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 Soolingen[2]

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 MTBCin 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. pinnipedii*, *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 unassembled and mapped genomes, comparison of de Brujin 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. pinnipedii*, 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

Author affiliations: University of Oxford, Oxford, UK (S. Lipworth, T. Peto, D. Crook, T. Walker); National Institute for Public Health and the Environment, Bilthoven, the Netherlands (R. Jajou, A. de Neeling, W. van der Hoek, D. van Soolingen); Wellcome Trust Centre for Human Genetics, Oxford (P. Bradley); National Reference Laboratory, Ministry of Health, Mbabane, Swaziland (G. Maphalala); Epicentre, Paris, France (M. Bonnet, E. Sanchez-Padilla); University of Kiel, Kiel, Germany (R. Diel); Borstel Research Centre, Borstel, Germany (S. Niemann); European Bioinformatics Institute, Cambridge, UK (Z. Iqbal); Public Health England, Birmingham, UK (G. Smith)

DOI: https://doi.org/10.3201/eid2503.180894

[1]These authors contributed equally to this article.

[2]Joint 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

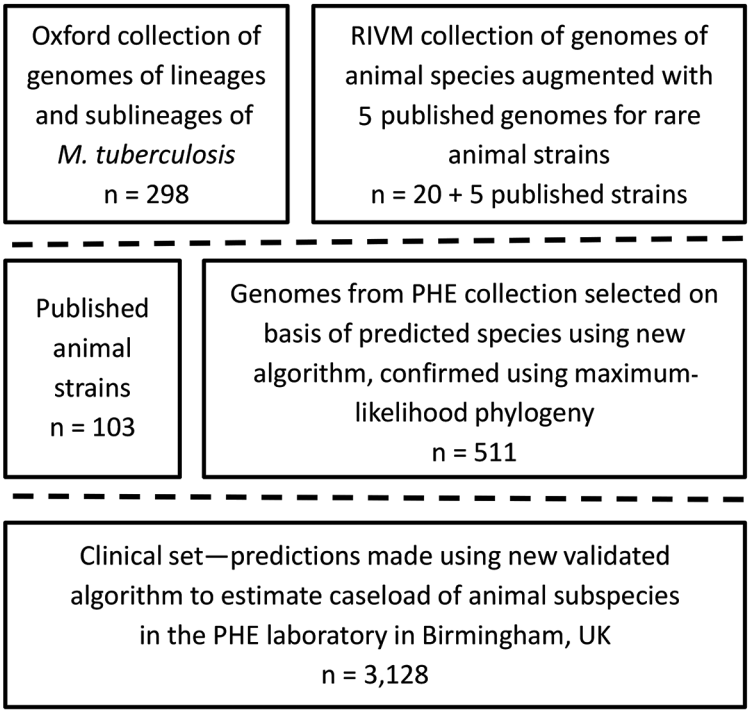


Figure 1. Description of the *Mycobacterium tuberculosis* complex datasets used in the 3 stages of calibration, validation, and application to a clinical set of the new SNPs to Identify TB tool. PHE, Public Health England (Birmingham, UK); RIVM, Netherlands National Institute for Public Health and the Environment (Bilthoven, the Netherlands); SNP, single-nucleotide polymorphism.

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. canettii* (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). In

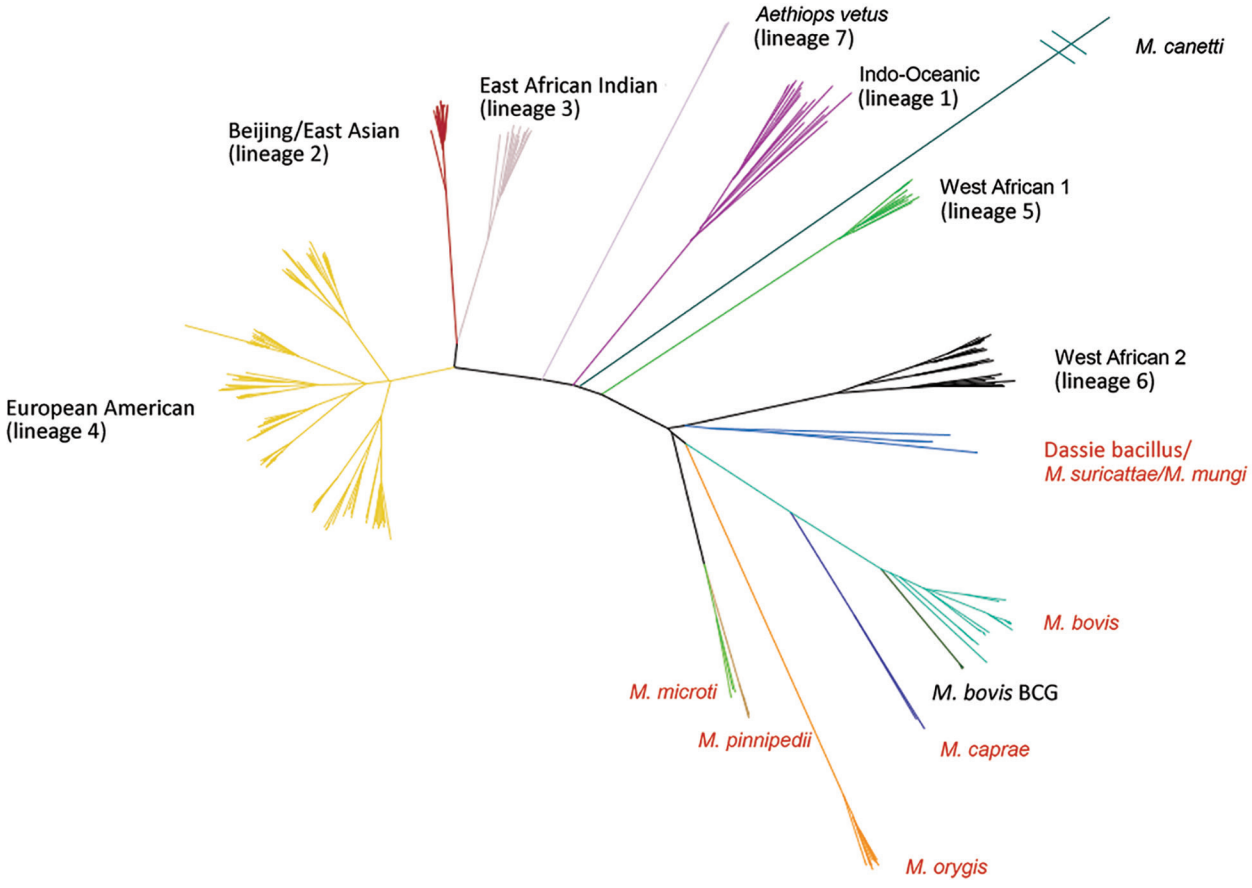


Figure 2. Maximum-likelihood tree built from 70,144 informative positions from whole-genome sequences of all 323 *Mycobacterium tuberculosis* complex samples in the calibration set for the new SNPs to Identify TB tool. Lineages are of *Mycobacterium tuberculosis*. Underlined text denotes animal subspecies. BCG, bacillus Calmette–Guérin; SNP, single-nucleotide polymorphism.

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

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

Species (lineage)	SNP-IT typing calls	Maximum-likelihood calls	% Correct
Mycobacterium bovis BCG	22	22	100%
M. bovis	15	15	100%
M. orygis†	31	31	100%
M. microti†	18	18	100%
M. canettii†	34	34	100%
M. pinnipedii†	8	8	100%
M. caprae†	15	15	100%
Dassie bacillus†	2	2	100%
M. suricattae†	2	2	100%
M. tuberculosis			
Indo Oceanic (lineage 1)	44	44	100%
Beijing/East Asian (lineage 2)	42	42	100%
East African Indian (lineage 3)	46	46	100%
European American (lineage 4)‡	18	18	100%
Lineage 4 sublineages			
Ghana (4.1)	12	12	100%
X-type (4.1.1)	45	45	100%
Haarlem (4.1.2.1)	45	45	100%
Ural (4.2.1)	19	19	100%
Tur (4.2.2.1)	25	25	100%
LAM (4.3)	37	37	100%
S-type (4.4.1.1)	45	45	100%
Uganda (4.6.1)	18	18	100%
Cameroon (4.6.2.2)	32	32	100%
West African 1 (lineage 5)	13	13	100%
West African 2 (lineage 6)	11	11	100%
M. aethiops vetus (lineage 7)	15	15	100%

*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 sublineage call made.

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. This recently described member of the complex has a host range that includes waterbucks, gazelles, rhesus monkeys, cows, and rhinoceri (*2*,*28*,*29*). Several human cases have been described in patients in the Netherlands of South/Southeast Asian origin (*2*), but no cases have been described 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 2. Speciation predictions for collection of 3,128 clinical MTBC isolates from 2,106 patients using SNP-IT*

Species and subspecies (lineage)	No. isolates (no. patients)
Mycobacterium tuberculosis	
Indo-Oceanic (lineage 1)	240 (208)
Beijing/East Asian (lineage 2)	242 (175)
East African Indian (lineage 3)	775 (644)
European American (lineage 4); no sublineage call made†	512 (+ 368 H37Rv‡) (336)
Lineage 4 sublineages	
Ghana (4.1)	5 (4)
X-type (4.1.1)	197 (159)
Haarlem (4.1.2.1)	266 (213)
Ural (4.2.1)	43 (34)
Tur (4.2.2.1)	41 (36)
LAM (4.3)	213 (159)
S-type (4.4.1.1)	60 (45)
Uganda (4.6.1)	13 (12)
Cameroon (4.6.2.2)	40 (32)
West African 1 (lineage 5)	4 (2)
West African 2 (lineage 6)	11 (9)
No call made	10 (10)
M. bovis BCG	26 (20)
M. bovis	34 (28)
M. orygis	24 (19)
M. microti	3 (2)
M. caprae	1 (1)
Total	3,128 (2,106)

*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.
†These samples belong to lineage 4 but not to one of the named sublineages.
‡SNP-IT correctly identified 880 isolates as being lineage 4; however, 368 of these were identified as being H37Rv when we unblinded ourselves to their laboratory records.

All the animal subspecies had phylogenetic SNPs in drug resistance–associated genes. When these genes are not known to be associated with drug resistance, they can be helpfully annotated as such by diagnostic algorithms and excluded for the purpose of predicting susceptibility. An unavoidable weakness of any SNP-based approach is its vulnerability to null-calls as a result of minor alleles at informative positions or a lack of coverage. SNP-IT uses the entire library of subspecies/lineage/sublineage-defining 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

1. Magee J, Ward A. Mycobacterium. In: Whitman WB, editor. Bergey's Manual of Systematics of Archaea and Bacteria. Hoboken (NJ): John Wiley & Sons; 2015. p. 2–4.
2. van Ingen J, Rahim Z, Mulder A, Boeree MJ, Simeone R, Brosch R, et al. Characterization of *Mycobacterium orygis* as *M. tuberculosis* complex subspecies. Emerg Infect Dis. 2012; 18:653–5. http://dx.doi.org/10.3201/eid1804.110888
3. Parsons SDC, Drewe JA, Gey van Pittius NC, Warren RM, van Helden PD. Novel cause of tuberculosis in meerkats, South Africa. Emerg Infect Dis. 2013;19:2004–7. http://dx.doi.org/10.3201/eid1912.130268
4. Mostowy S, Cousins D, Behr MA. Genomic interrogation of the dassie bacillus reveals it as a unique RD1 mutant within the *Mycobacterium tuberculosis* complex. J Bacteriol. 2004;186: 104–9. http://dx.doi.org/10.1128/JB.186.1.104-109.2003
5. Alexander KA, Laver PN, Michel AL, Williams M, van Helden PD, Warren RM, et al. Novel *Mycobacterium tuberculosis* complex pathogen, *M. mungi*. Emerg Infect Dis. 2010;16:1296–9. http://dx.doi.org/10.3201/eid1608.100314

6. Fabre M, Hauck Y, Soler C, Koeck J-L, van Ingen J, van Soolingen D, et al. Molecular characteristics of "*Mycobacterium canettii*" the smooth *Mycobacterium tuberculosis* bacilli. Infect Genet Evol. 2010;10:1165–73. http://dx.doi.org/10.1016/j.meegid.2010.07.016
7. Olea-Popelka F, Muwonge A, Perera A, Dean AS, Mumford E, Erlacher-Vindel E, et al. Zoonotic tuberculosis in human beings caused by *Mycobacterium bovis*—a call for action. Lancet Infect Dis. 2017;17:e21–5. http://dx.doi.org/10.1016/S1473-3099(16)30139-6
8. Müller B, Dürr S, Alonso S, Hattendorf J, Laisse CJM, Parsons SDC, et al. Zoonotic *Mycobacterium bovis*–induced tuberculosis in humans. Emerg Infect Dis. 2013;19:899–908. http://dx.doi.org/10.3201/eid1906.120543
9. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, et al. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. Proc Natl Acad Sci U S A. 2002;99:3684–9. http://dx.doi.org/10.1073/pnas.052548299
10. Jagielski T, van Ingen J, Rastogi N, Dziadek J, Mazur PK, Bielecki J. Current methods in the molecular typing of *Mycobacterium tuberculosis* and other mycobacteria. Biomed Res Int. 2014;2014:645802. http://dx.doi.org/10.1155/2014/645802
11. Steiner A, Stucki D, Coscolla M, Borrell S, Gagneux S. KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. BMC Genomics. 2014;15:881. http://dx.doi.org/10.1186/1471-2164-15-881
12. Coll F, McNerney R, Preston MD, Guerra-Assunção JA, Warry A, Hill-Cawthorne G, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med. 2015;7:51. http://dx.doi.org/10.1186/s13073-015-0164-0
13. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17:132. http://dx.doi.org/10.1186/s13059-016-0997-x
14. Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, et al. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis.* Nat Commun. 2015;6:10063. http://dx.doi.org/ 10.1038/ncomms10063
15. van Soolingen D, van der Zanden AGM, de Haas PEW, Noordhoek GT, Kiers A, Foudraine NA, et al. Diagnosis of *Mycobacterium microti* infections among humans by using novel genetic markers. J Clin Microbiol. 1998;36:1840–5.
16. Emmanuel FX, Seagar A-L, Doig C, Rayner A, Claxton P, Laurenson I. Human and animal infections with *Mycobacterium microti*, Scotland. Emerg Infect Dis. 2007;13:1924–7. http://dx.doi.org/10.3201/eid1312.061536
17. 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.
18. Stucki D, Malla B, Hostettler S, Huna T, Feldmann J, Yeboah-Manu D, et al. Two new rapid SNP-typing methods for classifying *Mycobacterium tuberculosis* complex into the main phylogenetic lineages. PLoS One. 2012;7:e41253. http://dx.doi.org/10.1371/journal.pone.0041253
19. Seemann T. Snippy: Rapid haploid variant calling and core SNP phylogeny; 2015 [cited 2017 May 1]. https://github.com/tseemann/snippy
20. Deatherage DE, Barrick JE. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using *breseq*. Methods Mol Biol. 2014;1151:165–88. http://dx.doi.org/10.1007/978-1-4939-0554-6_12
21. Maddison WP, Maddison DR. Mesquite: a modular system for evolutionary analysis; 2017 [cited 2017 May 1]. http://mesquiteproject.org
22. Bradley P, den Bakker H, Rocha E, McVean G, Iqbal Z. Real-time search of all bacterial and viral genomic data. bioRxiv; 2017 [cited 2017 Dec 18]. https://www.biorxiv.org/content/early/2017/12/18/234955.abstract
23. Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, et al. A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. Nat Commun. 2014;5:4812. http://dx.doi.org/10.1038/ncomms5812
24. Cingolani P, Platts A, Wang L, 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. http://dx.doi.org/10.4161/fly.19695
25. Walker TM, Kohl TA, Omar SV, Hedge J, Del Ojo Elias C, Bradley P, et al.; Modernizing Medical Microbiology (MMM) Informatics Group. Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. Lancet Infect Dis. 2015;15:1193–202. http://dx.doi.org/10.1016/S1473-3099(15)00062-6
26. Comas I, Homolka S, Niemann S, Gagneux S. Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. PLoS One. 2009;4:e7815. http://dx.doi.org/10.1371/journal.pone.0007815
27. Schleusener V, Köser CU, Beckert P, Niemann S, Feuerriegel S. *Mycobacterium tuberculosis* resistance prediction and lineage classification from genome sequencing: comparison of automated analysis tools. Sci Rep. 2017;7:46327. http://dx.doi.org/10.1038/srep46327
28. Thapa J, Paudel S, Sadaula A, Shah Y, Maharjan B, Kaufman GE, et al. *Mycobacterium orygis*–associated tuberculosis in free-ranging rhinoceros, Nepal, 2015. Emerg Infect Dis. 2016;22:570–2. http://dx.doi.org/10.3201/eid2203.151929
29. Rahim Z, Thapa J, Fukushima Y, van der Zanden AGM, Gordon SV, Suzuki Y, et al. Tuberculosis caused by *Mycobacterium orygis* in dairy cattle and captured monkeys in Bangladesh: a new scenario of tuberculosis in south Asia. Transbound Emerg Dis. 2017;64:1965–9. http://dx.doi.org/10.1111/tbed.12596
30. Dawson KL, Bell A, Kawakami RP, Coley K, Yates G, Collins DM. Transmission of *Mycobacterium orygis* (*M. tuberculosis* complex species) from a tuberculosis patient to a dairy cow in New Zealand. J Clin Microbiol. 2012;50:3136–8. http://dx.doi.org/10.1128/JCM.01652-12
31. Sanoussi CN, Affolabi D, Rigouts L, Anagonou S, de Jong B. Genotypic characterization directly applied to sputum improves the detection of *Mycobacterium africanum* West African 1, under-represented in positive cultures. PLoS Negl Trop Dis. 2017;11:e0005900. http://dx.doi.org/10.1371/journal.pntd.0005900
32. Dürr S, Müller B, Alonso S, Hattendorf J, Laisse CJM, van Helden PD, et al. Differences in primary sites of infection between zoonotic and human tuberculosis: results from a worldwide systematic review. PLoS Negl Trop Dis. 2013;7:e2399. http://dx.doi.org/10.1371/journal.pntd.0002399

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

Article DOI: https://doi.org/10.3201/eid2503.180894

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

Appendix

Post Hoc Analysis of Lineage 4 Genomes

Within the validation set we included 18 genomes from lineage 4 for which no further sublineage could be assigned by SNP-IT and which we therefore defined by the presence of conserved bases relative to the reference. Their position on a maximum likelihood tree revealed that these were phylogenetically distinct from other currently named sublineages of lineage 4. We therefore carried out a post hoc analysis of all genomes in the validation and clinical sets labeled by SNP-IT as being "lineage 4" with no further sublineage name given, comparing their position on the maximum likelihood tree to the nomenclature of Coll (*1*) and Stucki (*2*) (Appendix Figure 1). Of 886 genomes labeled "lineage 4" with no further sublineage identification by SNP-IT, most were lineage 4.10 (N = 817) (Stucki), also known as lineage 4.7 (N = 35), 4.8 (N = 247), and 4.9 (N = 535) (Coll). A minority belonged to lineage 4.5 (N = 35) and lineage 4.6 (N = 27) where they were phylogenetically distinct from the Uganda and Cameroon sublineages. The remainder (N = 7) were not further classifiable from lineage 4 by Coll nomenclature. We subsequently updated our SNP catalog for lineage 4 to include these unnamed sublineages using the same methods as we described for the original calibration set. Lineage 4.9 had 368 samples which were found to be H37Rv when we unblinded ourselves to their corresponding laboratory records.

References

1. Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, et al. A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. Nat Commun. 2014;5:4812. http://dx.doi.org/10.1038/ncomms5812

2. Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, et al. *Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages. Nat Genet. 2016;48:1535–43. http://dx.doi.org/10.1038/ng.3704

3. Somoskovi A, Dormandy J, Parsons LM, Kaswa M, Goh KS, Rastogi N, et al. Sequencing of the pncA gene in members of the *Mycobacterium tuberculosis* complex has important diagnostic applications: Identification of a species-specific pncA mutation in "*Mycobacterium canettii*" and the reliable and rapid predictor of pyrazinamide resistance. J Clin Microbiol. 2007;45:595–9. http://dx.doi.org/10.1128/JCM.01454-06

4. de Jong BC, Antonio M, Gagneux S. *Mycobacterium africanum*—review of an important cause of human tuberculosis in West Africa. PLoS Negl Trop Dis. 2010;4:e744. http://dx.doi.org/10.1371/journal.pntd.0000744

5. van Ingen J, Rahim Z, Mulder A, Boeree MJ, Simeone R, Brosch R, et al. Characterization of *Mycobacterium orygis* as *M. tuberculosis* complex subspecies. Emerg Infect Dis. 2012;18:653–5. http://dx.doi.org/10.3201/eid1804.110888

6. Bruker-Hain Diagnostics. Hain genotype Mtbc [cited 2017 Jan 1]. https://www.hain-lifescience.de/en/products/microbiology/mycobacteria/tuberculosis/genotype-mtbc.html

7. Ross BC, Raios K, Jackson K, Dwyer B. Molecular cloning of a highly repeated DNA element from *Mycobacterium tuberculosis* and its use as an epidemiological tool. J Clin Microbiol. 1992;30:942–6.

8. van Soolingen D, de Haas PE, Hermans PW, Groenen PM, van Embden JD. Comparison of various repetitive DNA elements as genetic markers for strain differentiation and epidemiology of *Mycobacterium tuberculosis*. J Clin Microbiol. 1993;31:1987–95.

9. van Soolingen D, de Haas PEW, Hermans PWM, van Embden JDA. DNA Fingerprinting of *Mycobacterium tuberculosis*. In: Methods in Enzymology. Abelson J, Simon M, Verdine G, Pyle A, editors. New York: Academic Press; 1994. p. 196–205.

10. Van Embden JD, Cave MD, Crawford JT. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J Clin Microbiol. 1993;31:406–9. http://jcm.asm.org/content/31/2/406.short

11. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, et al. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J Clin Microbiol. 1997;35:907–14.

12. 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. http://dx.doi.org/10.1128/JCM.01392-06

13. de Beer JL, Akkerman OW, Schürch AC, Mulder A, van der Werf TS, van der Zanden AGM, et al. Optimization of standard in-house 24-locus variable-number tandem-repeat typing for *Mycobacterium tuberculosis* and its direct application to clinical material. J Clin Microbiol. 2014;52:1338–42. http://dx.doi.org/10.1128/JCM.03436-13

14. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, et al. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat Genet. 2013;45:1176–82. http://dx.doi.org/10.1038/ng.2744

15. Nebenzahl-Guimaraes H, Yimer SA, Holm-Hansen C, de Beer J, Brosch R, van Soolingen D. Genomic characterization of *Mycobacterium tuberculosis* lineage 7 and a proposed name: 'Aethiops vetus'. Microb Genom. 2016;2:e000063. http://dx/doi.org/10.1099/mgen.0.000063

Appendix Table 1. Identification methods for training isolates acquired from the Netherlands National Institute for Public Health and the Environment (RIVM)*

Sample ID	Species	Human/Animal	Sex	Age at diagnosis	Date of isolation	Country of origin	Method of identification
7da4f935-2b2f-4640-9fd6-ea4dac70ef43	*M. microti*	Animal (mouse)	Unknown	Unknown	Unknown	UK	VNTR, spoligo, RFLP, PGRS
a5706cea-2eeb-4dc0-ae64-47e4d6d5fbb3	*M. microti*	Human	Male	57	2002 May 24	NL	VNTR, RFLP
b1fa7231-fa7a-4ba9-afbd-eb2b96444118	*M. microti*	Human	Female	59	2015 Oct 8	NL	VNTR, Hain MTBC
f206f3a5-d127-4698-a0ec-19a524c408ac	*M. canettii*	Human	Unknown	Unknown	1993 Feb 10	NL	VNTR, spoligo, RFLP, pncA (*3*)
46a56d43-ac15-48fd-81b5-846e12bc6d10	*M. canettii*	Human	Male	34	2002 Jun 10	NL	VNTR, RFLP, pncA (3)
fbd67644-f009-4f78-9f14–5a21683d30d5	*M. canettii*	Human	Male	25	2007 Oct 19	NL	VNTR, spoligo, RFLP, pncA (*3*)
7ea3d4d6-a416-49bb-8c01-edf4efbd2cd8	*Mtb* West African 1 (lineage 5)	Human	Unknown	Unknown	1995 Oct 18	NL	VNTR, RFLP
8307cd56-a417-4f33–8c65–335ba11cc0c5	*Mtb* West African 2 (lineage 6)	Human	Unknown	Unknown	1993	NL	VNTR, molecular classification (*4*)
6a63a9bb-6eef-4baa-8be1-b06bb0e31b97	*Mtb* West African 2 (lineage 6)	Human	Unknown	Unknown	1993	NL	VNTR, molecular classification (*4*)
25c63622-317b-4387-8484-da79bf7543a6	*M. pinnipedii*	Animal	Unknown	Unknown	1992–93	Argentina	VNTR, spoligo, RFLP, PGRS
f0d95ba8-52bf-46eb-a196-c60147186e84	*M. pinnipedii*	Animal	Unknown	Unknown	1992–93	Argentina	VNTR, spoligo, RFLP, PGRS
df23492d-6f79-475e-8654-a4256e79c5d4	*M. pinnipedii*	Animal (seal)	Unknown	Unknown	2006	NL	VNTR, spoligo
2fec2a96-cc9a-4895-a01f-24a6c30d4235	*M. caprae*	Human	Male	52	2006 Nov 30	NL	VNTR, Hain MTBC
ee2503c3–52aa-4b64-a5ba-eff58036a05a	*M. caprae*	Human	Female	22	2006 Dec 15	NL	VNTR, Hain MTBC

Sample ID	Species	Human/Animal	Sex	Age at diagnosis	Date of isolation	Country of origin	Method of identification
ed1e9f45-ace3-464d-b48e-b5038a68b227	*M. aethiops vetus* (Mtb lineage 7)	Human	Unknown	Unknown	Unknown	Ethiopia	VNTR
26af0867-691e-41da-aa79-3e531895ec33	*M. aethiops vetus* (Mtb lineage 7)	Human	Unknown	Unknown	Unknown	Ethiopia	VNTR
dbc3bd41-21a6-4164-b4d5-c1c9900d6e8a	*M. orygis*	Human	Male	42	2013 Mar 20	NL	VNTR, pncA (*5*)
e2f10303-774d-4b68-b3fa-d2b00f5eded0	*M. orygis*	Human	Female	81	2013 Jun 19	NL	VNTR, pncA (*5*)
c2140309-094d-46e0-8d18-8689c0475e38	*M. bovis BCG*	Human (vaccination)	Male	7	2015 Dec 10	NL	VNTR, Hain MTBC
c64bc667-ac80-4a90-9a40-685c13e542fb	*M. bovis BCG*	Human (bladder carcinoma)	Male	69	2016 May 19	NL	VNTR, Hain MTBC

*BCG, bacillus Calmette-Guerin; Hain MTBC, Hain Genotype MTBC line probe assay (*6*); NL, Netherlands; PGRS, polymorphic GC-rich repeat sequence (*7,8*); RFLP, restriction fragment length polymorphism (*9,10*); spoligo, spacer oligonucleotide typing (*11*); VNTR, variable number tandem repeat (*12,13*).

Appendix Table 2. Nomenclature for the *Mycobacterium tuberculosis* complex members adopted in this study; *Mtb* lineage numbering shown according to the system proposed by Comas/Gagneux (*14*).

Subspecies	Lineage
M. tuberculosis	Indo-Oceanic (*Mtb* lineage 1) Beijing/East Asian (*Mtb* lineage 2) East African Indian (*Mtb* lineage 3) European American (*Mtb* lineage 4) sublineages: Haarlem, LAM, Cameroon, Ghana, S-type, Tur, Uganda, Ural, X-type West African 1 (*Mtb* lineage 5) West African 2 (*Mtb* lineage 6) *M. aethiops vetus* (*Mtb* lineage 7) (*15*)
M. orygis	
M. bovis	
M. caprae	
M. bovis BCG	
M. pinnipedii	
M. suricattae	
M. mungi	
M. canettii	
Dassie bacillus	

Appendix Table 3. Unique phylogenetic single nucleotide polymorphisms (SNPs) per subspecies/lineage/sublineage

Subspecies/lineage/sublineage	Unique phylogenetic SNPs
Dassie bacillus	323
M. bovis	23
M. bovis BCG	223
M. orygis	781
M. microti	128
M. canettii	6837
M. pinipedii	301
M. caprae	294
M. mungi	609
M. suricattae	510
Indo-Oceanic (Mtb lineage 1)	425
Beijing/East Asian (Mtb lineage 2)	304
East African Indian (Mtb lineage 3)	235
European American (Mtb lineage 4*)	163
Haarlem	59

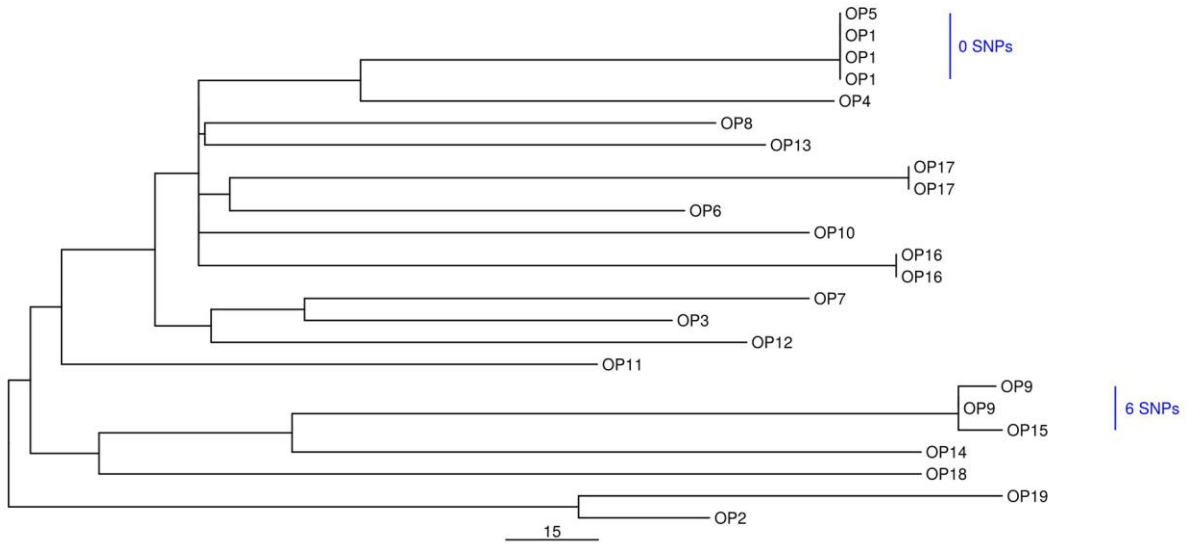
Subspecies/lineage/sublineage	Unique phylogenetic SNPs
LAM	102
Cameroon	172
Ghana	186
Uganda	100
S-type	165
Tur	43
Ural	33
X-type	60
West African 1 (Mtb lineage 5)	657
West African 2 (Mtb lineage 6)	343
M. aethiops vetus (Mtb lineage 7)	817

*Conserved bases with respect to the reference.

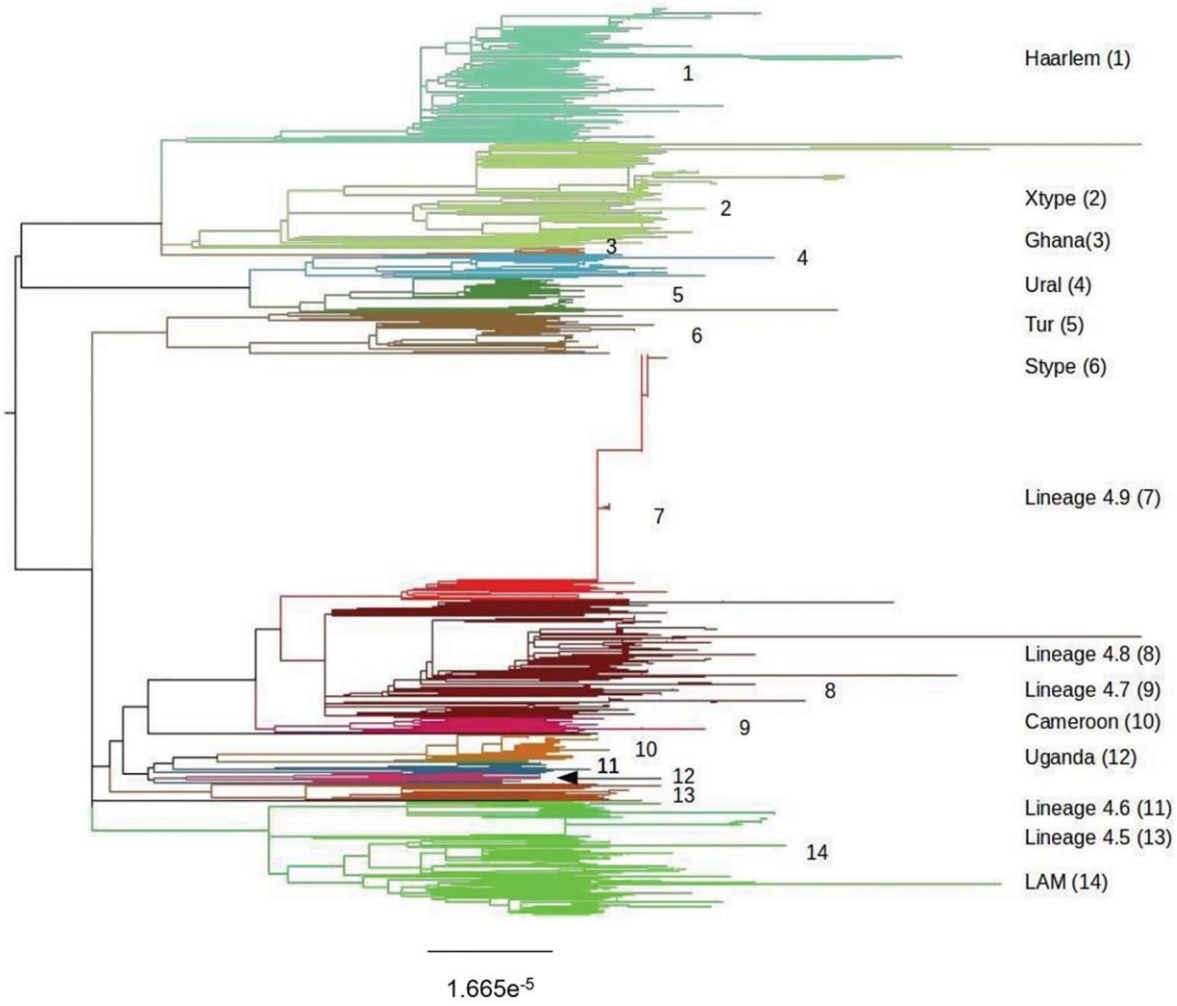
AppendixTable 4. Phylogenetic single-nucleotide polymorphisms in drug resistance-associated genes in members of the *Mtbc*; National Center for Biotechnology Information reference genome no. NC000962.2.

Position	Ref	Alt	Gene	Position	Ref	Alt	Gene
M. canettii				*M. bovis*			
1822859	A	G	cydD	9217	A	C	gyrA
4243690	T	C	embA	Dassie bacillus			
4244154	A	C	embA	1822202	G	A	cydD
4244312	T	G	embA	413654	G	T	iniC
4245023	A	G	embA	*M. caprae*			
4247590	A	G	embB	2714317	C	T	eis
4247815	C	T	embB	1673766	C	T	fabG1
4248206	A	G	embB	6307	T	G	gyrB
4240237	G	C	embC	*M. microti*			
1417294	T	C	embR	1823237	C	A	cydD
9377	A	G	gyrA	5671	C	T	gyrB
9716	T	C	gyrA	471630	G	C	ndhA
9740	G	A	gyrA	472502	T	C	ndhA
9776	T	C	gyrA	1473079	G	A	rrs
5452	G	A	gyrB	*M. mungi*			
2156055	G	A	katG	4242177	C	A	embC
3645738	T	C	manB	8134	T	C	gyrA
3645930	A	G	manB	1833589	A	G	rpsA
2101777	G	C	ndh	1918158	G	A	tlyA
471908	T	C	ndhA	*M. orygis*			
471944	T	C	ndhA	2726378	T	A	ahpC
472397	G	A	ndhA	1819985	G	C	cydC
2289104	T	C	pncA	1821891	G	C	cydD
1833554	A	G	rpsA	4244154	A	G	embA
1833568	G	C	rpsA	5516	A	G	gyrB
1834912	A	G	rpsA	6109	G	A	gyrB
M. bovis BCG				6717	T	C	gyrB
4247173	G	A	embB	412484	C	G	iniA
8624	G	T	gyrA	2154707	C	G	katG
2102106	C	G	ndh	1834363	G	A	rpsA
781568	C	T	rpsL	*M. suricattae*			
M. pinnipedii				1820784	C	T	cydC
1674520	C	T	inhA	1472059	T	G	rrs
1473094	T	C	rrs				
Indo-Oceanic Mtb lineage 1				Ural			
4245969	C	T	embA	3646964	C	G	rmlD
4241042	A	G	embC	X-type			
1417019	C	T	embR	4249408	G	A	embB
8452	C	T	gyrA	West African 1 Mtb lineage 5			
6112	G	C	gyrB	4244635	T	C	embA
471666	A	G	ndhA	4245147	C	T	embA
3647591	A	G	rmlD	9566	C	T	gyrA
Beijing/East Asian Mtb lineage 2				2101921	C	T	ndh

Position	Ref	Alt	Gene	Position	Ref	Alt	Gene
4243460	C	T	embA	West African 2 Mtb lineage 6			
1834177	A	C	rpsA	4244379	C	T	embA
East African Indian Mtb lineage 3				4241843	C	A	embC
4242075	G	A	embC	1674434	T	C	inhA
3645524	C	T	manB	760969	C	T	rpoB
762434	T	G	rpoB	761723	A	C	rpoB
European American Mtb lineage 4				*M. aethiops vetus* Mtb lineage 7			
Haarlem				4248073	C	T	embB
760115	C	T	rpoB	4240153	G	A	embC
S-type				1416977	T	C	embR
411371	T	C	iniA	8876	C	T	gyrA
2102990	A	G	ndh	412842	A	G	iniC
Uganda				2102218	G	A	ndh
7539	A	G	gyrA	1834916	A	C	rpsA
412017	C	G	iniA	1918281	A	C	tlyA



Appendix Figure 1. Maximum likelihood tree of *M. orygis* isolates from the clinical dataset. The 2 instances of patients with a pairwise single-nucleotide polymorphism (SNP) distances that could plausibly support person-to-person transmission or exposure to a common source are highlighted with blue bars. The accompanying SNP distance highlighted in blue is the pairwise distance between the patients. Scale bar indicates nucleotide substitutions per genome. OP, orygis patient.



Appendix Figure 2. Maximum likelihood tree of all lineage 4 genomes from the validation and clinical sets. Unnamed lineages are numbered per Coll et al. (*1*). There are a number of lineage 4 sublineages without a name, which are called simply “Lineage 4” in our validation set. We subsequently updated the SNP-IT database to include resolution of all major sublineages as derived by principal components analysis by Stucki et al. (*2*). Scale bar shows nucleotide substitutions per site.