SNP-IT Tool for Identifying Subspecies and Associated Lineages of *Mycobacterium tuberculosis* Complex

Samuel Lipworth,¹ Rana Jajou,¹ Albert de Neeling, Phelim Bradley, Wim van der Hoek, Gugu Maphalala, Maryline Bonnet, Elizabeth Sanchez-Padilla, Roland Diel, Stefan Niemann, Zamin Iqbal, Grace Smith, Tim Peto, Derrick Crook, Timothy Walker,² Dick van Soolingen²

The clinical phenotype of zoonotic tuberculosis and its contribution to the global burden of disease are poorly understood and probably underestimated. This shortcoming is partly because of the inability of currently available laboratory and in silico tools to accurately identify all subspecies of the *Mycobacterium tuberculosis* complex (MTBC). We present SNPs to Identify TB (SNP-IT), a single-nucleotide polymorphism–based tool to identify all members of MTBC, including animal clades. By applying SNP-IT to a collection of clinical genomes from a UK reference laboratory, we detected an unexpectedly high number of *M. orygis* isolates. *M. orygis* is seen at a similar rate to *M. bovis*, yet *M. orygis* cases have not been previously described in the United Kingdom. From an international perspective, it is possible that *M. orygis* is an underestimated zoonosis. Accurate identification will enable study of the clinical phenotype, host range, and transmission mechanisms of all subspecies of MTBC in greater detail.

*Mycobacterium tuberculosis* complex (MTBC) encompasses a group of organisms that cause tuberculosis (TB) in humans and animals. TB in humans is caused mainly by *M. tuberculosis* but also by other members of MTBC, including the less well understood animal-associated subspecies *M. bovis, M. caprae, M. pinnipedi,* *M. suricattae, M. orygis, M. microti,* and *M. mungi* (1–6). The global burden of zoonotic TB is thought to be both underestimated and increasing (7); however, accurate assessment of prevalence is made difficult by a lack of clinical diagnostic tools and surveillance (8).

Efforts to differentiate members of MTBC and study the phylogeny of the complex have thus far included analysis of large genomic deletions (9), variable-number tandem-repeats (VNTR), spacer oligonucleotide typing (spoligotyping), multilocus sequence typing, and, more recently, single-nucleotide polymorphism (SNP)–based phylogenies (10). Numerous tools now exist that make in silico predictions of lineages within the complex from whole-genome sequencing (WGS) data using a variety of approaches, including the detection of single SNPs from both assembled and mapped genomes, comparison of de Bruijn graphs, and MinHash-based comparisons (11–14). None of these tools has yet been calibrated to reliably differentiate among all subspecies, particularly the animal-associated ones, whose incidence and clinical significance are likely to be underestimated as a result.

The host ranges of the various MTBC subspecies differ, which has serious implications for contact investigations and source case finding. For example, *M. microti* is found in wild cats and rodents and causes human infection, usually in association with rodent contact (15). In contrast with infections caused by *M. bovis*, most *M. microti* infections have been reported to cause pulmonary TB, which raises the possibility of onward transmission, although such transmission has not yet been reported (16). *M. pinnipedi*, which causes TB in seals, is sometimes transmitted to humans during outbreaks in zoos or wildlife parks (17). Although isolated mostly from gazelle species, *M. orygis* has also been seen in humans in recent years, although how humans contract this bacterium is still unclear (2). *M. mungi* causes

¹These authors contributed equally to this article.
²Joint senior authors.
disease in banded mongooses, the dassie bacillus causes disease in rock hyraxes, and M. suricattae causes disease in meerkats, but none of these bacteria is currently known to cause disease in humans (3,5).

The spectrum of clinical phenotypes associated with human infection by MTBC animal lineages is largely unknown, partly because the identification of these organisms is currently difficult. Accurate identification of the causative subspecies in all cases would enable characterization of disease associated with animal lineages and of diversity in clinical phenotypes, which would contribute to better disease management. A higher level of knowledge on the spread and host range of the subspecies would also provide a better basis on which to study the history of the evolutionary development of the complex as a whole. Now that routine WGS is being performed by Public Health England (PHE) for all MTBC isolates, we sought to use these data to estimate the burden of animal-associated TB in England. Therefore, we identified a broad panel of SNPs that define each subspecies, lineage, and sublineage within the MTBC and assessed them using a new SNP-based tool, SNPs to Identify TB (SNP-IT).

Materials and Methods

Calibration Set
We defined a set of isolates (N = 323) from which to identify SNPs associated with subspecies, lineages, and sublineages within the MTBC (Figure 1). We identified isolates from the collection of the National Institute for Public Health and the Environment (RIVM; Bilthoven, Netherlands) using a combination of spoligotyping patterns, SNPs, restriction fragment length polymorphism (RFLP) patterns, the hybridization patterns in the HAIN Genotype MTBC assay, polymorphic GC-rich repeat sequence (PGRS) profiles, and VNTR patterns in accordance with current and previous standard practice (Appendix Table 1, http://wwwnc.cdc.gov/EID/article/25/3/18-0894-App1.pdf). We identified isolates from a whole-genome sequencing archive held at the University of Oxford (Oxford, UK) using the SNP typing system of Stucki et al. (18) for Mtb lineages 1–4 and by clustering on a maximum likelihood tree with the isolates from the Netherlands for lineages 5 and 6. In addition, we used published strains (n = 5) to be able to include M. suricattae, M. mungi, and the dassie bacillus, which were not present in the Oxford or RIVM collections. The nomenclature we adopted for this study is summarized in Appendix Table 2.

Bioinformatics
We applied parallel bioinformatics approaches to assess applicability across pipelines. As such, we independently mapped reads from Illumina platforms (https://www.illumina.com) to 2 different versions of the H37Rv reference genome. We mapped reads to NC000962.3 with Breseq version 0.28.1 (http://barricklab.org/twiki/bin/view/Lab/ToolsBacterialGenomeResequencing), using a minimum
allele frequency of 80% and minimum coverage of 5 times for SNP calls. Separately, we mapped reads again to NC000962.2, for which we used Snippy version 3.1 with default settings (minimum coverage 10 times, minimum allele frequency 90%) (19,20). We extracted all SNPs shared exclusively by isolates of each subspecies, lineage, and sublineage identified by both pipelines. The lineage-defining positions for lineage 4 are not variants with respect to the reference, itself lineage 4, but are uniquely conserved positions. We therefore identified these positions by mapping a core SNP alignment to a maximum-likelihood tree using Mesquite version 3.30 (21). These nucleotide loci were added to the catalog of phylogeny-determining SNPs.

All newly sequenced genomes are available from the National Center for Biotechnology Information under project accession no. PRJNA418900. The SNP-IT tool, including all relevant reference libraries used for this study, is available as an open-source package online (https://github.com/samlipworth/snpit).

Validation Set
To validate the algorithm, we compiled an independent collection of genomes (N = 511) using clinical isolates sequenced by PHE Birmingham, UK, identified as MTBC and not included in the calibration set (Figure 1). We augmented them with data from the European Nucleotide Archive and the National Center for Biotechnology Information Sequence Read Archive to increase the representation of animal subspecies (N = 103; Appendix Table 2). To maximize inclusion of animal isolates from the public archives, we used the new Colored Bloom Graph (CBG) software (22). Using CBG, we searched a snapshot of the Sequence Read Archive (to December 2016, N = 455,632) with our new set of reference kmers for Mykrobe predictor (see Comparison with Existing Tools).

We compared FASTA files of the whole genome (https://www.ebi.ac.uk/Tools/sss/fasta) to the catalog of phylogeny-determining SNPs to make predictions for PHE isolates, whereas for isolates downloaded from the nucleotide archive, we created only limited variant calling format files using Snippy (only SNPs with respect to the reference genome were included to increase computational efficiency). To ensure that genomic loci defining lineage 4 were included, we used a mutated reference genome to create these limited variant calling format files. We compared SNPs in the query sample with reference libraries of lineage-specific SNPs for each clade. We assigned query genomes to particular subspecies or lineages if ≥10% of lineage- or subspecies-specific SNPs were detected in the strain in question. We assessed all predictions against the maximum-likelihood phylogeny. For *M. mungi*, we could locate only 1 genome in the public sequence libraries, so we could not validate this subspecies.

Clinical Isolates
To assess the caseload across the different members of the MTBC seen by the PHE laboratory in Birmingham, we applied the algorithm to 3,128 MTBC genome sequences from consecutively obtained clinical isolates. H37Rv is routinely sequenced by the laboratory on WGS plates; these isolates were not removed, and their identification served as an internal control.

Comparison with Existing Tools
We first compared strain characterization by our new SNP-IT tool with those of KvarQ (https://github.com/kvarq/kvarq), TB-Profiler (http://tbdr.lshtm.ac.uk), and Mykrobe predictor (http://www.mykrobe.com/products/predictor) on default settings and then after integrating our updated SNP library. To enable our new data to be integrated with published SNP libraries (23) and for practical reasons when modifying existing tools, we created a minimal SNP dataset. We filtered our larger SNP catalog for synonymous SNPs that occurred in coding regions (as annotated by SnpEff version 4.3 [24]) and selected 1 representative SNP for each subspecies, lineage, and sublineage at random. We then modified the existing software packages to include reference SNPs (or kmers for Mykrobe predictor) for the subspecies, lineages, or sublineages that they initially failed to identify.

Results
Calibration and Validation
In total, we identified 13,893 SNPs (median of 229 SNPs per group, interquartile range 296) as predictive of taxonomic and phylogenetic groups of interest (Appendix Table 3). The greatest number of phylogenetic SNPs was seen in *M. canetti* (n = 6,837) and the fewest in *M. bovis* (n = 23). Subspecies that arise from common deep branches, such as *M. microti* and *M. pinnipedii* (Figure 2), have lower numbers of unique phylogenetic SNPs (n = 128 for *M. microti* and n = 301 for *M. pinnipedii*) than those that do not, such as *M. orygis* (n = 781). All predictions made by SNP-IT across all the subspecies, lineages, and sublineages were consistent with the maximum-likelihood phylogeny for all isolates in the validation set (Table 1).

Determining Prevalence of Animal Subspecies in a Collection of Clinical Isolates
We retrospectively applied SNP-IT to clinical isolates sequenced as part of the routine PHE diagnostic workflow in Birmingham to estimate the prevalence of the animal subspecies among MTBC samples. Of 3,128 samples from 2,106 patients for which there was a whole-genome sequence available, we identified 24 as *M. orygis*, 3 as *M. microti*, 34 as *M. bovis*, and 1 as *M. caprae* (Table 2).
the case of *M. orygis*, we further investigated whether there was any genomic signal of possible person-to-person transmission. We identified 2 such instances, 1 in which the pairwise genetic distance between 2 patients was 0 SNPs and a second in which it was 6 SNPs (Appendix Figure 1).

**Phylogenetic SNPs in Drug Resistance–Associated Genes**

Using a previously published list of drug resistance–associated genes for *M. tuberculosis* (25), we searched all subspecies for phylogenetic SNPs in drug resistance–associated genes (Appendix Table 4). All subspecies contain unique phylogenetic SNPs (N = 95 in total) in these genomic regions, but on the basis of our data, we were unable to determine whether any of these mutations are linked to lineage-specific resistance because we did not have the corresponding phenotypic drug susceptibility testing data.

**Comparison with Existing Software**

Compared with SNP-IT for the clinical set of isolates, Mykrobe predictor reported *M. orygis* as *M. tuberculosis* West African lineage and *M. pinnipedii* as *M. microti*. KvarQ identified all animal-associated subspecies only as “animal lineage.” TB-Profiler was unable to delineate among animal subspecies, which were all reported as *M. bovis/M. tuberculosis* West African lineage.

After we modified the KvarQ, TB-Profiler, and Mykrobe predictor databases with our minimal SNP catalog, all systems agreed on the identity of all the MTBC isolates in the clinical set. SNP-IT was unable to identify 10 samples because <10% of type-specific SNPs were present in these strains. This result was because our pipeline made no call at the lineage informative sites because of the presence of a minor allele, most likely the result of contamination or a mixture of 2 strains in the sample. However, TB-Profiler and Mykrobe predictor were both able to identify 2 of these isolates as mixed Beijing lineage/*M. orygis* using our new minimal SNP dataset.

**Discussion**

SNP typing is a powerful method for discriminating among members of MTBC, which are often not discernible...
by conventional laboratory methods. The SNP databases of Stucki, Coll, and Comas are currently used as the knowledge base for KvarQ, TB-Profiler, and Mykrobe predictor (18, 23, 26). None of these databases, however, provides adequate resolution for the animal subspecies. In contrast, SNP-IT was able to assign subspecies, lineages, and sublineages to all samples in the validation set with 100% accuracy compared with maximum-likelihood phylogeny. Implementing this fine-resolution algorithm into a routine diagnostic workflow would be a major step toward understanding the epidemiology and pathogenicity of the less common members of MTBC. All 3 existing systems tested (KvarQ, Mykrobe predictor, and TB-Profiler) were identical in performance when given the same SNP reference database, demonstrating that the clinically meaningful differences highlighted in a recent review are easily ameliorated (27).

By applying SNP-IT to a clinical dataset, we discovered an unexpectedly high number of animal subspecies among MTBC isolates, particularly M. orygis, from humans in the United Kingdom. Human-to-animal transmission has been described in 1 case in New Zealand (30). Given that zoonotic TB is associated with higher rates of extrapulmonary disease and may be less likely to grow in culture (31, 32), retrospective interrogation of WGS libraries, as in this study, is likely to underestimate the true burden of disease.

Given the large amount of resources aimed at controlling bovine TB, it is noteworthy that another zoonosis is seen at a similar rate in this collection of clinical isolates. This finding raises questions about the host range and transmission of M. orygis, with potential implications for TB control both in animals and humans. To recognize the particulars of the clinical phenotype, epidemiology, and optimal management strategy of M. orygis infection, it is first crucial to accurately distinguish these cases from M. tuberculosis West African lineages (5, 6). This discernment is currently not possible by either the Hain Genotype MTBC molecular probe or existing SNP-based platforms. We identified 2 pairs of nearly identical M. orygis isolates that could be compatible with either person-to-person transmission or, possibly, common exposure to the same infected animal. From an international perspective, the role of M. orygis in zoonotic transmission in Africa, Asia, and other high-prevalence settings with extensive animal contact is poorly understood and may warrant further investigation.

### Table 1. Comparison between speciation calls for 614 MTBC samples in validation set made by SNP-IT and position on maximum-likelihood phylogenetic tree*

<table>
<thead>
<tr>
<th>Species (lineage)</th>
<th>SNP-IT typing calls</th>
<th>Maximum-likelihood calls</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>M. tuberculosis</td>
<td>22</td>
<td>22</td>
<td>100%</td>
</tr>
<tr>
<td>M. bovis</td>
<td>15</td>
<td>15</td>
<td>100%</td>
</tr>
<tr>
<td>M. orygis†</td>
<td>31</td>
<td>31</td>
<td>100%</td>
</tr>
<tr>
<td>M. microti†</td>
<td>18</td>
<td>18</td>
<td>100%</td>
</tr>
<tr>
<td>M. canetti†</td>
<td>34</td>
<td>34</td>
<td>100%</td>
</tr>
<tr>
<td>M. pinnipedi†</td>
<td>8</td>
<td>8</td>
<td>100%</td>
</tr>
<tr>
<td>M. caprae†</td>
<td>15</td>
<td>15</td>
<td>100%</td>
</tr>
<tr>
<td>Dassie bacillus†</td>
<td>2</td>
<td>2</td>
<td>100%</td>
</tr>
<tr>
<td>M. suricatta†</td>
<td>2</td>
<td>2</td>
<td>100%</td>
</tr>
</tbody>
</table>

*Numbers in parentheses represent the lineage numbering scheme of Coll for the major lineages 1–7 and Stucki for the lineage 4 sublineages. BCG, bacillus Calmette–Guérin; MTBC, Mycobacterium tuberculosis complex; SNP, single-nucleotide polymorphism; SNP-IT, SNPs to Identify TB.
†Validation set for subspecies augmented with published strains.
‡No subspecies call made.
more healthcare systems begin to routinely use WGS, there may not be able to correctly identify isolates that originate from deeper phylogenetic branches than those in our calibration set. Since we have sought to calibrate SNP-IT using the most diverse collection of samples available to us, it is that, although we have sought to calibrate SNP-IT using SNPs such that this weakness is not an issue unless it occurs at most of these positions. An additional limitation is that, although we have sought to calibrate SNP-IT using the most diverse collection of samples available to us, it may not be able to correctly identify isolates that originate from deeper phylogenetic branches than those in our calibration set.

In conclusion, in this study we demonstrate a higher-than-expected burden of zoonotic TB in a large collection of clinical isolates from the United Kingdom. The SNP-IT tool we have developed will help researchers to examine the epidemiology of zoonotic TB in a global context, as well as optimizing the disease’s clinical management. As more healthcare systems begin to routinely use WGS, there is an opportunity to accurately diagnose the causative subspecies of TB in all cases, which will enable identification of previously underrecognized zoonoses and reverse zoonoses and implementation of control interventions in the interests of One Health.

Acknowledgments

We thank Phillip Fowler his work improving the SNP-IT Python package and Pretin Davda for his assistance in linking isolates to patient level.

This research was supported by the National Institute for Health Research (NIHR) Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance at the University of Oxford in partnership with Public Health England (PHE) and by Oxford NIHR Biomedical Research Centre. T.P. is a NIHR senior investigator. The report presents independent research funded by NIHR. The views expressed in this publication are those of the authors and not necessarily those of the National Health Service, NIHR, the Department of Health, or PHE. Z.I. is a Sir Henry Dale Fellow jointly funded by the Wellcome Trust and the Royal Society (grant no. 102541/Z/13/Z).

About the Author

Dr. Lipworth is a clinical fellow with the Modernising Medical Microbiology research group in the Nuffield Department of Medicine, University of Oxford. His research interests lie in the clinical applications of whole-genome sequencing, particularly diagnostics and molecular epidemiology. Ms. Jajou is a PhD candidate supervised by Dick van Soolingen at the National Institute for Public Health and the Environment in the Netherlands. Her primary research interest is the use of whole-genome sequencing to improve diagnostics and public health practice in tuberculosis.

References


19. Address for correspondence: Samuel Lipworth, University of Oxford, Nuffield Department of Medicine, John Radcliffe Hospital, Oxford OX3 9DU, UK; email: samuel.lipworth@medsci.ox.ac.uk