Genome sequencing of Mycobacterium pinnipedii strains: genetic characterization and evidence of superinfection in a South American sea lion (Otaria flavescens)

Taiana Silva Pereira
Universidade de Sao Paulo https://orcid.org/0000-0001-9313-0151
Cássia Y. Ikuta
Universidade de Sao Paulo
Cristina K. Zimpel
Universidade de Sao Paulo
Naila C. S. Camargo
Universidade de Sao Paulo
Antônio F. de Souza Filho
Universidade de Sao Paulo
José S. Ferreira Neto
Universidade de Sao Paulo
Marcos B. Heinemann
Universidade de Sao Paulo
Ana M. S. Guimarães (anamarcia@usp.br)
https://orcid.org/0000-0002-8261-5863

Research article

Keywords: Mycobacterium pinnipedii, genome, superinfection, comparative genomics, Mycobacterium tuberculosis Complex.

Posted Date: December 20th, 2019

DOI: https://doi.org/10.21203/rs.2.9779/v4

Version of Record: A version of this preprint was published on December 30th, 2019. See the published version at https://doi.org/10.1186/s12864-019-6407-5.

Abstract

Background: Mycobacterium pinnipedii , a member of the Mycobacterium tuberculosis Complex (MTBC), is capable of infecting several host species, including humans. Recently, ancient DNA from this organism was recovered from pre-Columbian mummies of Peru, sparking debate over the origin and frequency of tuberculosis in the Americas prior to European colonization.

Results: We present the first comparative genomic study of this bacterial species, starting from the genome sequencing of two M. pinnipedii isolates (MP1 and MP2) obtained from different organs of a stranded South American sea lion. Our results indicate that MP1 and MP2 differ by 113 SNPs (single nucleotide polymorphisms) and 46 indels, constituting the first report of a mixed-strain infection in a sea lion. SNP annotation analyses indicate that genes of the VapBC family, a toxin-antitoxin system, and genes related to cell wall remodeling are under evolutionary pressure for protein sequence change in these strains. OrthoMCL analysis with seven modern isolates of M. pinnipedii shows that these strains have highly similar proteomes. Gene variations were only marginally associated with hypothetical proteins and PE/PPE (proline-glutamate and proline-proline-glutamate, respectively) gene families. We also detected large deletions in ancient and modern M. pinnipedii strains, including a few occurring only in modern strains, indicating a process of genome reduction occurring over the past one thousand years. Our phylogenomic analyses suggest the existence of two modern clusters of M. pinnipedii associated with geographic location, and possibly host species, and one basal node associated with the ancient M. pinnipedii strains. Previously described MiD3 and MiD4 deletions may have occurred independently, twice, over the evolutionary course of the MTBC.

Conclusion: The presence of superinfection (i.e. mixed-strain infection) in this sea lion suggests that M. pinnipedii is highly endemic in this population. Mycobacterium pinnipedii proteomes of the studied isolates showed a high degree of conservation, despite being under genomic decay when compared to M. tuberculosis. This finding indicates that further genomes need to be sequenced and analyzed to increase the chances of finding variably present genes among strains or that M. pinnipedii genome remodeling occurred prior to bacterial speciation.

Background

Tuberculosis is a contagious and often severe infectious disease caused by members of the *Mycobacterium tuberculosis* Complex (MTBC). MTBC is a clonal bacterial group composed of 11 species that can be divided into two major groups: those highly adapted to human beings, named *Mycobacterium tuberculosis* and *Mycobacterium africanum* L5 and L6, and those adapted to animal hosts, named *Mycobacterium bovis, Mycobacterium caprae, Mycobacterium microti, Mycobacterium pinnipedii, Mycobacterium origys, Mycobacterium mungi, Mycobacterium suricattae*, "dassie bacillus", and “chimpanzee bacillus” [1–3]. These species share great genomic similarity, with more than 99.95% nucleotide identity over alignable regions, and no evidence of horizontal gene transfer or major recombination events [4]. Single nucleotide polymorphisms (SNPs) and deletions of genomic regions ranging from 2 to 12.7 Kb, denominated "regions of differences (RDs)”, allow for species differentiation. Despite this high genetic similarity, members of MTBC vary in their

host tropism and ability to cause disease, which are likely consequences of these discrete genetic differences in addition to the presence of duplications and mobile genetic elements [3].

Recently, special attention has been given to the MTBC species known to infect pinnipeds, *M. pinnipedii*, when its DNA was detected in three 1,000-year-old, pre-Columbian human skeletons from South America (Peru). This study was the first genetic evidence of tuberculous mycobacteria infecting humans prior to the Europeans' first contact with the New World [5]. Moreover, it was also the first report of the genome sequencing of modern *M. pinnipedii* strains, yet to be fully compared and analyzed. In modern days, tuberculosis in pinnipeds was first reported in 1913 [6] with *M. pinnipedii* being described as "seal bacillus" in the early 1990s [7, 8] and proposed as a new member of the MTBC in 2003 [2]. Strains of *M. pinnipedii* have been isolated from pinnipeds worldwide, especially in captivity, but also from free-living animals of the southern hemisphere [9, 10]. Surprisingly, since its first description, *M. pinnipedii* has been detected in a variety of host species, including six different pinniped species [11–13], bactrian camels, snow leopards, amur leopards, cattle, llamas, lowland gorillas, Malayan tapirs, Hector's dolphins [14], and possibly humans, which supports a generalist behavior for host tropism [9, 10, 15–18]. Pinnipeds infected with *M. pinnipedii* in zoos and marine parks are the main source of infection for other animals and humans. Furthermore, reports of infected cattle were associated with a contaminated water canal connecting directly to the ocean or beach grazing areas where pinnipeds were found [18]. The actual impact, however, of tuberculosis in pinniped species remains largely unknown, mainly because its prevalence in free-ranging seals and sea lions is completely unexplored.

The carcass of a South American sea lion (*Otaria flavescens*) has recently been recovered from the southern coast of the state of Rio Grande do Sul, Brazil. As reproductive colonies of these animals are absent in Brazil, specimens found on the coast are probably from Uruguay. At the animal necropsy, lesions compatible with tuberculosis in different organs were found, from which *M. pinnipedii* isolates were obtained, constituting the first report of this pathogen in the Brazilian coast [19]. Unfortunately, genome sequencing of the isolates was not undertaken at the time, precluding opportunities to better explore the genetic makeup of these bacteria, including the possibility of intra-host bacterial clonal variants (i.e. microevolution) or mixed-strain infection (i.e. superinfection). Until now, *M. pinnipedii* genome sequences have never been fully characterized by comparative genomics. Genomic reads of modern *M. pinnipedii* strains described by Bos et al. [5] were solely used to compare to the ancient *M. pinnipedii*-like genomes. During the development of this study, another strain of *M. pinnipedii* was sequenced [20], but only applied to evaluate the phylogenomic position of this species within the MTBC. Therefore, the objectives of the present study were to sequence and compare two *M. pinnipedii* isolates obtained from different organs of this sea lion carcass [19], and to perform a comparative genomic analysis with other *M. pinnipedii* strains available from public databases.

Results And Discussion

First description of M. pinnipedii mixed-strain infection in a pinniped

The genome sequencing and assembly of the two *M. pinnipedii* isolates (named MP1 and MP2) resulted in 9,766,154 reads (228x coverage), and 9,433,048 reads (220x coverage), assembled into 102 and 106 contigs (~4.28 Mb/genome), respectively. Genome annotation with Prokaryotic Genome Annotation Pipeline (PGAP)

indicates a GC content of 65.4% and 45 tRNA for both strains, and 4,269 and 4,276 genes, 3,993 and 3,996 coding DNA sequences (CDSs), and 152 and 158 pseudogenes for *M. pinnipedii* MP1 and MP2, respectively. Genomes of both strains showed similar characteristics to other genomes of the same species available in the database (Fig. 1).

Mapping of the *M. pinnipedii* MP1 and MP2 reads with the genome of the *M. tuberculosis* strain H37Rv resulted in a reference coverage of 99.08% and 99.03%, and 2,312 and 2,350 variants, respectively. Upon exclusion of false-positive prone areas, the isolates shared 1,912 SNPs (*M. pinnipedii* MP1) and 1,998 SNPs (*M. pinnipedii* MP2) with *M. tuberculosis* H37Rv. In addition, 181 (*M. pinnipedii* MP1) and 208 (*M. pinnipedii* MP2) indels (single-nucleotides or short indels ranging from 2 to 36 bp) were detected. When comparing *M. pinnipedii* MP1 and *M. pinnipedii* MP2, the distance between these strains, after excluding false-positive prone areas, was found to be 113 SNPs and 46 indels (single-nucleotides or short indels ranging from 2 to 36 bp). A total of 122 (76.73%) of these 159 SNPs and indels were located in CDSs with known functions, based on the COG (Cluster of Orthologous Groups) functional annotation (Additional file 1: Table S1). The main functional annotations were those related to energy production and conversion (11.95%, 19/159), lipid transport and metabolism (9.43%, 15/159) and cell wall/membrane/envelope biogenesis (7.55%, 12/159), which are known to play crucial roles in the pathogenicity of MTBC species. Interestingly, the majority of detected SNPs are non-synonymous (86.72%, 98/113), which means that related CDSs are under selective pressure for protein sequence change, having the potential to promote functional diversity.

CDSs of known function with non-synonymous SNPs identified between *M. pinnipedii* MP1 and *M. pinnipedii* MP2 were evaluated using GO enrichment analysis and STRING protein networks. The statistically significant biological processes identified were: trehalose metabolic process (P value = 3.07E-04), fatty acid biosynthetic process (P value = 2.70E-04) and cell wall biogenesis (P value = 2.20E-04). The identified molecular functions were ATPase activity (P value = 8.05E-05) and ATP binding (P value = 5.09E-05), while the identified cellular component was the cell wall class (P value = 9.95E-05). These findings suggest an evolutionary pressure for change in genes associated with cell wall remodeling. The unique lipid-rich mycobacteria cell wall is the first line of interaction between host and pathogen, with crucial implications in virulence, and is also important for environmental stress resistance (e.g. protection from osmotic stress, starvation, freezing, and desiccation).

When using STRING, we identified strong protein interactions between specific members of the VapBC protein family (Fig. 2). The fact that multiple proteins with STRING-detected relationships have non-synonymous SNPs also indicates that these Vap genes, in particular, are under evolutionary pressure for change. In *M. tuberculosis*, the VapBC family accounts for more than half of the proportion of toxin-antitoxin (TA) systems (47/88 putative TA systems) [21], and genomic regions containing VapBC are strongly associated with virulence and pathogenicity factors [22]. The identified CDSs are part of a type-II TA system related to mRNA translation regulation. More specifically, VapC25 and VapC29 are ribonucleases shown to inhibit *M. smegmatis* cell growth when induced (i.e. have bacteriostatic effects) [23]. Taken together, these results suggest that MP1 and MP2 may have different survival capacities.

Two distinct spoligotypes, SB0155 and SB2455, and two distinct Mycobacterial Interspersed Repetitive Unit-Variable Number of Tandem Repeat (MIRU-VNTR) profiles were identified in *M. pinnipedii* MP1 and *M.*

pinnipedii MP2, respectively (Additional file 2: Figures S1 and S2). The *in silico* predicted MIRU-VNTR matched the pattern identified using PCR. The *M. pinnipedii* strains MP1 and MP2 differed at MIRU04 and MIRU10 loci (Additional file 2: Figure S1). Variations at multiple loci between isolates of the same animal normally indicate mixed-strain infections [24]. In contrast to the spoligotype SB0155 [25], SB2455 profile has never been described in *M. pinnipedii* strains. Therefore, based on current parameters of distinction between microevolution and superinfection regarding SNPs/indels [26] and MIRU-VNTR [24], the sea lion analyzed herein was infected with two different strains of *M. pinnipedii*.

This is the first report of *M. pinnipedii* superinfection in a sea lion. As these animals normally live in dense colonies, they are highly vulnerable to epizootics of infectious diseases that can be transmitted by direct animal-to-animal contact. Superinfection caused by *M. tuberculosis* in humans is normally reported in highly endemic countries, in which people are exposed to multiple strains of *M. tuberculosis* throughout their lives and HIV plays an important role in shaping the disease incidence [24]. By tracing a parallel, it is possible that *M. pinnipedii* is highly endemic in the population from which this sea lion came, represented by different circulating strains, which may have unprecedented effects on the conservation of this species. Adverse environmental conditions, insufficient nutrition, and chronic stress due to disturbance or competition can act to suppress the immune system and, therefore, pinnipeds may be predisposed to diseases caused by several pathogens. In Brazil, sea lion stranding occurs normally due to compromised animal health when they travel long distances during the migratory period [27]. The weakened state in which these animals arrive at the coast, often infected by large bacterial loads, is the factor that most commonly leads to death [28, 29]. The contribution of *M. pinnipedii* infection in this context remains to be elucidated.

Although *M. pinnipedii* has been described in both captive and free-ranging animals, systematic population surveillance studies are still lacking. In 2011, a captive South American sea lion housed in a Czech Republic zoo that was imported from a German zoo (not specified) was found to be infected with *M. pinnipedii*. The animal's parents also died of tuberculosis in Germany and they were captured as juveniles from the coastal waters of Uruguay in 1992 [16], suggesting that the infection came from wildlife. Outbreaks of the disease in wild and/or captive-born South American sea lions housed at the Heidelberg Zoo, Germany and the Le Pal Zoo, France have also been described [30]. Added to this report of superinfection, it is thus clear that *M. pinnipedii* is endemic in wild populations of South American sea lions and is being introduced into zoos, warranting an urgent need to evaluate the extent of tuberculosis in free-ranging pinnipeds of the southern hemisphere.

Strains of M. pinnipedii have a highly conserved proteome

Figure 3 illustrates groups of orthologous proteins present in seven modern *M. pinnipedii* strains that were deposited in the National Center for Biotechnology Information (NCBI) (*M. pinnipedii* MP1, MP2, ATCC BAA-688, G01222, G01491, G01492, G01498). A total of 3,986 protein clusters (i.e. groups of ≥ 2 proteins) were identified among the analyzed strains, including 3,694 (92.67%) core protein clusters present in all seven genomes (Fig. 3A; Additional file 3: Table S2). According to COG classification, about a third (1,147/3,694; 31.05%) of the core protein clusters have unknown function, while the majority of these have functional annotation (2,547/3,694; 68.95%). The top five functional classes were: cell wall/membrane/envelope biogenesis (M; 426/3,694, 11.5%), lipid transport and metabolism (I; 284/3,694, 7.7%), energy production and

conversion (C; 280/3,694, 7.6%), transcription (K; 205/3,694, 5.5%) and amino acid transport and metabolism (E, 181/3,694, 4.9%) (Fig. 3B-C).

There are no protein clusters (i.e. groups of ≥ 2 proteins) unique to each *M. pinnipedii* strain (Fig. 3A). However, a number of single proteins not categorized into any protein cluster were considered unique to each *M. pinnipedii* strain (MP1 = 32, MP2 = 37, ATCC BAA-688 = 58, G01498 = 103, G01492 = 68, G01491 = 52, G01222 = 80). As the annotation tool used herein (RAST; see methods) does not differentiate between pseudogenes/broken genes and true genes, we used BLASTp and/or tBLASTn to search the NCBI's PGAP annotation of MP1, MP2 and ATCC BAA-688 strains using these “unique proteins” as input. From those with annotated function (i.e. not hypothetical or PE/PPE genes), our BLAST search revealed that all functional CDSs are considered pseudogenes, are non-existent or are present in contigs smaller than 700 bp (with doubtful sequencing/assembly quality) based on PGAP annotation (Additional file 5: Table S4). As RAST reports broken CDSs at the end of contigs as true genes, these CDS fragments (i.e. artefacts of draft genomes) failed to cluster into orthologous groups. Nevertheless, there is still a number of hypothetical and PE/PPE genes that we were not able to distinguish between pseudogenes and true genes using the proposed tools, indicating the possibility of distinct phenotypes within *M. pinnipedii* strains. These variations may be a consequence of gene truncation or pseudogenization, which may lead to neofunctionalization or gene loss, respectively, as well as duplication events, phenomena commonly observed in the MTBC [31, 32].

When considering non-core protein clusters (n = 292) shared between two and up to six *M. pinnipedii* strains, similar results were obtained, with all functional CDSs identified as pseudogenes or non-existent based on PGAP annotation, and an unknown number of possible hypothetical and PE/PPE genes that may be variably present among *M. pinnipedii* strains. Taken together, our results indicate that the analyzed *M. pinnipedii* strains present a high level of proteome conservation, which is in contrast with a recent pangenome analysis of *M. tuberculosis* strains that detected at least 1,122 CDSs in the accessory genome and 964 strain-specific CDSs in 36 genomes [33]. However, the *M. tuberculosis* pangenome in that study was also analyzed using RAST, and pseudogenes were not taken into consideration.

Modern M. pinnipedii strains have unique deletion markers

In the large sequence polymorphism (LSP) analysis, eight previously described deletions (RD2seal, MiD3, MiD4, RD3, RD7, RD8, RD9, RD10) were found in *M. pinnipedii* strains (Fig. 4). Although the genome coverage of the ancient *M. pinnipedii* reads in respect to *M. tuberculosis* H37Rv varied from 31.94% to 47.83% (after quality trimming), the deletions MiD3 and MiD4 were not found in the ancient strains of *M. pinnipedii*, while present in the eight analyzed modern strains of the same species. This finding contributes to the understanding of *M. pinnipedii* evolution over time, indicating an active process of genome reduction occurring for the past one thousand years.

We have also detected 19 regions with ambiguous read mapping (i.e. areas with low sequencing coverage composed of mapped reads with non-specific match) ranging from 548 bp to 5,329 bp (Additional file 6: Table S5). The vast majority (16/19; 84.21%) are associated with at least one PE/PPE gene, while the three remaining regions comprise a hypothetical protein with an oxireductase (Rv3530c-Rv3531c) and two transposases (Rv0797 and Rv3023c). Given the repetitive nature of these regions and possible sequencing

coverage bias, these findings should be interpreted with caution. Future validation using PCR assays are strongly indicated. The PE/PPE gene families correspond to approximately 10% of the coding capacity of the *M. tuberculosis* genome [34] and show a high degree of variation across members of the MTBC and between strains of the same species.

The finding of MiD3 and MiD4 being present only in modern strains of *M. pinnipedii* suggests that these deletions may have occurred independently, twice, over its evolutionary course (Fig. 5). MiD3 and MiD4 were first identified in *Mycobacterium microti* [35, 36]. This pathogen was initially described in voles [37], but has since been isolated from pigs, llamas, cats, wild boars, and immunosuppressed humans, being restricted to Eurasia. This raises questions about the true host range of these closely related pathogens, which may or may not have changed over the past one thousand years.

Two distinct clusters of modern M. pinnipedii strains

Figure 5 illustrates the SNP-based phylogenetic tree of the MTBC including eight modern and three ancient *M. pinnipedii* strains. A total of 1,698 polymorphic positions were found to be unique to *M. pinnipedii* strains, i.e. not present in any other of the MTBC species included in this study (Additional file 7:Table S6). This finding is not comprehensive of all possible MTBC strains distributed worldwide, which means that different strains not analyzed herein may also have mutations in these same polymorphic sites. In addition, strains of *M. pinnipedii* appeared distributed among three main clusters, with the ancient strains emerging from the most basal node (Fig. 5). Interestingly, modern *M. pinnipedii* strains appeared divided into two groups according to geographic locations, and possibly host species: modern cluster 1, comprising isolates of South American origin, and modern cluster 2, comprising isolates from Australia. These results are in accordance with findings of *M. tuberculosis* and *M. bovis* lineages/clonal complexes that are also associated with distinct geographic locations [32, 38, 39]. Whether or not these are also associated with different virulence phenotypes is unknown.

Conclusion

This is the first report of mixed-strain infection of *M. pinnipedii* in pinnipeds. This finding, coupled with previous reports in the literature, suggest that *M. pinnipedii* infection is endemic in free-ranging populations of South American sea lions. Genetic differences between these strains were found to be associated with virulence factors and enzymes necessary for intracellular maintenance and membrane-shaping of mycobacteria. The actual effect of these phenomena for the disease outcome and conservation of the animal species, and potential population-level implications are unknown. Nevertheless, this prompts an urgent need to evaluate the extent of the disease in these animals and to consider tuberculosis in pinnipeds as one of the main health concerns for these species. Investigations of the infection and clinical disease should be conducted when introducing animals into zoo facilities and studies involving free-ranging populations are encouraged.

As with other species of MTBC, *M. pinnipedii* genomes are under evolutionary decay through the loss of specific genomic regions. These results were further supported by the finding of LSPs not present in the ancient genomes of *M. pinnipedii* compared to modern *M. pinnipedii* strains. The genome remodeling of *M.*

pinnipedii affects well-known MTBC virulence factors, with potential impact on host adaptability and disease outcome, such as the deletion of MiD3 and MiD4 and possible regions containing PE/PPE genes. However, the proteome of the studied strains showed high degree of conservation, indicating that further genomes need to be sequenced and analyzed to increase the chance of finding significant differences or that *M. pinnipedii* genome remodeling occurred prior to bacterial speciation. In addition, particular attention should be given to possible “pseudogenes” and/or truncated CDSs, as to standardize genome annotations and guide future gene-based analyses. And finally, *M. pinnipedii* strains, as with other MTBC species, are likely to cluster based on geographic occurrence, which will coincide with pinniped species, as these animal populations are already segregated by geography.

Methods

Selection of genomes

Sequence reads from six modern *M. pinnipedii* strains [G01222 from Argentina; G01491, G01492, G01498 from Australia; and 7739 and 7011 from a zoo in Germany (SRR1239336, SRR1239337, SRR1239338, SRR1239339, SRR1239341, and SRR1239340)] and three ancient *M. pinnipedii* strains from 1,000-year-old Peruvian mummies were selected for this study [58, 54 and 64 (SRR1238557, SRR1238558, and SRR1238559)] [5]. All read sets were retrieved from the Sequence Read Archive (SRA) of the NCBI. Assembled contigs of *M. pinnipedii* strain ATCC BAA-688 (NZ_MWXB01000000) from Australia deposited in RefSeq were also selected [20]. Read sets from this ATCC strain (BAA-688) are not deposited in public databases and were not included in any analysis requiring reads. Also, *Mycobacterium pinnipedii* strains 7739 and 7011 had only single read sets publicly available, and these were not appropriate for genome assembly and not included in analyses requiring assemblies (Table 1). Whole genome of *M. tuberculosis* H37Rv (NC_000962.3) was used as reference. For the phylogenomic analysis, reads of MTBC representatives [chimpanzee bacillus (ERR150046), dassie bacillus (SRR3745458), *M. africanum* GM041182 (ERR234255), *M. africanum* MAL010070 (SRR998578), *M. bovis* SP38 (SRR6705904), *M. caprae* EPDC02 (DRR120409), *M. microti* 94-2272 (ERR027298), *M. mungi* BM22813 (SRR3500411), *M. orygis* IDR1100020842 (SRR5642712), *M. suricattae* ERS798580 (ERR970409)] were also included.

Isolation and bacterial DNA extraction

Two *M. pinnipedii* isolates (MP1 and MP2) obtained from the lung and a mesenteric lymph node of a South American sea lion (*Otaria flavescens*) found dead in Capão da Canoa, Rio Grande do Sul, Brazil [19] were kept at -80ºC in 7H9 broth with 20% glycerol. For this study, both isolates were reactivated in Stonebrink media and a single colony from each isolate was sub-cultured for DNA extraction as described previously [40, 41]. The quality and quantity of the DNA was measured by Nanodrop 2000c (Thermo Fisher Scientific, Waltham, MA, USA) following the absorbance ratios of 260/280 and 260/230 and in 0.8% agarose gel with a mass ladder. A final analysis was performed with an Agilent 2100 Bioanalyzer High Sensitivity Chip DNA (Agilent Technologies, Santa. Clara, CA, USA) for sample concentration and fragmentation. All procedures involving infectious material were performed in a Biosafety Level 3+ Laboratory (BSL-3+ Prof. Dr. Klaus Eberhard Stewien) located at the Department of Microbiology, Institute of Biomedical Sciences, University of São Paulo, Brazil.

Sequencing of M. pinnipedii isolates

A paired-end genomic library was constructed using a TruSeq DNA PCR-free sample preparation kit (Illumina, San Diego, CA, USA) according to the manufacturer's instructions. HiSeq2500 with Illumina v3 chemistry was used to sequence the genomic library (100 bp). These procedures were performed at the Central Laboratory of High Performance Technologies in Life Sciences (LaCTAD), State University of Campinas (UNICAMP), Campinas, Brazil. Illumina sequencing reads were deposited in the SRA, NCBI under accession numbers: SRR7693584 and SRR7693090.

Assembly and annotation of the genomes

The 100 bp paired-end reads from each library were first filtered by quality and presence of adaptors using the Trimmomatic software version 0.36 [42]. *Mycobacterium pinnipedii* MP1 and MP2 genomes were *de novo* assembled using the SPAdes software version 3.13.0 [43]. Gene identification and annotation were performed automatically by the NCBI's PGAP and draft genomes deposited as GCA_003027795.2 (MP1) and GCA_003027895.2 (MP2).

Circular map of M. pinnipedii genomes

In order to visualize the similarity among the predicted proteins of different *M. pinnipedii* strains, the genomes of strains G01222, G01491, G01492 and G01498 were assembled using CLC Genomics WorkBench 11 (Qiagen, Netherlands) and SPAdes software version 3.13.0 [43]. The best assemblies based on the number of contigs and N50 were selected (Table 1). Assembled contigs were then annotated with RAST version 2.0 [44]. Resulting contigs from strains G01222, G01491, G01492 and G01498 and contigs from *M. pinnipedii* strains MP1, MP2, and ATCC BAA-688 deposited in GenBank were reordered with MAUVE [45] using *M. tuberculosis* H37Rv (GenBank NC000962.3) as the reference genome. Subsequently, a circular genomic map was constructed using the Proteome Comparison toll implemented in Patric 3.5.18 web resources [46].

Mapping of reads and variant calling

The Burrows-Wheeler Aligner (BWA) software version 0.7.17 – r1188 [47] was used to map the reads of *M. pinnipedii* MP1 and MP2 against the reference genome *M. tuberculosis* H37Rv, separately, and the reads of *M. pinnipedii* MP2 against the assembled contigs of *M. pinnipedii* MP1. The BWA outputs were analyzed with SAMtools version 1.9 [48] to sort and index .bam files, according to filters of mapping quality of Phred's scale 20 and removal of duplicated reads. Detection of SNPs and insertions and deletions (indels) were performed with FreeBayes version 1.3.1-1 [49] with the following parameters: minimum mapping quality of 10, minimum base quality at a position of 10, minimum read depth at a position of 5 and without strand bias. The maximum indel (insertion or deletion) size reported by FreeBayes was 36 bp. Identified SNPs and indels were then annotated using SNPEff version 4.0 [50]. Variants annotated in regions related to transposable elements, 13E12 family, phages or PE/PPE family of proteins were removed to avoid false positives.

To identify unique SNPs of *M. pinnipedii*, reads of MTBC representatives (chimpanzee bacillus, dassie bacillus, *M. africanum* GM041182, *M. africanum* MAL010070, *M. bovis*, *M. caprae*, *M. microti*, *M. mungi*, *M. orygis*, *M. suricattae*) were included in the analysis and mapped against *M. tuberculosis* H37Rv. The variants present only in the 11 strains of *M. pinnipedii* were considered unique to these bacterial species.

Protein interaction networks and GO enrichment analysis

CDSs identified with non-synonymous mutations comparing MP1 and MP2 strains were analyzed using STRING database version 11.0 with default settings for the prediction of network associations between proteins [51]. Results were used to identify specific metabolic pathways or protein interactions under evolutionary pressure. We also performed a GO (Gene Ontology) enrichment analysis (biological processes, molecular function and cellular component) of the same CDSs using PANTHER version 14 according to *M. tuberculosis* H37Rv reference annotation [52]. Results were considered significant when P value ≤ 0.05. When needed, protein sequences were searched against the Virulence Factor Database (VFDB) [53] to identify virulence factors.

In silico spoligotyping and Mycobacterial Interspersed Repetitive Unit-Variable Number of Tandem Repeat (MIRU-VNTR)

For the identification of spoligotypes, the reads of *M. pinnipedii* isolates were analyzed in SpoTyping version 2.0 [54]. The identified patterns were submitted to the Spoligotyping database of *M. bovis* (www.mbovis.org) for the identification of the pattern number. In order to identify the variable number tandem repeats (VNTRs) of genetic elements named mycobacterial interspersed repetitive units (MIRUs), the 24 MIRU-VNTR loci were individually identified *in silico* using previously described primers [55] and the resulting patterns were analyzed in the MIRU-VNTRplus database [56, 57]. We also performed 24-loci MIRU-VNTR PCR using extracted DNA from *M. pinnipedii* MP1 and *M. pinnipedii* MP2 as previously described [55]. PCR products were separated by electrophoresis in a 1.5% agarose gel stained with SYBR Safe DNA Gel Stain (Thermo Fisher Scientific) and visualized under ultraviolet light.

Mycobaterium pinnipedii orthologous and paralogous protein clusters

To identify groups of orthologous genes, seven *M. pinnipedii* strains were used (G01222, G01491, G01492, G01498, MP1, MP2 and ATCC BAA-688). The *M. pinnipedii* strains G01498, G01492, G01491 and G01222 are not available as assembled genomes in GenBank, which precluded PGAP annotation. As to avoid biases in orthology clustering due to discrepancies in annotation from different platforms, we then annotated or re-annotated all genomes analyzed herein using RAST version 2.0 [44]. Predicted CDS $\leq$150 bp were excluded. Proteins were then clustered using OrthoMCL [58] available from the KBase platform [59]. Briefly, homologous pairs of sequences were found using the all-against-all BLASTp algorithm with an E-value <1e-5. OrthoMCL then converted the BLASTp results into a normalized similarity matrix that was analyzed by a Markov Cluster algorithm (MCL) for clustering of orthologous sequences. The inflation index of 1.5 was used to regulate cluster tightness. Since the RAST algorithm reports pseudogenes and broken genes located at the end of contigs as true genes, the number of strain-specific CDSs may be overestimated. This overestimation was manually checked using BLASTp and/or tBLASTn [60] by comparing obtained results with the

annotation of *M. pinnipedii* strains MP1, MP2 and ATCC BAA-688, which are deposited in Genbank and have been annotated with PGAP from NCBI. Groups of orthologous proteins were classified with eggNOG [61] using a graph-based unsupervised clustering algorithm extending the COG methodology [62].

Large sequence polymorphisms

LSPs were detected with CLC Genomics Workbench 11 (QIAGEN, Venlo, Netherlands) by mapping reads of ancient *M. pinnipedii* (samples 54, 58, 64) and modern *M. pinnipedii* (G01222, G01491, G01492, G01498, 7739, 7011, MP1, MP2) strains against *M. tuberculosis* H37Rv. When mapping the ancient samples against the reference genome, special treatment was given to the reads prior to mapping. Specifically, reads were trimmed based on quality parameters of 0.05 base-calling error probability and presence of two or more ambiguous nucleotides at the ends using modified-Mott trimming algorithm. Reads were mapped against the reference genome in two ways: by excluding or not non-specific matches. A non-specific match is given when a “read aligns at more than one position of the genome with an equally good score”. Obtained results were subsequently compared and contrasted. In both scenarios, mapped regions with less than 10 aligned reads and/or containing low mapping quality reads were considered absent in the query genome. LSPs were called when this region spanned more than 500 bp.

Phylogenetic analysis

Members of MTBC and eleven of the *M. pinnipedii* strains described above were used to construct a phylogenetic tree. Briefly, all read sets were mapped against the reference *M. tuberculosis* H37Rv genome and true variants were identified as described above, with exclusion of SNPs located in repetitive regions (transposable elements, 13E12 family, phages or PE/PPE family of proteins). A customized script in Python language (software version 3.6.3) was then used to build a positional matrix of SNPs identified in genomes of all strains/species. The matrix was used as input for the RAxML software version 7.3 [63] for the construction of the phylogenetic tree using Maximum Likelihood (ML) algorithm with GTRCAT model and autoMRE for best-scoring ML tree and a maximum of 1,000 bootstrap inferences.

Abbreviations

MTBC: *Mycobacterium tuberculosis* Complex; SNPs: Single nucleotide polymorphisms; RDs: Regions of differences; PGAP: Prokaryotic Genome Annotation Pipeline; CDSs: Coding Sequences; COGs: Clusters of Orthologous Groups; MIRU-VNTR: Mycobacterial Interspersed Repetitive Unit-Variable Number of Tandem Repeat;

Indels: Insertions and/or deletions; NCBI: National Center for Biotechnology Information; PE/PPE: Proline-Glutamate and Proline-Proline-Glutamate; LSP: Large Sequence Polymorphisms; SRA: Sequence Read Archive; TA: toxin-antitoxin.

Declarations

Availability of data and material

The datasets generated during the current study are available in RefSeq and SRA: https://www.ncbi.nlm.nih.gov/assembly/GCA_003027895.2,

https://www.ncbi.nlm.nih.gov/assembly/GCA_003027795.2,

https://www.ncbi.nlm.nih.gov/sra/?term=SRR7693584,

https://www.ncbi.nlm.nih.gov/sra/?term=SRR7693090.

Competing Interests

The authors declare that they have no competing interests.

Authors Contribution

Conceived and designed experiments: TTSP, CI, CKZ, NC, AF, JSFN, MBH, AG. Analyzed the data: TTSP, CKZ, NC, AG. Contributed reagents/materials/analysis tools: JSFN, MBH, AG. Wrote the paper: TTSP, CKZ, AG. Approved the paper: TTSP, CI, CKZ, NC, AF, JSFN, MBH, AG. All authors have read and approved the manuscript.

Acknowledgements

The authors are in debt to Gisele Oliveira de Souza and Carolina Bertelli de Souza Ferreira from the University of São Paulo, São Paulo, Brazil for invaluable technical assistance. We also thank the LaCTAD, UNICAMP, Campinas, Brazil for aiding in the genome sequencing of *M. pinnipedii* isolates, and CEFAP, University of São Paulo, for computer core services. We are in debt to Dr Paulo Eduardo Brandão for continuous mentoring support throughout this study and for the availability of the CLC Genomics WorkBench software.

Funding

Fellowships for TTPS, CKZ, NCSC, AFSF, CYI, MBH were provided by CNPq (Conselho Nacional de Pesquisa Científica), Ministry of Science, Brazil (134266/2017-0; 140003/2019-3), CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), Ministry of Education, Brazil (1539669) and São Paulo Research Foundation (FAPESP) (2017/04617-3; 2017/20147-7). This study was financed in part by CAPES (Finance Code 001 and 1841/2016). The main research funding was provided by Morris Animal Foundation (D17ZO-307). Jim Hesson revised the manuscript (https://www.academicenglishsolutions.com). Funding agencies had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

References

1. Cousins D V, Peet RL, Gaynor WT, Williams SN, Gow BL. Tuberculosis in imported hyrax (*Procavia capensis*) caused by an unusual variant belonging to the *Mycobacterium tuberculosis* complex. Vet Microbiol. 1994;42:135–45.
2. Cousins D V., Bastida R, Cataldi A, Quse V, Redrobe S, Dow S, et al. Tuberculosis in seals caused by a novel member of the *Mycobacterium tuberculosis* complex: *Mycobacterium pinnipedii* sp. nov. Int J Syst Evol Microbiol. 2003;53:1305–14.
3. Coscolla M, Gagneux S. Consequences of genomic diversity in *Mycobacterium tuberculosis*. Semin Immunol. 2014;26:431–44.
4. Brites D, Gagneux S. Co-evolution of *Mycobacterium tuberculosis* and Homo sapiens. Immunol Rev. 2015;264:6–24.
5. Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I, et al. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. Nature. 2014;514:494–7.
6. Blair WR. Report of the veterinarian. 1913.
7. Cousins D V, Francis BR, Gow BL, Collins DM, McGlashan CH, Gregory A, et al. Tuberculosis in captive seals: bacteriological studies on an isolate belonging to the *Mycobacterium tuberculosis* complex. Res Vet Sci. 1990;48:196–200.
8. Cousins D V., Williams SN, Reuter R, Forshaw D, Chadwick B, Coughran D, et al. Tuberculosis in wild seals and characterisation of the seal bacillus. Aust Vet J. 1993;70:92–7.
9. Cousins D V. ELISA for detection of tuberculosis in seals. Vet Rec. 1987;121:305.
10. Bastida R, Loureiro J, Quse V, Bernardelli A, Rodríguez D, Costa E. Tuberculosis in a Wild Subantarctic Fur Seal from Argentina. J Wildl Dis.1999;35:796–8.
11. Forshaw D, Phelps GR. Tuberculosis in a captive colony of pinnipeds. J Wildl Dis. 1991;27:288–95.
12. Bernardelli Bastida AR, Loureiro J, Michelis H, Romano M, Cataldi A, Costa E, et al. Tuberculosis in sea lions and fur seals from the south-western Atlantic coast. Rev Sci Tech. 1996;15:985-1005.
13. Kiers A, Klarenbeek A, Mendelts B, Van Soolingen D, Koëter G. Transmission of Mycobacterium pinnipedii to humans in a zoo with marine mammals. Int J Tuberc Lung Dis. 2008;12:1469–73.
14. Id WDR, Lenting B, Kokosinska A, Hunter S, Duignan J, Gartrell B, et al. Pathology and molecular epidemiology of *Mycobacterium pinnipedii* tuberculosis in native New Zealand marine mammals. PLoS One. 2019; 14:e0212363 .
15. Thorel MF, Karoui C, Varnerot A, Fleury C, Vincent V. Isolation of *Mycobacterium bovis* from baboons, leopards and a sea-lion. Vet Res. 1998;29:207–12.
16. Kriz P, Kralik P, Slany M, Slana I, Svobodova J, Parmova I, et al. *Mycobacterium pinnipedii* in a captive Southern sea lion (*Otaria flavescens*): a case report. Veterinární medicína. 2011;56:307-313.
17. Boardman WSJ, Shephard L, Bastian I, Globan M, Fyfe JAM, Cousins D V, et al. *Mycobacterium pinnipedii* tuberculosis in a free-ranging australian fur seal (*Arctocephalus pusillus doriferus*) in South Australia. J Zoo Wildl Med. 2014;45:970–2.

18. Loeffler SH, de Lisle GW, Neill MA, Collins DM, Price-Carter M, Paterson B, et al. The seal tuberculosis agent, *Mycobacterium pinnipedii*, infects domestic cattle in New Zealand: epidemiologic factors and DNA strain typing. J Wildl Dis. 2014;50:180–7.
19. de Amorim DB, Casagrande RA, Alievi MM, Wouters F, De Oliveira LGS, Driemeier D, et al. *Mycobacterium pinnipedii* in a Stranded South American Sea Lion (*Otaria byronia*) in Brazil. J Wildl Dis. 2014;50:419–22.
20. Riojas MA, McGough KJ, Rider-Riojas CJ, Rastogi N, Hazbón MH. Phylogenomic analysis of the species of the *Mycobacterium tuberculosis* complex demonstrates that *Mycobacterium africanum*, *Mycobacterium bovis*, *Mycobacterium caprae*, *Mycobacterium microti* and *Mycobacterium pinnipedii* are later heterotypic synonyms of Mycob. Int J Syst Evol Microbiol. 2017;68:324–32.
21. Ramage HR, Connolly LE, Cox JS. Comprehensive Functional Analysis of *Mycobacterium tuberculosis* Toxin-Antitoxin Systems: Implications for Pathogenesis, Stress Responses, and Evolution. PLoS Genet. 2009;5:e1000767.
22. Arcus VL, Rainey PB, Turner SJ. The PIN-domain toxin–antitoxin array in mycobacteria. Trends Microbiol. 2005;13:360–5.
23. Winther K, Tree JJ, Tollervey D, Gerdes K. VapCs of *Mycobacterium tuberculosis* cleave RNAs essential for translation. Nucleic Acids Res. 2016;44:9860–71.
24. Cohen T, van Helden PD, Wilson D, Colijn C, McLaughlin MM, Abubakar I, et al. Mixed-Strain *Mycobacterium tuberculosis* Infections and the Implications for Tuberculosis Treatment and Control. Clin Microbiol Rev. 2012;25:708–19.
25. Cousins D V., Bastida R, Cataldi A, Quse V, Redrobe S, Dow S, et al. Tuberculosis in seals caused by a novel member of the *Mycobacterium tuberculosis* complex: *Mycobacterium pinnipedii* sp. nov. Int J Syst Evol Microbiol. 2003;53:1305–14.
26. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. Lancet Infect Dis. 2013;13:137–46.
27. Ruoppolo V. Comparative pathology of cetaceans and pinnipeds. Digital Library of Theses and Dissertations of the University of São Paulo; 2003.
28. Higgins R. Bacteria and fungi of marine mammals: a review. Can Vet J. 2000;41:105–16.
29. Gonzales-Viera O, Marigo J, Ruoppolo V, Rosas FCW, Kanamura CT, Takakura C, et al. Toxoplasmosis in a Guiana dolphin (*Sotalia guianensis*) from Paraná, Brazil. Vet Parasitol. 2013;191:358–62.
30. Jurczynski K, Lyashchenko KP, Gomis D, Moser I, Greenwald R, Moisson P. Pinniped Tuberculosis in Malayan Tapirs (*Tapirus indicus*) and its Transmission to Other Terrestrial Mammals. J Zoo Wildl Med. 2011;42:222–7.
31. Bolotin E, Hershberg R. Gene Loss Dominates As a Source of Genetic Variation within Clonal Pathogenic Bacterial Species. Genome Biol Evol. 2015;7:2173–87.
32. Zimpel CK, Brandão PE, de Souza Filho AF, de Souza RF, Ikuta CY, Ferreira Neto JS, et al. Complete Genome Sequencing of *Mycobacterium bovis* SP38 and Comparative Genomics of *Mycobacterium bovis* and *M. tuberculosis* Strains. Front Microbiol. 2017;8:2389.

33. Yang T, Sheng Y, Yue L, Ding N, Wang G, Zhong J, et al. Pan-Genomic Study of *Mycobacterium tuberculosis* Reflecting the Primary/Secondary Genes, Generality/Individuality, and the Interconversion Through Copy Number Variations. Front Microbiol. 2018;9:1886.
34. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature. 1998;393:537–44.
35. Brodin P, Eiglmeier K, Marmiesse M, Billault A, Garnier T, Niemann S, et al. Bacterial Artificial Chromosome-Based Comparative Genomic Analysis Identi es. Infect Immun. 2002;70:5568–78.
36. Garcia-Pelayo MC, Caimi KC, Inwald JK, Hinds J, Bigi F, Romano MI, et al. Microarray analysis of *Mycobacterium microti* reveals deletion of genes encoding PE-PPE proteins and ESAT-6 family antigens. Tuberculosis. 2004;84:159–66.
37. Wells AQ, Oxon DM. Tuberculosis in Wild Voles. Lancet. 1937;229:1221.
38. Filliol I, Motiwala AS, Cavatore M, Qi W, Hazbón MH, Bobadilla del Valle M, et al. Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. J Bacteriol. 2006;188:759–72.
39. Gagneux S. Host-pathogen coevolution in human tuberculosis. Philos Trans R Soc B Biol Sci. 2012;367:850–9.
40. van Soolingen D, de Haas PE, Haagsma J, Eger T, Hermans PW, Ritacco V, et al. Use of various genetic markers in differentiation of *Mycobacterium bovis* strains from animals and humans and for studying epidemiology of bovine tuberculosis. J Clin Microbiol. 1994;32:2425–33.
41. Bemer-Melchior P, Drugeon HB. Inactivation of *Mycobacterium tuberculosis* for DNA typing analysis. J Clin Microbiol. 1999;37:2350–1.
42. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
43. 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;19:455–77.
44. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics. 2008;9:75.
45. Rissman AI, Mau B, Biehl BS, Darling AE, Glasner JD, Perna NT. Reordering contigs of draft genomes using the Mauve Aligner. Bioinformatics. 2009;25:2071–3.
46. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, et al. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. Nucleic Acids Res. 2017;45:D535–42.
47. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
48. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
49. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. 2012. arXiv:1207.3907.

50. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin). 2012;6:80–92.
51. Jensen LJ, Simonovic M, Bork P, von Mering C, Muller J, Stark M, et al. STRING 8–a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res. 2008;37:D412–6.
52. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. 2019;47:D419–26.
53. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, et al. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res. 2004;33:D325–8.
54. Xia E, Teo Y-Y, Ong RT-H. SpoTyping: fast and accurate in silico *Mycobacterium* spoligotyping from sequence reads. Genome Med. 2016;8:19.
55. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rüsch-Gerdes S, Willery E, et al. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. J Clin Microbiol. 2006;44:4498–510.
56. Allix-Béguec C, Harmsen D, Weniger T, Supply P, Niemann S. Evaluation and strategy for use of MIRU-VNTRplus, a multifunctional database for online analysis of genotyping data and phylogenetic identification of *Mycobacterium tuberculosis* complex isolates. J Clin Microbiol. 2008;46:2692–9.
57. Weniger T, Krawczyk J, Supply P, Niemann S, Harmsen D. MIRU-VNTRplus: a web tool for polyphasic genotyping of *Mycobacterium tuberculosis* complex bacteria. Nucleic Acids Res. 2010;38 Web Server:W326–31.
58. Li L, Stoeckert CJ, Roos DS. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13:2178–89.
59. Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, et al. KBase: The United States Department of Energy Systems Biology Knowledgebase. Nat Biotechnol. 2018;36:566–9.
60. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
61. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 2016;44:D286-93.
62. Tatusov RL, Koonin E V, Lipman DJ. A genomic perspective on protein families. Science. 1997;278:631–7.
63. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.

Table

Table 1. Assembly statistics, genotyping and demographics of modern *Mycobacterium pinnipedii* strains used in this study.

Strains		Assembly statistics					Genotyping	Demographics	
	Available data	Coverage	Number of contigs	N50	Contig size (range)	Software	Spoligotypes	Source (species, country)[a]	Ref.
M. pinnipedii MP1	Paired-end reads and draft genome	228x	102	61,148	1,020-184,975	CLC Genomics Workbench 11	SB0155	*Otaria flavescens*, Brazil	This study
M. pinnipedii MP2	Paired-end reads and draft genome	220x	106	60,933	1,010-184,294	CLC Genomics Workbench 11	SB2455	*Otaria flavescens*, Brazil	This study
M. pinnipedii G01222	Paired-end reads	32x	138	38,346	1,020-121,089	CLC Genomics Workbench 11	SB0155	Argentina	[5]
M. pinnipedii G01491	Paired-end reads	242x	194	45,251	1,135-150.198	SPAdes 3.13.0	SB0155	Australia	[5]
M. pinnipedii G01492	Paired-end reads	484x	128	64,538	1,146-193.250	SPAdes 3.13.0	SB0155	Australia	[5]
M. pinnipedii G01498	Paired-end reads	113x	201	42,094	1,026-122,203	SPAdes 3.13.0	SB0155	Australia	[5]
M. pinnipedii 7739[b]	Single reads	86x	ND	ND	ND	ND	SB0155	Germany (zoo)	[5]
M. pinnipedii 7011[b]	Single reads	87x	ND	ND	ND	ND	SB0155	Germany (zoo)	[5]
M. pinnipedii ATCC BAA-688	Draft genome	NA	162	113,505	500-281,603	Velvet 1.2.10	Unknown[c]	Australia	[20]

ND: not done. NA: not available. [a]where animal species are not listed, it is because they were not informed by the authors. [b]Reads of *M. pinnipedii* 7739 and 7011 were not assembled because they were available as single reads. These reads were only used in the phylogenomic analysis. Spoligotypes were defined using SpoTyping [54]. [c]Spoligotype pattern not found in Mbovis.org database (Bincode: 000000100000000000000001010001000100000000). Sequencing coverage was calculated by the number of bases (after adaptors and quality trimming) divided by the average size of an MTBC genome (i.e. 4.3 Mb).

Figures

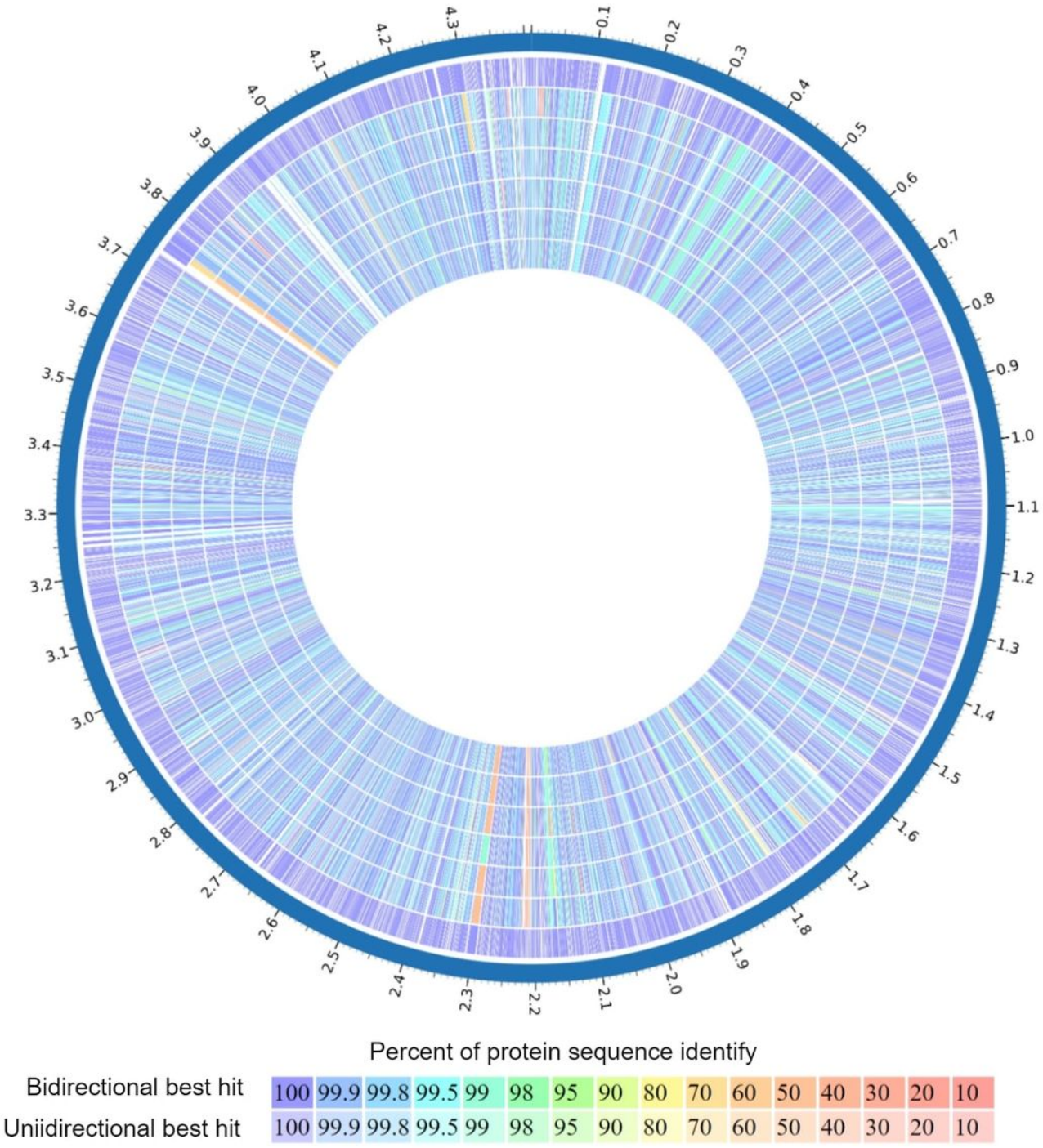

Figure 1

Circular map of proteome of Mycobacterium pinnipedii genomes compared to the reference genome of Mycobacterium tuberculosis H37Rv. From the outer ring to the inner ring: Genome position, M. tuberculosis H37Rv, M. pinnipedii MP1, M. pinnipedii MP2, M. pinnipedii G01222, M. pinnipedii G01491, M. pinnipedii G01492, M. pinnipedii G01498, M. pinnipedii ATCC BAA-688. Color scale represents protein identity. Contigs were ordered using MAUVE and map generated in Patric 3.5.18 web resources.

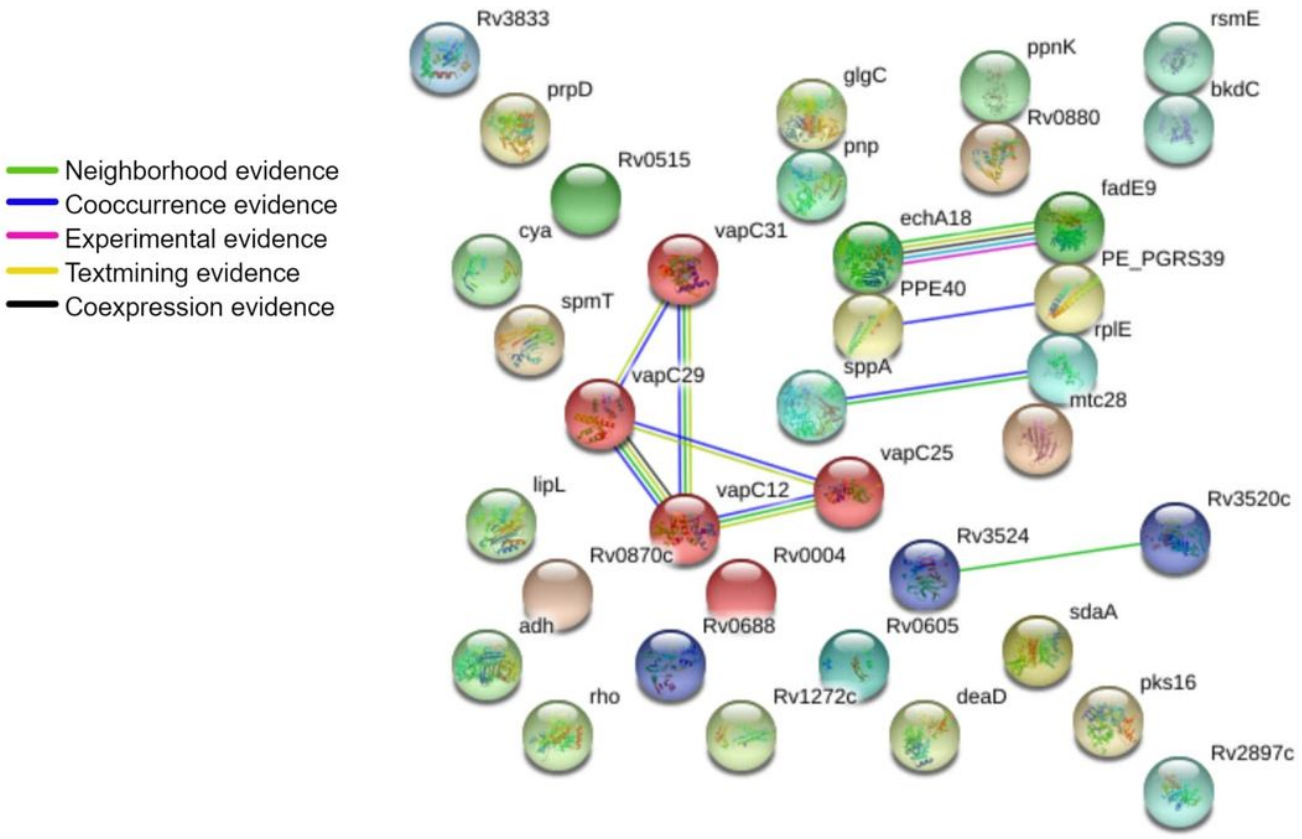

Figure 2

Protein interaction network of CDS (coding DNA sequences) with known function containing non-synonymous SNPs (single nucleotide polymorphisms). CDSs with mutations were identified between strains MP1 and MP2 of Mycobacterium pinnipedii. Network was generated using STRING database, with default parameters.

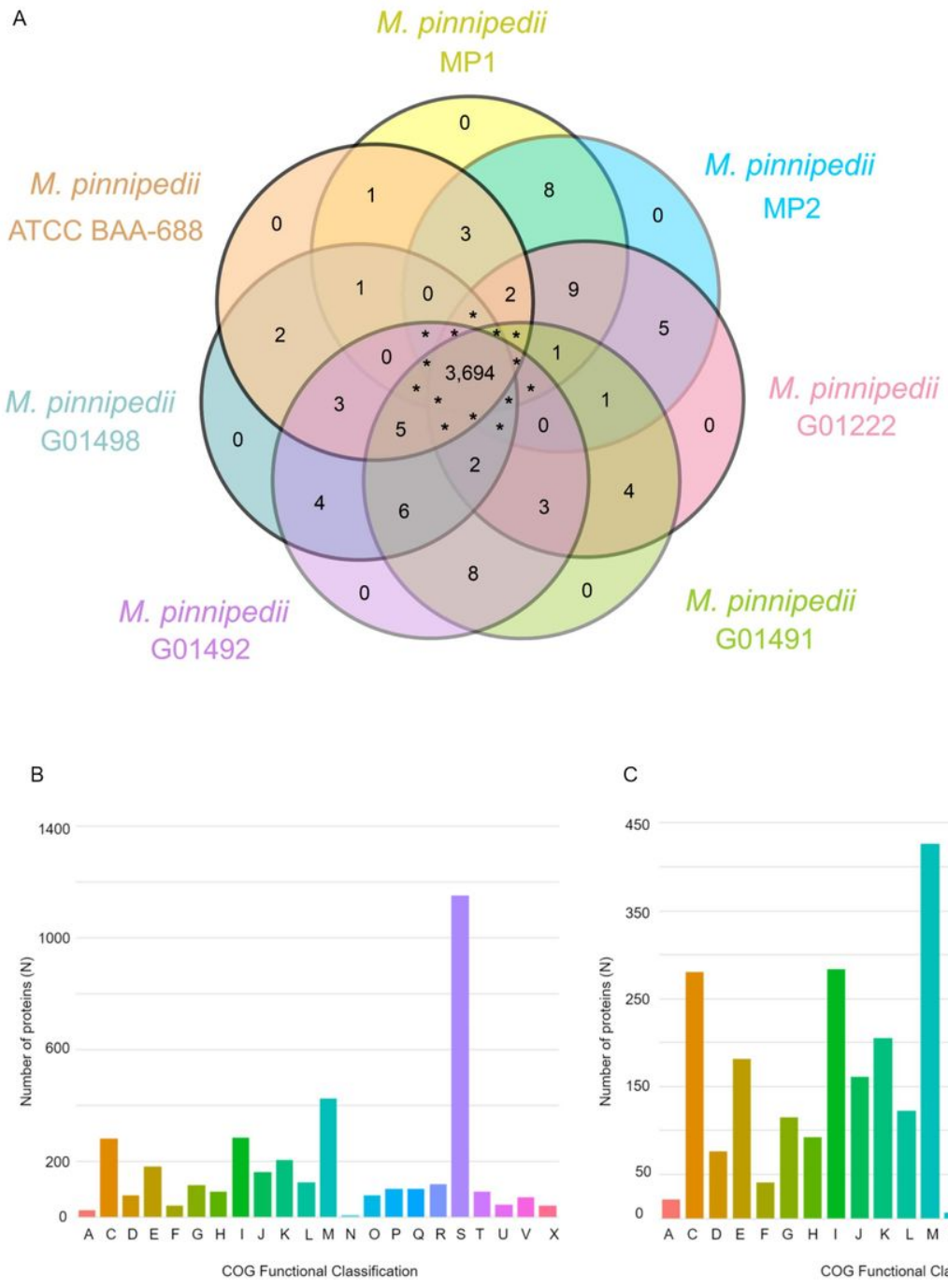

Figure 3

Clusters of orthologous proteins of Mycobacterium pinnipedii strains. (A) Venn diagram of orthologous proteins. (B) Cluster of Orthologous Groups (COG) classification based on function of the core protein clusters of the M. pinnipedii strains. (C) COG classification without unknown function class (S). Protein clustering was performed using OrthoMCL as available in Kbase plataform. *The number of orthologous gene proteins shared between these groups are available in Additional file 4: Table S3. Cluster size varied from two to forty-three proteins.

Start	End	Sample 58*	Sample 54*	Sample 64*	Strain G01222	Strain G01491	Strain G01492	Strain G01498	Strain 7011	Strain 7739	Strain MP1	Strain MP2	Affected Gene	RD
264755	266656	1901	1901	1901	1901	1901	1901	1901	1901	1901	1901	1901	Rv0221-echA1-Rv0223c	RD10
1779279	1789536	10257	10257	10257	10257	10257	10257	10257	10257	10257	10257	10257	Rv1573-Rv1588c (phage)	RD3
2208006	2220724	12718	12718	12718	12718	12718	12718	12718	12718	12718	12718	12718	yrbE3A/B-mce3A/B/C/D-lprM-mceF-Rv1972-Rv1977	RD7
2220948	2222975	2027	2027	2027	2027	2027	2027	2027	2027	2027	2027	2027	Rv1978-Rv1979c (permease)	RD2seal
2330074	2332101	2027	2027	2027	2027	2027	2027	2027	2027	2027	2027	2027	cobL-Rv2073c-Rv2075c	RD9
3377696	3380674	0	0	0	2978	2978	2978	2978	2978	2978	2978	2978	PPE46-esxR-esxS-PPE47-48	MiD4
3741146	3755776	0	0	0	14630	14630	14630	14630	14630	14630	14630	14630	PE_PGRS50-PPE55-Rv3348-Rv3349c	MiD3
4056841	4062730	5889	5889	5889	5889	5889	5889	5889	5889	5889	5889	5889	ephA-Rv3618-esxV-esxW-PPE65-PE32-lpqG	RD8

Size (bp) 0 1000 2000 3000 4000 5000 6000 7000 8000 9000 10000 11000 12000 13000 14000 15000

Figure 4

Large sequence polymorphisms (LSPs) of Mycobacterium pinnipedii strains. Values indicate number of nucleotides spanning each deleted region (zero means the region is not deleted in the corresponding bacterial strain). Start and end: nucleotide positions according to Mycobacterium tuberculosis H37Rv reference genome. Genes are annotated according to reference genome. *Ancient South American Mycobacterium pinnipedii strains. RD: regions of difference. All listed deletions have been already described in modern M. pinnipedii strains.

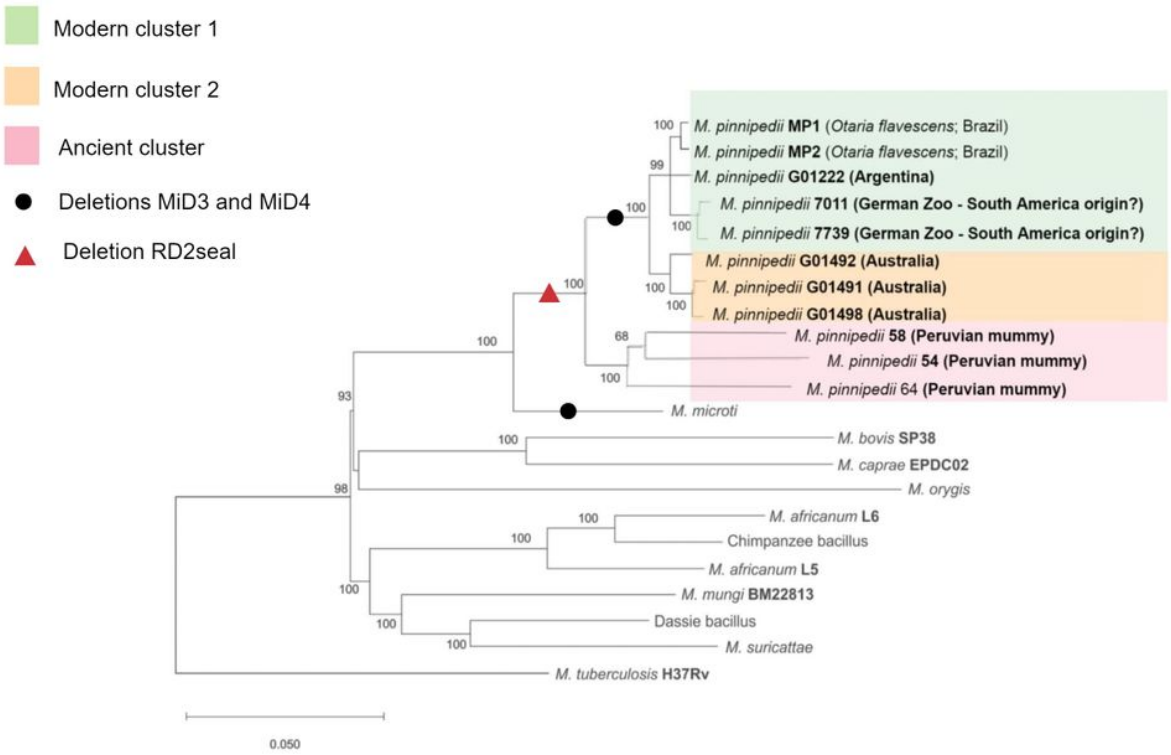

Figure 5

Phylogenetic tree based on SNPs (single nucleotide polymorphisms) of Mycobacterium tuberculosis Complex (MTBC) using Maximum Likelihood (ML) model. Green box: modern cluster 1; orange box: modern cluster 2; pink box: ancient cluster; black circle: deletions MiD3 and MiD4; red triangle: deletion RD2seal. Mycobacterium tuberculosis H37Rv was used as outgroup. A customized script in Python version 3.6.3 was used to build a matrix of SNPs, which was used to infer a ML tree using RAxML with GTRCAT model and

autoMRE for best-scoring ML tree and a maximum of 1,000 bootstrap inferences using RaxML software version 7.3.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile7TableS6.xlsx
- Additionalfile3TableS2.xlsx
- Additionalfile1TableS1.xlsx
- Additionalfile4TableS3.xlsx
- Additionalfile5TableS4.xlsx
- Additionalfile2FigureS1andS2.pdf
- Additionalfile6TableS5.xlsx