Whole Genome Sequencing based Identification of Human Tuberculosis caused by Animal-lineage *Mycobacterium orygis*

Md Rashedul Islam[a], Meenu K. Sharma[a,b], Rupinder KhunKhun[c], Cary Shandro[d], Inna Sekirov[c], Gregory J. Tyrrell[d], and Hafid Soualhine[a,b*]

[a]National Reference Centre for Mycobacteriology, National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg, Manitoba, Canada

[b]Department of Medical Microbiology and Infectious Diseases, Max Rady College of Medicine, University of Manitoba, Winnipeg, Manitoba, Canada

[c]BC Center for Disease Control Public Health Laboratory, Vancouver, British Columbia, Canada

[d]Provincial Laboratory for Public Health, Alberta Precision Labs. Edmonton, Alberta, Canada

[*]Corresponding author's contact information:

E-mail: Hafid.Soualhine@phac-aspc.gc.ca; Phone: 204.789.6051

National Reference Centre for Mycobacteriology, National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg, Manitoba, Canada

ABSTRACT A recently described member of the *Mycobacterium tuberculosis* complex (MTBC) is *Mycobacterium orygis,* which can cause disease primarily in animals but also in humans. Although *M. orygis* has been reported from different geographic regions around the world, due to a lack of proper identification techniques, the contribution of this emerging pathogen to the global burden of zoonotic tuberculosis is not fully understood. In the present work, we report single nucleotide polymorphisms (SNPs) analysis using whole genome sequencing (WGS) that can accurately identify *M. orygis* and differentiate it from other members of MTBC species. WGS-based SNPs analysis was performed for 61 isolates from different Provinces in Canada that were identified as *M. orygis.* A total of 56 *M. orygis* sequences from the public databases were also included in the analysis. Several unique SNPs in *gyrB, PPE55, Rv2042c, leuS, mmpL6,* and *mmpS6* genes were used to determine their effectiveness as genetic markers for the identification of *M. orygis.* To the best of our knowledge, five of these SNPs, *viz., gyrB*^{277} (A→G), *gyrB*^{1478} (T→C), *leuS*^{1064} (A→T), *mmpL6*^{486} (T→C), and *mmpS6*^{334} (C→G) are reported for the first time in this study. Our results also revealed several SNPs specific to other species within MTBC. The phylogenetic analysis shows that studied genomes were genetically diverse and clustered with *M. orygis* sequences of human and animal origin reported from different geographic locations. Therefore, the present study provides a new insight into the high confidence identification of *M. orygis* from MTBC species based on WGS data, which can be useful for reference and diagnostic laboratories.

KEYWORDS *Mycobacterium orygis, Mycobacterium tuberculosis* complex, Zoonotic tuberculosis, Whole genome sequencing, Single nucleotide polymorphism, Phylogeny

Tuberculosis is a severe and complex infectious disease caused by the *Mycobacterium tuberculosis* complex (MTBC) and remains a public health concern leading to the death of 1·6 million people per year worldwide, predominantly in low-income and middle income countries (1). *Mycobacterium orygis*, a species belonging to MTBC, was first described by van Ingen et al. (2). *M. orygis* is capable of causing infection in both animal and human hosts. This species has received considerable interest in recent years, and has been reported to be isolated from dairy cattle and captive monkeys (3), captive wild animals (4), deer (5), rhinoceros (6), black buck (5), and humans (2, 7, 8). *M. orygis* infection has been recognized as a zoonotic source of human tuberculosis (9). Moreover, in New Zealand, a presumptive transmission of *M. orygis* from human to animal was reported, with the original infection being mapped out from contacts with domestic animals in India (10). *M. orygis* is endemic in Southeast Asian countries including Bangladesh, India, Nepal, and Pakistan (3, 7, 10). Although tuberculosis incidence in Canada is relatively low (11), zoonotic tuberculosis is progressively being recognized as a significant menace to the public health. Hence, zoonotic tuberculosis could pose a significant challenge in controlling this disease and meeting global tuberculosis elimination goals (12).

Along with *M. orygis*, other phylogenetically related MTBC species are *M. tuberculosis*, *M. bovis* and its variant BCG, *M. africanum*, *M. caprae*, *M. pinnipedii*, *M. microti*, *M. canettii*, and members of animal adapted clade A1 such as "Dassie" bacillus, Chimpanzee bacillus, *M. suricattae*, and *M. mungi* (13-15). To improve human and animal health surveillance, it is important to implement proper identification methods and analysis tools for quickly and accurately discriminating MTBC species. Probes hybridization-based assays have been used to differentiate the causative agents of tuberculosis, however, studies have reported a limitation of this method in differentiating *M. orygis* from *M. africanum* (16). The mutations in genes

including *gyrB* and *Rv2042c*, and the regions of difference (RDs) deletion have been used to discriminate *M. orygis* from MTBC species, using PCR-based approaches (2, 3, 17). Since *M. orygis* strains share $gyrB^{1450}$ (G→T) mutation with MTBC members including *M. africanum* and *M. pinnipedii*, they may have previously been mislabeled as *M. africanum* (3, 17, 18).

Whole genome sequencing (WGS) technologies are now increasingly being used in clinical and research laboratories to investigate tuberculosis surveillance, outbreak detection, antimicrobial resistance prediction, characterization and diversity of MTBC species (14, 19). One of the key challenges for adopting WGS for these applications is data analysis that requires bioinformatics support and data interpretation (20). Consequently, a number of analytical tools have been developed to detect pathogenic bacterial strains using WGS data, with single nucleotide polymorphisms (SNP)-based methods being the common ones used in public health laboratories (21-23). BioHansel, for example, performs high-resolution genotyping by detecting phylogenetically informative SNPs in WGS data (24).

In order to improve our ability to accurately identify *M. orygis* and other species within MTBC, whole genome analysis was performed in this investigation. We report the identification of 61 new *M. orygis* from Canada by WGS-based SNP analysis, and to validate the results, publicly available 56 *M. orygis* genomes were added to the analysis. We performed molecular markers characterization on these genomes that demonstrate a clear differentiation of *M. orygis* from members of other MTBC species. Furthermore, the newly sequenced *M. orygis* isolates were phylogenetically analyzed to determine their diversity and global distribution. This study may improve our understanding of this poorly monitored emerging zoonotic pathogen and address its burden in Canada and globally.

MATERIALS AND METHODS

Sample collection, and project background. Cultures from across Canada were received at the National Reference Centre for Mycobacteriology (NRCM), National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg for *M. tuberculosis* testing between 2009 and 2022. The cultures were grown on mycobacteria growth indicator tube (MGIT) media and Middlebrook 7H10 plates using standard and aerobic growth conditions. At NRCM, the presence of the insertion sequence IS*6110* and region of difference RD9 was confirmed by RT-PCR analysis to detect MTBC and *M. tuberculosis*, respectively. Classical methods such as *gyrB*, RD1, RD4, RD7 and mycobacterial interspersed repetitive unit-variable number of tandem repeat (MIRU-VNTR) were also used for identification and genotyping purpose. These methods can distinguish between mostly isolated MTBC species, *M. tuberculosis*, *M. bovis* and its variant BCG, *M. caprae* and *M. africanum*, but they are unable to identify some rare species of the complex such as *M. orygis* and *M. pinnipedii*. Since 2018, our group started using routinely WGS technologies and BioHansel (24) program for identification and differentiation between MTBC species. BioHansel performs high-resolution genotyping of MTBC by detecting phylogenetically informative SNPs in WGS data. However, the SNPs in the *gyrB* gene (at positions 432, 513, 870, 1068, 1167, and 1207) currently incorporated in BioHansel cannot differentiate *M. orygis*, and the pipeline eventually identifies this species, with low confidence, as a member of an animal lineage of the MTBC (probable *M. tuberculosis* var. *orygis*) based on SNP typing (2). We then separated all isolates identified as animal lineage "probable *M. orygis*", all isolates with same MIRU-VNTR profile and *M. africanum* or *M. pinnipedii* or *M. bovis*/BCG from the inventory and investigated in this study. Therefore, a total of 137 isolates were interrogated (Supplementary Data S1), from which 61 were identified to be

M. orygis. Most of the isolates were from British Columbia 'BC' (n = 38), followed by Alberta 'AB' (n = 19), Manitoba 'MB' (n = 2), Nova Scotia 'NS' (n = 1), and Saskatchewan 'SK' (n = 1). The detailed information (*e.g.*, source Province, specimen type, and identification) on *M. orygis* isolates from this study is depicted in Table 1.

Whole genome sequencing (WGS). (i) DNA extraction, sequencing and data processing. Genomic DNA was extracted from the isolates using the InstaGene™ Matrix comprising 6% (*w/v*) Chelex resin for PCR-ready DNA purification (Bio-Rad #7326030; California, United States). DNA quality was checked by measuring the *A260/280* ratio using NanoDrop, and quantified fluorometrically using the Qubit 3.0 (Thermofisher, Massachusetts, United States). The sequencing library was prepared using the Illumina DNA Prep Kit (Illumina, California, United States), according to the manufacturer's protocol. Paired-end sequencing was conducted on an Illumina MiSeq platform (Illumina Inc.) using 300-cycle MiSeq Reagent Kit v2, 500-cycle MiSeq Reagent Kit v2 or 600-cycle MiSeq Reagent Kit v3 (Illumina, California, United States), creating 2 × 150, 2 × 250 or 2 × 300 bp paired-end reads, respectively.

Raw sequencing reads were processed in the IRIDA platform (25) (http://irida.ca) using Shovill for assembly (26), Prokka for annotation (27), and QUAST for assembly assessment (28). Sequencing Report workflow v.2.3, where FastQC (29) and SMALT (https://www.sanger.ac.uk/tool/smalt-0/) are the two main tools, was implemented in the Galaxy platform (30) for all studied isolates by aligning the reads against the reference genome of *M. tuberculosis* H37Rv (accession # NC_000962.3). Only sequences with at least 60% of mapped reads and a >28x average coverage were filtered for further analysis. While 54 out of 61 studied *M. orygis* sequences had mapped reads of over 80%, the remaining seven sequences were found

to have 60-80% of mapped reads, and only one sequence (2100725) had average coverage below 30x (28.85).

(ii) Genomic comparison of the new sequences with public database. The newly sequenced genomes were compared with 56 publicly available *M. orygis* genomes reported from around the world, downloaded from the National Center for Biotechnology Information (NCBI) GenBank (31), the NCBI Sequence Read Archive (32), and the European Nucleotide Archive (33). The list of reference *M. orygis* genomes with their accession numbers, collection centres, country of origin, host species and specimen source, is presented in Supplementary Data S2A. Aside from *M. orygis*, a total of 190 previously published genome sequences (complete/draft/SRA) representing all MTBC species and lineages (L1 to L9), including its animal lineages, were also downloaded from the above repositories and added to the analysis (Supplementary Data S2B).

Gene selection and R analysis: Unique SNPs for *M. orygis* and other MTBC species/lineages. To develop a simple, quick, effective and yet powerful identification approach, a set of six genes, *viz., gyrB, PPE55, Rv2042c, leuS, mmpL6,* and *mmpS6* were selected for SNPs analysis, and these genes were chosen based on the previous publications (2, 17, 34). The sequence for *gyrB, PPE55, Rv2042c* and *leuS* genes were obtained from H37Rv (accession # NC_00962.3), and *mmpL6* and *mmpS6* gene sequences were recovered from Mtb-specific deletion 1 region 'TbD1' (accession # AJ426486.1). The numerical positions for the new and known SNPs in the *gyrB, Rv2042c, PPE55,* and *leuS* genes are relative to *M. tuberculosis* H37Rv reference genome (accession # NC_000962.3, (35), while the SNPs positions within *mmpL6* and *mmpS6* genes are provided following TbD1 region (AJ426486.1, (36)). All

identified SNPs within the six genes are numbered according to gene position (5□ to 3□), and the complementary sequences were used for genes *Rv2042c* and *PPE55* (Table 2).

To extract the above-mentioned six gene sequences from *M. orygis* and MTBC species/lineage genomes, a high-throughput analysis was performed using our custom in-house-developed R-scripts in RStudio v.2021.9.2.382 (37), which used Basic Local Alignment Search Tool (BLAST) (38), with the e-value cutoff option set to 10e-100. The scripts interrogated the assembled genomes using a reference gene to identify and extract the sequence corresponding to the SNPs of interest. NCBI BLAST search was also done for all genes, with any unique sequences investigated. In order to identify *M. orygis* or other MTBC species-specific new and known SNPs, the extracted gene sequences were aligned using MUSCLE (39), and sequence variations of the unique SNPs were detected using MEGA-X v.10.0.5 (40) and Geneious Prime 2022 (Geneious Prime 2022.0.2, https://www. geneious.com).

The SNPs-based phylogenetic reconstruction of *M. orygis*. To compare *M. orygis* sequences from the present study with those available in the public repositories, a whole genome phylogeny was analyzed using the single nucleotide variant phylogenomics (SNVPhyl) pipeline v.1.2.3 in IRIDA (23). The snvAlignment file obtained from SNVPhyl was then used to construct a maximum likelihood tree using RAxML v.8.2.12 with the GTRGAMMA model and 100 rapid bootstrap replicates (41). The output phylogenetic tree was visualized using the online tool iTOL (Interactive Tree Of Life) v.6 (42). *M. tuberculosis* H37Rv (NC_00962.3) was chosen as a reference genome with parameters: minimum SNV coverage of 15, SNV abundance ratio of 0.75, and minimum mapping quality of 30. The SNV density filtering was enabled to remove high-density SNV regions that could suggest possible recombination. The identified SNVs were

used in the construction of a phylogeny by calculating genetic distance between isolates by a generalized time reversible (GTR) model with PhyML 3.0 (43).

Data availability. The whole genome sequences of *M. orygis* isolates from this project have been deposited in GenBank and are available through the SRA under BioProject no. PRJNA934340 with the submission ID SUB12537725. Accession numbers for all publicly available genomes used in this study are listed in Supplementary Data S2.

RESULTS

Data collection and genome analysis. In the present study, a total of 61 *M. orygis* isolates were identified retrospectively from cultures submitted by five different Provinces in Canada over a period of 14 years (2009-2022), with more than 90% from British Columbia and Alberta. The cultures were recovered from pulmonary (65%) and extra-pulmonary samples. Forty-five *M. orygis* cases were female patients (74%), 14 were male patients (23%) and two cases did not provide gender on the requisition. The PCR, MIRU-VNTR testing, BioHansel and Mykrobe Predictor v.0.7.0 (44) analyses identified the isolates as *M. africanum* or *M. africanum/M. pinnipedii* or MTBC animal lineages (Table 1). We then performed WGS for all 61 isolates and determined their identification as *M. orygis* using SNPs analysis. The sequencing data were augmented with a total of 56 publicly available *M. orygis* genome sequences to validate the identification and their discrimination from other MTBC species. Of these 56 sequences, 35 were reported from India, followed by eight from The United States, five from Norway, four from Switzerland, two from The United Kingdom, one from The Netherlands, and one from Canada (Supplementary Data S2A).

Whole genome-based SNPs analysis. We performed the WGS-based SNPs analysis for identification and genetic differentiation of *M. orygis* from other MTBC members. Several new and known SNPs in *gyrB* (2028 bp), *PPE55* (2405 bp), *Rv2042c* (798 bp), *leuS* (2910 bp), *mmpL6* (2904 bp), and *mmpS6* (447 bp) genes were evaluated to ascertain their effectiveness as molecular markers for *M. orygis*. By gene sequence analysis of multiple sequence alignment from 117 *M. orygis* genomes (61 studied and 56 public repositories) and 190 sequences of MTBC species and lineages, we identified 10 unique SNPs across six selected genes that can identify and discriminate *M. orygis* from all members of MTBC. These *M. orygis* specific SNPs along with their corresponding genomic positions, mutation type, locus, and function are listed in Table 2.

A multiple alignment of the *gyrB* gene showed the presence of two novel nonsynonymous mutations at positions 277 (A→G) and 1478 (T→C), along with a known synonymous mutation at position 870 (G→A) (Table 2, Table 3). Within *PPE55* gene, our study identified two *M. orygis* specific SNPs at positions 2162 (C→T) and 2163 (T→G). The SNPs analysis also identified a unique SNP in *Rv2042c* gene at position 113 (T→G). In addition, we found two *M. orygis* specific SNPs within *leuS* gene; one of which was a novel SNP ($leuS^{1064}$ A→T) and associated with synonymous mutation, and another one ($leuS^{1251}$ G→T) was previously reported. The sequence of the TbD1 region (AJ426486.1) from the ancestral *M. tuberculosis* contains genes *mmpS6* and *mmpL6*. Two more novel SNPs were detected in this region and were found to be associated with a synonymous ($mmpL6^{486}$ T→C) and a nonsynonymous ($mmpS6^{334}$ C→G) mutation (Table 2, Table 3). The genes *mmpL6* and *mmpS6* encode membrane proteins belonging to a large and a small family, respectively. The above 10

unique SNPs situated within a set of six genes were identified in all 117 *M. orygis* strains screened in the present study.

In order to define the discriminatory power of the WGS-based SNPs analysis in identifying all MTBC species/lineages, the above 117 *M. orygis* (Table 1, Supplementary Data S2A) and 190 MTBC (Supplementary Data S2B) genome sequences were also analyzed, and results of representative sequence from each species and lineage are shown in Table 3. Apart from *M. orygis* specific SNPs, we identified some new and previously reported SNPs that are unique for all MTBC members and lineages. These include 11 SNPs in *gyrB*, 10 SNPs in *leuS*, nine SNPs in *PPE55*, six SNPs in *mmpL6*, three SNPs in *Rv2042c*, and two SNPs in *mmpS6* genes.

In our study, we detected a series of SNPs specific for *M. canettii* in all six selected genes, including *gyrB* (213G→A, 492C→T, 7681C→T, 1215C→T and 1854T→C), *PPE55* (1445G→A), *Rv2042c* (134A→G), *mmpL6* (1879G→C), and *leuS* (321C→A and 2199T→C). Species-specific polymorphisms were also found for *M. caprae* (*gyrB*^{1068} T→G and *leuS*^{1262} A→C), *M. microti* (*gyrB*^{432} C→T) and *M. mungi* (*leuS*^{2343} G→A). Interestingly, one SNP identified in *leuS* gene (564A→G) of *M. microti* was overlapped in *M. pinnipedii* strains. The gene sequence analysis identified a SNP (*mmpL6*^{306} C→T) that can differentiate *M. africanum* lineage 6 from other MTBC members and lineage. While the SNPs in *gyrB* gene at position 1167 (C→T) and in *leuS* at position 1632 (G→A) were found to discriminate *M. bovis* and *M. bovis* BCG from all others, the SNP *PPE55*^{1701} A→G could even separate *M. bovis* from BCG. Furthermore, *M. bovis* and BCG shared SNPs *gyrB*^{513} G→A and *PPE55*^{556} C→A with *M. caprae* strains (Table 3). In addition to the above, we detected SNP markers in the selected six genes

that differentiated the strains belonging to different lineages from other MTBC members examined in this study.

Single nucleotide polymorphisms-based phylogenetic inference of *M. orygis*. A whole genome SNP-based phylogenetic tree was built to determine the genetic similarity of the newly sequenced *M. orygis* isolates in Canada with *M. orygis* sequences from the public repositories (Fig. 1). The SNV distance matrix generated from the whole genome alignments lists the pairwise SNV distances between every sequence (Supplementary Data S3). The genome sequences from this study clustered with *M. orygis* sequences reported from all five geographic locations.

M. orygis sequences from different samples of the same patient clustered together in the SNVPhyl tree, as expected. For example, BC isolate 1801225 clustered with 1801535, 1900059 with 1900097, 2001127 with 2001229, 2001439 with 2001480 and 2200196 with 2200263, and AB isolate 2101205 with 2101375 appeared to have a SNVs distance of 0-5. The isolates 2001439 and 2001480 also clustered (29-52 SNVs distance) with two *M. orygis* sequences from Norway (ERR5336158) and The United States (SRR5642712), and one previously reported *M. orygis* sequence 51145 (SRR16643349) from Quebec Canada. Similarly, 11 *M. orygis* sequences from the present study formed a large cluster with three sequences (ERR2659153, ERR2659154, ERR2659156) from Switzerland, one sequence (ERR5336157) from Norway, and another sequence (ERR015582) reported from The United Kingdom, with two sequences (1800965 from BC and 1900228 from AB) showing a distance of 92-96 SNVs with the *M. orygis* sequences from Switzerland (Fig. 1, Supplementary Data S3).

The studied isolates 1901282 and 2100725 from AB clustered with 11 *M. orygis* sequences of animal origin from India and one sequence of human origin from The United States

with a pairwise SNV distance of 34-88. Interestingly, the isolate 2100725 clustered with seven of these animal originated sequences (SRR10251185, SRR10251189-92, SRR10251197, and SRR10251200) by 34 SNV distances. However, the SNV numbers of 34-88 in MTBC are quite a large distance and are not indicative of spillover or direction of spillover.

DISCUSSION

The present study is the first to report a greater number of *M. orygis* from a single geographic area. *M. orygis* is a causative agent of tuberculosis in both animal and human hosts and it was first described in 2012 (2), and later by others (3-5, 7, 8). The MTBC species *M. bovis* has long been believed to be the only agent that causes zoonotic tuberculosis, however, recovering a larger number of *M. orygis* from both animals and humans in recent years from different areas of the world highlights the need for considering this bacterium as a zoonotic pathogen (45).

Since tuberculosis cases caused by *M. orygis* are often identified as MTBC or misidentified and published as *M. africanum* or *M. bovis*, the actual number of infections associated with this bacterium may have been underreported (3, 16, 46). From our culture collection, the investigation showed that since 2009, *M. orygis* was misidentified as *M. africanum*. Rahim et al. (18) initially reported *M. africanum* from four dairy cows that were later identified to be *M. orygis* by refined analysis (3). A part of the confusion towards this misidentification is that *M. orygis* shares the $gyrB^{1207}$ (G→T) ($gyrB^{1450}$ as per accession no. L27512.1, (47)) mutations with *M. africanum*, *M. bovis*, *M. microti*, *M. caprae*, *M. mungi*, and *M. pinnipedii* ((17), Table 3). Moreover, MTBC is genetically highly clonal, and thus without proper identification tools and analysis approaches the species differentiation could be challenging.

Since currently available laboratory tests are struggling to differentiate animal lineages, in the present investigation we used WGS-based SNPs analysis targeting *gyrB, PPE55, Rv2042c, leuS, mmpL6,* and *mmpS6* genes that can accurately identify *M. orygis* and unambiguously differentiate it from all members of MTBC species. The *gyrB* gene encodes for the *β*-subunit of the DNA gyrase and has been used as a molecular marker for identification of MTBC members. The discriminatory power of polymorphisms in *gyrB* gene in identifying *M. orygis* has been evaluated in previous studies (3, 17). While these authors described only one unique SNP ($gyrB^{870}$ G→A) ($gyrB^{1113}$ according to accession no. L27512.1), we detected two more novel and useful genetic markers within *gyrB* gene ($gyrB^{277}$ A→G, and $gyrB^{1478}$ T→C) in the screened *M. orygis* genomes. The *PPE55* (Rv3347c) gene is specific to MTBC species and play a major role in host-pathogen interactions. Huard et al. (17) earlier reported two *M. orygis* specific SNPs in *PPE55* gene ($PPE55^{2162}$ C→T, and $PPE55^{2163}$ T→G), that was also revealed in our study. The results suggest that these SNPs could be used as genetic markers for the identification of *M. orygis*. We also evaluated *Rv2042c* gene for possible unique SNPs to be used as a molecular marker to identify *M. orygis*. In accordance with van Ingen et al. (2), our SNPs analysis identified a nonsynonymous mutation in the 38th codon of *Rv2042c* gene at position 113 (T→G). In mycobacterial species, *leuS* gene encodes for L-leucyl-tRNA synthetase, which is involved in translation. In this study, in addition to the SNP $leuS^{1251}$ (G→T) described by Napier et al. (34), we showed that another novel SNP $leuS^{1064}$ (A→T) associated with nonsynonymous mutation is also present in *M. orygis* strains (Table 2, Table 3). Mtb-specific deletion 1 region 'TbD1' region comprises the *mmpL6* and *mmpS6* genes, which encode for mycobacterial membrane protein families. In modern *M. tuberculosis* strains, the *mmpS6* gene is fully deleted

and the *mmpL6* gene is trimmed (48). Our result shows two SNP *mmpL6*^{486} T→C, and *mmpS6*^{334} C→G, as two novel and distinct genetic markers for the identification of *M. orygis*.

We included previously described *M. orygis* specific polymorphisms in the WGS-based SNPs analysis, and they were reconfirmed in a larger number (61 studied and 56 from public repositories) of *M. orygis* genomes tested in this work. Thus, the results suggest WGS-based SNPs analysis is a useful tool to rapidly identify *M. orygis*, and to clearly differentiate this emerging pathogen from all MTBC species and lineages. To the best of our knowledge, the unique SNPs *gyrB*^{277} (A→G), *gyrB*^{1478} (T→C), *leuS*^{1064} (A→T), *mmpL6*^{486} (T→C), and *mmpS6*^{334} (C→G) are the first to be reported in the present study.

WGS-based SNPs analysis has been employed for the identification of MTBC species (9), and this approach has also become pertinent to MTBC genotyping (36, 49). In this study, we further identified the species-specific polymorphisms for members of MTBC species (excluding *M. orygis* here, as discussed above) and lineages, and results from *gyrB* gene confirm the discrimination of several MTBC members including *M. tuberculosis*, *M. bovis*, *M. canettii*, *M. microti*, and *M. caprae*, (Table 3). These results are in agreement with Huard et al. (17), who reported the identification of the above species (along with *M. orygis*) using a PCR-based SNPs analysis. Our study supports that many genomic characteristics could be shared between *M. bovis* and *M. caprae* strains, for example, the *gyrB* mutation at position 513, and *PPE55* mutation at position 556 (17).

Furthermore, we detected unique polymorphisms for lineage 1 in *gyrB*^{873} gene (G→C); for lineage 2 in *Rv2042c*^{597} (C→G) gene; for lineage 3 in *PPE55* gene at positions 1173 (G→C), 1177-1178 (A→G), 1179 (C→T) and 1182 (C→G); for lineage 5 in *PPE55*^{923} (G→T), *mmpS6*^{360} (C→T) and *leuS*^{2736} (T→C) genes; for lineage 7 in *mmpL6*^{1780} (G→C) and *mmpS6*^{339} (C→T)

genes; and for lineage 9 in *leuS*^{1363} (C→T) gene. The reason for not being able to extract some SNPs in *mmpL6* and *mmpS6* genes for *M. tuberculosis* H37Rv and lineages 1-4 in Table 3 is that these genes are partly or fully deleted from the modern *M. tuberculosis*, as discussed above. Our results indicate that WGS-based SNPs analysis could be successfully used to distinguish all members of MTBC species and lineages. Since we tested a limited number of sequences due to availability and data quality, particularly for *M. pinnipedii*, *M. caprae*, and *M. mungi*, further investigation with a large dataset would be useful for their specific identification with high confidence.

The use of whole genome SNPs-based phylogenetic tree allowed us to inspect the genetic relation of *M. orygis* recovered from Canada and those reported from other geographic regions. The *M. orygis* isolates of human origin from this study were found to be distributed across the phylogeny (Fig. 1). Phylogenetic analysis of *M. orygis* genomes from the same patient on the same episode (two isolates from each of the six patients) showed a difference of 0-5 SNV.

We sought to determine whether *M. orygis* identified in this study showed a close phylogenetic relation with animal strains. The result shows that the isolate 2100725 from an AB patient clustered (34 SNVs apart) with several *M. orygis* sequences of animal origin reported from India (Fig. 1, Supplementary Data S3). This result may be indicative of the adaptation of animal origin *M. orygis* strains to a human host. The zoonotic or zooanthroponosis potentials of *M. orygis* has been discussed in a previous work (45).

Furthermore, two isolates from BC were related (5-39 SNVs apart) to *M. orygis* sequences of human origin reported from Norway (ERR5336158, (8) and The United States (SRR5642712, (7), and segregated by 49-52 SNVs from previously published Canadian strain 51145 (50). Thus, the newly sequenced *M. orygis* from Canada phylogenetically clustered with

M. orygis sequences reported from different geographic locations, placing it within the global context.

In conclusion, the WGS analysis in the present study evaluated 10 novel and known unique SNPs within a set of six genes that could be used as molecular genetic markers to accurately identify *M. orygis* and unambiguously discriminate it from all members of MTBC species and lineages. As WGS technologies are increasingly being used by the healthcare systems, our approach will be helpful to diagnosis and surveillance of *M. orygis* associated tuberculosis, and optimizing the clinical management of this disease. The analysis of *M. orygis* sequences will improve our understanding of molecular characteristics and phylogenetic diversity of this emerging pathogen and its implication as a zoonosis. The ever-increasing evidence of *M. orygis* linked endemicity, and the identification of a greater number of *M. orygis* from animals and humans around the world highlight the urgency for a multi-sectoral collaboration linking the clinical and veterinary sectors towards One Health approach. The origin, epidemiology and transmission dynamics of *M. orygis* within Canada is currently under investigation.

ACKNOWLEDGMENTS

The authors would like to thank Darrell Johnstone, Debra Janella and members of the National Reference Centre for Mycobacteriology (NRCM) and DNA Sequencing Core Services for their technical assistance with this project. The authors also acknowledge Walter Demczuk for help with R-scripting, and Aaron Petkau for valuable suggestions with phylogenetic analysis. This work is supported under the Genomics Research and Development Initiative (GRDI) program funded by Public Health Agency of Canada.

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest.

ETHICS STATEMENT

This study was exempted from Health Canada and Public Health Agency of Canada Research Ethics Board (REB) since no clinical samples or human biological materials were used for this study.

REFERENCES

1. WHO. 2021. Global tuberculosis report 2021. World Health Organization, Geneva.
2. van Ingen J, Rahim Z, Mulder A, Boeree MJ, Simeone R, Brosch R, van Soolingen D. 2012. Characterization of *Mycobacterium orygis* as *M. tuberculosis* complex subspecies. *Emerg Infect Dis* 18:653-5.
3. Rahim Z, Thapa J, Fukushima Y, van der Zanden AGM, Gordon SV, Suzuki Y, Nakajima C. 2017. Tuberculosis caused by *Mycobacterium orygis* in dairy cattle and captured monkeys in Bangladesh: a new scenario of tuberculosis in South Asia. *Transbound Emerg Dis* 64:1965-1969.
4. Thapa J, Nakajima C, Maharjan B, Poudell A, Suzuki Y. 2015. Molecular characterization of *Mycobacterium orygis* isolates from wild animals of Nepal. *Jpn J Vet Res* 63:151-8.
5. Refaya AK, Ramanujam H, Ramalingam M, Rao GVS, Ravikumar D, Sangamithrai D, Shanmugam S, Palaniyandi K. 2022. Tuberculosis caused by *Mycobacterium orygis* in wild ungulates in Chennai, South India. *Transbound Emerg Dis* 69:e3327-e3333.
6. Thapa J, Paudel S, Sadaula A, Shah Y, Maharjan B, Kaufman GE, McCauley D, Gairhe KP, Tsubota T, Suzuki Y, Nakajima C. 2016. *Mycobacterium orygis*-Associated Tuberculosis in Free-Ranging Rhinoceros, Nepal, 2015. *Emerg Infect Dis* 22:570-2.
7. Marcos LA, Spitzer ED, Mahapatra R, Ma Y, Halse TA, Shea J, Isabelle M, Lapierre P, Escuyer VE. 2017. *Mycobacterium orygis* Lymphadenitis in New York, USA. *Emerg Infect Dis* 23:1749-1751.
8. Eldholm V, Rønning JO, Mengshoel AT, Arnesen T. 2021. Import and transmission of *Mycobacterium orygis* and *Mycobacterium africanum*, Norway. *BMC Infect Dis* 21:562.

9. Lipworth S, Jajou R, de Neeling A, Bradley P, van der Hoek W, Maphalala G, Bonnet M, Sanchez-Padilla E, Diel R, Niemann S, Iqbal Z, Smith G, Peto T, Crook D, Walker T, van Soolingen D. 2019. SNP-IT tool for identifying subspecies and associated lineages of *Mycobacterium tuberculosis* Complex. *Emerg Infect Dis* 25:482-488.

10. Dawson KL, Bell A, Kawakami RP, Coley K, Yates G, Collins Desmond M. 2012. Transmission of *Mycobacterium orygis* (*M. tuberculosis* Complex species) from a tuberculosis patient to a dairy cow in New Zealand. *J Clin Microbiol* 50:3136-3138.

11. Mounchili A, Perera R, Lee RS, Njoo H, Brooks J. 2022. Chapter 1: Epidemiology of tuberculosis in Canada. *Can J Respir Crit Care Sleep Med* 6:8-21.

12. WHO. 2016. Towards TB elimination in low-incidence countries: TB treatment and care 2016.

13. Boritsch EC, Khanna V, Pawlik A, Honoré N, Navas VH, Ma L, Bouchier C, Seemann T, Supply P, Stinear TP, Brosch R. 2016. Key experimental evidence of chromosomal DNA transfer among selected tuberculosis-causing mycobacteria. *Proc Natl Acad Sci U S A* 113:9876-9881.

14. Brites D, Loiseau C, Menardo F, Borrell S, Boniotti MB, Warren R, Dippenaar A, Parsons SDC, Beisel C, Behr MA, Fyfe JA, Coscolla M, Gagneux S. 2018. A new phylogenetic framework for the animal-adapted *Mycobacterium tuberculosis* Complex. *Front Microbiol* 9.

15. Landolt P, Stephan R, Stevens MJA, Scherrer S. 2019. Three-reaction high-resolution melting assay for rapid differentiation of *Mycobacterium tuberculosis* complex members. *Microbiologyopen* 8:e919.

16. Loiseau C, Brites D, Moser I, Coll F, Pourcel C, Robbe-Austerman S, Escuyer V, Musser Kimberlee A, Peacock Sharon J, Feuerriegel S, Kohl Thomas A, Niemann S, Gagneux S, Köser Claudio U. 2019. Revised Interpretation of the Hain Lifescience GenoType MTBC To Differentiate *Mycobacterium canettii* and members of the *Mycobacterium tuberculosis* Complex. *Antimicrob Agents Chemother* 63:e00159-19.

17. Huard Richard C, Fabre M, de Haas P, Claudio Oliveira Lazzarini L, van Soolingen D, Cousins D, Ho John L. 2006. Novel genetic polymorphisms that further delineate the phylogeny of the *Mycobacterium tuberculosis* Complex. *J Bacteriol* 188:4271-4287.

18. Rahim Z, Möllers M, te Koppele-Vije A, de Beer J, Zaman K, Matin MA, Kamal M, Raquib R, van Soolingen D, Baqi MA, Heilmann FG, van der Zanden AG. 2007. Characterization of *Mycobacterium africanum* subtype I among cows in a dairy farm in Bangladesh using spoligotyping. *Southeast Asian J Trop Med Public Health* 38:706-13.

19. Meehan CJ, Goig GA, Kohl TA, Verboven L, Dippenaar A, Ezewudo M, Farhat MR, Guthrie JL, Laukens K, Miotto P, Ofori-Anyinam B, Dreyer V, Supply P, Suresh A, Utpatel C, van Soolingen D, Zhou Y, Ashton PM, Brites D, Cabibbe AM, de Jong BC, de Vos M, Menardo F, Gagneux S, Gao Q, Heupink TH, Liu Q, Loiseau C, Rigouts L, Rodwell TC, Tagliani E, Walker TM, Warren RM, Zhao Y, Zignol M, Schito M, Gardy J, Cirillo DM, Niemann S, Comas I, Van Rie A. 2019. Whole genome sequencing of *Mycobacterium tuberculosis*: current standards and open issues. *Nat Rev Microbiol* 17:533-545.

20. Deurenberg RH, Bathoorn E, Chlebowicz MA, Couto N, Ferdous M, García-Cobos S, Kooistra-Smid AMD, Raangs EC, Rosema S, Veloo ACM, Zhou K, Friedrich AW,

Rossen JWA. 2017. Application of next generation sequencing in clinical microbiology and infection prevention. *J Biotechnol* 243:16-24.

21. Nadon C, Van Walle I, Gerner-Smidt P, Campos J, Chinen I, Concepcion-Acevedo J, Gilpin B, Smith AM, Man Kam K, Perez E, Trees E, Kubota K, Takkinen J, Nielsen EM, Carleton H. 2017. PulseNet International: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. *Euro Surveill* 22.

22. Dallman T, Ashton P, Schafer U, Jironkin A, Painset A, Shaaban S, Hartman H, Myers R, Underwood A, Jenkins C, Grant K. 2018. SnapperDB: a database solution for routine sequencing analysis of bacterial isolates. *Bioinformatics* 34:3028-3029.

23. Petkau A, Mabon P, Sieffert C, Knox NC, Cabral J, Iskander M, Iskander M, Weedmark K, Zaheer R, Katz LS, Nadon C, Reimer A, Taboada E, Beiko RG, Hsiao W, Brinkman F, Graham M, Van Domselaar G. 2017. SNVPhyl: a single nucleotide variant phylogenomics pipeline for microbial genomic epidemiology. *Microb Genom* 3:e000116.

24. Labbé G, Kruczkiewicz P, Robertson J, Mabon P, Schonfeld J, Kein D, Rankin MA, Gopez M, Hole D, Son D, Knox N, Laing CR, Bessonov K, Taboada EN, Yoshida C, Ziebell K, Nichani A, Johnson RP, Van Domselaar G, Nash JHE. 2021. Rapid and accurate SNP genotyping of clonal bacterial pathogens with BioHansel. *Microb Genom* 7:000651.

25. Matthews TC, Bristow FR, Griffiths EJ, Petkau A, Adam J, Dooley D, Kruczkiewicz P, Curatcha J, Cabral J, Fornika D, Winsor GL, Courtot M, Bertelli C, Roudgar A, Feijao P, Mabon P, Enns E, Thiessen J, Keddy A, Isaac-Renton J, Gardy JL, Tang P, João A Carriço TIC, Chindelevitch L, Chauve C, Graham MR, McArthur AG, Taboada EN,

Beiko RG, Brinkman FSL, Hsiao WWL, Domselaar GV. 2018. The Integrated Rapid Infectious Disease Analysis (IRIDA) Platform. *bioRxiv* doi:10.1101/381830:381830.

26. Seemann T. 2019. Shovill. https://github.com/tseemann/shovill.

27. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30:2068-2069.

28. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29:1072-1075.

29. Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

30. The Galaxy C. 2022. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res* 50:W345-W351.

31. NCBI. 2022a. https://www.ncbi.nlm.nih.gov/genbank/. Accessed December 9, 2022.

32. NCBI. 2022b. https://www.ncbi.nlm.nih.gov/sra. Accessed December 9, 2022.

33. ENA. 2022. https://www.ebi.ac.uk/ena. Accessed December 9, 2022.

34. Napier G, Campino S, Merid Y, Abebe M, Woldeamanuel Y, Aseffa A, Hibberd ML, Phelan J, Clark TG. 2020. Robust barcoding and identification of *Mycobacterium tuberculosis* lineages for epidemiological and clinical studies. *Genome Med* 12:114.

35. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Krogh A, McLean J, Moule S, Murphy L, Oliver K, Osborne J, Quail MA, Rajandream MA, Rogers J, Rutter S, Seeger K, Skelton J, Squares R, Squares S, Sulston

JE, Taylor K, Whitehead S, Barrell BG. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 396:190-190.

36. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, Garnier T, Gutierrez C, Hewinson G, Kremer K, Parsons LM, Pym AS, Samper S, van Soolingen D, Cole ST. 2002. A new evolutionary scenario for the *Mycobacterium tuberculosis* Complex. *Proc Natl Acad Sci U S A* 99:3684-3689.

37. RStudio T. 2020. RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. URL http://www.rstudio.com/.

38. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403-410.

39. Edgar RC. 2004. Local homology recognition and distance measures in linear time using compressed amino acid alphabets. *Nucleic Acids Res* 32:380-385.

40. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* 35:1547-1549.

41. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312-1313.

42. Letunic I, Bork P. 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research* 49:W293-W296.

43. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate Maximum-Likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst Biol* 59:307-321.

44. Hunt M, Bradley P, Lapierre SG, Heys S, Thomsit M, Hall MB, Malone KM, Wintringer P, Walker TM, Cirillo DM, Comas I, Farhat MR, Fowler P, Gardy J, Ismail N, Kohl TA, Mathys V, Merker M, Niemann S, Omar SV, Sintchenko V, Smith G, van Soolingen D, Supply P, Tahseen S, Wilcox M, Arandjelovic I, Peto TEA, Crook DW, Iqbal Z. 2019. Antibiotic resistance prediction for *Mycobacterium tuberculosis* from genome sequence data with Mykrobe. *Wellcome Open Res* 4:191.

45. Duffy SC, Srinivasan S, Schilling MA, Stuber T, Danchuk SN, Michael JS, Venkatesan M, Bansal N, Maan S, Jindal N, Chaudhary D, Dandapat P, Katani R, Chothe S, Veerasami M, Robbe-Austerman S, Juleff N, Kapur V, Behr MA. 2020. Reconsidering *Mycobacterium bovis* as a proxy for zoonotic tuberculosis: a molecular epidemiological surveillance study. *Lancet Microbe* 1:e66-e73.

46. Tortoli E. 2014. Microbiological features and clinical relevance of new species of the genus Mycobacterium. *Clin Microbiol Rev* 27:727-52.

47. Kasai H, Ezaki T, Harayama S. 2000. Differentiation of phylogenetically related slowly growing mycobacteria by their gyrB sequences. *J Clin Microbiol* 38:301-8.

48. Bottai D, Frigui W, Sayes F, Di Luca M, Spadoni D, Pawlik A, Zoppo M, Orgeur M, Khanna V, Hardy D, Mangenot S, Barbe V, Medigue C, Ma L, Bouchier C, Tavanti A, Larrouy-Maumus G, Brosch R. 2020. TbD1 deletion as a driver of the evolutionary success of modern epidemic Mycobacterium tuberculosis lineages. *Nat Commun* 11:684.

49. Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, Portugal I, Pain A, Martin N, Clark TG. 2014. A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat Commun* 5:4812.

50. Rufai Syed B, McIntosh F, Poojary I, Chothe S, Sebastian A, Albert I, Praul C, Venkatesan M, Mahata G, Maity H, Dandapat P, Michael Joy S, Katani R, Kapur V, Behr Marcel A. 2021. Complete genome sequence of *Mycobacterium orygis* strain 51145. *Microbiol Resour Announc* 10:e01279-20.

Figure legend

FIG 1 Phylogenetic tree of *M. orygis* based on single nucleotide variants (SNVs) analysis. The whole genome phylogeny was analyzed using SNVPhyl pipeline v.1.2.3, and the snvAlignment file was used to infer the tree using RAxML v.8.2.12 with the GTRGAMMA model and 100 rapid bootstrap replicates. The output tree was visualized using iTOL v.6. A maximum likelihood tree of 61 newly sequenced *M. orygis* isolates from this study (labelled in red), and the *M. orygis* genomes collected from public repositories. The left bar adjacent to the tree nodes shows the source of *M. orygis* (human in purple, and animal in orange), and the right bar denotes shows the geographic location of *M. orygis* (Canada in green, India in blue, Norway in pink, Switzerland in cyan, UK in red, and USA in yellow). *M. tuberculosis* H37Rv (NC_000962.3) used as a reference in this analysis. The filled circles on nodes represent a bootstrap value between 80 and 100. The studied isolates mostly clustered with *M. orygis* of human origin recovered from different geographic regions. The isolate 2100725 showed a good phylogenetic relation (pairwise distances of 34 SNVs) with seven *M. orygis* of animal origin from India.

569 **TABLE 1** Newly identified 61 *M. orygis* isolates from this study

Isolate	Province	Specimen	Identification (PCR/MIRU-VNTR or BioHansel)
900123	BC	Bronchoalveolar lavage LLL	*M. africanum* by PCR, MTBC by BioHansel
900124	BC	Bronchoalveolar lavage LUL	*M. africanum* by PCR, MTBC by BioHansel
900125	BC	Bronchoalveolar lavage RUL	*M. africanum* by PCR, MTBC by BioHansel
900942	SK	Bronchial wash	*M. africanum* by PCR, MTBC by BioHansel
1501282	BC	Sputum	*M. africanum* by PCR/MIRU-VNTR, MTBC by BioHansel
1600127	BC	Bronchial wash	*M. africanum* by PCR/MIRU-VNTR, MTBC by BioHansel
1601131	MB	Sputum	*M. africanum* by PCR/MIRU-VNTR, MTBC by BioHansel
1601221	AB	Sputum	*M. africanum* by PCR/MIRU-VNTR, MTBC by BioHansel
1601335	MB	Endotracheal	*M. africanum* by PCR/MIRU-VNTR, MTBC by BioHansel
1601614	BC	Bronchial wash	*M. africanum* by PCR/MIRU-VNTR, MTBC by BioHansel
1700074	AB	Stool	*M. africanum* by PCR/MIRU-VNTR, MTBC by BioHansel
1701554	AB	Sputum	*M. africanum/M. pinnipedii* by PCR/MIRU-VNTR, MTBC by BioHansel
1701642	AB	Sputum	*M. africanum/M. pinnipedii* by PCR/MIRU-VNTR, MTBC by BioHansel
1701797	BC	Urine	*M. africanum/M. pinnipedii* by PCR/MIRU-VNTR, MTBC by BioHansel
1701798	BC	Bronchial wash	*M. africanum/M. pinnipedii* by PCR/MIRU-VNTR, MTBC by BioHansel
1800159	BC	Ascitic fluid	*M. africanum/M. pinnipedii* by PCR/MIRU-VNTR, MTBC by BioHansel
1800300	AB	Sputum	*M. africanum/M. pinnipedii* by PCR/MIRU-VNTR, MTBC by BioHansel
1800446	BC	Pericardial fluid	*M. africanum/M. pinnipedii* by PCR/MIRU-VNTR, MTBC by BioHansel
1800448	BC	Bronchial wash	*M. africanum/M. pinnipedii* by PCR/MIRU-VNTR, MTBC by BioHansel

1800722	AB	Sputum	*M. africanum/M. pinnipedii* by PCR/MIRU-VNTR, MTBC by BioHansel
1800965	BC	Sputum	*M. africanum/M. pinnipedii* by PCR/MIRU-VNTR, MTBC by BioHansel
1801127	BC	Sputum	*M. africanum/M. pinnipedii* by PCR/MIRU-VNTR, MTBC by BioHansel
1801225	BC	Bronchial wash RUL	*M. africanum/M. pinnipedii* by PCR/MIRU-VNTR, MTBC by BioHansel
1801256	BC	Bronchial wash	*M. africanum/M. pinnipedii* by PCR/MIRU-VNTR, MTBC by BioHansel
1801535	BC	Sputum	*M. africanum/M. pinnipedii* by PCR/MIRU-VNTR, MTBC by BioHansel
1900059	BC	Thoracentesis fluid	*M. africanum/M. pinnipedii* by PCR/MIRU-VNTR, MTBC by BioHansel
1900097	BC	Sputum	*M. africanum/M. pinnipedii* by PCR/MIRU-VNTR, MTBC by BioHansel
1900228	AB	Right chest wall	*M. africanum/M. pinnipedii* by PCR/MIRU-VNTR, MTBC by BioHansel
1900323	BC	Bronchial wash	*M. africanum/M. pinnipedii* by PCR/MIRU-VNTR, MTBC by BioHansel
1900674	BC	Abscess L fallopian tube	MTBC by BioHansel
1900698	AB	Sputum	MTBC by BioHansel
1900817	BC	Bronchial wash RUL	MTBC by BioHansel
1901041	BC	Bronchial wash	MTBC by BioHansel
1901282	AB	Sputum	MTBC by BioHansel
1901423	BC	Abscess deep R psoas	MTBC by BioHansel
1901424	BC	Bronchial washing LUL	MTBC by BioHansel
2000368	BC	Bronchial wash RUL	MTBC by BioHansel
2000728	AB	Lymph node fluid	MTBC by BioHansel
2001127	BC	Pleural fluid	MTBC by BioHansel
2001229	BC	Tissue - diaphramatic nodule	MTBC by BioHansel

2001439	BC	Pleural fluid	MTBC by BioHansel
2001463	AB	Lymph node FNA	MTBC by BioHansel
2001464	AB	Sputum	MTBC by BioHansel
2001465	AB	Blood	MTBC by BioHansel
2001479	BC	Bronchial wash R lung	MTBC by BioHansel
2001480	BC	Sputum	MTBC by BioHansel
2100119	AB	Sputum	MTBC by BioHansel
2100187	BC	Bronchial wash	MTBC by BioHansel
2100376	BC	Pleural fluid	MTBC by BioHansel
2100430	AB	Sputum	MTBC by BioHansel
2100725	AB	Lymph node	MTBC by BioHansel
2100976	BC	Sputum	MTBC by BioHansel
2100977	BC	Bronchial wash	MTBC by BioHansel
2101097	BC	Tissue - lymph node mediastinal	MTBC by BioHansel
2101205	AB	Sputum	MTBC by BioHansel
2101352	BC	Sputum	MTBC by BioHansel
2101375	AB	Urine	MTBC by BioHansel
2101452	NS	Lymph node	MTBC by BioHansel
2200196	BC	Urine	MTBC by BioHansel
2200263	BC	Sputum	MTBC by BioHansel
2201129	AB	Pleural fluid	MTBC by BioHansel

PCR, Polymerase chain reaction.

MIRU-VNTR, Mycobacterial interspersed repetitive unit - Variable number tandem repeat.

LLL, Left lower lobe; LUL, Left upper lobe; RUL, Right upper lobe; FNA, Fine needle aspiration.

AB, Alberta; BC, British Columbia; MB, Manitoba; NS, Nova Scotia; SK, Saskatchewan.

MTBC, *Mycobacterium tuberculosis* complex.

576 **TABLE 2** Signature single nucleotide polymorphisms (SNPs) in *M. orygis*, as referenced to H37Rv (NC_000962.3) and Mtb-

577 specific *deletion* 1 region 'TbD1' (AJ426486.1*)

Gene	Gene position	Genomic position of mutation	Nucleotide change	Mutation type	Locus	Transcript biotype/function	Reference
gyrB	5240-7267	5516	A>G	Nonsynonymous	Rv0005	Protein coding	This study
gyrB	5240-7267	6109	G>A	Synonymous	Rv0005	Protein coding	Huard et al. (17)
gyrB	5240-7267	6717	T>C	Nonsynonymous	Rv0005	Protein coding	This study
Rv2042c	c2287884-2288681	2288569	T>G	Nonsynonymous	Rv2042c	Hypothetical protein	van Ingen et al. (2)
PPE55	c3750780-3753184	3751022	T>G	Synonymous	Rv3347c	PPE family protein	Huard et al. (17)
PPE55	c3750780-3753184	3751023	C>T	Nonsynonymous	Rv3347c	PPE family protein	Huard et al. (17)
leuS	43562-46471	44625	A>T	Nonsynonymous	Rv0041	Protein coding	This study
leuS	43562-46471	44812	G>T	Synonymous	Rv0041	Protein coding	Napier et al. (34)
*mmpL6**	735-3638	1220	T>C	Synonymous	TbD1 flanking region	Putative transporter	This study
*mmpS6**	292-738	625	C>G	Nonsynonymous	TbD1 flanking region	Putative small membrane protein	This study

578 'c' in 'Gene position' column for genes *Rv2042c* and *PPE55* denotes complementary sequence.

579

580

TABLE 3 Summary of single nucleotide polymorphisms (SNPs) analysis for *M. orygis* differentiation from other MTBC species and lineages. Fourteen SNPs on *gyrB*, eleven SNPs on *PPE55*, four SNPs on *Rv2042c*, seven SNPs on *mmpL6*, three SNPs on *mmpS6*, and twelve SNPs on *leuS* genes are shown

MTBC species/lineage	**Gene name and reference position**																							
	gyrB*														***PPE55***									
	213	**277**	**432**	**492**	**513**	**768**	**870**	**873**	**1068**	**1167**	**1207**	**1215**	**1478**	**1854**	**556**	**923**	**1173**	**1177-1178**	**1179**	**1182**	**1445**	**1701**	**2162**	**2163**
M. tuberculosis H37Rv / TbD1	G	A	C	C	G	C	G	G	T	C	G	C	T	T	C	G	G	A	C	C	G	A	G	A
M. bovis AF2122/97	G	A	C	C	A	C	G	G	T	T	T	C	T	T	A	G	G	A	C	C	G	A	G	A
M. bovis BCG 1173P2	G	A	C	C	A	C	G	G	T	T	T	C	T	T	A	G	G	A	C	C	G	G	G	A
M. africanum GM041182	G	A	C	C	G	C	G	G	T	C	T	C	T	T	C	G	G	A	C	C	G	A	G	A
M. orygis 51145	G	G	C	C	G	C	A	G	T	C	T	C	C	T	C	G	G	A	C	C	G	A	A	C
M. canettii CIPT140010059	A	A	C	T	G	T	G	G	T	C	G	T	T	C	C	G	G	A	C	C	A	A	G	A
M. caprae ATCC BAA-824	G	A	C	C	A	C	G	G	G	C	T	C	T	T	A	G	G	A	C	C	G	A	G	A
M. microti ATCC 35782	G	A	T	C	G	C	G	G	T	C	T	C	T	T	-	-		-	-	-	-	-	-	-
M. mungi BM22813	G	A	C	C	G	C	G	G	T	C	T	C	T	T	C	G	G	A	C	C	G	A	G	A
M. pinnipedii MP1	G	A	C	C	G	C	G	G	T	C	T	C	T	T	-	-	-	-	-	-	-	-	-	-
Lineage 1 (NC_021740.1)	G	A	C	C	G	C	G	C	T	C	G	C	T	T	C	G	G	A	C	C	G	A	G	A
Lineage 2 (CP007803.1)	G	A	C	C	G	C	G	G	T	C	G	C	T	T	C	G	G	A	C	C	G	A	G	A
Lineage 3 (CP010340.1)	G	A	C	C	G	C	G	G	T	C	G	C	T	T	C	G	C	G	T	G	G	A	G	A
Lineage 4 (NC_020559.1)	G	A	C	C	G	C	G	G	T	C	G	C	T	T	C	G	G	A	C	C	G	A	G	A
Lineage 5 (SRR998618)	G	A	C	C	G	C	G	G	T	C	T	C	T	T	C	T	G	A	C	C	G	A	G	A
Lineage 6 (SRR1577831)	G	A	C	C	G	C	G	G	T	C	T	C	T	T	C	G	G	A	C	C	G	A	G	A
Lineage 7 (ERR1971865)	G	A	C	C	G	C	G	G	T	C	G	C	T	T	C	G	G	A	C	C	G	A	G	A
Lineage 8 (SRR10828835)	G	A	C	C	G	C	G	G	T	C	G	C	T	T	C	G	G	A	C	C	G	A	G	A
Lineage 9 (ERR4162024)	G	A	C	C	G	C	G	G	T	C	T	C	T	T	C	G	G	A	C	C	G	A	G	A

TABLE 3 Continued …

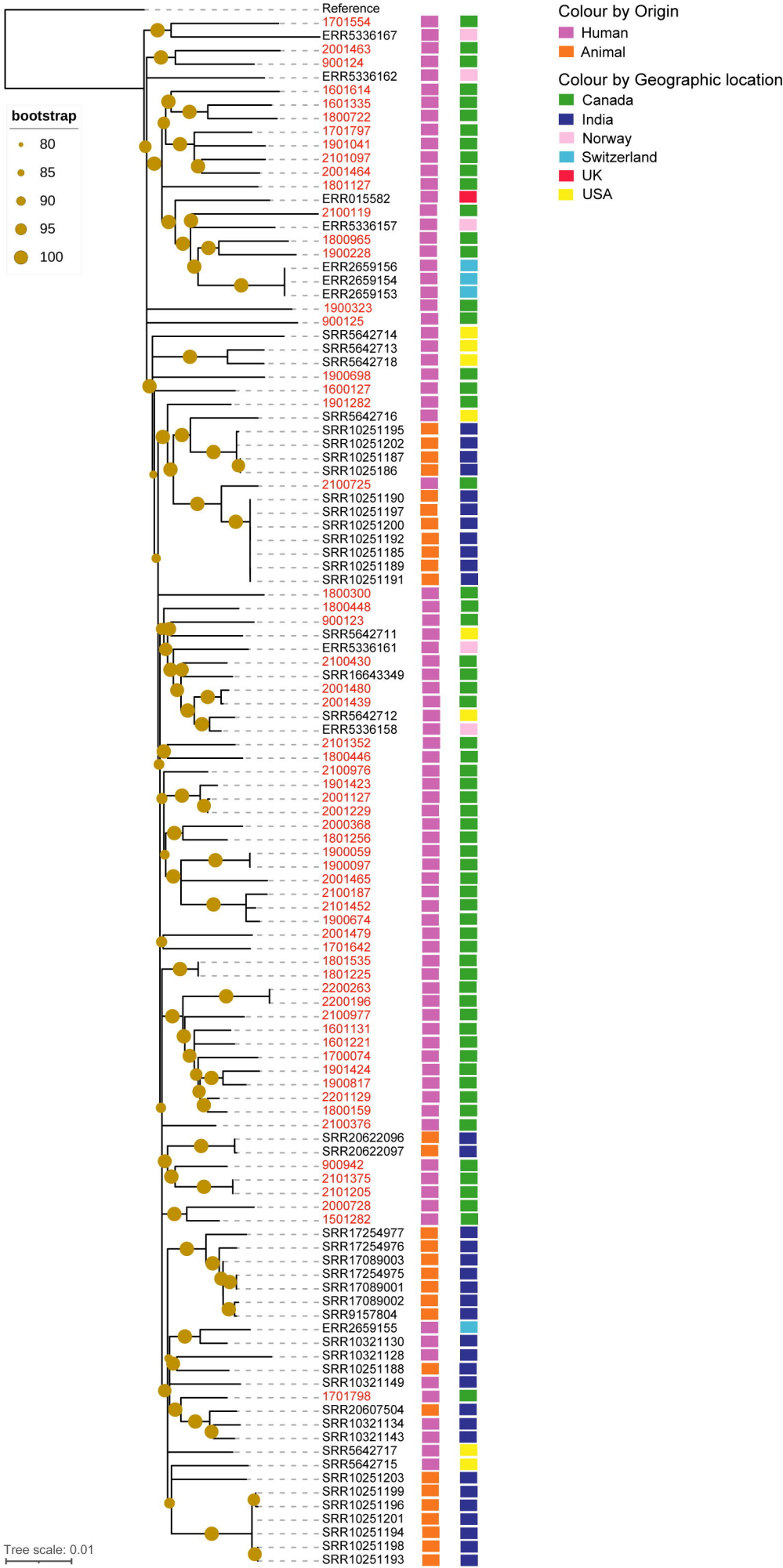
MTBC species/lineage	Gene name and reference position																									
	Rv2042c				*mmpL6*							*mmpS6*			*leuS*											
	113	**134**	**399**	**597**	**306**	**486**	**969**	**1653**	**1780**	**1879**	**2633**	**334**	**339**	**360**	**321**	**384**	**564**	**1064**	**1251**	**1262**	**1363**	**1632**	**2199**	**2343**	**2682**	**273**
M. tuberculosis H37Rv / TbD1	T	A	C	C	C	T	T	C	G	G	T	C	C	C	C	A	A	A	G	A	C	G	T	G	C	T
M. bovis AF2122/97	T	A	C	C	C	T	T	G	G	G	T	C	C	C	C	G	A	A	G	A	C	A	T	G	C	T
M. bovis BCG 1173P2	T	A	C	C	C	T	T	G	G	G	T	C	C	C	C	G	A	A	G	A	C	A	T	G	C	T
M. africanum GM041182	T	A	C	C	T	T	T	C	G	G	T	C	C	C	C	G	A	A	G	A	C	G	T	G	C	T
M. orygis 51145	G	A	C	C	C	C	T	G	G	G	T	G	C	C	C	G	A	T	T	A	C	G	T	G	C	T
M. canettii CIPT140010059	T	G	T	C	C	T	C	C	G	C	C	C	C	C	A	G	A	A	G	A	C	G	C	G	T	T
M. caprae ATCC BAA-824	T	A	C	C	C	T	T	G	G	G	T	C	C	C	C	G	A	A	G	C	C	G	T	G	C	T
M. microti ATCC 35782	T	A	C	C	C	T	T	G	G	G	T	C	C	C	C	G	G	A	G	A	C	G	T	G	C	T
M. mungi BM22813	T	A	C	C	C	T	T	C	G	G	T	C	C	C	C	G	A	A	G	A	C	G	T	A	C	T
M. pinnipedii MP1	T	A	C	C	C	T	T	G	G	G	T	C	C	C	C	G	G	A	G	A	C	G	T	G	C	T
Lineage 1 (NC_021740.1)	T	A	C	C	-	-	-	-	G	G	T	-	-	-	C	G	A	A	G	A	C	G	T	G	C	T
Lineage 2 (CP007803.1)	T	A	C	G	-	-	-	-	G	G	T	-	-	-	C	A	A	A	G	A	C	G	T	G	C	T
Lineage 3 (CP010340.1)	T	A	C	C	-	-	-	-	G	G	T	-	-	-	C	A	A	A	G	A	C	G	T	G	C	T
Lineage 4 (NC_020559.1)	T	A	C	C	-	-	-	-	G	G	T	-	-	-	C	A	A	A	G	A	C	G	T	G	C	T
Lineage 5 (SRR998618)	T	A	C	C	C	T	T	C	G	G	T	C	C	T	C	G	A	A	G	A	C	G	T	G	C	C
Lineage 6 (SRR1577831)	T	A	C	C	T	T	T	C	G	G	T	C	C	C	C	G	A	A	G	A	C	G	T	G	C	T
Lineage 7 (ERR1971865)	T	A	C	C	C	T	T	C	C	G	T	C	T	C	C	A	A	A	G	A	C	G	T	G	C	T
Lineage 8 (SRR10828835)	T	A	C	C	-	-	T	C	G	G	T	C	C	C	C	G	A	A	G	A	C	G	T	G	C	T
Lineage 9 (ERR4162024)	T	A	C	C	C	T	T	C	G	G	T	C	C	C	C	G	A	A	G	A	T	G	T	G	C	T

The nucleotides in red color show unique SNPs for a particular species/lineage, represented in a different colored block.

'-' indicates that the nucleotides for those positions were not extracted in the analysis.

The reference nucleotide positions for the SNPs within *gyrB*, *PPE55*, *Rv2042c* and *leuS* genes are relative to *M. tuberculosis* H37Rv (accession # NC_000962.3, Cole et al. (35)), while the positions within *mmpL6* and *mmpS6* genes are provided according to TbD1 region (accession # AJ426486.1, Brosch et al. (36)).

*It is noted that the numerical positions of 213, 277, 432, 492, 513, 768, 870, 873, 1068, 1167, 1207, 1215, 1478 and 1854 for the *gyrB* gene in this study are corresponding to 456, 520, 675, 735, 756, 1011, 1113, 1116, 1311, 1410, 1450, 1458, 1721 and 2097, respectively, as per accession # L27512.1 (Kasai et al. (46)).


Colour by Origin
Human
Animal
Colour by Geographic location
Canada
India
Norway
Switzerland
UK
USA
bootstrap
80
85
90
95
100
Reference
1701554
ERR5336167
2001463
900124
ERR5336162
1601614
1601335
1800722
1701797
1901041
2101097
2001464
1801127
ERR015582
2100119
ERR5336157
1800965
1900228
ERR2659156
ERR2659154
ERR2659153
1900323
900125
SRR5642714
SRR5642713
SRR5642718
1900698
1600127
1901282
SRR5642716
SRR10251195
SRR10251202
SRR10251187
SRR1025186
2100725
SRR10251190
SRR10251197
SRR10251200
SRR10251192
SRR10251185
SRR10251189
SRR10251191
1800300
1800448
900123
SRR5642711
ERR5336161
2100430
SRR16643349
2001480
2001439
SRR5642712
ERR5336158
2101352
1800446
2100976
1901423
2001127
2001229
2000368
1801256
1900059
1900097
2001465
2100187
2101452
1900674
2001479
1701642
1801535
1801225
2200263
2200196
2100977
1601131
1601221
1700074
1901424
1900817
2201129
1800159
2100376
SRR20622096
SRR20622097
900942
2101375
2101205
2000728
1501282
SRR17254977
SRR17254976
SRR17089003
SRR17254975
SRR17089001
SRR17089002
SRR9157804
ERR2659155
SRR10321130
SRR10321128
SRR10251188
SRR10321149
1701798
SRR20607504
SRR10321134
SRR10321143
SRR5642717
SRR5642715
SRR10251203
SRR10251199
SRR10251196
SRR10251201
SRR10251194
SRR10251198
SRR10251193
Tree scale: 0.01