Resequencing of *Corynebacterium pseudotuberculosis* Cp162 genome and the search for host tropism mechanisms (#89860)

First submission

Guidance from your Editor

Please submit by **8 Sep 2023** for the benefit of the authors (and your token reward) .

Structure and Criteria
Please read the 'Structure and Criteria' page for general guidance.

Raw data check
Review the raw data.

Image check
Check that figures and images have not been inappropriately manipulated.

If this article is published your review will be made public. You can choose whether to sign your review. If uploading a PDF please remove any identifiable information (if you want to remain anonymous).

Files
Download and review all files from the materials page.

4 Figure file(s)
7 Table file(s)

Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

- Clear, unambiguous, professional English language used throughout.
- Intro & background to show context. Literature well referenced & relevant.
- Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- Figures are relevant, high quality, well labelled & described.
- Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

- Original primary research within Scope of the journal.
- Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- Rigorous investigation performed to a high technical & ethical standard.
- Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- Impact and novelty not assessed. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
- All underlying data have been provided; they are robust, statistically sound, & controlled.
- Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

The best reviewers use these techniques

Tip	*Example*
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult. I suggest you have a colleague who is proficient in English and familiar with the subject matter review your manuscript, or contact a professional editing service.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

Resequencing of *Corynebacterium pseudotuberculosis* Cp162 genome and the search for host tropism mechanisms

Enrico Giovanelli Tacconi Gimenez [1], **Thiago de Jesus Sousa** [2], **Flávia Aburjaile** [3], **Bertram Brenig** [4], **Artur Silva** [5], **Marcus Vinicius Canário Viana** [Corresp., 1], **Vasco Azevedo** [Corresp. 1]

[1] Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

[2] Laboratório Central do Espírito Santo (LACEN-ES), Vitória, Espírito Santo, Brazil

[3] Veterinary School, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

[4] Institute of Veterinary Medicine, University of Göttingen, Göttingen, Niedersachsen, Germany

[5] Institute of Biological Sciences, Federal University of Pará, Belém, Pará, Brazil

Corresponding Authors: Marcus Vinicius Canário Viana, Vasco Azevedo
Email address: canarioviana@gmail.com, vascoariston@gmail.com

Background. *Corynebacterium pseudotuberculosis* is a zoonotic Gram-positive bacterial pathogen known to cause different diseases in many mammals, including lymph node abscesses in camels. Strains from biovars equi and ovis of *C. pseudotuberculosis* can infect camels. Comparative genomics could help to identify features related to host adaptation, and currently strain Cp162 from biovar equi is the only one from camel with a sequenced genome.

Methods. In this work, we compared the quality of three genome assemblies of strain Cp162 that used data from the DNA sequencing platforms SOLiD v3 Plus, IonTorrent PGM, and Illumina HiSeq 2500 with an optical map and investigate the unique features of this strain. For this purpose, we applied comparative genomic analysis on the different Cp162 genome assembly vesions and included other 129 genomes from the same species.

Results. Since the first version of the genome, there was an increase of 88 Kbp and 121 protein-coding sequences, a decrease of pseudogenes from 139 to 53, and two inversions and one rearrangement corrected. We identified the virulence genes *cpp, DIP_RS14950, mprA, nanH, pknG, pld, sodC, spaC,* and *tufA.* In comparison to 129 genomes of the same species, strain Cp162 has four genes exclusively present, two of them code transposases and two truncated proteins, and the three exclusively absent genes *lysG,* NUDIX domain protein, and Hypothetical protein. All 130 genomes had genes *rpob2* and *rbpA* predicted to confer resistance to rifampin. Our results found no unique gene that could be associated with tropism to camel host, and further studies should include more genomes and genome-wide association studies testing for genes and SNPs.

Resequencing of *Corynebacterium pseudotuberculosis* Cp162 genome and the search for host tropism mechanisms

Enrico Giovanelli Tacconi Gimenez[1], Thiago de Jesus Sousa[2], Flávia Aburjaile[3], Bertram Brenig[4], Artur Silva[5], Marcus Vinicius Canário Viana[1], Vasco Azevedo[2*]

[1] Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
[2] Laboratório Central do Espírito Santo (LACEN-ES), Vitória, Espírito Santo, Brazil
[3] Veterinary School, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
[4] Institute of Veterinary Medicine, University of Göttingen, 37077 Göttingen, Niedersachsen, Germany
[5] Laboratory of Genomics and Bioinformatics, Center of Genomics and Systems Biology, Institute of Biological Sciences, Federal University of Pará, Belém, Brazil

Corresponding Author:
Vasco Azevedo[1]
Av. Pres. Antônio Carlos, Belo Horizonte, Minas Gerais, 31270-901, Brazil
Email address: vasco@icb.ufmg.br

Abstract

Background. *Corynebacterium pseudotuberculosis* is a zoonotic Gram-positive bacterial pathogen known to cause different diseases in many mammals, including lymph node abscesses in camels. Strains from biovars equi and ovis of *C. pseudotuberculosis* can infect camels. Comparative genomics could help to identify features related to host adaptation, and currently strain Cp162 from biovar equi is the only one from camel with a sequenced genome.
Methods. In this work, we compared the quality of three genome assemblies of strain Cp162 that used data from the DNA sequencing platforms SOLiD v3 Plus, IonTorrent PGM, and Illumina HiSeq 2500 with an optical map and investigate the unique features of this strain. For this purpose, we applied comparative genomic analysis on the different Cp162 genome assembly versions and included other 129 genomes from the same species.
Results. Since the first version of the genome, there was an increase of 88 Kbp and 121 protein-coding sequences, a decrease of pseudogenes from 139 to 53, and two inversions and one rearrangement corrected. We identified the virulence genes *cpp, DIP_RS14950, mprA, nanH, pknG, pld, sodC, spaC,* and *tufA.* In comparison to 129 genomes of the same species, strain Cp162 has four genes exclusively present, two of them code transposases and two truncated proteins, and the three exclusively absent genes *lysG,* NUDIX domain protein, and Hypothetical protein. All 130 genomes had genes *rpob2* and *rbpA* predicted to confer resistance to rifampin. Our results found no unique gene that could be associated with tropism to camel host, and further

studies should include more genomes and genome-wide association studies testing for genes and SNPs.

Introduction

Corynebacterium pseudotuberculosis is a zoonotic Gram-positive bacterium that causes caseous lymphadenitis (CLA) in various animals, including small ruminants, cattle, camelids, and other host disease manifestations. In this species, biovar equi is nitrate positive and biovar ovis is nitrate negative (Dorella et al., 2006). Economic losses due to CLA can be severe, particularly in camelids, which are often valued as companion animals. In Australia, the wild dromedary population in the interior has frequently exhibited unsightly lymph node abscesses. Similarly, in East Africa, a high prevalence of swollen external lymph nodes has been observed in almost all dromedaries, and it is believed that this may be linked to Cutaneous Lymphadenitis (CLA) resulting from the consumption of thorny plants. CLA mortality rates in dromedaries in countries other than Europe, where it can reach 15%, are unknown. However, death always occurs when the pathogen spreads into internal organs, mainly the lung and liver (Wernery & Kinne, 2016).

In this context, genomic data can be used for identification and taxonomy (Parks et al., 2022), evolutionary studies (Sheppard, Guttman & Fitzgerald, 2018), epidemiology (Gardy & Loman, 2018), and the development of control mechanisms such as drugs (Serral et al., 2021) and vaccines (Goodswen, Kennedy & Ellis, 2023). An ideal genome assembly should be complete, closed, and artifacts-free to avoid bias in analysis that relies on gene content, variant calling (Di Marco et al., 2023), and synteny. The highly accurate but short reads from recent second-generation DNA sequencing platforms result in assembly gaps with long repetitive sequences (Loman et al., 2012). Two strategies are used to solve this issue using NGS data: scaffolding using an optical map (Lehri, Seddon & Karlyshev, 2017) and a hybrid assembly, in which longer reads with lower accuracy from a third-generation DNA sequencing platform are used to generate an assembly that is error corrected using reads from a second-generation platform (Craddock et al., 2022; Di Marco et al., 2023).

In *C. pseudotuberculosis,* it is known that horses and buffalo are only reported as hosts of the nitrate-positive biovar equi, but little is known about the mechanisms related to host tropism, besides the suggestion of diphtheria toxin as a requirement to infect buffalo (Viana et al., 2017). Strain Cp162 from the camel is currently the only strain from the camel with a sequenced genome. It was initially isolated from an external neck abscess of a camel in 1999 and was first sequenced in 2012 using the platform SOLiD v3 Plus (RefSeq accession NC_018019.1) (Hassan et al., 2012). The genome was then resequenced in 2017 using the Ion Torrent PGM platform (NC_018019.2) to improve genome assembly quality, and in 2019 using Illumina HiSeq 2500 with assembly using an optical map to improve the accuracy of the genome assembly (NC_018019.3) (Sousa et al., 2019).

In this study, we aimed to evaluate the improvements in the genome assemblies of strains Cp162, search for genes that could be related to tropism for camels and update the pangenome analysis of the species.

Materials & Methods

Samples, quality assessment, and taxonomy

The genome sequences of 142 *C. pseudotuberculosis* strains were obtained from the NCBI RefSeq Database (https://www.ncbi.nlm.nih.gov/genome/browse/#!/prokaryotes/2411/), which includes all genome sequences from the database. From those, we removed 10 mutant strains (SigH, SigmaE, SigM, sigC, T1, Cp13, sigB, SigD, phoP, SigK) and strains 1002 and DSM 20689 due to those being the same strains as 1002B and ATCC19410, respectively. A total of 130 strains remained for downstream analysis (*Data S1*). The genome assemblies were retrieved using NCBI Datasets v15.6.1 (https://github.com/ncbi/datasets). The current assembly of Cp162 is on version 3. The first and second assemblies of Cp162 (GCF_000265545.1 and GCF_000265545.2) were added for completeness of the dataset, but they were not included in the species analysis due to potential sequencing errors and misassemblies. CheckM2 v1.0.2 (https://github.com/chklovski/CheckM2) was used to evaluate the completeness and contamination of the genome sequences, while GUNC v1.0.5 (Orakov et al., 2021) was used to identify chimeric contigs. Taxonomic classification was performed using GTDBtk v2.3.0 with database R214 (Chaumeil et al., 2022).

Analysis of Cp162 genome assemblies

We compared the three versions of Cp162 assemblies for completeness and contamination, size, gene content, and synteny. The number of genes was collected from each genome RefSeq annotation. Differences in gene content were identified using Panaroo v1.3.3 (Tonkin-Hill et al., 2020) for gene clustering and an in-house script for identifying exclusive genes. Synteny was verified using Mauve v20150226 (Darling et al., 2004). Genome characterization was performed on the latest genome assembly (GCF_000265545.3). For mobile elements, prophages were predicted using PHASTER (Arndt et al., 2016), while genomic islands (GEIs) were predicted using GIPSy v1.1.3 (Soares et al., 2016) and *C. glutamicum* (NZ_CP025533.1) as a reference genome. Virulence genes were predicted using PanViTa v1.1.3 (Rodrigues et al., 2023) with the VFDB database (Liu et al., 2022) and manually using BLASTp against the sequences of the known virulence factors Phospholipase D (*pld*), Neuraminidase H (*nanH)*, CP40 (*cpp*), Diphtheria Toxin (*tox*), pili tip protein (*spaC*), Superoxide Dismutase (*sodC*) and Protein kinase G (*pknG*) (Trost et al., 2010; Santana-Jorge et al., 2016; Viana et al., 2017). Antimicrobial resistance genes were predicted using PanViTa with the CARD database (Alcock et al., 2023). CRISPR-Cas systems were predicted using CRISPRCasFinder v1.1.2 (Couvin et al., 2018). A circular map comparing the three assemblies of Cp162 was built using BRIG v0.95 (Alikhan et al., 2011).

Species-level analysis

To build a phylogenomic tree, we used Panaroo to identify the shared protein-coding genes across the 130 *C. pseudotuberculosis* isolates, and the outgroup *C. ulcerans* NCTC 7910 (GCF_900187135.1) and perform a multiple sequence alignment (MSA) using MAFFT (Katoh et al., 2005). The phylogenetic inference was made from the MSA using IQ-TREE2 v2.0.7 (Minh et al., 2020) with the maximum-likelihood (ML) method in which the best-fit model of nucleotide substitution was selected by ModelFinder (Kalyaanamoorthy et al., 2017), and support values were calculated using ultrafast bootstrap approximation with 1,000 replicates (Hoang et al., 2018). The tree was visualized and annotated using the Interactive Tree of Life (iTOL) v6.8 (https://itol.embl.de/).

A pangenome is a set of non-redundant genes that composes the repertoire of all genomes of a species (Tettelin et al., 2005). The pangenome and distribution of genes across all the 130 strains were identified using Panaroo due to its feature of "refinding" genes that were not annotated due to annotation artifacts. In this software, the genes are classified by frequency in the categories core genes (99% <= strains <= 100%), softcore genes (95% <= strains < 99%), shell genes (15% <= strains < 95%), cloud genes (0% <= strains < 15%) and total genes (0% <= strains <= 100%) (Tonkin-Hill et al., 2020). The pangenome development was calculated using Heap's Law formula, implemented in the R package Micropan v2.1 (https://github.com/larssnip/micropan) to estimate whether it is an open or closed pangenome using 10,000 permutations (nper = 10,000). The gene clusters were annotated using eggNOG-mapper v2.1.9 with database v5.0.2 (Huerta-Cepas et al., 2017).

Results

Quality assessment and taxonomy

All genomes were classified as *C. pseudotuberculosis*. Completeness and contamination ranged between 97.37% and 100% and 0.14% and 6.43%, respectively. No evidence of chimerism was detected by GUNC (*Data S1*).

Analysis of Cp162 genome assemblies

The comparisons between the three assembly versions showed increased genome size and the number of coding sequences (CDS) (*Table 1*). Across all three versions, we identified 2,128 genes, 2,010 of them shared. About virulence genes, the first assembly version lacks the virulence genes *nanH* while *spaC* is fragmented as CP162_RS09080 and CP162_RS09085). The synteny analysis showed two inversions and one rearrangement in the first version, which were later corrected in the subsequent versions using the optical map (*Fig. 1*). Since the first assembly, the genome had an increase of 88 Kbp and 121 CDSs and a decrease of pseudogenes from 139 to 53. The third version has 18 exclusively present genes (*Table 1, Data S2*).

In the third assembly version, we predicted one incomplete prophage with 8.9Kb and 18 CDSs (*Data S3, Data S4* and *S5*), 13 GEIs, five of them pathogenicity islands (*Fig. 2*, *Data S5*), three CRISPR arrays, and a Type I-E CRISPR-Cas system (*Data S5*). PanViTa identified three

virulence factors: Elongation Factor Tu (*tufA,* protein id WP_013241194.1), UPF0182 family protein (DIP_RS14950, WP_014800143.1), and Microcin-production regulation, locus A (*mprA,* WP_014800209.1). Using a BLASTp, we identified other six virulence factors: *pld* (WP_014799831.1), *cpp* (WP_072577955.1), *nanH* (WP_072577765.1), *spaC* (WP_231131414.1), *sodC* (WP_013241467.1) and *pknG* (WP_041481395.1). PanViTa identified two antimicrobial resistance proteins: Rifampin-resistant beta-subunit of RNA polymerase (*rpoB2*, WP_041481489.1) and RbpA bacterial RNA polymerase-binding protein (*rpbA*, WP_014800420.1) (*Table 1*).

Phylogeny and pangenome

A tree was generated by IQ-TREE2 using core genome alignment from Panaroo and MAFFT. The species tree in *Fig. 3* shows two main clades. The first one contains a subclade composed of strains Cp162 (camel), G1 (alpaca), and I37 (cow, Israel) and another subclade containing strains isolated from horses and buffalo, separated by a host. The second one contains strain 262 (cow, Belgium) and all biovar ovis strains (collapsed) as its sister group. In the pangenome analysis, 2,332 genes were identified in the pangenome ($\alpha = 1.27$), 1,877 as core, 68 as softcore, 173 as shell, and 214 as cloud. Of 2,332 genes, 2,181 were scanned by eggNOG-mapper, and 1,953 had a functional annotation (*Data S6*).

Exclusive genes of Cp162

From the pangenome analysis, we also identified four proteins exclusively present in Cp162 and three exclusively absent in this lineage (*Table 2*). In the exclusively present group, two were predicted as the same transposase (WP_048653436.1). The other two are truncated (41 and 45 aa), none showed conserved domains, and one is in GEI 6. Analysis against the BLAST database using WP_275060758.1 as a query showed 92% of coverage with 52% of identity to an ATP-dependent helicase of *Streptomyces spp.* The same analysis with WP_231131458.1 showed 97.8% of identity with other truncated proteins from CpE19_1664 (AKS14002.1).

From the group of proteins exclusively absent in Cp162, one was recognized by eggNOG-mapper as a Transcriptional Regulator named *lysG* (COG category: K), another as an enzyme from NUDIX superfamily (COG category: L), and the last one as a hypothetical protein with no domains.

Discussion

Our results showed that using Illumina HiSeq and an optical map increased the genome size and number of CDSs, corrected misassembles, and reduced the number of pseudogenes (*Table 1*). Concerning synteny, the correct sequence order and content are required to study the genome plasticity events such as inversion, translocation, insertion, and deletions (Lehri, Seddon & Karlyshev, 2017). As shown in *Fig. 1*, optical mapping could correctly order contigs from sequencing. The rearranged regions are strictly between transposase genes, which could explain possible rearrangements (Hickman & Dyda, 2016). Some transposase sequences were found only

in the third version of the Cp162 genome, within its exclusive 18 genes (*Data S2*). With frameshifts, the first and second genome versions were sequenced using SOLiD and Ion Torrent platforms, known for indel sequencing errors (Loman et al., 2012), leading to CDS frameshifts. The correct identification of pseudogenes is required for gene evolution analysis and for gene content studies such as pangenomics. In NCBI's PGAP annotation pipeline (Li et al., 2021), a pseudogene will not have a CDS annotation, while in the RAST-Tk pipeline (Brettin et al., 2015), each fragment of a pseudogene can be annotated as a CDS; this can lead to erroneous estimations of gene content across genomes and suggests that data generated using SOLiD and Ion Torrent should be used with caution.

In relation to virulence factors of strain Cp162, we identified three others using PanViTa with the VFDB database and six others described in the literature for the species (*pld*, *cpp*, *nanH*, *spaC*, *sodC*, *pknG*) (Trost et al., 2010; Santana-Jorge et al., 2016; Viana et al., 2017) using BLASTp (*Table 1*). This different result is because those six ones are not included as virulence factors of *Corynebacterium* in VFDB (http://www.mgc.ac.cn/cgi-bin/VFs/genus.cgi?Genus=Corynebacterium), probably because most of them are niche factors rather than *stricto sensu* virulence factors (Tauch & Burkovski, 2015), or because the query sequences had an identity value below the standard cutoff of PanViTa; this highlights the necessity of updating the database for *Corynebacterium* by including more representative sequences. Within antimicrobial resistance genes, the identified genes *rpoB2* and *rbpA* confer resistance to rifampin according to the CARD database (CARD accessions ARO:3000501 and ARO:3000245). We predicted incomplete prophage (*Data S3* and *S4*) and 13 GEIs (*Fig. 3*, *Data S5*). GEI 5 is exclusive to the clade composed of Cp162 (camel), I37 (cow), and G1 (alpaca) (*Fig. 2*), but it may not be due to a common ancestor rather than host tropism because strain 262 (equi) and I19 (ovis) also infect cows. The genome has three CRISPR arrays and a Type I-E CRISPR-Cas system (*Table 1*). Type I-E was previously found only in biovar equi in *C. pseudotuberculosis*, while proteins from Type III restriction-modification systems were exclusive from biovar ovis (Parise et al., 2018).

Cp162 is the only strain from a camel with a sequenced genome, and we looked for genes that could be involved in the tropism of this host by comparing its genome to 129 others from the same species. The exclusively present genes are transposases and truncated proteins, while the exclusively absent are *lysG*, an enzyme from the NUDIX superfamily and a hypothetical protein with no domains (*Table 2*). There is no clear relation between those genes and host tropism for camels. If there are any genome features related to tropism, they could be verified by sequencing the genomes of more strains from this host and performing a genome-wide association study (GWAS) testing for gene presence/absence or SNPs.

The phylogeny of 130 genomes (*Fig. 3*) supports the previous assumption that biovar ovis is a clade that originated from biovar equi, with its exclusive adaptations, and biovar equi as paraphyletic with two exclusive hosts (horse and buffalo) (Viana et al., 2018). Sampling more strains from camels could show they form exclusive clades in biovar equi and ovis that could suggest clonal expansion after host adaptation. The species pan-genome was estimated as closed ($\alpha > 1.00$), which means that sequencing more genomes will not reveal new genes (Tettelin et al.,

2008). The rifampin resistance genes identified in Cp162 (*rpoB2* and *rpbA*) are present in all 130 genomes, which suggests this antimicrobial should be avoided for infection treatment.

Conclusions

The genome resequencing of strain Cp162 and assembly using an optical map resulted in corrections of synteny and fewer pseudogenes caused by sequencing artifacts. The comparative analysis suggests that there are no genes related to the tropism for camels, but this could be tested again using more genomes from this host and performing GWAS testing for genes and SNPs.

References

Alcock BP, Huynh W, Chalil R, Smith KW, Raphenya AR, Wlodarski MA, Edalatmand A, Petkau A, Syed SA, Tsang KK, Baker SJC, Dave M, McCarthy MC, Mukiri KM, Nasir JA, Golbon B, Imtiaz H, Jiang X, Kaur K, Kwong M, Liang ZC, Niu KC, Shan P, Yang JYJ, Gray KL, Hoad GR, Jia B, Bhando T, Carfrae LA, Farha MA, French S, Gordzevich R, Rachwalski K, Tu MM, Bordeleau E, Dooley D, Griffiths E, Zubyk HL, Brown ED, Maguire F, Beiko RG, Hsiao WWL, Brinkman FSL, Van Domselaar G, McArthur AG. 2023. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Research* 51:D690–D699. DOI: 10.1093/nar/gkac920.

Alikhan N-F, Petty NK, Ben Zakour NL, Beatson SA. 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12:402. DOI: 10.1186/1471-2164-12-402.

Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, Wishart DS. 2016. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic acids research* 44:W16–W21. DOI: 10.1093/nar/gkw387.

Brettin T, Davis JJ, Disz T, Edwards R a, Gerdes S, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD, Shukla M, Thomason J a, Stevens R, Vonstein V, Wattam AR, Xia F. 2015. RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific Reports* 5:1–6. DOI: 10.1038/srep08365.

Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2022. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* 38:5315–5316. DOI: 10.1093/bioinformatics/btac672.

Couvin D, Bernheim A, Toffano-Nioche C, Touchon M, Michalik J, Néron B, Rocha EPC, Vergnaud G, Gautheret D, Pourcel C. 2018. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Research* 46:W246–W251. DOI: 10.1093/NAR/GKY425.

Craddock HA, Motro Y, Zilberman B, Khalfin B, Bardenstein S, Moran-Gilad J. 2022. Long-Read Sequencing and Hybrid Assembly for Genomic Analysis of Clinical *Brucella melitensis* Isolates. *Microorganisms* 10:619. DOI: 10.3390/microorganisms10030619.

Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Research* 14:1394–1403. DOI: 10.1101/gr.2289704.

279 Dorella FA, Carvalho Pacheco L, Oliveira SC, Miyoshi A, Azevedo V. 2006. *Corynebacterium*
280 *pseudotuberculosis* : microbiology, biochemical properties, pathogenesis and molecular studies of
281 virulence. *Veterinary Research* 37:201–218. DOI: 10.1051/vetres:2005056.

282 Gardy JL, Loman NJ. 2018. Towards a genomics-informed, real-time, global pathogen surveillance
283 system. *Nature Reviews Genetics* 19:9–20. DOI: 10.1038/nrg.2017.88.

284 Goodswen SJ, Kennedy PJ, Ellis JT. 2023. A state-of-the-art methodology for high-throughput in silico
285 vaccine discovery against protozoan parasites and exemplified with discovered candidates for
286 *Toxoplasma gondii*. *Scientific Reports* 13:8243. DOI: 10.1038/s41598-023-34863-9.

287 Hassan SS, Schneider MPC, Ramos RTJ, Carneiro AR, Ranieri A, Guimarães LC, Ali A, Bakhtiar SM,
288 Pereira U de P, Santos AR dos, Soares S de C, Dorella F, Pinto AC, Ribeiro D, Barbosa MS,
289 Almeida íntia, Abreu V, Aburjaile F, Fiaux K, Barbosa E, Diniz C, Rocha FS, Saxena R, Tiwari S,
290 Zambare V, Ghosh P, Pacheco LGC, Dowson CG, Kumar A, Barh D, Miyoshi A, Azevedo V, Silva
291 A. 2012. Whole-genome sequence of *Corynebacterium pseudotuberculosis* strain Cp162, isolated
292 from camel. *Journal of Bacteriology* 194:5718–5719. DOI: 10.1128/JB.01373-12.

293 Hickman AB, Dyda F. 2016. DNA Transposition at Work. *Chemical Reviews* 116:12758–12784. DOI:
294 10.1021/acs.chemrev.6b00003.

295 Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast
296 Bootstrap Approximation. *Molecular Biology and Evolution* 35:518–522. DOI:
297 10.1093/molbev/msx281.

298 Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, Von Mering C, Bork P. 2017. Fast
299 genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular*
300 *Biology and Evolution* 34:2115–2122. DOI: 10.1093/molbev/msx148.

301 Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model
302 selection for accurate phylogenetic estimates. *Nature Methods* 14:587–589. DOI:
303 10.1038/nmeth.4285.

304 Katoh K, Kuma KI, Toh H, Miyata T. 2005. MAFFT version 5: Improvement in accuracy of multiple
305 sequence alignment. *Nucleic Acids Research* 33:511–518. DOI: 10.1093/nar/gki198.

306 Lehri B, Seddon AM, Karlyshev A V. 2017. The hidden perils of read mapping as a quality assessment
307 tool in genome sequencing. *Scientific Reports* 7:43149. DOI: 10.1038/srep43149.

308 Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvernin V, Badretdin A, Coulouris G, Chitsaz F,
309 Derbyshire MK, Durkin AS, Gonzales NR, Gwadz M, Lanczycki CJ, Song JS, Thanki N, Wang J,
310 Yamashita RA, Yang M, Zheng C, Marchler-Bauer A, Thibaud-Nissen F. 2021. RefSeq: expanding
311 the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic*
312 *Acids Research* 49:D1020–D1028. DOI: 10.1093/nar/gkaa1105.

313 Liu B, Zheng D, Zhou S, Chen L, Yang J. 2022. VFDB 2022: a general classification scheme for bacterial
314 virulence factors. *Nucleic Acids Research* 50:D912–D917. DOI: 10.1093/nar/gkab1107.

315 Loman NJ, Misra R V., Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ. 2012.
316 Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*
317 30:434–439. DOI: 10.1038/nbt.2198.

Di Marco F, Spitaleri A, Battaglia S, Batignani V, Cabibbe AM, Cirillo DM. 2023. Advantages of long- and short-reads sequencing for the hybrid investigation of the *Mycobacterium tuberculosis* genome. *Frontiers in Microbiology* 14. DOI: 10.3389/fmicb.2023.1104456.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*:6–10. DOI: 10.1093/molbev/msaa015.

Orakov A, Fullam A, Coelho LP, Khedkar S, Szklarczyk D, Mende DR, Schmidt TSB, Bork P. 2021. GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biology* 22:178. DOI: 10.1186/s13059-021-02393-0.

Parise D, Parise MTD, Viana MVC, Muñoz-Bucio A V, Cortés-Pérez YA, Arellano-Reynoso B, Díaz-Aparicio E, Dorella FA, Pereira FL, Carvalho AF, Figueiredo HCP, Ghosh P, Barh D, Gomide ACP, Azevedo VAC. 2018. First genome sequencing and comparative analyses of *Corynebacterium pseudotuberculosis* strains from Mexico. *Standards in Genomic Sciences* 13:21. DOI: 10.1186/s40793-018-0325-z.

Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil P-A, Hugenholtz P. 2022. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research* 50:D785–D794. DOI: 10.1093/nar/gkab776.

Rodrigues DLN, Ariute JC, Rodrigues da Costa FM, Benko-Iseppon AM, Barh D, Azevedo V, Aburjaile F. 2023. PanViTa: Pan Virulence and resisTance analysis. *Frontiers in Bioinformatics* 3. DOI: 10.3389/fbinf.2023.1070406.

Santana-Jorge KTO, Santos TM, Tartaglia NR, Aguiar EL, Souza RFS, Mariutti RB, Eberle RJ, Arni RK, Portela RW, Meyer R, Azevedo V. 2016. Putative virulence factors of *Corynebacterium pseudotuberculosis* FRC41: vaccine potential and protein expression. *Microbial Cell Factories* 15:83. DOI: 10.1186/s12934-016-0479-6.

Serral F, Castello FA, Sosa EJ, Pardo AM, Palumbo MC, Modenutti C, Palomino MM, Lazarowski A, Auzmendi J, Ramos PIP, Nicolás MF, Turjanski AG, Martí MA, Fernández Do Porto D. 2021. From Genome to Drugs: New Approaches in Antimicrobial Discovery. *Frontiers in Pharmacology* 12. DOI: 10.3389/fphar.2021.647060.

Sheppard SK, Guttman DS, Fitzgerald JR. 2018. Population genomics of bacterial host adaptation. *Nature Reviews Genetics* 19:549–565. DOI: 10.1038/s41576-018-0032-z.

Soares SC, Geyik H, Ramos RTJ, de Sá PHCG, Barbosa EGV, Baumbach J, Figueiredo HCP, Miyoshi A, Tauch A, Silva A, Azevedo V. 2016. GIPSy: Genomic island prediction software. *Journal of Biotechnology* 232:2–11. DOI: 10.1016/j.jbiotec.2015.09.008.

Sousa T de J, Parise D, Profeta R, Parise MTD, Gomide ACP, Kato RB, Pereira FL, Figueiredo HCP, Ramos R, Brenig B, Costa da Silva AL da, Ghosh P, Barh D, Góes-Neto A, Azevedo V. 2019. Re-sequencing and optical mapping reveals misassemblies and real inversions on Corynebacterium pseudotuberculosis genomes. *Scientific Reports* 9. DOI: 10.1038/s41598-019-52695-4.

Tauch A, Burkovski A. 2015. Molecular armory or niche factors: virulence determinants of Corynebacterium species. *FEMS Microbiology Letters* 67:fnv185. DOI: 10.1093/femsle/fnv185.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli S V., Crabtree J, Jones AL, Durkin AS, DeBoy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae* : Implications for the microbial "pan-genome." *Proceedings of the National Academy of Sciences* 102:13950–13955. DOI: 10.1073/pnas.0506758102.

Tettelin H, Riley D, Cattuto C, Medini D. 2008. Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology* 11:472–477. DOI: 10.1016/j.mib.2008.09.006.

Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, Gladstone RA, Lo S, Beaudoin C, Floto RA, Frost SDW, Corander J, Bentley SD, Parkhill J. 2020. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology* 21:180. DOI: 10.1186/s13059-020-02090-4.

Trost E, Ott L, Schneider J, Schröder J, Jaenicke S, Goesmann A, Husemann P, Stoye J, Dorella FA, Rocha FS, Soares SDC, D'Afonseca V, Miyoshi A, Ruiz J, Silva A, Azevedo V, Burkovski A, Guiso N, Join-Lambert OF, Kayal S, Tauch A. 2010. The complete genome sequence of *Corynebacterium pseudotuberculosis* FRC41 isolated from a 12-year-old girl with necrotizing lymphadenitis reveals insights into gene-regulatory networks contributing to virulence. *BMC genomics* 11:728. DOI: 10.1186/1471-2164-11-728.

Viana MVC, Figueiredo H, Ramos R, Guimarães LC, Pereira FL, Dorella FA, Selim SAK, Salaheldean M, Silva A, Wattam AR, Azevedo V. 2017. Comparative genomic analysis between *Corynebacterium pseudotuberculosis* strains isolated from buffalo. *PLOS ONE* 12:e0176347. DOI: 10.1371/journal.pone.0176347.

Viana MVC, Sahm A, Góes Neto A, Figueiredo HCP, Wattam AR, Azevedo V. 2018. Rapidly evolving changes and gene loss associated with host switching in *Corynebacterium pseudotuberculosis*. *PloS one* 13:e0207304. DOI: 10.1371/journal.pone.0207304.

Wernery U, Kinne J. 2016. Caseous Lymphadenitis (Pseudotuberculosis) in Camelids: A Review. *Austin Journal of Veterinary Science & Animal Husbandry*.

Table 1(on next page)

Comparison between the three versions of the strain Cp162 genome assembly.

1

	Version 1	Version 2	Version 3
Deposit date	01/31/2014	07/15/2017	12/16/2019
Platform	SOLiD	IonTorret	Illumina HiSeq 2500
Coverage	686x	200x	713x
Size (bp)	2,293,464	2,365,874 (+72,410)	2,382,183 (+88,639)
Completeness (%)	99.9	99.9	99.9
Contamination (%)	0.21	0.2	0.23
CDSs	2,043	2,112 (+69)	2,164 (+121)
Exclusively present CDSs	3	2	18
Exclusively absent CDSs	88	1	6
Pseudo Genes (total)	139	80	53
Pseudo Genes (ambiguous residues)	0	0	0
Pseudo Genes (frameshifted)	116	67	41
Pseudo Genes (incomplete)	13	9	7
Pseudo Genes (internal stop)	17	9	6
Pseudo Genes (multiple problems)	7	5	1
tRNA	49	49	63
5S rRNA	4	4	4
16S rRNA	4	4	4
23S rRNA	4	4	4
ncRNA	3	3	3
Virulence genes	*cpp, DIP_RS14950, mprA, pknG, pld, sodC, tufA*	*cpp, DIP_RS14950, mprA, nanH, pknG, pld, sodC, spaC, tufA*	*cpp, DIP_RS14950, mprA, nanH, pknG, pld, sodC, spaC, tufA*
Antimicrobial resistance genes	*rpoB2*, *rbpA*	*rpoB2*, *rbpA*	*rpoB2*, *rbpA*

2

Table 2(on next page)

Exclusively present and absent genes in *Corynebacterium pseudotuberculosis* strain Cp162 (camel) in comparison to other 129 genomes of the same species.

COG – Cluster of Orthologous Genes, GEI – Genomic Island, KEGG – Kyoto Encyclopedia of Genes and Genomes.

1

Locus Tag (protein ID)	Gene	Product	GEI	Functional annotation
Exclusively present				
CP162_RS04525 (WP_048653436.1)	tnp3510a	IS110 family transposase	-	COG: L, Pfam: DEDD_Tnp_IS110, Transposase_20
CP162_RS09445 (WP_048653436.1)	tnp3510a	IS110 family transposase	-	-
P162_RS11030 (WP_231131458.1)	-	Hypothetical protein	GEI 6	-
CP162_RS11150 (WP_275060758.1)	-	Hypothetical protein	-	-
Exclusively absent				
(WP_014366903.1)	*lysG*	Transcriptional regulator	-	COG: K, KEGG: K05596, PFAM: HTH_1, LysR_substrate
(AKC74244.1)	-	NTP pyrophosphohydrolases, including oxidative damage repair enzymes	-	COG: L, PFAM: NUDIX
(WP_038617038.1)	-	-	-	-

2

Figure 1

Alignment of the three genome assembly versions of Corynebacterium pseudotuberculosis strain Cp162.

(A) Two inversions and one rearrangement in the first assembly. (B) Increase in genome size throughout the assemblies.

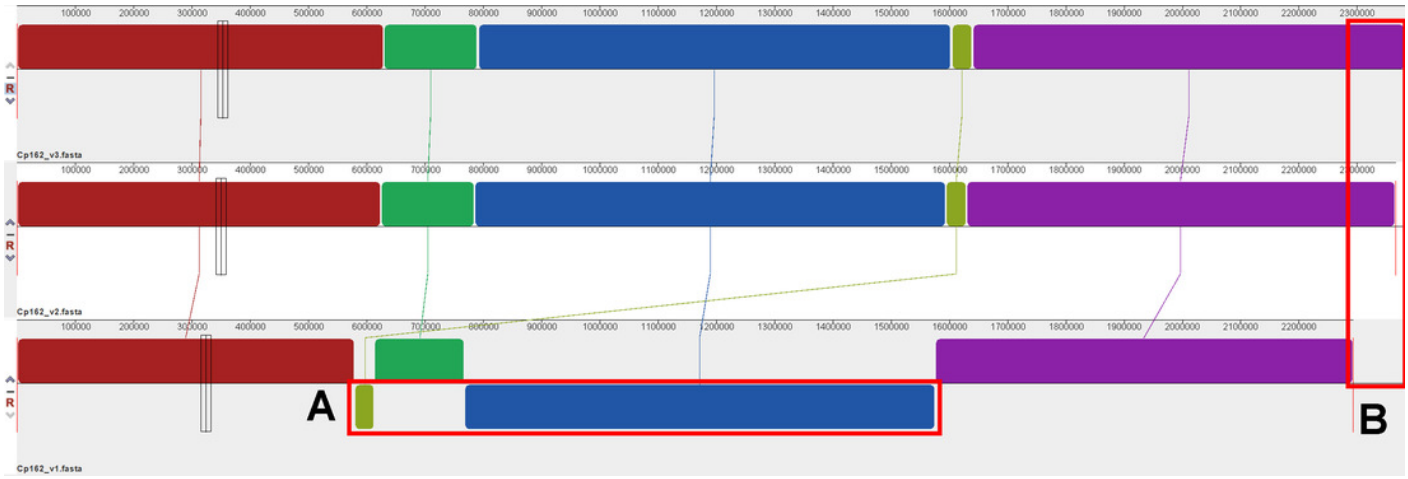

Figure 2

Circular map of *Corynebacterium pseudotuberculosis* Cp162 isolated from camel.

From inner to outer circle: Cp162 v3 (equi, camel), CG Content, GC Skew, Cp162 v2, Cp162 v1, I31 (equi, cow), G1 (equi, alpaca), 31 (equi, buffalo), 258 (equi, cow), 262 (equi, cow), I19 (ovis, cow), 1002B (ovis, goat), genomic islands (GEI) and pathogenicity islands (PAI), prophage, and exclusive genes of Cp162 v3 in comparison to v2 and v1, and exclusive genes of Cp162 v3.

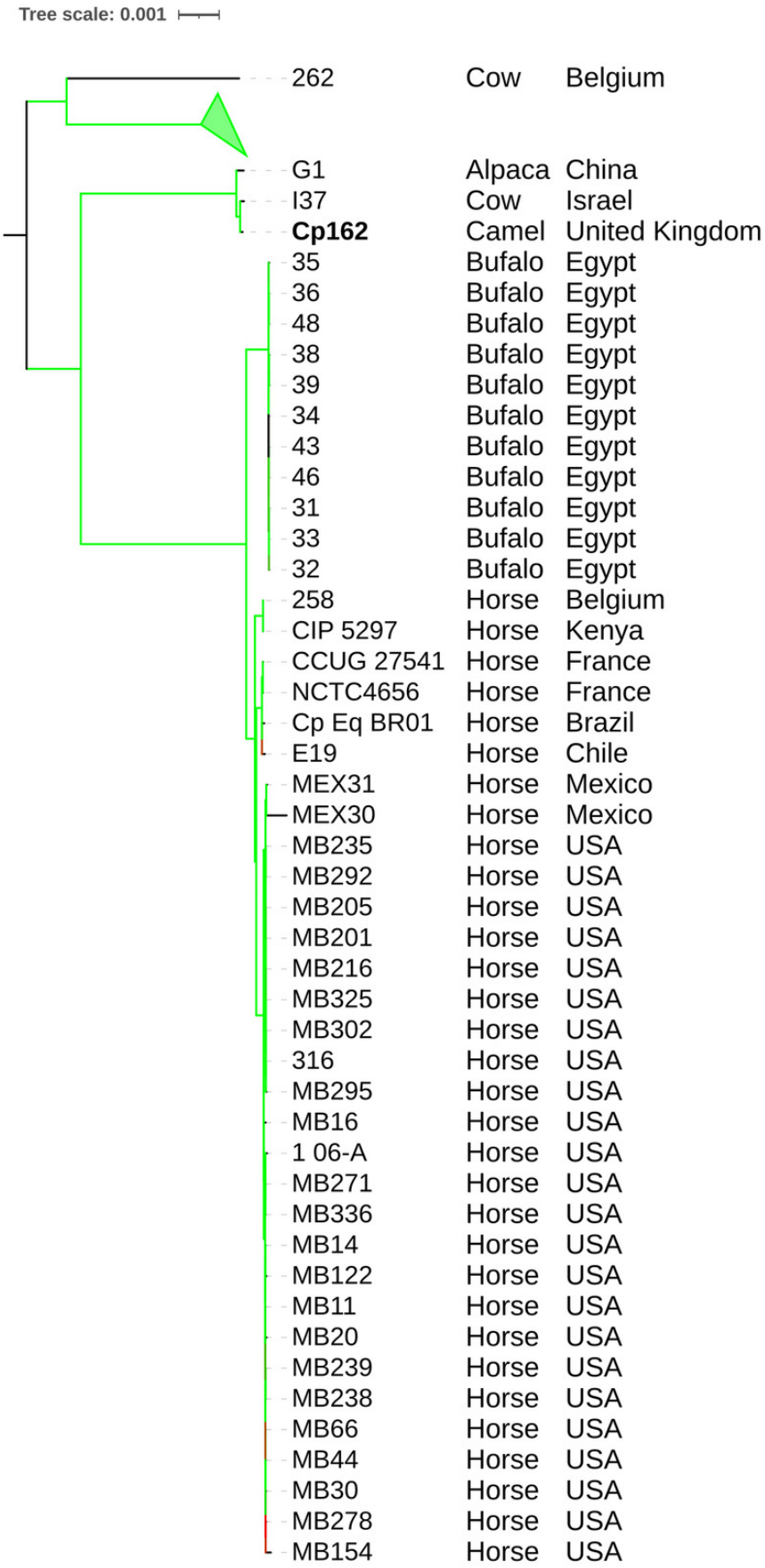
Tree scale: 0.001
262 Cow Belgium
G1 Alpaca China
I37 Cow Israel
Cp162 Camel United Kingdom
35 Bufalo Egypt
36 Bufalo Egypt
48 Bufalo Egypt
38 Bufalo Egypt
39 Bufalo Egypt
34 Bufalo Egypt
43 Bufalo Egypt
46 Bufalo Egypt
31 Bufalo Egypt
33 Bufalo Egypt
32 Bufalo Egypt
258 Horse Belgium
CIP 5297 Horse Kenya
CCUG 27541 Horse France
NCTC4656 Horse France
Cp Eq BR01 Horse Brazil
E19 Horse Chile
MEX31 Horse Mexico
MEX30 Horse Mexico
MB235 Horse USA
MB292 Horse USA
MB205 Horse USA
MB201 Horse USA
MB216 Horse USA
MB325 Horse USA
MB302 Horse USA
316 Horse USA
MB295 Horse USA
MB16 Horse USA
1 06-A Horse USA
MB271 Horse USA
MB336 Horse USA
MB14 Horse USA
MB122 Horse USA
MB11 Horse USA
MB20 Horse USA
MB239 Horse USA
MB238 Horse USA
MB66 Horse USA
MB44 Horse USA
MB30 Horse USA
MB278 Horse USA
MB154 Horse USA

Figure 3

Phylogenomic tree of *Corynebacterium pseudotuberculosis* genomes.

The tree was built using the core genome identified and aligned using Panaroo and MAFFT, respectively. A phylogeny using the Maximum Likelihood method was built using IQ-TREE2 with 1,000 replicates of ultrafast bootstrap approximation and *C. ulcerans* NCTC7910 (not shown) as an outgroup. Bootstrap values are represented as a branch color scale that ranges from 85% (red) to 100% (green). The biovar ovis clade is collapsed.