and prophages, and SNP-based phylogeny. In the phylogenomic tree, *C. pseudotuberculosis* was divided into two distinct clades, one formed by nitrate non-reducing species (biovar ovis) and another formed by nitrate-reducing species (biovar equi). In the latter group, the strains isolated from California were more related to each other, while the strains CIP 52.97 and 1/06-A formed the outermost clade of the biovar equi. A total of 1,355 core genes were identified, corresponding to 42.5% of the pangenome. This pangenome has one of the smallest core genomes described in the literature, suggesting a high genetic variability of biovar equi of *C. pseudotuberculosis*. The analysis of the similarity between the resistance islands identified a higher proximity between the strains that caused more severe infectious conditions (infection in the internal organs). Pathogenicity islands were largely conserved between strains. Several genes that modulate the pathogenicity of *C. pseudotuberculosis* were described including peptidases, recombination enzymes, micoside synthesis enzymes, bacteriocins with antimicrobial activity and several others. Finally, no genotypic differences were observed between the strains that caused the three different types of infection (external abscess formation, infection with abscess formation in the internal organs, and ulcerative lymphangitis). Instead, it was noted that there is a higher phenetic correlation between strains isolated at California compared to the other strains. Additionally, high variability of resistance islands suggests gene acquisition through several events of horizontal gene transfer.

RESEARCH ARTICLE

Assessing the Genotypic Differences between Strains of *Corynebacterium pseudotuberculosis* biovar *equi* through Comparative Genomics

Rafael A. Baraúna^{1*}, Rommel T. J. Ramos¹, Adonney A. O. Veras¹, Kenny C. Pinheiro¹, Leandro J. Benevides², Marcus V. C. Viana², Luís C. Guimarães¹, Judy M. Edman³, Sharon J. Spier³, Vasco Azevedo², Artur Silva¹

1 Laboratory of Genomics and Bioinformatics, Center of Genomics and Systems Biology, Institute of Biological Sciences, Federal University of Pará, Belém, Pará, Brazil, **2** Laboratory of Cellular and Molecular Genetics, Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, **3** School of Veterinary Medicine, Department of Medicine and Epidemiology, University of California Davis, Davis, California, United States of America

* rabarauna@ufpa.br



OPEN ACCESS

Citation: Baraúna RA, Ramos RTJ, Veras AAO, Pinheiro KC, Benevides LJ, Viana MVC, et al. (2017) Assessing the Genotypic Differences between Strains of *Corynebacterium pseudotuberculosis* biovar *equi* through Comparative Genomics. PLoS ONE 12(1): e0170676. doi:10.1371/journal.pone.0170676

Editor: Ulrike Gertrud Munderloh, University of Minnesota, UNITED STATES

Received: August 25, 2016

Accepted: January 9, 2017

Published: January 26, 2017

Copyright: © 2017 Baraúna et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The Whole Genome Shotgun project of each genome has been deposited at GenBank under the accession numbers MCO000000000, MCOB000000000, MCOA000000000, MCNZ000000000, MCNY000000000, MCNX000000000, and MCNW000000000. The versions described in this paper are versions MCO001000000, MCOB010000000, MCOA010000000, MCNZ010000000, MCNY010000000, MCNX010000000, and MCNW010000000.

Abstract

Seven genomes of *Corynebacterium pseudotuberculosis* biovar *equi* were sequenced on the Ion Torrent PGM platform, generating high-quality scaffolds over 2.35 Mbp. This bacterium is the causative agent of disease known as “pigeon fever” which commonly affects horses worldwide. The pangenome of biovar *equi* was calculated and two phylogenomic approaches were used to identify clustering patterns within *Corynebacterium* genus. Furthermore, other comparative analyses were performed including the prediction of genomic islands and prophages, and SNP-based phylogeny. In the phylogenomic tree, *C. pseudotuberculosis* was divided into two distinct clades, one formed by nitrate non-reducing species (biovar *ovis*) and another formed by nitrate-reducing species (biovar *equi*). In the latter group, the strains isolated from California were more related to each other, while the strains CIP 52.97 and 1/06-A formed the outermost clade of the biovar *equi*. A total of 1,355 core genes were identified, corresponding to 42.5% of the pangenome. This pangenome has one of the smallest core genomes described in the literature, suggesting a high genetic variability of biovar *equi* of *C. pseudotuberculosis*. The analysis of the similarity between the resistance islands identified a higher proximity between the strains that caused more severe infectious conditions (infection in the internal organs). Pathogenicity islands were largely conserved between strains. Several genes that modulate the pathogenicity of *C. pseudotuberculosis* were described including peptidases, recombination enzymes, micoside synthesis enzymes, bacteriocins with antimicrobial activity and several others. Finally, no genotypic differences were observed between the strains that caused the three different types of infection (external abscess formation, infection with abscess formation in the internal organs, and ulcerative lymphangitis). Instead, it was noted that there is a higher phenetic correlation between strains isolated at California compared to the other strains. Additionally, high variability of resistance islands suggests gene acquisition through several events of horizontal gene transfer.

VIII. Appendix

B. Book chapter and review

Book Chapter and Review
(My contributions to these manuscripts)

Participation in book chapters and reviews gives the opportunity to review and organize the relevant information about the studied issues and to identify gaps in the student knowledge. In these manuscripts, my contribution was to write about structural genomics, including generations of sequencing platforms, genome assembly and annotation, and pangenomics.

Book chapter

Aburjaile FF, Santana MP, **Viana MVC**, Silva WM, Folador EL, Silva A and Azevedo V. Genomics. In: **Zahoorullah S MD. (Org.). A Textbook of Biotechnology. 1ed.: SM Online Publishers LLC, 2015, p. 1-19.**

Abstract

In the last few years, the technologies have revolutionized new areas in science especially those involving genomics, proteomics and transcriptomics. This chapter approached each of these areas showing their concepts, importance and applications.

Genomics

Aburjaile FF¹, Santana MP¹, Viana MVC¹, Silva WM¹, Folador EL¹, Silva A² and Azevedo V¹

¹Universidade Federal de Minas Gerais, UFMG, Belo Horizonte, Brazil

²Universidade Federal do Pará, Bélem, Brazil

***Corresponding author:** Vasco Azevedo, Universidade Federal de Minas Gerais/Instituto de Ciências Biológicas, Minas Gerais, Brazil, Email: vasco@icb.ufmg.br

Published Date: February 15, 2015

ABSTRACT

In the last few years, the technologies have revolutionized new areas in science especially those involving genomics, proteomics and transcriptomics. This chapter approach each of these areas showing their concepts, importance and applications.

Keywords: Genomics; Structural genomics; Functional genomics; Transcriptomics; Proteomics

STRUCTURAL GENOMICS

Introduction

Structural genomics is the analysis of sequence and structure of genome elements such as genes, regulators and mobile elements. Sequencing is necessary to obtain this information. The identification of variants among genomes has applications in evolution studies, health, biotechnology and the comprehension of relation between genotype and phenotype. Knowledge of genome sequence is fundamental to obtain this kind of information. Sequencing is the process of characterization of nucleotide sequence at a region (Targeted Genome Sequencing, TGS) or in the entire genome (Whole Genome Sequencing, WGS).

Review Article

Luis Carlos Guimarães, Leandro Benevides de Jesus, **Marcus Vinícius Canário Viana**, Artur Silva, Rommel Thiago Jucá Ramos, Siomar de Castro Soares, and Vasco Azevedo. Inside the Pan-genome - Methods and Software Overview. **Current Genomics**, v. 16, p. 1-1, 2015. doi: **10.2174/1389202916666150423002311**.

Abstract

The number of genomes that have been deposited in databases has increased exponentially after the advent of Next-Generation Sequencing (NGS), which produces high-throughput sequence data; this circumstance has demanded the development of new bioinformatics software and the creation of new areas, such as comparative genomics. In comparative genomics, the genetic content of an organism is compared against other organisms, which helps in the prediction of gene function and coding region sequences, identification of evolutionary events and determination of phylogenetic relationships. However, expanding comparative genomics to a large number of related bacteria, we can infer their lifestyles, gene repertoires and minimal genome size. In this context, a powerful approach called Pan-genome has been initiated and developed. This approach involves the genomic comparison of different strains of the same species, or even genus. Its main goal is to establish the total number of non-redundant genes that are present in a determined dataset. Pan-genome consists of three parts: core genome; accessory or dispensable genome; and species-specific or strain-specific genes. Furthermore, pan-genome is considered to be “open” as long as new genes are added significantly to the total repertoire for each new additional genome and “closed” when the newly added genomes cannot be inferred to significantly increase the total repertoire of the genes. To perform all of the required calculations, a substantial amount of software has been developed, based on orthologous and paralogous gene identification.

Inside the Pan-genome - Methods and Software Overview

Luis Carlos Guimarães^{1,2*}, Leandro Benevides de Jesus¹, Marcus Vinícius Canário Viana¹, Artur Silva², Rommel Thiago Jucá Ramos², Siomar de Castro Soares³ and Vasco Azevedo^{1*}

¹Department of General Biology, Institute of Biological Sciences, Federal University of Minas Gerais, Avenue Antônio Carlos, 6627, Belo Horizonte, Minas Gerais, Brazil; ²Department of Genetics, Institute of Biological Sciences, Federal University of Pará, Belém, Pará, Brazil; ³Department of Immunology, Microbiology and Parasitology, Institute of Biological Sciences and Natural Sciences Federal University of Triângulo Mineiro, Uberaba, Minas Gerais, Brazil



V. Azevedo

Abstract: The number of genomes that have been deposited in databases has increased exponentially after the advent of Next-Generation Sequencing (NGS), which produces high-throughput sequence data; this circumstance has demanded the development of new bioinformatics software and the creation of new areas, such as comparative genomics. In comparative genomics, the genetic content of an organism is compared against other organisms, which helps in the prediction of gene function and coding region sequences, identification of evolutionary events and determination of phylogenetic relationships. However, expanding comparative genomics to a large number of related bacteria, we can infer their lifestyles, gene repertoires and minimal genome size. In this context, a powerful approach called Pan-genome has been initiated and developed. This approach involves the genomic comparison of different strains of the same species, or even genus. Its main goal is to establish the total number of non-redundant genes that are present in a determined dataset. Pan-genome consists of three parts: core genome; accessory or dispensable genome; and species-specific or strain-specific genes. Furthermore, pan-genome is considered to be "open" as long as new genes are added significantly to the total repertoire for each new additional genome and "closed" when the newly added genomes cannot be inferred to significantly increase the total repertoire of the genes. To perform all of the required calculations, a substantial amount of software has been developed, based on orthologous and paralogous gene identification.

Keywords: Pan-genome, Core genome, Accessory genome, Species-specific genome, Comparative genome.

BACKGROUND

The advent of Next-Generation Sequencing (NGS) has allowed the reduction in the time and cost per genome sequenced [1-3]; with the use of this tool, we have observed an exponential increase in the number of whole genome sequences that have been deposited in public databases (<http://www.genomesonline.org>). In this context, the large number of genomes available boosted the development of comparative genomics and, consequently, the rise of the pan-genomic area [4, 5].

Comparative genomics is the direct comparison of the genetic content of an organism against another, and its main aim is to obtain a better biological understanding of many species [6]. This approach could help to determine gene function and coding region sequences of genomes as well as to characterize the frequency of evolutionary events, such as genome plasticity, and to establish phylogenetic relationships [7, 8]. Most of the comparative analyses have as an objective to identify similarities and differences among the organisms [9].

A comparative genomics approach is used often in many different aspects of science, such as in the comparison of the *Drosophila melanogaster* (fruit fly - model organism) genes versus human genes, where 548 human genes were identified as homologous in the fly genome. All of these genes are linked to human diseases of different natures (cardiovascular, visual, auditory, endocrine and skeletal diseases) [10]. Thus, the finding of homologous genes that are commonly shared between humans and model organisms has opened the possibility of testing new therapies in model organisms [6].

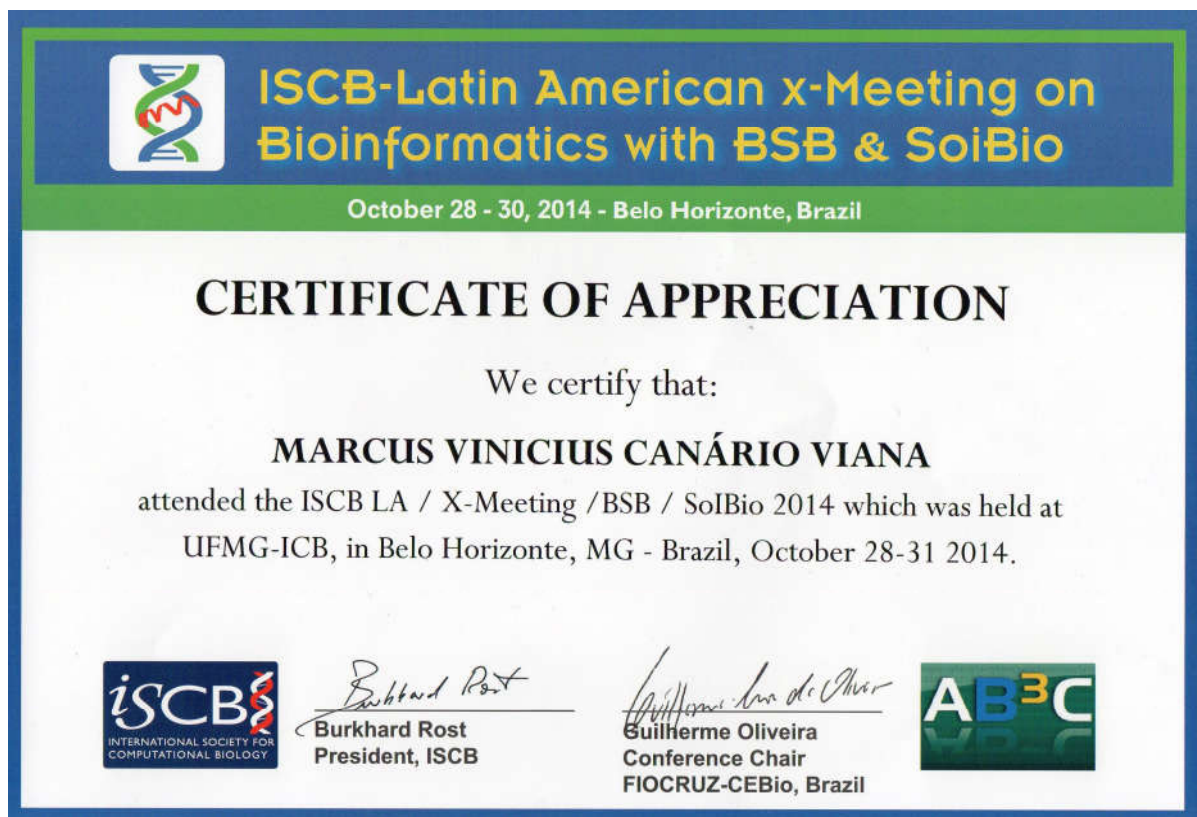
Similarly, comparative genetics can be used in prokaryotic organisms, e.g., in the comparison of *Bacillus licheniformis*, which is a gram-positive bacterium of biotechnology and pharmaceutical interest and is used for the expression of proteins and antibiotic production, in two related species (*Bacillus subtilis* and *Bacillus halodurans*). The comparison among these three bacteria not only enabled the assembly of the *Bacillus licheniformis* genome but also helped in evolutionary studies and the identification of horizontal gene transfer between them [11]. Furthermore, comparative genomics analyses in related species have shown an extensive genomic intra-species diversity and highlighted the associated bacterial promiscuity [12].

However, comparative genomics can be used with a large number of bacteria with distinct lifestyles. A study that used three hundred and seventeen genomes was performed, aim-

*Address correspondence to these authors at the Department of General Biology, Institute of Biological Sciences, Federal University of Minas Gerais, Avenue Antônio Carlos, 6627, Belo Horizonte, Minas Gerais, Brazil; Tel/Fax: +55 (31) 3409-2610; E-mails: luisguimaraes.bio@gmail.com; vascoariston@gmail.com

VIII. Appendix

C. National and international workshop, course and poster presentations



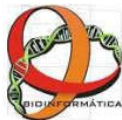


UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

Certifico que *Marcus Vinícius Canário Viana* participou do workshop internacional **"PATRIC: recursos integrados para estudo de sistemas patogênicos"**, com carga horária de 15 horas, ministrado pelos doutores Rebecca Wattam e Maulik Shukla, no dia 27 de outubro de 2014.

Prof. Dr. Vasco Ariston de Carvalho Azevedo
Coordenador do Programa de Pós-Graduação em Bioinformática

Certificado



UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

Certifico que *Marcus Vinícius Canário Viana* participou do workshop **"Anotação avançada de sequências e uso de pipelines com a plataforma EGene2"**, com carga horária de 15 horas, ministrado pelos doutores Arthur Gruber e Robson Francisco de Souza, no dia 31 de outubro de 2014.

Prof. Dr. Vasco Ariston de Carvalho Azevedo
Coordenador do Programa de Pós-Graduação em Bioinformática

Certificado



ISCB-Latin American x-Meeting on Bioinformatics with BSB & SoiBio

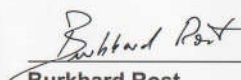
October 28 - 30, 2014 - Belo Horizonte, Brazil

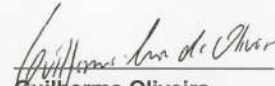
CERTIFICATE

We certify that the poster K02 - COMPLETE GENOME SEQUENCE OF *Corynebacterium ulcerans* STRAIN 210932 - Authors: Marcus vinicius Canário Viana, Leandro Benevides, Diego Mariano, Flávia Rocha, Edson Folador, Felipe Pereira, Fernanda Dorella, Carlos Leal, Alex Carvalho, Artur Silva, Siomar Soares, Henrique Figueiredo, Vasco Azevedo, Luis Guimarães

was presented at "ISCB-LA / X-meeting / BSB / SoiBio" which was held at Belo Horizonte-MG – Brasil – October 28-31 2014




Burkhard Rost
President, ISCB


Guilherme Oliveira
Conference Chair
FIOCRUZ-CEBio, Brazil



ISCB-Latin American x-Meeting on Bioinformatics with BSB & SoiBio

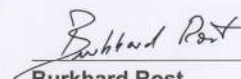
October 28 - 30, 2014 - Belo Horizonte, Brazil

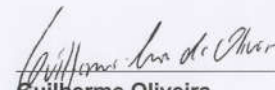
CERTIFICATE

We certify that the poster O09 - Complete genome sequence of *Corynebacterium ulcerans* 210931 - Authors: Leandro de Jesus Benevides, Marcus Viana, Diego Mariano, Flávia Rocha, Edson Folador, Felipe Pereira, Fernanda Dorella, Alex Carvalho, Carlos Leal, Artur Silva, Siomar Soares, Henrique Figueiredo, Vasco Azevedo, Luis Guimarães,

was presented at "ISCB-LA / X-meeting / BSB / SoiBio" which was held at Belo Horizonte-MG – Brasil – October 28-31 2014




Burkhard Rost
President, ISCB


Guilherme Oliveira
Conference Chair
FIOCRUZ-CEBio, Brazil





ISCB-Latin American x-Meeting on Bioinformatics with BSB & SoiBio

October 28 - 30, 2014 - Belo Horizonte, Brazil

CERTIFICATE

We certify that the poster G08 - Funcional genomics of *Propionibacterium freudereichii* -
 Authors: FF Aburjaile, R Ramos, L Benevides, M Viana, A Silva, V Azevedo, YL Loir,
 H Falentin3
 was presented at "ISCB-LA / X-meeting / BSB / SoiBio" which was held at Belo Horizonte-
 MG – Brasil – October 28-31 2014



Burkhard Rost
 Burkhard Rost
 President, ISCB

Guilherme Oliveira
 Guilherme Oliveira
 Conference Chair
 FIOCRUZ-CEBio, Brazil



tances, for each element (call it n) of the permutation, between the current position and the position of $n-1$ and $n+1$ elements is minimum.

All the algorithms were implemented using C++ language and executed in the supercomputer of the Technological Center of Electronics and Computer (Centro Tecnológico de Eletrônica e Informática – CTEI), at the Federal University of Mato Grosso do Sul (UFMS). For this work, it was used a set of random entries with different sizes.

17. SIMBA: a simple way to make complete assemblies of bacterial genomes

Diego C. B. Mariano¹, Leticia C. Oliveira², Edson L. Folador¹, Edgar L. Aguiar¹, Leandro Benevides¹, Felipe L. Pereira¹, Marcus Canário¹, Thiago J. Sousa¹, Rommel T. J. Ramos², Vasco A. C. Azevedo²

¹UFMG,

²UFFPA

Background

The evolution of large-scale sequencing platforms has reduced the time and spent cost on the process of DNA fingerprinting. However, sequencers still have limitations, such as the capacity to read the maximum size of DNA fragments, leading to the need to fragment the DNA into small pieces before sequencing. Through this approach, it is necessary, after this step, to rearrange the fragments read (reads), for then can be possible to represent the original genome. This process is known as genome assembly. The process is very complex and dependent of limitations of sequencers, so several computer programs are necessary to work with the data. Nowadays, a lot of strategies for genome assembly have been proposed, but there isn't consensus on the best approach yet. One of the problems detected was the large amount of programs that should be implemented in assembly processes, requiring a large computational domain from the bioinformaticians. The adoption of a tool with good usability could reduce the spent time on hand labor training. In this context we present SIMBA (Simple Manager for bacterial genomes assembly), a Web tool designed to manage strategies for a hybrid pipe-

line that aims to facilitate the implementation of the processes of genome assembly.

Results

SIMBA was developed using PHP and SQLite database. The software allows data processing and conversion extensions through various scripts. To de novo assembly, SIMBA allows to use four separated softwares, which three are based on overlap-layout-consensus algorithm (Mira3, Mira4, and Newbler), and one are based on De Bruijn graph (Minia) algorithm. To finish the installation, SIMBA utilizes two approaches: the first one is based on ordering contigs by a reference genome and the second one is based on reports of optical mapping generated by Opgen MapSolver. Finally, SIMBA allows the download of the results generated and the manual curation of the data with some other software.

Conclusions

The friendly interface of SIMBA allows an easiest performing of the genome assembly process and also facilitates the ordering of the contigs and closing gaps. Any bioinformatician without a great specific knowledge of hardware and software can do all the steps for assemble a genome. The application all the developed scripts and the source code were made available in <http://github.com/dcbmariano/simba>.

18. The complete genome sequence of *Streptococcus agalactiae* strain GBS85147

Edgar Lacerda de Aguiar¹, Diego C. B. Mariano¹, Leticia Castro Oliveira¹, Lucas Amorim Gonçalves¹, Alberto F. Oliveira¹, Marcus Canário², Flávia de Souza Rocha¹, Felipe Luiz Pereira¹, Siomar de Castro Soares¹, Fernanda Alves Dorella¹, Carlos Augusto Gomes Leal², Henrique Cesar Pereira Figueiredo¹, Vasco Ariston de Carvalho Azevedo²

¹Universidade Federal de Minas Gerais

Streptococcus agalactiae (Lancefield group B, GBS), is a gram-positive and cocci, bacterial patho-gen. This species can causes diseases in humans, cattles and fishes. In humans, it is associated with neonatal sepsis and meningi-



FUNDAÇÃO OSWALDO CRUZ
CENTRO DE PESQUISAS RENÉ RACHOU

CERTIFICADO

Certificamos que **Marcus Vinicius Canário Viana** participou do Curso Internacional **"Metagenomics Course"** do Programa de Pós-Graduação em Ciências da Saúde do Centro de Pesquisas René Rachou/FIOCRUZ no período de 02 a 06 de fevereiro de 2015, com duração de 30 horas.

Belo Horizonte, 06 de fevereiro de 2015

Dr. Edelberto Santos Dias
Coordenador do Programa



28º Congresso Brasileiro de Microbiologia

De 18 a 22 de Outubro de 2015 | Centro Sul - Centro de Convenções de Florianópolis | Florianópolis - Santa Catarina - Brasil

Certificado

Certificamos que o trabalho intitulado GENOME SEQUENCING OF TWELVE STRAINS OF THE ANIMAL PATHOGEN CORYNEBACTERIUM PSEUDOTUBERCULOSIS BV. EQUI com a autoria de: BARAÚNA, R.A., LOBATO, A.R.F., COSTA, S.S., BARRETO, D.F., SILVA, Y.R.O., CAVALCANTE, A.L.Q., MAUÉS, D.B., SÁ, P.H.C.G., VERAS, A.A.O., GRAÇAS, D.A., GUIMARÃES, L.C., CANÁRIO, M.V., BENEVIDES, L., SCHNEIDER, M.P.C., RAMOS, R.T.J., SPIER, S., AZEVEDO, V., SILVA, A. foi apresentado na forma de pôster durante o 28º Congresso Brasileiro de Microbiologia realizado no Centro de Convenções de Florianópolis, na cidade de Florianópolis, SC, no período de 18 a 22 de outubro de 2015.

Profa. Dra. Marina Baquerizo
Presidente da SBM

Prof. Dr. Gustavo Henrique Goldmann
1º Secretário

Apoio:



Organização:



GENÉTICA 2016
 Brazilian-International Congress of Genetics
 11 a 14 de setembro de 2016 - Hotel Glória, Caxambu, MG



MARCUS VINICIUS CANÁRIO VIANA participou do curso Introdução à bioinformática genômica e à filogenômica, realizado durante o 62º CONGRESSO BRASILEIRO DE GENÉTICA, em Caxambu, MG, no período de 11 a 14 de setembro de 2016, com carga horária de 3 horas.


 Fabrício Rodrigues dos Santos
 Presidente da SBG




 Célia Maria de Almeida Soares
 Primeira Secretária da SBG

GENÉTICA 2016
 Brazilian-International Congress of Genetics
 11 a 14 de setembro de 2016 - Hotel Glória, Caxambu, MG



MARCUS VINICIUS CANÁRIO VIANA participou do curso Introdução à bioinformática genômica e à filogenômica, realizado durante o 62º CONGRESSO BRASILEIRO DE GENÉTICA, em Caxambu, MG, no período de 11 a 14 de setembro de 2016, com carga horária de 3 horas.


 Fabrício Rodrigues dos Santos
 Presidente da SBG




 Célia Maria de Almeida Soares
 Primeira Secretária da SBG

GENÉTICA 2016
 Brazilian-International Congress of Genetics
 11 a 14 de setembro de 2016 - Hotel Glória, Caxambu, MG



Certificamos que **MARCUS VINICIUS CANÁRIO VIANA** apresentou o trabalho intitulado **GENOME SEQUENCE OF CORYNEBACTERIUM PSEUDOTUBERCULOSIS 32** de autoria de **VIANA, MVC, PARISE, D, SOUSA, TJ, BENEVIDES, LJ, MARIANO, D, ROCHA, FS, BAGANO, P, GUIMARAES, LC, PEREIRA, FL, DORELLA, FA, RAMMOS, R, SILVA, A, SELIM, S, MOAWAD, MS, FIGUEIREDO, H, AZEVEDO, V**, no **62º CONGRESSO BRASILEIRO DE GENÉTICA**, no período de 11 a 14 de setembro de 2016, em Caxambu, MG, na área de **GENÉTICA DE MICROORGANISMOS**.


 Fabrice Rodrigues dos Santos
 Presidente da SBG

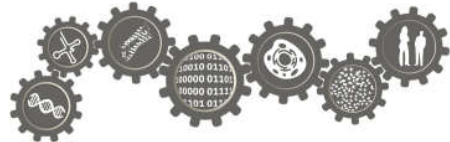



 Célia Maria de Almeida Soares
 Primeira Secretária da SBG

Verifique o código de autenticidade 14919.876094.88655 em even3.com.br/documentos

X-meeting 2016

Belo Horizonte | November 16th to 18th



12th International Conference of the AB3C

This certifies that **Marcus Vinicius Canário Viana** has attended the 6 hours Study group entitled **"RSG Brazil - 1st Student Council Symposium"** during the X-Meeting 2016 - 12th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held in Belo Horizonte - Brazil between November 16 and 18 of 2016.



Glória Franco
 AB³C President



Alan Durham
 AB³C Vice-President



AB³C

Verifique o código de autenticidade 14914.876094.6 em even3.com.br/documentos



Certificate of Participation

This certifies that **Marcus Vinicius Canário Viana** has participated in the X-Meeting 2016 - 12th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held in Belo Horizonte - Brazil between November 16 and 18 of 2016.

Belo Horizonte, 18th November 2016

Glória Franco

Glória Franco
AB³C President

Alan Durham

Alan Durham
AB³C Vice-President



AB³C

Verifique o código de autenticidade 14913.876094.26810 em even3.com.br/documentos



Certificate of poster presentation

This certifies that the work entitled **Complete genome sequence of *Corynebacterium pseudotuberculosis* 33**, authored by **Marcus Vinicius Canário Viana, Douglas Parise, Thiago De Jesus Sousa, Leandro De Jesus Benevides, Diego César Batista Mariano, Flávia De Souza Rocha, Priscilla Bagano, Luís Carlos Guimarães, Felipe Luiz Pereira, Fernanda Alves Dorella, Rommel Ramos, Salah Abdel Karim Selim, Mohammad Salaheldean, Artur Silva, Alice Rebecca Wattam and Vasco A De C Azevedo** was presented during the poster session of the X-Meeting 2016 - 12th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held in Belo Horizonte - Brazil between November 16 and 18 of 2016.

Nicole Scherer

Nicole Scherer
Poster Session Chair

Mainá Bitar

Mainá Bitar
Poster Session Co-Chair

Glória Franco

Glória Franco
AB³C President



AB³C

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

Instituto Oswaldo Cruz

DECLARAÇÃO

Declaro que Marcus Vinicius Canário Viana participou do curso: “Análise de Dados Genômicos utilizando Computação de Alto Desempenho” no período de 22 a 24 de novembro de 2017, ministrado na Biominas Brasil (Belo Horizonte, MG), com carga horária de 20 horas, realizado pela Laboratório de Biologia Computacional e Sistemas do Instituto Oswaldo Cruz.

Rio de Janeiro, 26 de novembro de 2017.



Dr. Alberto M. R. Dávila
Laboratório de Biologia Computacional e Sistemas
Instituto Oswaldo Cruz

Verifique o código de autenticidade 183908.876094.6.8 em <https://www.even3.com.br/documentos>



X-meeting 2017

13th International Conference of the AB3C

Certificate of Participation

This certifies that **Marcus Vinicius Canário Viana** has participated in the X-Meeting 2017 - 13th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held in São Pedro - Brazil between October 4 and 6 of 2017.

São Pedro, October 6th 2017.

Alan M. Durham
AB3C President

Ney Lemke
AB3C Vice President

Verifique o código de autenticidade 183947.876094.08007.8 em <https://www.even3.com.br/documentos>



X-meeting 2017

13th International Conference of the AB3C

Certificate of Poster presentation

This certifies that the work entitled **Identification of genes under positive selection in *Corynebacterium pseudotuberculosis***, authored by *Marcus Vinicius Canário Viana, Henrique Figueiredo, Felipe Luiz Pereira, Fernanda Alves Dorella, anne cybelle pinto gomide, Alice Rebecca Wattam and Vasco A de C Azevedo* was presented by Marcus Vinicius Canário Viana during the Poster session of the X-Meeting 2017 - 13th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held in São Pedro - Brazil between 4th and 6th October of 2017.

São Pedro, 6th October 2017.

Alan M. Durham
AB3C President

Robson Francisco de Souza
X-meeting 2017 Poster chair

Nicole Scherer
X-meeting 2017 Poster Co-chair

Verifique o código de autenticidade 183947.876094.45141.8 em <https://www.even3.com.br/documentos>



X-meeting 2017

13th International Conference of the AB3C

Certificate of Poster presentation

This certifies that the work entitled **Biovar equi versus ovis: What genetically differentiate them?**, authored by *Doglas Parise, Mariana Teixeira Dornelles Parise, Marcus Vinicius Canário Viana, Elma Lima Leite, anne cybelle pinto gomide and Vasco Ariston de Carvalho Azevedo* was presented by Doglas Parise during the Poster session of the X-Meeting 2017 - 13th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held in São Pedro - Brazil between 4th and 6th October of 2017.

São Pedro, 6th October 2017.

Alan M. Durham
AB3C President

Robson Francisco de Souza
X-meeting 2017 Poster chair

Nicole Scherer
X-meeting 2017 Poster Co-chair

Verifique o código de autenticidade 183947.876094.45740.8 em <https://www.even3.com.br/documentos>



X-meeting 2017

13th International Conference of the AB3C

Certificate of Poster presentation

This certifies that the work entitled **Output Organizer – a software to facilitate POTION results interpretation**, authored by *Mariana Teixeira Dornelles Parise, Douglas Parise, Marcus Vinicius Canário Viana, anne cybelle pinto gomide and Vasco Ariston de Carvalho Azevedo* was presented by Mariana Teixeira Dornelles Parise during the Poster session of the X-Meeting 2017 - 13th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held in São Pedro - Brazil between 4th and 6th October of 2017.

São Pedro, 6th October 2017.

Alan M. Durham
AB3C President

Robson Francisco de Souza
X-meeting 2017 Poster chair

Nicole Scherer
X-meeting 2017 Poster Co-chair

Verifique o código de autenticidade 183947.876094.08073.8 em <https://www.even3.com.br/documentos>



X-meeting 2017

13th International Conference of the AB3C

Certificate of Poster presentation

This certifies that the work entitled **Genome assembly completeness and its effect on phylogenetic estimation**, authored by *Rafael Cabus Gantois, Raquel Enma Hurtado Castillo, Rodrigo Profeta Silveira Santos, Thiago de Jesus Sousa, Marcus Vinicius Canário Viana, anne cybelle pinto gomide, Artur Silva, Rafael Azevedo Baraúna and Vasco A de C Azevedo* was presented by Rafael Cabus Gantois during the Poster session of the X-Meeting 2017 - 13th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held in São Pedro - Brazil between 4th and 6th October of 2017.

Alan M. Durham
AB3C President

São Pedro, 6th October 2017.

Robson Francisco de Souza
X-meeting 2017 Poster chair

Nicole Scherer
X-meeting 2017 Poster Co-chair

Verifique o código de autenticidade 183947.876094.08022.8 em <https://www.even3.com.br/documentos>



X-meeting 2017

13th International Conference of the AB3C

Certificate of Poster presentation

This certifies that the work entitled **Genomic analysis of *Corynebacterium pseudotuberculosis* strain 262**, authored by *Raquel Enma Hurtado Castillo, Marcus Vinicius Canário Viana, anne cybelle pinto gomide, Vasco A de C Azevedo, Rommel Thiago Jucá Ramos and Artur Silva* was presented by Raquel Enma Hurtado Castillo during the Poster session of the X-Meeting 2017 - 13th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held in São Pedro - Brazil between 4th and 6th October of 2017.

São Pedro, 6th October 2017.

Alan M. Durham
AB3C President

Robson Francisco de Souza
X-meeting 2017 Poster chair

Nicole Scherer
X-meeting 2017 Poster Co-chair