


## RESEARCH

## Open Access



# Robust barcoding and identification of *Mycobacterium tuberculosis* lineages for epidemiological and clinical studies

Gary Napier<sup>1</sup>, Susana Campino<sup>1</sup>, Yared Merid<sup>2,3,4</sup>, Markos Abebe<sup>2</sup>, Yimtubezinash Woldeamanuel<sup>3</sup>, Abraham Aseffa<sup>2</sup>, Martin L. Hibberd<sup>1</sup>, Jody Phelan<sup>1</sup> and Taane G. Clark<sup>1,5\*</sup> 

## Abstract

**Background:** Tuberculosis, caused by bacteria in the *Mycobacterium tuberculosis* complex (MTBC), is a major global public health burden. Strain-specific genomic diversity in the known lineages of MTBC is an important factor in pathogenesis that may affect virulence, transmissibility, host response and emergence of drug resistance. Fast and accurate tracking of MTBC strains is therefore crucial for infection control, and our previous work developed a 62-single nucleotide polymorphism (SNP) barcode to inform on the phylogenetic identity of 7 human lineages and 64 sub-lineages.

**Methods:** To update this barcode, we analysed whole genome sequencing data from 35,298 MTBC isolates (~ 1 million SNPs) covering 9 main lineages and 3 similar animal-related species (*M. tuberculosis* var. *bovis*, *M. tuberculosis* var. *caprae* and *M. tuberculosis* var. *oryzidis*). The data was partitioned into training ( $N = 17,903$ , 50.7%) and test ( $N = 17,395$ , 49.3%) sets and were analysed using an integrated phylogenetic tree and population differentiation ( $F_{ST}$ ) statistical approach.

**Results:** By constructing a phylogenetic tree on the training MTBC isolates, we characterised 90 lineages or sub-lineages or species, of which 30 are new, and identified 421 robust barcoding mutations, of which a minimal set of 90 was selected that included 20 markers from the 62-SNP barcode. The barcoding SNPs (90 and 421) discriminated perfectly the 86 MTBC isolate (sub-)lineages in the test set and could accurately reconstruct the clades across the combined 35k samples.

**Conclusions:** The validated 90 SNPs can be used for the rapid diagnosis and tracking of MTBC strains to assist public health surveillance and control. To facilitate this, the SNP markers have now been incorporated into the *TB-Profiler* informatics platform (<https://github.com/jodyphelan/TBProfiler>).

**Keywords:** Tuberculosis, Diagnostics, Profiling, SNPs, Barcoding, Mycobacteria tuberculosis complex

\* Correspondence: [taane.clark@lshtm.ac.uk](mailto:taane.clark@lshtm.ac.uk)

<sup>1</sup>Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, UK

<sup>5</sup>Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London WC1E 7HT, UK

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

Tuberculosis, caused by bacteria in the *Mycobacterium tuberculosis* complex (MTBC), is a major global burden causing approximately ten million active cases and killing 1.5 million people in 2018 ([www.who.int/tb](http://www.who.int/tb)). The MTBC consists of *Mycobacterium tuberculosis* sensu stricto (*Mtb*) (lineages 1, 2, 3, 4 and 7) and *M. tuberculosis* var. *africanum* (lineages 5 and 6; *M. africanum*), which cause human disease, but others including *M. tuberculosis* var. *bovis* affect predominantly animals [1]. Recently, new *Mtb* lineages (8, 9) have been proposed [2, 3]. The MTBC lineages vary in their geographic distribution and spread, being endemic in different locations around the globe, leading to the hypothesis that the strain types are specifically adapted to different human populations [4]. Lineage 2 is particularly mobile with evidence of recent spread from Asia to Europe and Africa. Lineage 4 is common in Europe and southern Africa, with regions of high TB incidence and high levels of HIV co-infection, whilst lineages 5, 6 and 7 appear isolated within West Africa and Ethiopia, respectively [1].

There is some evidence to suggest that MTBC lineages can determine the transmission, control, and clinical outcome of pulmonary and extra-pulmonary tuberculosis. In particular, variational phenotypes include differences in the emergence of drug resistance, transmissibility, virulence, host response, disease site and severity [5, 6]. Such phenotypes confer advantages for those MTBC lineages and may lead to an increased likelihood of disease spread and poorer prognosis for patients. Whether increased virulence is associated with poorer prognosis is unclear, with some studies reporting increased mortality risk with strains thought to be less virulent [7]. Of particular concern are the emergence of drug-resistant, multidrug-resistant (MDR-TB) and extensively drug-resistant (XDR-TB) strains, where Beijing strains show strong linear-resistance associations [8]. However, there is considerable inter-strain variation within lineages. For example, when comparing two different Beijing sub-lineages, the “ancient” (atypical) and “modern” (typical) strains show differences in geographical distribution, drug resistance and virulence patterns [9]. In particular, the “modern” sub-lineage is distributed worldwide and has been largely associated with MDR-TB and XDR-TB and hypervirulence [9].

Tracking the spread of lineages is of great importance in tuberculosis research and control. Rapid lineage identification enables the analysis of phenotypic associations, informs on likely provenance and can assist in the prediction of potential future outbreaks. The molecular barcoding of lineages and sub-lineages can be used to classify clinical isolates to aid in the evaluation of tools to control the disease, including therapeutics and vaccines, whose effectiveness may vary by strain type [1, 5]. Historically, strain identification has involved the genotyping of

tandem repeats (e.g. spoligotypes) and large deletions (regions of difference (RDs)) [10], but these approaches are being replaced by methods analysing data from whole genome sequencing (WGS) technologies. These approaches include in silico spoligotyping and RD detection, the characterisation of lineage-associated single nucleotide polymorphisms (SNPs) and higher resolution methods such as core genome MLST [11]. SNP-based approaches can be applied in silico or implemented within a laboratory typing assay [12, 13]. Although the SNP-defined lineages do not offer the same resolution as using the whole genome, they provide a valuable insight into the epidemiology of circulating strains. A 62-SNP barcode was developed using WGS data for 1601 MTBC isolates and was the first to position samples within clades of a global phylogeny of 7 human lineages and 64 sub-lineages, covering all common strain types [1].

Here, we update the 62-SNP barcode using WGS for 35,298 MTBC isolates. In particular, we use WGS data for 17,903 (50.7%) isolates to reconstruct a global phylogeny, resulting in 30 new (sub-)lineages. This analysis led to the 62-SNP barcode being modified and extended to ninety robust SNPs to cover 90 MTBC (sub-)lineages or species, including animal-related *M. tuberculosis* var. *bovis* (*M. bovis*), *M. tuberculosis* var. *caprae* (*M. caprae*) and *M. tuberculosis* var. *orygis* (*M. orygis*), which are similar and sometimes misclassified. The new barcode was validated on the 17,395 (49.3%) remaining MTBC isolates. The ninety SNP markers have been incorporated into the *TB-Profiler* software (<https://github.com/jodyphelan/TBProfiler>) [14], which has been used to profile more than fifty thousand MTBC for strain types and drug resistance, and will thereby assist with barcode implementation for research and infection control activities.

## Methods

### Sample, raw data and sequence analysis

Illumina whole genome sequencing data was publicly available across 35,298 MTBC isolates, which encompassed *Mtb* lineages (1, 2, 3, 4 and 7), *M. africanum* (lineages 5 and 6), *M. bovis*, *M. caprae* and *M. orygis* [14], and the recently proposed lineages 8 [2] and 9 [3] (Additional file 1: Table S1). The data were convenience sampled with the first processed set ( $n = 17,903$ ; 50.7%) serving as a training dataset, and the second set collated subsequently ( $n = 17,395$ ; 49.3%) serving as a testing dataset (Additional file 1: Table S1). The test set covers all the sub-lineages in the training set with at least 10 isolates (range 10–917), except (sub-)lineages 3.1.2.2, 4.6.2.1, 8 and 9, but for these the number of training samples is relatively small.

All raw sequences were trimmed using *trimmomatic* software [15] (v0.36, parameters: PE -phred33 LEAD ING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:

36). Trimmed reads were then aligned with *BWA-MEM* software [16] (v0.7.17-r1188, default parameters) using the H37Rv reference sequence (Genbank accession number: NC\_000962.3). Alignments from *BWA-MEM* were converted to “bam” format and sorted using *samtools* software [17] (v1.9, default parameters). SNPs were identified by applying *BCFtools* [17] (v1.9, mpileup parameters: default, call parameters: -mv) and *GATK* software [18] (version: 4.1.3.0) using the HaplotypeCaller function (parameters: -ERC GVCF). Individual sample “vcf” files were merged using *GATK* GenomicsDBImport (default parameters) and *GATK* CombineGVCFs (default parameters) to perform joint calling using all samples. The resulting multi-sample vcf file was filtered to remove indels and heterozygous calls and monomorphic SNPs. A multi-FASTA file containing all isolates was generated from the filtered SNP file ( $N = 1,014,762$  SNPs; training 620,652 SNPs; test 533,152 SNPs) and H37Rv reference genome using *bedtools* (v2.28.0) [19] and in-house python scripts. The regions of difference (RDs) were detected using *delly* software [20] and confirmed using de novo assembly by applying *Spades* software [21]. Spoligotypes were called using *spolpred* software [22].

#### Principal component analysis and phylogenetic tree

Distance matrices and the principal components of the multi-FASTA files were computed with *Plink* software (v1.90b4; <https://www.cog-genomics.org/plink2>) [23]. The distance matrices were used for the new cluster identification. Maximum likelihood phylogenetic trees were constructed from the multi-FASTA file using *IQ-TREE* (v1.6.12) (<http://www.iqtree.org/>) [24]. A general time reversible model with rate heterogeneity set to a discrete Gamma model and an ascertainment bias correction were used (parameters -m GTR+G+ASC), with 1000 bootstrap samples used to measure branch quality and robustness. Phylogenetic trees were generated for all MTBC isolates, as well as for each main lineage separately. The resulting Newick-formatted tree files were visualised and annotated with metadata in *iTOL* (v5.2; <https://itol.embl.de/>) [25]. These metadata included the 62-SNP barcode sub-lineage predictions [1], allowing for the rapid identification of outliers. By annotating the branches with ancestral mutations, it was possible to inform on SNP markers for barcoding.

#### Lineage revision and new sub-lineage identification

The visual inspection of the phylogenetic trees (and principal component analysis plots) revealed that some pre-existing (sub-)lineages (as defined using the 62-SNP barcode) could be merged or split, as well as new ones created. The original 62-SNP barcode was constructed to reflect the original strain-type families used by researchers based on spoligotypes and RDs. We sought to

analyse the phylogenetic tree to further divide these clades where obvious splits in the phylogeny existed. To aid in old lineage revision and new lineage identification, phylogenetic trees relating to lineages 1 to 9 and animal strains were analysed using a semi-automated procedure. Each tree was traversed (and each clade inspected) from root to tip using the *ETE3 Toolkit* (v3.1.1) package in Python3 (<http://etetoolkit.org/>) [26]. We identified metrics and parameters such as branch bootstrap support values and intra/inter-cluster SNP distances to determine splits in the tree, which led to clusters that are separated by long branch lengths from other isolates. Whilst traversing, the following criteria had to be met to establish clades leading to new or revised sub-lineages: (1) a minimum clade size of 20, with a branch supported by a bootstrap value of  $> 95$ ; (2) differences in the distributions of SNP distances where comparing the isolates within and outside the clade, using a Welch  $t$  test assuming unequal variances [27] ( $P < 0.05$ ) and a Cohen's  $d$  effect size [28] ( $d > 0.5$ ); (3) the ratio of the branch length of the clade compared to the mean branch length of its descendants (ratio  $> 1$ ); (4) estimation of the number of clade-informative SNPs, requiring at least 10 SNPs with a fixation index ( $F_{ST}$ ) [29] value of 1; (5) confirmation of the clade through visual inspection of the tree. Each of the parameter thresholds was based on established cut-offs or determined using standard point of inflection methods [1]. The population differentiation  $F_{ST}$  statistic assigns a strength of association between each SNP and (sub-)lineage, with a score of 1 indicating that the SNP allele is fixed in the sub-lineage of interest and not present outside that group. Using the five criteria led to the addition of 87 (27 new) sub-lineages or lineages (including 8 and 9), or changing the branch position of established others (e.g. 1.2 and 1.1.1.1) (see Additional file 1: Fig. S1). The *SNP-IT* tool for identifying species in MTBC [30] was applied to the *M. bovis*, *M. orygis* and *M. caprae* isolates ( $N = 110$ ; test set), and three barcoding SNPs were required for these mycobacteria. The overall number of (sub-)lineages or species covered was 90.

#### Barcoding SNPs

To ensure that the required 90 clade-specific mutations (“potential barcoding SNPs”, all with  $F_{ST} = 1$ ) were robust, where possible, we retained synonymous SNPs in essential genes [31], and excluded those in drug resistance loci (from *TB-Profler* [14]) and non-essential PE/PPE gene families [32]. From those retained “robust” SNPs ( $n = 421$ ), a minimal set of one per lineage included preferentially those already present in the 62-SNP barcode [1] and, if not possible, (arbitrarily) the lowest position was chosen. The gene functional categories were extracted from *Tuberculist* ([tuberculist.epfl.ch](http://tuberculist.epfl.ch)), and the frequency of

ontologies across all potential barcoding, robust and minimal SNPs, was assessed for differences across lineage using the chi-squared tests.

### Validation of lineage barcode

To validate the final set of robust 421 clade-defining SNPs (Additional file 1: Table S2), the 17,395 samples in the testing set (with 572,021 SNPs) were used. The (sub-)lineage of these samples was predicted with *TB-Profler* [14]. At the same time, a phylogenetic tree was reconstructed of the training and test samples together using *FastTree2* software [33]. To assess the sensitivity and specificity of the predictions, this tree was traversed in the *ETE3 Toolkit*, and test samples were examined for their presence in the clades defined by the training dataset.

## Results

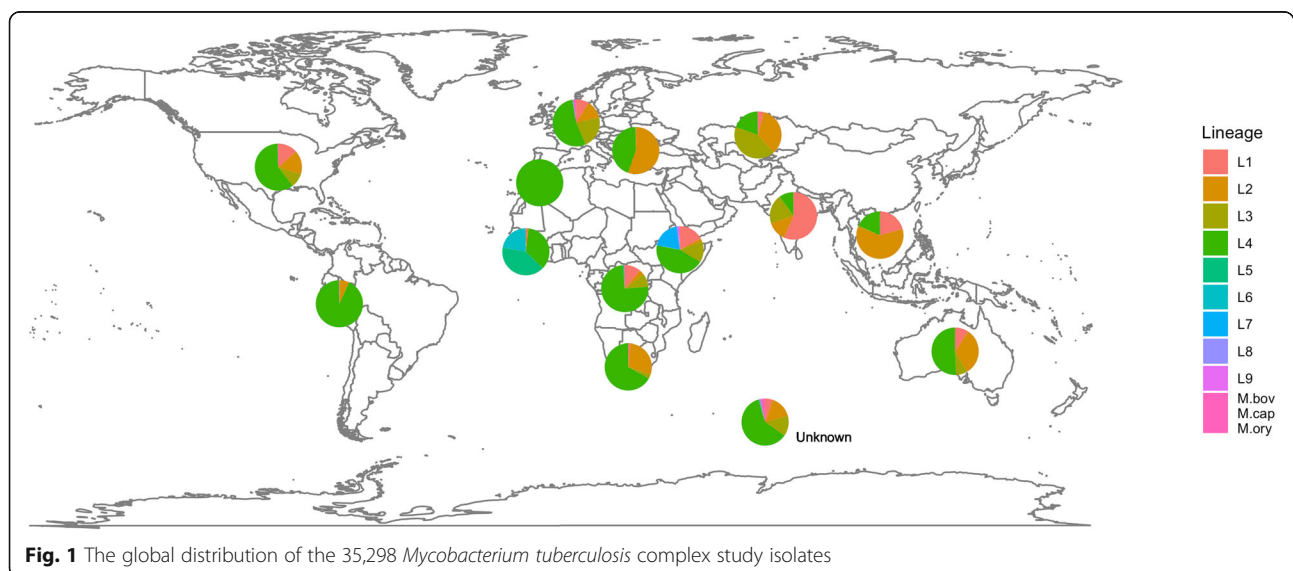
### MTBC isolates, SNPs and phylogeny

Across a total of 35,298 MTBC isolates with sequencing data, we identified 1,014,762 high-quality SNPs. The isolates represented all MTBC lineages (1–9), *M. bovis*, *M. orygis* and *M. caprae*, but the majority were from lineages 4 (51.6%), 2 (25.2%), 3 (11.1%) and 1 (9.5%), with the frequency of others being at most 1% (Additional file 1: Table S1). Whilst it is a convenience set of sampled isolates, the geographical distribution of the lineages was as expected, with lineage 2 dominating in Southeast Asia, lineages 1 and 3 predominant in South Asia, lineage 4 abundant in Europe, Americas and Africa and lineages 5 and 6 present in West Africa (Fig. 1). The East Asian lineage 2 had the highest frequency of MDR-TB isolates (36.2%), driven by a higher prevalence in the Beijing sub-lineage (lineage 2.2; 36.5%) compared to the Manu ancestor or proto-Beijing strain type (lineage 2.1, 19.8%) (Table 1).

The 35k isolates were split into training ( $N = 17,903$ , 50.7%; all MTBC; 620,652 SNPs) and test ( $N = 17,395$ , 49.3%, all MTBC except lineages 8 and 9; 572,021 SNPs) datasets (Table 1; Additional file 1: Table S1). A phylogenetic tree was constructed on the training isolates and confirmed the clustering by lineage and sub-lineages (Fig. 2). Similarly, a principal component analysis of the 35k isolates using the ~1 million SNPs revealed the expected clustering by lineage or species (Additional file 1: Fig. S1(a)). Phylogenetic trees were constructed for each lineage separately and confirmed the sub-lineage and strain-type clustering (Additional file 1: Fig. S1(b)-(f)). However, by assessing the fine-scale clustering of sub-lineages predicted by the 62-SNP barcode, outlying samples were revealed and suggested a need for the repositioning of mutations underlying the clades or, alternatively, the creation of new sub-lineages that were on long branches (Additional file 1: Fig. S12(b, c)). In some cases, new sub-lineages reflected existing RD- or spoligotype-based strain classifications which were imperfectly or not captured using the 62-SNP barcode (see Additional file 1: Fig. S2 (d,e)).

### Barcoding SNPs

By traversing the whole MTBC and lineage-based phylogenetic trees using a semi-automated algorithm, it was possible to modify sub-lineages within the flexible nomenclature structure of the previous barcode [1], as well as define clade-informative SNPs. The phylogenetic analyses characterised 27 additional (sub-)lineages covering lineages 1 (8), 3 (2), 4 (15), 8 (1) and 9 (1). The final number of (sub-)lineages in *Mtb* was 85 (L (ineage)1 16, L2 7, L3 7, L4 52, L7 1, L8 1, L9 1) and *M. africanum* was 2 (L5 1, L6 1) (Table 1; Fig. 2), requiring 87 SNP markers. A further three SNP markers were required to



**Fig. 1** The global distribution of the 35,298 *Mycobacterium tuberculosis* complex study isolates

**Table 1** *Mycobacterium tuberculosis* complex lineages and sub-lineages across the 35,298 isolates

Lineage	No. training (test)	No. countries train (test)	% MDR-TB	No. transmission [clusters]	Potential barcoding SNPs*	Robust SNPs**
1	2162 (1203)	25 (42)	7.8	354 [130]	344	17
1.1	1487 (530)	19 (36)	5.5	218 [82]	23	2
1.1.1	706 (170)	8 (16)	3.8	60 [25]	41	5
1.1.1.1	358 (120)	5 (9)	2.1	28 [11]	52	3
1.1.2	459 (278)	15 (25)	9.0	83 [31]	109	3
1.1.3	299 (80)	11 (16)	3.2	73 [25]	42	2
<b>1.1.3.1</b>	84 (31)	7 (13)	3.5	10 [4]	68	2
<b>1.1.3.2</b>	155 (33)	7 (7)	1.1	57 [18]	113	6
<b>1.1.3.3</b>	32 (7)	5 (4)	10.3	4 [2]	36	2
<b>1.2</b>	309 (550)	13 (21)	7.5	40 [16]	60	2
1.2.1	28 (44)	3 (7)	6.9	6 [2]	78	5
1.2.2	277 (505)	13 (18)	7.5	34 [14]	159	8
<b>1.2.2.1</b>	244 (453)	12 (18)	6.9	34 [14]	34	1
<b>1.3</b>	366 (122)	16 (19)	18.0	96 [32]	71	2
<b>1.3.1</b>	88 (25)	7 (11)	10.6	20 [7]	50	4
<b>1.3.2</b>	278 (97)	16 (17)	20.3	76 [25]	83	4
2	4556 (4322)	45 (56)	36.2	1778 [413]	72	4
2.1	95 (41)	6 (9)	19.8	27 [10]	172	4
2.2	4461 (4281)	45 (56)	36.5	1751 [403]	79	17
2.2.1	4239 (4007)	45 (56)	35.1	1632 [389]	17	2
2.2.1.1	338 (443)	19 (18)	28.0	98 [40]	6	2
2.2.1.2	29 (21)	6 (9)	36.0	10 [3]	5	1
2.2.2	222 (273)	16 (15)	59.0	119 [14]	54	4
3	2654 (1271)	24 (31)	13.4	847 [242]	166	8
<b>3.1</b>	715 (362)	15 (22)	9.5	372 [80]	1	1
3.1.1	387 (280)	11 (16)	6.2	243 [43]	17	2
3.1.2	295 (69)	13 (8)	14.3	124 [35]	8	2
3.1.2.1	98 (25)	8 (7)	19.5	25 [12]	15	7
3.1.2.2	48 (0)	3 (0)	0	36 [2]	85	6
<b>3.2</b>	89 (31)	6 (9)	10.0	31 [7]	85	2
4	8320 (9883)	44 (99)	18.5	3109 [731]	94	3
4.1	2594 (2325)	35 (64)	18.5	1043 [191]	58	3
4.1.1	889 (482)	20 (27)	18.1	403 [72]	30	13
4.1.1.1	210 (158)	14 (16)	9.5	92 [20]	39	2
4.1.1.2	55 (44)	4 (6)	2.0	33 [3]	92	2
4.1.1.3	579 (247)	18 (23)	22.4	266 [44]	58	3
<b>4.1.1.3.1</b>	207 (13)	3 (3)	9.6	158 [5]	46	3
4.1.2	1612 (1743)	32 (61)	17.3	622 [113]	13	1
4.1.2.1	1383 (1087)	32 (60)	22.5	563 [96]	49	3
<b>4.1.2.1.1</b>	231 (18)	1 (1)	97.6	221 [2]	73	3
<b>4.1.3</b>	28 (70)	7 (10)	57.1	4 [2]	124	3
<b>4.1.4</b>	24 (12)	8 (7)	38.9	10 [2]	60	4
4.2	481 (532)	23 (26)	28.0	87 [32]	116	8

**Table 1** *Mycobacterium tuberculosis* complex lineages and sub-lineages across the 35,298 isolates (Continued)

Lineage	No. training (test)	No. countries train (test)	% MDR-TB	No. transmission [clusters]	Potential barcoding SNPs*	Robust SNPs**
4.2.1	206 (240)	13 (20)	28.3	34 [13]	26	2
<b>4.2.1.1</b>	54 (148)	9 (10)	6.9	2 [1]	36	2
4.2.2	274 (288)	20 (18)	28.1	53 [19]	20	2
4.2.2.1	74 (41)	10 (6)	45.2	22 [7]	26	2
<b>4.2.2.2</b>	120 (139)	11 (14)	27.8	15 [7]	31	10
4.3	2507 (2928)	30 (75)	23.1	993 [244]	38	2
4.3.1	58 (67)	7 (15)	6.4	40 [3]	28	1
<b>4.3.1.1</b>	37 (2)	3 (1)	0.0	36 [1]	52	2
4.3.2	409 (1200)	16 (21)	7.2	75 [32]	75	1
4.3.2.1	291 (917)	6 (7)	3.7	50 [23]	55	4
4.3.3	648 (810)	25 (57)	41.3	210 [66]	33	1
4.3.4	1366 (807)	23 (45)	24.1	664 [142]	8	1
4.3.4.1	194 (170)	14 (30)	28.9	49 [14]	19	4
4.3.4.2	1170 (635)	22 (34)	23.1	614 [128]	26	1
4.3.4.2.1	877 (287)	13 (18)	5.6	457 [103]	11	1
4.4	560 (1059)	24 (29)	15.7	190 [63]	37	2
4.4.1	420 (861)	22 (25)	16.0	149 [48]	38	4
4.4.1.1	379 (755)	21 (24)	17.8	136 [44]	16	1
<b>4.4.1.1.1</b>	75 (206)	5 (4)	19.6	22 [9]	60	3
4.4.1.2	39 (106)	8 (6)	1.4	13 [4]	95	9
4.4.2	112 (181)	7 (9)	14.7	33 [13]	7	2
4.5	293 (357)	17 (17)	15.7	49 [22]	50	1
4.6	340 (442)	21 (25)	22.1	139 [39]	12	1
4.6.1	73 (296)	9 (12)	29.8	24 [8]	53	3
4.6.1.1	29 (126)	6 (7)	1.3	14 [3]	22	1
4.6.1.2	40 (154)	9 (11)	54.6	10 [5]	37	1
4.6.2	164 (89)	16 (17)	15.4	65 [20]	22	1
4.6.2.1	2 (0)	1 (0)	0	2 [1]	45	2
4.6.2.2	150 (89)	14 (17)	15.9	60 [18]	106	6
<b>4.6.3</b>	23 (9)	3 (4)	0	20 [3]	135	3
<b>4.6.4</b>	23 (7)	5 (4)	50.0	10 [2]	49	1
<b>4.6.5</b>	23 (18)	5 (5)	19.5	9 [3]	8	2
4.7	158 (200)	18 (23)	10.3	56 [20]	10	3
4.8	1051 (1807)	29 (55)	7.8	419 [88]	17	1
<b>4.8.1</b>	63 (90)	7 (4)	22.2	21 [5]	46	3
<b>4.8.2</b>	116 (5)	3 (2)	0	113 [1]	42	2
<b>4.8.3</b>	21 (3)	1 (1)	0	19 [1]	34	1
4.9	243 (141)	14 (22)	12.5	114 [24]	37	3
<b>4.9.1</b>	74 (15)	6 (3)	5.6	44 [1]	49	3
5	26 (255)	6 (12)	14.6	2 [1]	460	13
6	32 (135)	6 (13)	3.6	5 [2]	214	10
7	38 (26)	3 (2)	0	3 [1]	837	38
<b>8</b>	2 (0)	1 (0)	0	0 [0]	888	43

**Table 1** *Mycobacterium tuberculosis* complex lineages and sub-lineages across the 35,298 isolates (Continued)

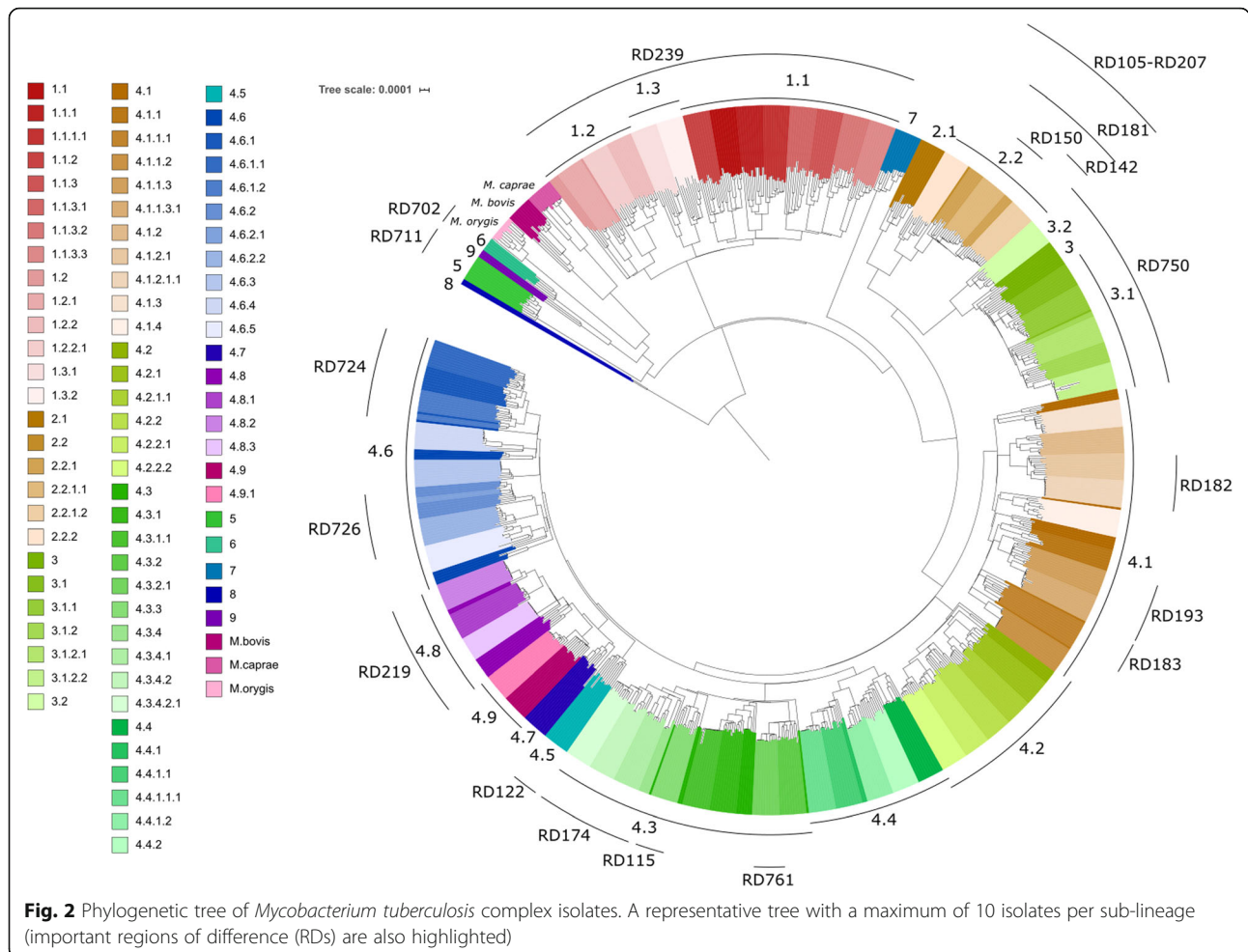
Lineage	No. training (test)	No. countries train (test)	% MDR-TB	No. transmission [clusters]	Potential barcoding SNPs*	Robust SNPs**
<b>9</b>	3 (0)	1 (0)	0	0 [0]	160	5
<b><i>M. bovis</i></b>	81 (281)	9 (12)	0.8	42 [11]	93	3
<b><i>M. caprae</i></b>	3 (7)	2 (3)	0	0 [0]	225	5
<b><i>M. orygis</i></b>	26 (12)	4 (4)	0	0 [0]	743	28
Totals	17,903 (17,395)	165 (269)	21.0	6140 [1531]	8128	421

Bolded are changes from the barcode in reference [1]—either new sub-lineages or new barcoding SNPs; MDR-TB multidrug-resistant TB, which is resistant to at least rifampicin and isoniazid drugs. \*All potential barcoding SNPs ( $F_{ST} = 1$ ). \*\*Final robust SNP set, based on synonymous changes in essential and non-drug resistance genes only (except 12 sub-lineages which had no informative SNPs in essential genes; see Additional file 1: Table S2)

discriminate *M. bovis*, *M. caprae* and *M. orygis*, which have highly similar mycobacterial genomes, and therefore, their accurate typing will greatly assist with the misclassification of *M. bovis* infections.

To find informative SNPs for each of the 90 MTBC clades, we used the population differentiation metric  $F_{ST}$  to identify mutations that were only present in the isolates in the selected (sub-)lineage of interest ( $F_{ST} = 1$ ). We identified 8128 potential barcoding SNPs (with  $F_{ST} =$

1) across the 90 clades (Table 1). These barcoding SNPs were distributed evenly genome-wide, with no visible clustering of informative mutations for individual lineages (Additional file 1: Fig. S3). Of these SNPs, 7282 (89.6%) were in genic regions, with mutations leading to 4699 non-synonymous (NS) and 2564 synonymous (S) amino acid changes, as well as 20 changes in non-coding genes. By focusing on essential genes, 889 (10.9%) SNPs remained (499 NS, 390 S). Furthermore, variants in



drug-resistance-associated genes were removed, leaving 824 SNPs (464 NS and 360 S mutations). Across all lineages, except lineages 8 ( $N = 2$ ) and 9 ( $N = 3$ ) which had small sample sizes, we compared the distribution of gene functions for all potential barcoding SNPs in all characterised genes (7060/7282 SNPs) with only those in essential (and non-drug resistance) loci (790/824 SNPs) (Additional file 1: Fig. S4). The distribution of gene function for all potential barcoding SNPs is similar across all lineages. However, after filtering for essential and non-drug-resistant genes, lineage 2 has a relatively high proportion of non-synonymous SNP mutations in cell wall and cell process genes, whilst for lineage 6, *M. bovis*, *M. caprae* and *M. orygis*, there are relatively higher proportions of non-synonymous SNP mutations in intermediary metabolism and pathway genes. For 11 (sub-)lineages, there were no potential barcoding SNPs lying within essential and non-drug resistance genes, so they were identified in non-essential and non-PE/PPE loci (Additional file 1: Table S3) (180 SNPs, 61 synonymous mutations).

By considering only the SNPs with synonymous changes, similar to the selection strategy applied in [1], a total of 421 SNPs were considered suitable for barcoding the 90 (sub-)lineages (Table 1; Additional file 1: Table S2). Of these, 20 SNPs represented (sub-)lineages in the 62-SNP barcode [1] and were therefore retained, leading to 70 new SNPs chosen for final (sub-)lineage classification (Additional file 1: Table S3). Across the 60 (sub-)lineages common to the 62- and 90-SNP barcodes, the 40 new SNPs had higher  $F_{ST}$  values than those in the old barcode (Additional file 1: Fig. S5). Using the test set ( $N = 17,395$ ) which had representation of 86 of the 90 (sub-)lineages, we found that the minimal set of 90 SNPs had perfect predictive performance for all clades (all sensitivities and specificities of value 1). This analysis excluded four (sub-)lineages (3.1.2.2, 4.6.2.1, 8 and 9), which had no test samples.

### Comparisons to other software

The barcode was compared to lineage predictions from SNP-IT [30] software, a 27 strain-type system covering MTBC, including 6 animal lineages that are not present in our large dataset. First, we assessed the assigned major MTBC lineages (1–6) by both barcodes and found complete concordance. Second, we quantified how the increased number of strain types in our barcode ( $n = 90$ ) improved the resolution of sub-lineage assignment over the SNP-IT tool. For 14 of the 21 SNP-IT strain types present in our data, the 90-SNP approach provides higher resolution of clades (range 2 to 15 sub-lineages per SNP-IT clade) (Additional file 1: Fig. S6). Six other strain types have direct mapping between our barcode and SNP-IT, and there is one instance where isolates

classified as *M. bovis* with our barcode are further classified into *M. bovis BCG* and *M. bovis bovis* using SNP-IT.

### Discussion

MTBC strain types and lineages are distributed phylogeographically and have been associated with differences in the emergence of drug resistance, transmissibility, virulence, host response, vaccine efficacy, disease site and severity [5, 6, 34]. However, further research into lineage, genotype–phenotype associations are required. Such research needs to be underpinned by molecular barcodes of MTBC (sub-)lineages, strain types and species. Here, we updated a 62-SNP barcode that forms a highly resolved phylogenetic identification system that determines 7 lineages, 64 sub-lineages and *M. bovis*, but was constructed using ~1600 MTBC isolates with WGS data [1]. Using twenty-fold more MTBC isolates with WGS data, we identified and validated a set of 90 robust SNPs (of 421 alternatives) to cover a global phylogeny of 9 lineages, 87 sub-lineages, *M. bovis*, *M. caprae* and *M. orygis*. These SNPs can be used to construct high-resolution and reproducible phylogenies, which can be incorporated within diagnostic assays and assess genotype–phenotype associations. By extending an established 62-SNP barcode system with a flexible nomenclature [1], it was possible to update and add seamlessly (sub-)lineages and species and in the future include potentially novel strain types should they be reported. Such modifications could involve inclusion of SNPs to barcode other MTBC animal lineages or partitioning of *M. africanum* lineages 5 and 6 into sub-lineages [3]. Further, incorporating drug resistance loci will further enhance the usefulness of the 90-SNP barcode as an important tool for tuberculosis control and elimination activities worldwide. To assist this, the 90-SNP variants have been incorporated into the publicly available *TB-Profiler* informatics tool [14], which predicts resistance to 14 anti-tuberculosis drugs from WGS data.

Our barcode development focused on SNPs, but future work could include other types of strain-specific polymorphisms (e.g. insertions, deletions and large structural variants), which are less common than SNPs, but may have major functional consequences. An analysis of the gene ontologies of the barcoding SNPs revealed some differences across lineages, but there is a need to characterise functional effects of the lineage-specific SNP variants, as these could provide insights into disease control measures. Overall, we have provided an updated molecular barcode for MTBC strain types, with ninety robust markers that can be detected from applications of WGS or integrated within high-throughput genotyping or sequencing (e.g. amplicon) platforms to inform ongoing TB surveillance and control.



## Conclusions

The use of molecular barcoding of MTBC bacteria causing tuberculosis can provide insights into outbreaks and help to reveal strain types that are more virulent and prone to drug resistance. In an analysis of 35,298 isolates from MTBC, we update an established 62-SNP barcode with a minimal set of 90 genetic markers, which now cover *M. tuberculosis* (7 lineages, 85 sub-lineages), *M. africanum* (2 lineages), *M. bovis*, *M. caprae* and *M. orygis* bacteria. The new barcode has been implemented within the publicly available *TB-Profiler* informatics tool, to assist the rapid, simple and reliable phylogenetic identification of individual MTBC isolates, thereby aiding clinical studies in the tracking, maintenance and phenotypic determination of MTBC pathogens.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13073-020-00817-3>.

**Additional file 1: Table S1.** The study samples ( $N=35,298$ ) used and their lineages. **Table S2.** Robust barcoding SNPs (421 SNPs, including the 90 SNPs in **Table S3**). **Table S3.** The ninety minimal barcoding SNPs. **Figure S1.** Population structure of the *Mycobacterium tuberculosis* complex isolates by lineage. **Figure S2.** Examples of discrepancies using the 62-SNP barcode. **Figure S3.** The genome-wide distribution of barcoding SNPs ( $F_{ST}=1$ ) for each lineage. **Figure S4.** Functional differences between genes containing lineage-barcoding ( $F_{ST}=1$ ) SNPs. **Figure S5.** Differentiation of sub-lineages when comparing the 62- versus 90-SNP barcodes. **Figure S6.** The increased resolution of our 90-SNP barcode (implemented in *TB-Profiler* software) over the comparable (sub-)lineages of the *SNP-IT* tool.

## Abbreviations

MDR-TB: Multidrug-resistant TB; MTBC: *Mycobacterium tuberculosis* complex; RD: Region of difference; SNP: Single nucleotide polymorphism; TB: Tuberculosis; WGS: Whole genome sequencing; XDR-TB: Extensively drug-resistant TB

## Acknowledgements

The MRC eMedLab computing resource was used for bioinformatics and statistical analysis.

## Authors' contributions

JP and TGC conceived and directed the project. GN and JP performed bioinformatic and statistical analyses under the supervision of SC, MLH and TGC. GN, SC, JP and TGC interpreted results. YM, MA, AA, and YW contributed sequence data. GN, SC, JP and TGC wrote the first draft of the manuscript. All authors commented and edited on various versions of the draft manuscript. GN, JP and TGC compiled the final manuscript. All authors read and approved the final manuscript.

## Funding

GN is supported by a BBSRC LiDO PhD studentship. TGC is funded by the Medical Research Council UK (Grant no. MR/M01360X/1, MR/N010469/1, MR/R025576/1, and MR/R020973/1) and BBSRC UK (Grant no. BB/R013063/1). SC is funded by Medical Research Council UK (MR/M01360X/1, MR/R025576/1, and MR/R020973/1) and BBSRC UK (Grant no. BB/R013063/1) grants. The study was funded in part from the core AHRI budget (NORAD and SIDA grants) and the National Institutes of Health (NIH) Fogarty International Center Global Infectious Diseases grant entitled "Ethiopia-Emory TB Research Training Program" (D43TW009127). These funding bodies did not have a role in the design of the study and collection, analysis and interpretation of data and in writing the manuscript.

## Availability of data and materials

All raw sequence data is available from the EBI short read archive. A dedicated GitHub repository (<https://github.com/GaryNapier/tb-lineages>) [35] contains the list of accession numbers and code. The new (sub-)lineages have been implemented within the TB-Profiler tool <https://github.com/jodyphelan/TBProfiler> [14].

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Competing interests

The authors declare that they have no competing interests.

## Author details

<sup>1</sup>Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, UK. <sup>2</sup>Armauer Hansen Research Institute, Addis Ababa, Ethiopia. <sup>3</sup>Department of Microbiology, Immunology and Parasitology, College of Health Sciences, Addis Ababa University, Addis Ababa, Ethiopia. <sup>4</sup>Hawassa University College of Medicine and Health Sciences, Hawassa, Ethiopia. <sup>5</sup>Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London WC1E 7HT, UK.

Received: 19 July 2020 Accepted: 3 December 2020

Published online: 14 December 2020

## References

- Coll F, McNeerney 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 [cited 2017 Jul 17] Available from: <http://www.nature.com/articles/ncomms5812>.
- Ngabonziza JCS, Loiseau C, Marceau M, Jouet A, Menardo F, Tzfadia O, et al. A sister lineage of the *Mycobacterium tuberculosis* complex discovered in the African Great Lakes region. *Nat Commun*. 2020;11:1–11.
- Coscolla M, Brites D, Menardo F, Loiseau C, Darko Otchere I, Asante-Poku A, et al. Phylogenomics of *Mycobacterium africanum* reveals a new lineage and a complex evolutionary history. *bioRxiv*. 2020;17:19.
- Brites D, Gagneux S. Co-evolution of *Mycobacterium tuberculosis* and *Homo sapiens*. *Immunol Rev* 2015;264:6–24. [cited 2018 Sep 3] Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25703549>.
- Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T, et al. *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nat Genet* ; 2013;45:784–90. [cited 2020 Oct 26] Available from: <https://pubmed.ncbi.nlm.nih.gov/25776166/>report=abstract.
- Reiling N, Homolka S, Walter K, Brandenburg J, Niwinski L, Ernst M, et al. Clade-specific virulence patterns of *Mycobacterium tuberculosis* complex strains in human primary macrophages and aerogenically infected mice. *MBio*. American Society for Microbiology; 2013;4. [cited 2020 Oct 26] Available from: <https://pubmed.ncbi.nlm.nih.gov/25776166/>report=abstract.
- Smittipat N, Miyahara R, Juthayothin T, Billamas P, Dokladda K, Imsanguan W, et al. Indo-Oceanic *Mycobacterium tuberculosis* strains from Thailand associated with higher mortality. *Int J Tuberc Lung Dis*. 2019;23:972–9 [cited 2020 Oct 26] Available from: <https://pubmed.ncbi.nlm.nih.gov/31615603/>.
- Oppong YEA, Phelan J, Perdigão J, MacHado D, Miranda A, Portugal I, et al. Genome-wide analysis of *Mycobacterium tuberculosis* polymorphisms reveals lineage-specific associations with drug resistance. *BMC Genomics*; 2019;20.
- Forrellad MA, Klepp LI, Gioffré A, García JS, Morbidoni HR, de la Paz Santangelo M, et al. Virulence factors of the *Mycobacterium tuberculosis* complex. *Virulence*. Taylor and Francis Inc; 2013. p. 3–66.
- 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. <https://doi.org/10.1155/2014/645802>. Epub 2014 Jan 5.
- Kohl TA, Harmsen D, Rothgänger J, Walker T, Diel R, Niemann S. Harmonized genome wide typing of tubercle bacilli using a web-based gene-by-gene nomenclature system. *EBioMedicine*; 2018;34:131–8. [cited 2020 Oct 26] Available from: <https://pubmed.ncbi.nlm.nih.gov/31647575/>report=abstract.

12. Conceição EC, Refregier G, Gomes HM, Olessa-Daragon X, Coll F, Ratovonirina NH, et al. *Mycobacterium tuberculosis* lineage 1 genetic diversity in Pará, Brazil, suggests common ancestry with east-African isolates potentially linked to historical slave trade. *Infect Genet Evol.* 2019;73:337–41 [cited 2019 Jul 29] Available from: <https://www.sciencedirect.com/science/article/pii/S1567134819301030?via%3Dihub>.
13. Cancino-Muñoz I, Gil-Brusola A, Torres-Puente M, Mariner-Llicer C, Dogba J, Akinseye V, et al. Development and application of affordable SNP typing approaches to genotype *Mycobacterium tuberculosis* complex strains in low and high burden countries. *Sci Rep.* 2019;9:1–12 [cited 2020 Oct 29] Available from: <https://doi.org/10.1038/s41598-019-51326-2>.
14. Phelan JE, O'Sullivan DM, Machado D, Ramos J, Oppong YEA, Campino S, et al. Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Med.* 2019; 11:41. [cited 2019 Jun 28] Available from: <http://www.ncbi.nlm.nih.gov/pubmed/31234910>.
15. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30:2114–20 [cited 2018 Sep 5] Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24695404>.
16. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013; [cited 2017 Sep 6] Available from: <http://arxiv.org/abs/1303.3997>.
17. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics.* 2011;27:2987–93. [cited 2017 Sep 6] Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21903627>.
18. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20:1297–303.
19. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26:841–2.
20. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics.* 2012;28:i333–9.
21. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 2012;19:455–77.
22. Coll F, Mallard K, Preston MD, Bentley S, Parkhill J, McNERNEY R, et al. SpolPred: rapid and accurate prediction of *Mycobacterium tuberculosis* spoligotypes from short genomic sequences. *Bioinformatics.* 2012;28:2991–3 [cited 2017 Sep 7] Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23014632>.
23. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81:559–75 [cited 2017 Sep 6] Available from: <http://www.ncbi.nlm.nih.gov/pubmed/17701901>.
24. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* 2015;32:268–74.
25. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 2019;47:W256–9 [cited 2019 Sep 12] Available from: <https://academic.oup.com/nar/article/47/W1/W256/5424068>.
26. Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol.* 2016;33:1635–8 [cited 2020 Mar 16] Available from: <http://www.ncbi.nlm.nih.gov/pubmed/26921390>.
27. Welch BL. The generalization of 'Student's' problem when several different population variances are involved. *Biometrika.* 1947;34:28.
28. Cohen J. *Statistical power analysis for the behavioral sciences.* New York: Routledge Academic; 1988.
29. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. *Evolution.* 1984;38:1358.
30. Lipworth S, Jajou R, De Neeling A, Bradley P, Van Der Hoek W, Maphalala G, et al. SNP-IT tool for identifying subspecies and associated lineages of *Mycobacterium tuberculosis* complex. *Emerg Infect Dis.* 2019;25:482–8.
31. Dejesus MA, Gerrick ER, Xu W, Park SW, Long JE, Boutte CC, et al. Comprehensive essentiality analysis of the *Mycobacterium tuberculosis* genome via saturating transposon mutagenesis. *mBio.* 2017;8(1):e02133-16. <https://doi.org/10.1128/mBio.02133-16>.
32. Phelan JE, Coll F, Bergval I, Anthony RM, Warren R, Sampson SL, et al. Recombination in pe/ppe genes contributes to genetic variation in *Mycobacterium tuberculosis* lineages. *BMC Genomics.* 2016;17:151 [cited 2017 Jul 17] Available from: <http://www.ncbi.nlm.nih.gov/pubmed/26923687>.
33. Price MN, Dehal PS, Arkin AP. FastTree 2 – approximately maximum-likelihood trees for large alignments. Poon AFY, editor. *PLoS One*; 2010;5:e9490.
34. Coscolla M, Gagneux S. Consequences of genomic diversity in *Mycobacterium tuberculosis*. *Semin. Immunol*; 2014; 431–44. [cited 2020 Oct 26] Available from: <https://pubmed.ncbi.nlm.nih.gov/25453224/>. Accessed 1 Nov 2020.
35. Napier G. *tb-lineages*. GitHub. <https://github.com/GaryNapier/tb-lineages> (2020). Accessed 1 Nov 2020.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

