

The relationship between transmission time and clustering methods in *Mycobacterium tuberculosis* epidemiology

Conor J Meehan^{1,*}, Pieter Moris^{1,2,3}, Thomas A. Kohl^{4,5}, Jūlija Pečerska⁶, Suriya Akter¹, Matthias Merker^{4,5}, Christian Utpatel^{4,5}, Patrick Beckert^{4,5}, Florian Gehre^{1,7,8}, Pauline Lempens¹, Tanja Stadler⁶, Michel K. Kaswa^{1,10}, Denise Kühnert⁹, Stefan Niemann^{4,5,&}, Bouke C de Jong^{1,&}

Affiliations

¹Unit of Mycobacteriology, Biomedical Sciences, Institute of Tropical Medicine, Antwerp, Belgium

²Advanced Database Research and Modelling (ADReM), Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium

³Biomedical Informatics Research Network Antwerp (biomina), University of Antwerp, Antwerp, Belgium

⁴German Center for Infection Research, Partner Site Hamburg-Lübeck-Borstel-Riems, Borstel, Germany

⁵Molecular and Experimental Mycobacteriology, Priority Area Infections, Research Center Borstel, Borstel, Germany

⁶Department of Biosystems Science and Engineering, ETH Zürich, Switzerland

⁷Vaccines and Immunity Theme, Medical Research Council Unit The Gambia

⁸Bernhard Nocht Institute for Tropical Medicine, Hamburg, Germany

⁹University Hospital Zurich, Zurich, Switzerland

¹⁰National Tuberculosis Program, Kinshasa, Democratic Republic of Congo

[&]equal contribution; *corresponding author: cmeehan@itg.be

Abstract

Tracking recent transmission is a vital part of controlling widespread pathogens such as *Mycobacterium tuberculosis*. Multiple approaches exist for detecting recent transmission chains, usually by clustering strains based on the similarity of their genotyping results. However, each method gives varying estimates of transmission cluster sizes and inferring when transmission events within these clusters occurred is almost impossible. This study combines whole genome sequence (WGS) data derived from a high endemic setting with phylodynamics to unveil the timing of transmission events posited by a variety of standard genotyping methods. Our results suggest that clusters based on spoligotyping could encompass transmission events that occurred hundreds of years prior to sampling while 24-loci-MIRU-VNTR often represented decades of transmission. Instead, WGS based genotyping applying a low SNP thresholds allows for estimation of recent transmission events. These findings can guide the selection of appropriate clustering methods for uncovering relevant transmission chains within a given time-period.

Introduction

Despite the large global efforts at curbing the spread of *Mycobacterium tuberculosis* complex (Mtb) strains, 10.4 million new patients develop tuberculosis (TB) every year (WHO, 2018). In addition, the prevalence of multidrug resistant Mtb strains (MDR-TB) is increasing (WHO, 2018), predominantly through ongoing transmission within large populations (Kendall et al., 2015; Merker et al., 2015). The tracking and timing of recent transmission chains allows TB control programs to effectively pinpoint transmission hotspots and employ targeted intervention measures. This is especially important for the transmission of drug resistant strains as it appears that drug resistance may be transmitted more frequently than acquired

(Kendall et al., 2015; Trauer, Denholm, & McBryde, 2014). Thus, interrupting transmission is key for the control of MDR-TB (Klopper et al., 2013; Merker et al., 2015; Shah et al., 2017). For the development of the most effective control strategies, there is a strong need for (i) appropriate identification of relevant transmission chains, risk factors and hotspots and (ii) robust timing of when outbreaks first arose.

Epidemiological TB studies often apply genotyping methods to *Mtbc* strains to determine whether two or more patients are linked within a transmission chain (molecular epidemiology) (Merker, Kohl, Niemann, & Supply, 2017). Contact tracing is a primary epidemiological method for investigating transmission networks of TB, mainly based on patient interviews (Fox, Barry, Britton, & Marks, 2012). Although this method is often seen as a gold standard of transmission linking, it does not always match the true transmission patterns, even in low incidence settings (Behr et al., 1998; Diel et al., 2002; Roetzer et al., 2011, 2013; Small et al., 1994) and misses many connections (Bjorn-Mortensen et al., 2017; Vluggen et al., 2017). The implementation of molecular epidemiological approaches has overcome these limitations and is often used as the main approach for cluster analysis. Classical genotyping has involved *IS6110* DNA fingerprinting (Thierry et al., 1990; van Embden et al., 1993), spoligotyping (Goguet de la Salmonière et al., 1997; Guernier, Sola, Brudey, Guégan, & Rastogi, 2008; Kamerbeek et al., 1997), and variable-number tandem repeats of mycobacterial interspersed repetitive units (MIRU-VNTR) (Supply, Magdalena, Himpens, & Locht, 1997) which is the most common method at the moment (Merker et al., 2017). The latter method is based on copy numbers of a sequence in tandem repeat patterns derived from 24 distinct loci within the genome (Supply et al., 2006). If two patients have the same classical genotyping pattern such as a 24-loci MIRU-VNTR pattern (or up to one locus

difference (Jonsson et al., 2014; Supply et al., 2006)) they are considered to be within a local transmission chain. The combination of spoligotyping and MIRU-VNTR-typing, where patterns must match in both methods to be considered a transmission link, is often considered the molecular gold standard for transmission linking and genotyping (Supply et al., 2006). However, examples of unlinked patients with identical patterns have been observed, suggesting that this threshold covers too broad a genetic diversity and timespan between infections (Gardy et al., 2011; Roetzer et al., 2013).

The application of (whole genome) sequence-based approaches for similarity analysis of Mtb isolates and cluster determination is known to have high discriminatory power when assessing transmission dynamics (Comas, 2017; Roetzer et al., 2013; Walker et al., 2013, 2018; Wyllie et al., 2018). Single nucleotide polymorphisms (SNPs) in the *pncA* gene are associated with resistance to pyrazinamide (PZA) and can be used to improve the discriminatory power of spoligotyping in a method referred to as SpoNC (Said et al., 2016). However, this is limited by the low occurrence of PZA resistance, even in MDR-TB isolates (Kurbatova, Cavanaugh, Dalton, S. Click, & Cegielski, 2013; Ngabonziza et al., 2017; Xu et al., 2016; Zignol et al., 2016). The advent of widespread whole genome sequencing (WGS) capabilities has allowed for highly discriminatory analyses of Mtb strains either using core genome multi-locus sequence typing (cgMLST) (Thomas A. Kohl et al., 2014) or SNP distances (Bjorn-Mortensen et al., 2016; Gardy et al., 2011; Roetzer et al., 2013; Walker et al., 2013, 2018). WGS-based approaches compare the genetic relatedness of the genomes of the clinical strains under consideration, albeit usually excluding large repetitive portions of the genome, with the assumption that highly similar strains are linked by a recent transmission event (Roetzer et al., 2013; Walker et al., 2013). Although many SNP cut-offs for linking isolates have been proposed (Hatherell

et al., 2016), the most commonly employed is based on the finding that a 5 SNP cut-off will cluster the genomes of strains from the majority of epidemiologically linked TB patients, with an upper bound of 12 SNPs between any two linked isolates (Walker et al., 2013). The widespread use of WGS has quickly pushed these cut-offs to be considered the new molecular gold standard of recent transmission linking, although SNP distances may vary for technical reasons (e.g. assembly pipelines or filter criteria (Guthrie & Gardy, 2017)) and between study populations e.g. high and low incidence settings (Bjorn-Mortensen et al., 2016).

In addition to cluster detection, uncovering the timing of transmission events within a given cluster is highly useful information for TB control e.g. for assessing the impact of interventions on the spread of an outbreak. Accordingly, knowledge of the rate change associated with different genotyping methods is essential for correct timing. The whole genome mutation rate of Mtb strains has been estimated by several studies as between 10^{-7} and 10^{-8} substitutions per site per year or ~0.3-0.5 SNPs per genome per year (Bryant et al., 2013; Duchêne et al., 2016; Eldholm et al., 2015; Eldholm & Balloux, 2016; Roetzer et al., 2013; Walker et al., 2013) while the rate of change in the MIRU-VNTR loci specifically is known to be quicker ($\sim 10^{-3}$) (Ragheb et al., 2013; Wirth et al., 2008). Since these mutation rates have been shown to also vary by lineage (Duchêne et al., 2016; Ford et al., 2013) and over short periods of time (Bryant et al., 2013), such variation needs to be accounted for, e.g. in Bayesian phylogenetic dating techniques (Bryant et al., 2013; Merker et al., 2015; Wirth et al., 2008).

Considering the multiple genotyping methods currently available, many of them proposed as a “gold standard”, there is an urgent need to precisely define the individual capacity of each method to accurately detect recent transmission events and perform timing of outbreaks. To

provide this essential information, this study harnesses the power of WGS-based phylogenetic dating methods to assign timespans onto Mtbc transmission chains encompassed by the different genotypic clustering methods commonly used in TB transmission studies.

Results

In this study, we assessed 20 different approaches for generating putative *M. tuberculosis* transmission clusters (see methods for approaches and naming schemes) using a dataset of 324 phenotypically rifampicin resistant isolates collected 2005-2010 from retreatment cases in Kinshasa, Democratic Republic of Congo (DRC). These 20 sets of clustering patterns were then characterised using whole genome sequence data and the propensity for convergence of clustering patterns was estimated (see methods). Bayesian phylodynamic approaches implemented in BEAST-2 (Bouckaert et al., 2014) were then utilised to assign timespans to the transmission events estimated by each genotyping method.

As expected, both the genome- and membrane-based spoligotyping approaches (named Gen-Spo and Mem-Spo respectively), clustered the most strains, with the lowest resolution (i.e. highest clustering rate) (Figure 1, Table 1). Convergent evolution (defined as the same pattern observed in unrelated strains; see methods) was found to affect 39% (12) of Mem-Spo clusters and 25% (7) of Gen-Spo clusters. Additionally, some discrepancies between the Mem-Spo and Gen-Spo patterns of each isolate were observed, with 291 isolates (90%) having the same pattern in both Mem-Spo and Gen-Spo approaches with 1 mismatch allowed (Supplementary table 1). The remaining 33 isolates mismatched with 2 to 17 spacers (average of 5 spacers). Although MIRU-VNTR performed far better than spoligotyping, 16% (6) of

clustering patterns were influenced by convergence in this study (see methods) (Table 1, Figure 1). Mixed MIRU-VNTR patterns were observed in 18 isolates although this mixing was not observed in the WGS data.

WGS-based methods had by far the highest discriminatory power and low SNP cut-offs grouped isolates into smaller clusters (e.g. 2-10 isolates per cluster for a 5 SNP cut-off) (Table 1, Figure 1). When the clusters were expanded to better represent transmission chains using the novel phylogenetic inclusion method implemented here (see methods), the resulting SNP clusters often did not increase dramatically in size (Table 1). Discriminatory power and cluster sizes based on cgMLST alleles were similar to the SNP-based clusters (Table 1, Figure 1).

155 Table 1: Clustering method overview.

156 For each clustering method, the general features are outlined in the tables. a) All clusters for each method affected by convergence. b) Clusters
157 derived only from non-convergent patterns. c) SNP- and cgMLST-based methods Mean ages and 95% HPD ranges are based upon the BEAST2
158 estimates of clade mean heights.

a)

Method	Strains in clusters	Number of clusters	Percent of strains in clusters	Cluster sizes	Maximum SNP distances	Clustering rate
Gen-Spo	293	29	90.43	2-42	1-653	0.8148
GenSpo-MIRU	190	39	58.64	2-27	0-48	0.466
Gen-SpoNC	76	23	23.46	2-10	0-195	0.1636
Mem-Spo	276	33	85.19	2-39	1-685	0.75
MemSpo-MIRU	174	36	53.7	2-25	0-611	0.4259
Mem-SpoNC	64	18	19.75	2-10	0-21	0.142
MIRU-VNTR	207	38	63.89	2-30	0-611	0.5216
MIRU-NC	59	17	18.21	2-9	0-21	0.1296

b)

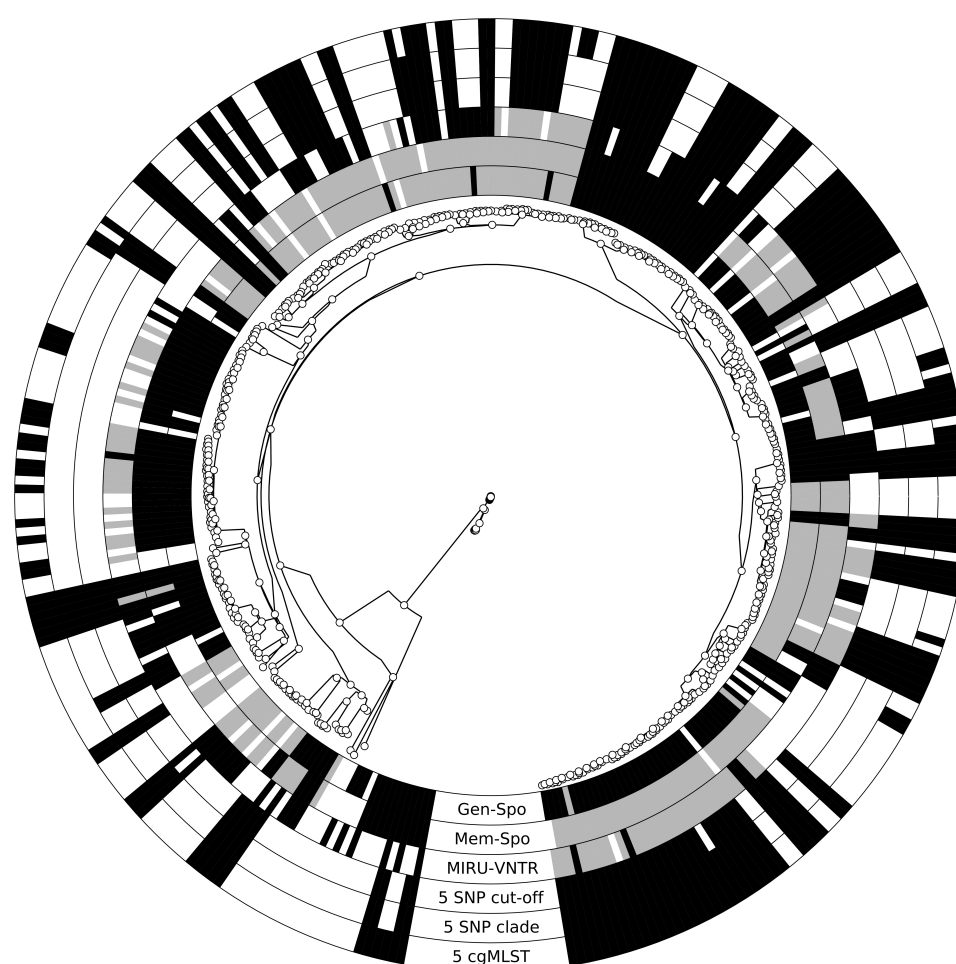
Method	Strains in clusters	Number of clusters	Percent of strains in clusters	Cluster sizes	Maximum SNP distances	Clustering rate	Mean Timespan	Timespan 95% HPD
Gen-Spo	191	22	58.95	2-37	1-322	0.5216	382.8101	0.96 - 1893.15
GenSpo-MIRU	77	22	23.77	2-10	0-48	0.1698	63.91188	0 - 278.77
Gen-SpoNC	34	11	10.49	2-6	0-14	0.071	21.52556	0.16 - 94.95
Mem-Spo	118	21	36.42	2-28	0-189	0.2994	141.1556	0.81 - 823.21
MemSpo-MIRU	50	12	15.43	2-10	2-48	0.1173	48.80688	0.8 - 216.31
Mem-SpoNC	15	5	4.63	2-4	0-14	0.0309	21.38239	1.03 - 97.91
MIRU-VNTR	121	32	37.35	2-11	0-48	0.2747	37.97812	0 - 162.27
MIRU-NC	25	9	7.72	2-3	1-11	0.0494	15.45935	0.77 - 58.38

c)

Method	Strains in clusters	Number of clusters	Percent of strains in clusters	Cluster sizes	Maximum SNP distances	Clustering rate	Mean Timespan	Timespan 95% HPD
0 SNP cluster	54	25	16.67	2-4	0	0.0895	4.309937	0 - 15.9
1 SNP cluster	74	29	22.84	2-6	0-2	0.1389	5.698197	0 - 23.54
5 SNP cluster	147	40	45.37	2-27	0-10	0.3302	13.4115	0 - 47.07
12 SNP cluster	242	47	74.69	2-34	0-23	0.6019	28.95219	0 - 102.58
0 SNP clade	66	21	20.37	2-9	0-9	0.1389	5.746077	0 - 23.96
1 SNP clade	80	27	24.69	2-9	0-9	0.1636	6.104103	0 - 25.74
5 SNP clade	149	40	45.99	2-28	0-11	0.3364	13.48716	0 - 47.41
12 SNP clade	253	45	78.09	2-39	0-27	0.642	29.73941	0 - 104.64
0 allele cgMLST	51	24	15.74	2-4	0-1	0.0833	4.231405	0.03 - 15.48
1 allele cgMLST	80	31	24.69	2-6	0-4	0.1512	6.371668	0 - 24.65
5 allele cgMLST	173	42	53.4	2-28	0-22	0.4043	17.54352	0 - 68.53
12 allele cgMLST	254	45	78.4	2-39	0-51	0.6451	30.08732	0 - 112.25

Figure 1: Clustering of *M. tuberculosis* isolates.

For a representative approach of each of the main methods (Mem-Spo, Gen-Spo, MIRU-VNTR, 5 SNP cut-off, 5 SNP clade and 5 cgMLST) the inclusion of an isolate into a cluster is outlined in the surrounding circles using GraPhlAn(Asnicar, Weingart, Tickle, Huttenhower, & Segata, 2015). If an isolate is in a cluster not affected by convergence, it is highlighted in black for the given method. If an isolate is in a cluster affected by convergence, it is shown in grey. The clustering based on all 20 approaches is shown in Supplementary Figure 1.



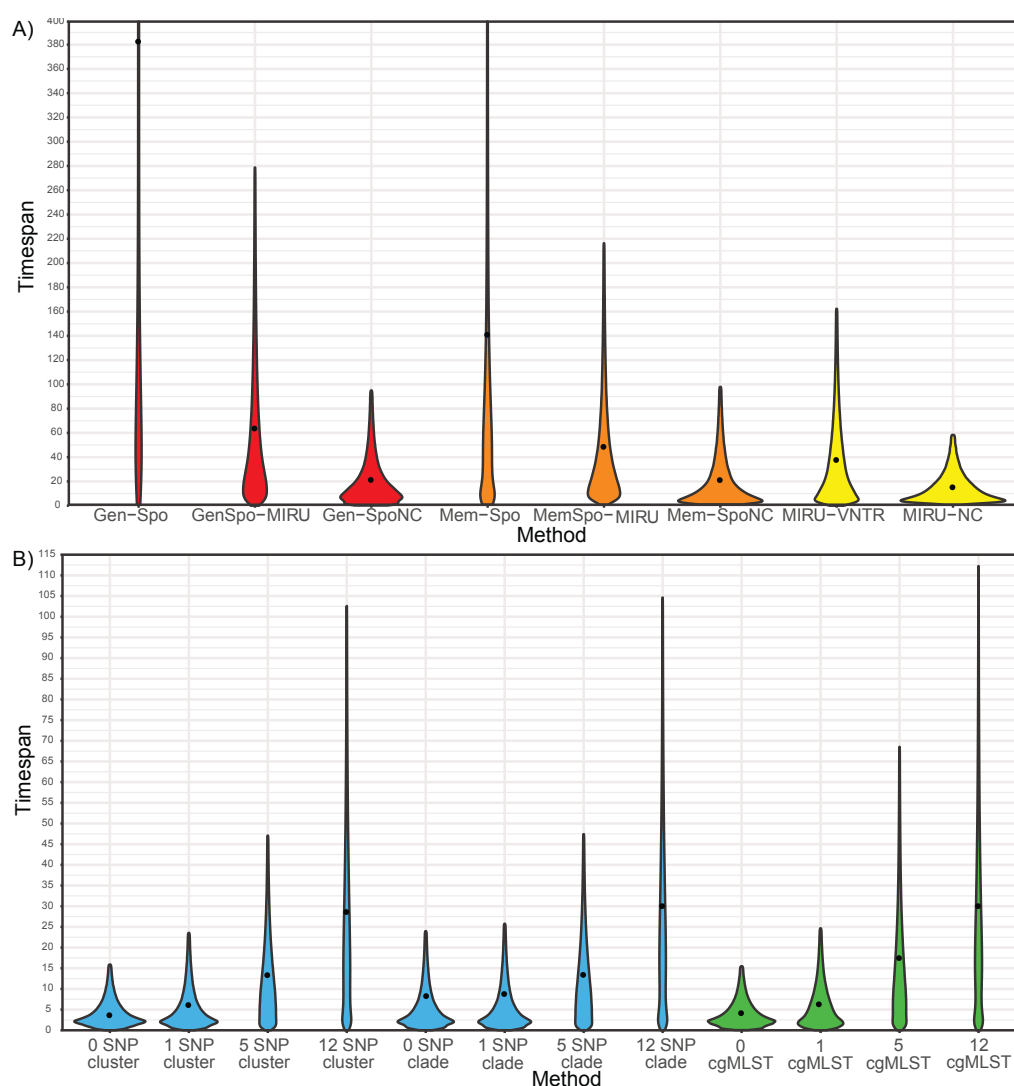
Statistical estimation of the timeframe associated with particular transmission chains showed large differences in estimated cluster ages between the genotyping approaches used (Table 1, Figure 2), correlating well with the difference in discriminatory power. Cluster ages are defined here as the most ancient transmission event that links any two isolates within a specific cluster. Thus, in phylogenetic terms, the cluster age is the difference in time between when the most recent common ancestor (MCRA) of the entire cluster existed and the date of isolation of the furthest isolate from this ancestor. The aggregate mean ages of clusters derived from spoligotyping approaches were found to often be several hundreds of years old (Gen-Spo: 383 years ago (95% HPD: 1-1893); Mem-Spo: 141 years ago (95% HPD: 1-823)) (Table 1b, Figure 2a). The addition of MIRU-VNTR or *pncA* mutation data to spoligotyping resulted in clusters that, on average, originated less than 100 years ago (Table 1b, Figure 2a). MIRU-VNTR alone gave similar cluster ages as to when combined with spoligotyping (MIRU-VNTR: 38 (0-162); GenSpo-MIRU: 64 (0-279); MemSpo-MIRU: 49 (1-216)) (Table 1b, Figure 2a).

Clusters based on SNP cut-offs correlated to 4 years of transmission using a 0 SNP cut-off (95% HPD: 0-16), 6 years using a 1 SNP cut-off (95% HPD: 0-24), 13 years using a 5 SNP cut-off (95% HPD: 0-47), and 29 years using a 12 SNP cut-off (95% HPD: 0-103) (Table 1c, Figure 2b). Extension on the tree using the phylogenetic inclusion approach to form SNP clades did not greatly increase the lengths of transmissions encompassed by clusters (one year increase, on average) (Table 1c). Similar findings were obtained when clusters were based on allele differences in the cgMLST method: 4 years of transmission using a 0 cgMLST cut-off (95% HPD: 0-15), 6 years using a 1 cgMLST cut-off (95% HPD: 0-25), 18 years using a 5 cgMLST cut-off

(95% HPD: 0-68), and 30 years using a 12 cgMLST cut-off (95% HPD: 0-112) (Table 1c, Figure 2b)

Figure 2: Timespans associated with transmission clusters

For each clustering method, the timespan associated with a cluster was estimated using BEAST-2. The ages of each cluster (Y-axis) was aggregated per clustering method (X-axis). Violin plots show the mean (black dot) for timespans along with the proportion of clusters with a given age (coloured kernel plots). Methods are split as follows: A) Spoligotype-based (Gen-Spo-based (red), Mem-Spo-based (orange)) and MIRU-VNTR-based (yellow), B) SNP-based (blue) and cgMLST-based (green). Note the y-axis is different for each and panel A) is cut at 400 years.



Discussion

The term 'recent transmission' is often applied to gain a better understanding of the current transmission dynamics of pathogens in a given population. However, little data is available on how recent a likely transmission event occurred when measured with different genotyping methods. To get a better understanding of the discriminatory power of different classical genotyping techniques and WGS-based approaches in relation to outbreak timing, this study has performed an in-depth comparison of clustering rates and dated phylogenies obtained in a collection of 324 Mtb strains from a high incidence setting (Kinshasa, DRC). With a whole genome phylodynamic approach employed as a gold standard, our study demonstrates that each genotyping method was associated with a specific discriminatory power resulting in clusters representing vastly different time periods of transmission events (Table 1 and Figure 2). This has significant implications for data interpretations e.g. when selecting and utilising different genotyping methods/clustering approaches for epidemiological studies and assessing the effectiveness of public health intervention strategies.

As the most extreme example, spoligotyping-derived clusters were associated with transmission events that can be hundreds of years old. This low discriminatory power coupled with the high rate of convergent evolution (the same spoligotype pattern found in phylogenetically distant isolates) in both Mem-Spo and Gen-Spo add weight to the previous suggestion that these techniques are not suitable for recent transmission studies (Comas, Homolka, Niemann, & Gagneux, 2009), although Mem-Spo may be of use as a low-cost method of sorting Mtb strains into the seven primary lineages (Filliol et al., 2006; Kato-Maeda et al., 2011). Differences between Mem-Spo and Gen-Spo patterns from the same isolate were observed for 10% of isolates in this study, even after rechecking of patterns,

requiring more investigation into which method is closer to the ‘true’ spoligotyping pattern within a genome (Coll et al., 2012; Mokrousov et al., 2016; Warren et al., 2002; Xia, Teo, & Ong, 2016).

In line with previous findings (Comas et al., 2009; Scott et al., 2005), convergent evolution of 24-loci MIRU-VNTR patterns was rarer than observed for spoligotyping, but did occur in 16% of MIRU-VNTR-based clusters. Additionally, the transmission times encompassed by MIRU-VNTR clusters spanned several decades (Table 1b, Figure 2a), confirming previous studies showing over-estimation of recent transmission with this method (Bjorn-Mortensen et al., 2016; Roetzer et al., 2013; Stucki et al., 2016; Wyllie et al., 2018).

The combination of MIRU-VNTR or spoligotyping with *pncA* mutations (MIRU-NC and Gen-SpoNC/Mem-SpoNC) appeared to reflect true clusters of PZA resistance transmission based on the relatively young ages of such transmission clusters (Table 1b). Thus, as discussed before (den Hertog, Sengstake, & Anthony, 2015; Sengstake et al., 2017), although transmission of *pncA* mutations seems to occur, further investigation is needed to find out whether *pncA* mutants are less transmissible than those with a wildtype gene.

For defining transmission events that occurred in more recent time frames before sampling, WGS-based methods (SNP or cgMLST) were found to be better suited than classical genotyping methods (Table 1, Figure 2). The 12 SNP cut-off, currently the recommended upper bound for clustering isolates, likely defines transmission events that occurred on average three decades prior to sampling, similar in age to clusters estimated by MIRU-VNTR. This suggests that the 12 SNP cluster method may be a good replacement for MIRU-VNTR as

it detects larger transmission networks spanning similar transmission time periods but is less affected by convergent evolution. Isolates clustered at identical (0 SNP) or nearly identical (1 SNP) cut-offs were found to represent transmission events occurring four to six years previous. These findings correlate well with previous studies where confirmed contact tracing-based epidemiological links were found between patients that were two (Walker et al., 2014), three (Roetzer et al., 2013) or five (Walker et al., 2013) SNPs apart. Indeed, a recent study of a cross-country MDR-TB outbreak found only a maximum of two SNP differences between all 29 isolates involved in the origin of the outbreak (Walker et al., 2018). Although this supports their use for detection or exclusion of very recent transmission, this low variability between isolates makes robust identification of transmission direction impossible, especially during short timespans.

Comparisons between the SNP-based (using almost all genomic differences) and the cgMLST-based cluster detection (using a defined core set of genes) demonstrated that the latter approach gives similar estimations to full SNP approaches. However, as current SNP assembly pipelines for Illumina data exclude repetitive region such as PE/PPE genes, larger differences between cgMLST and full SNP estimation may be seen once all aspects of the genome can be utilised.

Different clustering approaches can be applied when grouping isolates by SNP distance. Two partitional clustering methods are primarily utilised: either the creation of tight clusters (where the maximum pairwise distance between isolates in a cluster is less than the SNP cut-off; e.g. (Thomas A. Kohl et al., 2014)) or loose clusters (where each isolate is less than the SNP cut-off distance from at least one other isolate in the cluster; e.g. (Walker et al., 2014)).

Tight clusters ensure high connectivity within clusters, but may result in isolates belonging to multiple groups, making interpretation and delineation of transmission events difficult. Loose clusters (the definition used in this study), separate isolates into non-overlapping clusters, but may result in low connectivity within clusters. Here we present an extension of the loose cluster, termed the phylogenetic inclusion method, which adds all other isolates with the same phylogenetically defined common ancestor to the cluster, potentially identifying larger circulating genotypes. Tight, loose and phylogenetic inclusion clusters each aim to define different levels of connectivity through time, an aspect that should be considered when selecting the appropriate clustering approach.

The mutation rate of *M. tuberculosis* has been estimated to be between 10^{-7} and 10^{-8} substitutions per site per year (Duchêne et al., 2016; Merker et al., 2015; Roetzer et al., 2013). Within the Bayesian analysis employed here, the mutation rate was free to vary between these values but was found to strongly favour $\sim 3 \times 10^{-8}$ (ESS > 1000 for all runs), translating to approximately 0.3 SNPs per genome per year. While the mutation rate used here is primarily applicable for lineage 4 (which most of this dataset is comprised of) and in line with previous estimates for this lineage (Duchêne et al., 2016), it may be similar in other lineages, although this has only been shown for lineage 2 (Duchêne et al., 2016; Merker et al., 2015). Thus, per-lineage estimates are required for all seven lineages to ensure similar transmission times are linked to genotyping methods across the whole population diversity of the Mtb.

While this study has many advantages due to its five year population based design in an endemic setting coupled with the application of three different genotyping methods (membrane based spoligotyping analysis, 24-locus MIRU-VNTR and WGS), future

confirmatory studies could address the following drawbacks that are inherent to genomic epidemiology (Comas, 2017; Guthrie & Gardy, 2017): 1) studies employing contact tracing and/or digital epidemiology (Salathé et al., 2012) in conjunction with these genotyping methods can help confirm transmission times associated with different clusters; 2) as outlined above, strains of other lineages of the Mtb should be analysed in a similar fashion to ensure transferability of findings across the entire complex; 3) a broad range of drug resistance profiles should be included to fully assess the impact of such mutations on transmission estimates; 4) improved WGS methods, such as directly from clinical samples to help reduce culture biases (Sanoussi, Affolabi, Rigouts, Anagonou, & de Jong, 2017) and longer reads (e.g. PacBio SMRT or Nanopore MinION) to capture the entire genome, including repetitive regions such as PE/PPE genes known to impact genome remodelling (Ates et al., 2018; Phelan et al., 2016), will ensure that the maximum diversity between isolates is captured and 5) standardised SNP calling pipelines appropriate across all lineages, with high true positive/low false negative rates, will ensure that Mtb molecular epidemiology can be uniformly implemented and comparable across studies.

In conclusion, since each method was found to represent different timespans and clustering definitions, they can be used in a stratified manner in an integrated epidemiological and public health investigation addressing the transmission of Mtb strains. For instance, although spoligotyping clusters represented potentially very old transmission events, the low associated cost and its ability to be applied directly on sputum helps reduce culture bias and thus robustly assign lineages. Thus, spoligotyping and/or MIRU-VNTR would serve well as first-line surveillance of potential transmission events in the population, guiding further investigations and resource allocations.

322

323 These potential transmission hotspots could be further investigated with contact tracing
 324 and/or WGS. Employment of different cut-offs and clustering approaches to WGS data can
 325 then address several questions. The 12 SNP cluster/clade or 12 allele cgMLST approaches
 326 serve well for high level surveillance targeting larger (older) transmission networks, akin to
 327 what is currently often done using MIRU-VNTR (e.g. (Guthrie et al., 2018; Walker et al., 2018)).
 328 Recent transmission events can then be detected through employment of low SNP or cgMLST-
 329 based cut-offs (e.g. 5 SNPs for transmission in the past 15 years or 0-1 SNPs for transmission
 330 in the past 5 years). These clusters can then be linked to historical isolates or other clusters
 331 through employment of the phylogenetic inclusion method to resolve the local circulating
 332 genotypes. This is especially useful if bursts of sampling are undertaken such as in drug
 333 resistance surveys (WHO, 2015), which are increasingly employing WGS approaches (Cabibbe
 334 & Cirillo, 2016; Zignol et al., 2016, 2018). Alternatively, in high incidence/low diversity settings
 335 where amalgamation of clusters may inadvertently obscure distinct hotspots of transmission
 336 at different time points, subdivision into distinct time-dependant clusters can be undertaken
 337 using the algorithm presented in such a study in East Greenland (Bjorn-Mortensen et al.,
 338 2016).

339

340 Overall, phylodynamic approaches applied to whole genome sequences, as undertaken here,
 341 are recommended to fully investigate the specific transmission dynamics within a study
 342 population to account for setting-specific conditions, such as low/high TB incidence, low/high
 343 pathogen population diversity, sampling fractions and social factors influencing transmission.
 344 Thus, each genotyping method can be employed as part of an overall evidence gathering

program for transmission, placing molecular epidemiological approaches as an integral part in tracking and stopping the spread of TB.

Materials and Methods

Dataset and sequencing

A set of 324 isolates from Kinshasa, Democratic Republic of Congo were collected from consecutive retreatment TB patients between 2005 and 2010 at TB clinics, servicing an estimated 30% of the population of Kinshasa. All isolates were phenotypically resistant to rifampicin (RR-TB) and the majority are also isoniazid resistant (i.e. MDR-TB). Use of the stored isolates without any linked personal information was approved by the health authorities of the DRC and the Institutional Review Board of the ITM in Antwerp (ref no 945/14). Libraries for whole genome sequencing were prepared from extracted genomic DNA with the Illumina Nextera XT kit, and run on the Illumina NextSeq platform in a 2x151bp run according to manufacturer's instructions. Illumina read sets will be available at ReSeqTB (platform.reseqtb.org) upon publication.

Genome reconstruction and maximum likelihood phylogeny estimation

The MTBseq pipeline(Thomas Andreas Kohl et al., 2018) was used to detect the SNPs for each isolate using the H37Rv reference genome (NCBI accession number NC000962.3) (Lew, Kapopoulou, Jones, & Cole, 2011; Médigue, Cole, Camus, & Pryor, 2002). Sites known to be involved in drug resistance (as outlined in the PhyResSE list of drug mutations v27 (Feuerriegel et al., 2015)) were excluded from the alignment and additional filtering of sites with ambiguous calls in >5% of isolates and those SNPs within a 12bp window of each other was also applied.

369

370 The SNP alignment of all isolates was used as the basis for creating a maximum likelihood
371 (ML) phylogeny. RAxML-NG version 0.5.1b (Kozlov, 2017) was used to reconstruct the
372 phylogeny from this alignment using a GTR+GAMMA model of evolution, accounting for
373 ascertainment bias (Lewis, 2001) with the Stamatakis reconstituted DNA approach (Leaché et
374 al., 2015) and site repeat optimisation (Kobert, Stamatakis, & Flouri, 2016) with 20 different
375 starting trees and 100 bootstraps. All subsequent topology visualisation was undertaken using
376 FigTree version 1.4.3 (Rambaut, 2016) and GraPhlAn (Asnicar et al., 2015).

377

378 Transmission cluster estimation methods

379 Several standard transmission clustering approaches were chosen for comparison and
380 analysis. For each method, the total SNP distances were calculated to investigate the range
381 of variability encompassed within each cluster. Maximum SNP distances were derived from
382 pairwise comparisons of isolates within the SNP alignment using custom python scripts. A
383 clustering rate was calculated for each method using the formula $(n_c - c)/n$, where n_c is the
384 total number of isolates clustered by a given method, c is the number of clusters, and n is the
385 total number of isolates in the dataset ($n=324$).

386

387 Spoligotyping

388 Spoligotype patterns were estimated by 2 methods: membrane-based and genome-based.
389 Membrane-based patterns were obtained following the previously published protocol
390 (Kamerbeek et al., 1997). This method is referred to as Mem-Spo. Genome-based
391 spoligotyping was derived from the Illumina reads of each isolate using SpoTyping v2.1 (Xia
392 et al., 2016). Reads (both forward and reverse) were input to SpoTyping with default

parameters and the 43 spacer values were extracted from the output. This method is referred to as Gen-Spo. For both methods, isolates were said to be clustered if all 43 spacers matched.

MIRU-VNTR

Genotyping by MIRU-VNTR was undertaken as previously described (Supply et al., 2006). 2 µl of DNA was extracted from cultures and amplified using the 24 loci MIRU-VNTR typing kit (Genoscreen, Lille, France). Analysis of patterns was undertaken using the ABI 3500 automatic sequencer (Applied Biosystems, California, USA) and Genemapper software (Applied Biosystems). Isolates were said to be clustered if all 24 loci matched. MIRU-VNTR patterns were also combined with spoligotyping patterns for additional refinement of clusters. Isolates were clustered if both the spoligotyping pattern and the 24 loci MIRU-VNTR pattern matched. These clustering methods are referred to as MemSpo-MIRU and GenSpo-MIRU.

SpoNC

Transmission estimation using spoligotyping has been shown to be improved if combined with *pncA* mutations (Said et al., 2016). This method, referred to as SpoNC, was applied to both Mem-Spo (Mem-SpoNC) and Gen-Spo (Gen-SpoNC). Mutations in *pncA* were extracted from the MTBseq tabular output for each isolate. All mutations were selected, regardless of drug resistance association, as is done in the SpoNC approach. The upstream promoter region of *pncA* did not reveal any mutations in this dataset. Isolates were said to be clustered if all 43 spacers matched and the *pncA* mutation was the same in both isolates. MIRU-VNTR patterns were combined with *pncA* mutations in a similar manner. This is referred to as MIRU-NC.

SNP cut-off clustering

The advent of whole genome reconstruction has allowed for genome-based comparisons for transmission clustering. Previous work has suggested that linked and recent transmission can be estimated by comparison of SNP differences between isolates. The cut-offs proposed by Walker *et al.* (Walker et al., 2013) are the most widely used and have been employed in multiple studies (Casali et al., 2016; Tessema et al., 2017; Witney et al., 2016). In this study, we employed both the 5 SNP (proposed by Walker *et al.* as the likely boundary for linked transmission) and 12 SNP cut-offs (proposed maximum boundary) for cluster definition. Additionally, we employed lower cut-offs of 0 and 1 SNPs to look for clusters of very highly related isolates. Pairwise SNP distances were calculated between all isolates. A loose cluster definition was used, where every isolate in a cluster at most the SNP cut-off from at least 1 other isolate in the cluster.

Phylogenetic information was used to extend these SNP-based clusters to include any other isolates that share the same most recent common ancestor (MRCA). These isolates may exceed the SNP cut-off but should be included as, through sharing an MRCA, they are intrinsically within the same putative transmission chain. The MRCA is defined here as the internal node in a phylogenetic tree that is shared by all the isolates within the putative SNP-based cluster. This extension was achieved by mapping each SNP cluster onto the ML phylogenetic tree and the MRCA (shared internal node) of all isolates was found using DendroPy v4.0.3 (Sukumaran & Holder, 2010). Any additional taxa with the same MRCA were then added to the transmission cluster (Supplemental Figure 2). In other words, all leaf nodes of the MRCA internal node were labelled as being part of the putative transmission cluster. We call this approach the phylogenetic inclusion method and extended clusters are hereafter

referred to as extended SNP clades to distinguish them from SNP clusters as created by the standard non-phylogenetic method above. The python script that implements this method can be found at <https://github.com/conmeehan/pathophy>.

cgMLST

An alternative approach to clustering using WGS data is the concept of core genome MLST (cgMLST) patterns (Thomas A. Kohl et al., 2014). Since SNP detection can be variable between assembly pipelines, SNP clusters between studies may be difficult to compare. The cgMLST approach standardises comparisons by ensuring the same core genes are always compared. BAM files for all isolates are input into Ridom SeqSphere⁺ software (Ridom GmbH, Münster, Germany) to compile an allelic distance matrix based on the cgMLST v2 scheme consisting of 2,891 core Mtb genes. Loose clusters were then defined as above using allelic differences of 0, 1, 5 and 12 as cut-offs. These methods are referred to as 0/1/5/12 cgMLST respectively.

Detection of convergent evolution

Convergent evolution towards identical patterns may occur for Spoligotyping, MIRU-VNTR and *pncA* mutations (Driscoll, 2009; Miotto et al., 2014; Scott et al., 2005; Warren et al., 2002). Convergence was detected and cross-checked with two methods. Firstly, Mtb lineage and sub-lineage numbering (Coll et al., 2014) was applied to all isolates based on the PhyResSE lineage-defining SNP list v27 (Feuerriegel et al., 2015). If the same clustering pattern was observed in two different sub-lineages, with other patterns seen in-between, this was flagged as potential convergence. Additional convergence confirmation was also undertaken using phylogenetic distances, as estimated by DendroPy. If the phylogenetic distance (combined branch lengths that separate 2 isolates) between two isolates with identical clustering

patterns was greater than 0.0005, this was flagged as potential convergence. Any isolates flagged by both methods (lineage-based and distance-based) were marked as clustered by convergence. For example, if isolates with the same spoligotyping pattern appeared in lineage 4,1 and 4,6 with different patterns in-between and these isolates were distant on the tree (distance greater than 0.0005), this was confirmed as a convergent pattern. Convergence was checked for all approaches except the SNP cut-off clusters/clades, which, by definition, could not be convergent. Clustering methods that combined two other methods (e.g. Gen-SpoNC) were first checked separately for convergence and then combined to create the final clusters.

Estimation of transmission times

To estimate the age and timespan of potential transmission clusters, SNP alignments were created from the convergence-free version of the five primary clustering types: Gen-Spo, Mem-Spo, MIRU-VNTR, extended 12 SNP clades and 12 allele cgMLST. All other methods are sub-clustering methods of at least one of these five methods (e.g. Mem-SpoNC clusters are inherently included in any Mem-Spo clusters, and all SNP-based clusters are sub-clusters of the 12 SNP clades).

A Bayesian approach to transmission time estimation was then undertaken. The SNP alignments were created as above for the five high-level clustering types. Each cluster method alignment was separately input to BEAST-2 v2.4.7 (Bouckaert et al., 2014) to create a time tree for those isolates. These phylogenies were built using the following priors: GTR+GAMMA substitution model, a log-normal relaxed molecular clock model to account for variation in mutation rates (Alexei J Drummond, Ho, Phillips, Rambaut, & Rambaut, 2006) and coalescent constant size demographic model (A. J. Drummond, Rambaut, Shapiro, & Pybus, 2005), both

of which have been found to be suitable for lineage 4 isolates in a previous study (Bjorn-Mortensen et al., 2016). The MCMC chain was run six times independently per alignment with a length of at least 400 million, sampled every 40,000th step (Gen-Spo: 400 million; extended 12 SNP & cgMLST: 500 million; MIRU & MemSpo: 600 million). A log normal prior (mean 1.5×10^{-7} ; variance 1.0) was used for the clock model to reflect the previously estimated mutation rate of *M. tuberculosis* lineage 4 (Bryant et al., 2013; Duchêne et al., 2016; Eldholm et al., 2015; Eldholm & Balloux, 2016; Roetzer et al., 2013; Walker et al., 2013), while allowing for variation as previously suggested (Bryant et al., 2013). A 1/X non-informative prior was selected for the population size parameter of the demographic model. Isolation dates were used as informative heterochronous tip dates and the SNP alignment was augmented with a count of invariant sites for each of the four nucleotide bases to avoid ascertainment bias (Leaché et al., 2015). Tracer v1.6 (Rambaut, Suchard, Xie, & Drummond, 2013) was used to determine adequate mixing and convergence of chains (ESS >150) after a 25% burn-in. The chains were combined via LogCombiner v2.4.8 (Bouckaert et al., 2014) to obtain a single chain for each clustering type with high (>1000) effective sample sizes. The tree samples were combined in the same manner and resampled at a lower frequency to create thinned samples of (minimum) 20,000 trees.

The timespan of transmission events estimated by each method was then calculated as follows: for each cluster created by the given method, we defined the MRCA node as the internal node that connects all taxa in that cluster. The youngest node was then defined as the tip that is furthest from this MRCA within the clade (i.e. the tip descendant from that node that was sampled closest to the present time). For each retained tree in the MCMC process, the difference in age between the MRCA node and youngest node was calculated. This gave

a distribution of likely maximum transmission event times within that cluster. For each method, these per-cluster aggregated ages were then combined across all clusters to give a per-method distribution of transmission event times represented by the clusters. The 95% HPD interval of these distributions was calculated with the LaplacesDemon p.interval function (Statisticat, 2016) in R v3.4.0 (R Core Team, 2017) and the distribution within this interval for each method along with the mean based upon this interval were then visualized in violin plots per clustering method using ggplot2 (Wickham, 2009) in R.

Acknowledgements/Funding

The authors would like to thank Armand Van Deun and Koen Vandelannoote for valuable discussion and input and Cecile Uwizeye for aid with spoligotyping. This work was supported by an ERC grant (INTERRUPTB; no. 311725) to BdJ, FG and CJM; an ERC grant to TS (PhyPD; no. 335529); an FWO PhD fellowship to PM (grant number 1141217N); the German Centre for Infection Research (DZIF) for TAK, MM and SN; a SNF SystemsX grant (TBX) to JP and TS and a Marie Heim-Vögtlin fellowship granted to DK by the Swiss National Science Foundation. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government – department EWI;

Competing interests

The authors declare there are no competing interests attached to this work.

References

Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C., & Segata, N. (2015). Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ*, 3, e1029. <https://doi.org/10.7717/peerj.1029>

536 Ates, L. S., Dippenaar, A., Ummels, R., Piersma, S. R., van der Woude, A. D., van der Kuij, K.,
537 ... Bitter, W. (2018). Mutations in ppe38 block PE_PGRS secretion and increase
538 virulence of Mycobacterium tuberculosis. *Nature Microbiology*, 3(2), 181–188.
539 <https://doi.org/10.1038/s41564-017-0090-6>

540 Behr, M. A., Hopewell, P. C., Antonio Paz, E., Masae Kawamura, L., Schechter, G. F., & Small,
541 P. M. (1998). Predictive Value of Contact Investigation for Identifying Recent
542 Transmission of Mycobacterium tuberculosis. *American Journal of Respiratory and*
543 *Critical Care Medicine*, 158(2), 465–469.
544 <https://doi.org/10.1164/ajrccm.158.2.9801062>

545 Bjorn-Mortensen, K., Lillebaek, T., Koch, A., Soborg, B., Ladefoged, K., Sørensen, H. C. F., ...
546 Andersen, A. B. (2017). Extent of transmission captured by contact tracing in a
547 tuberculosis high endemic setting. *European Respiratory Journal*, 49(3). Retrieved from
548 <http://erj.ersjournals.com/content/49/3/1601851>

549 Bjorn-Mortensen, K., Soborg, B., Koch, A., Ladefoged, K., Merker, M., Lillebaek, T., ... Kohl, T.
550 A. (2016). Tracing Mycobacterium tuberculosis transmission by whole genome
551 sequencing in a high incidence setting: a retrospective population-based study in East
552 Greenland. *Scientific Reports*, 6, 33180. <https://doi.org/10.1038/srep33180>

553 Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., ... Drummond, A. J.
554 (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS*
555 *Computational Biology*, 10(4), e1003537.
556 <https://doi.org/10.1371/journal.pcbi.1003537>

557 Bryant, J. M., Schürch, A. C., van Deutekom, H., Harris, S. R., de Beer, J. L., de Jager, V., ... van
558 Soolingen, D. (2013). Inferring patient to patient transmission of Mycobacterium
559 tuberculosis from whole genome sequencing data. *BMC Infectious Diseases*, 13(1), 110.

<https://doi.org/10.1186/1471-2334-13-110>

Cabibbe, A. M., & Cirillo, D. M. (2016). Best approaches to drug-resistance surveillance at the country level. *International Journal of Mycobacteriology*, 5, S40–S41.

<https://doi.org/10.1016/j.ijmyco.2016.09.010>

Casali, N., Broda, A., Harris, S. R., Parkhill, J., Brown, T., & Drobniewski, F. (2016). Whole Genome Sequence Analysis of a Large Isoniazid-Resistant Tuberculosis Outbreak in London: A Retrospective Observational Study. *PLOS Medicine*, 13(10), e1002137.

<https://doi.org/10.1371/journal.pmed.1002137>

Coll, F., Mallard, K., Preston, M. D., Bentley, S., Parkhill, J., McNerney, R., ... Clark, T. G.

(2012). SpolPred: rapid and accurate prediction of Mycobacterium tuberculosis spoligotypes from short genomic sequences. *Bioinformatics (Oxford, England)*, 28(22), 2991–3. <https://doi.org/10.1093/bioinformatics/bts544>

Coll, F., McNerney, R., Guerra-Assunção, J. A., Glynn, J. R., Perdigão, J., Viveiros, M., ... Clark, T. G. (2014). A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nature Communications*, 5, 4812. <https://doi.org/10.1038/ncomms5812>

Comas, I. (2017). Genomic Epidemiology of Tuberculosis (pp. 79–93). Springer, Cham. https://doi.org/10.1007/978-3-319-64371-7_4

Comas, I., Homolka, S., Niemann, S., & Gagneux, S. (2009). Genotyping of Genetically Monomorphic Bacteria: DNA Sequencing in Mycobacterium tuberculosis Highlights the Limitations of Current Methodologies. *PLoS ONE*, 4(11), e7815.

<https://doi.org/10.1371/journal.pone.0007815>

den Hertog, A. L., Sengstake, S., & Anthony, R. M. (2015). Pyrazinamide resistance in Mycobacterium tuberculosis fails to bite? *Pathogens and Disease*, 73(6), ftv037.

<https://doi.org/10.1093/femspd/ftv037>

Diel, R., Schneider, S., Meywald-Walter, K., Ruf, C.-M., Rusch-Gerdes, S., & Niemann, S. (2002). Epidemiology of Tuberculosis in Hamburg, Germany: Long-Term Population-Based Analysis Applying Classical and Molecular Epidemiological Techniques. *Journal of Clinical Microbiology*, 40(2), 532–539. <https://doi.org/10.1128/JCM.40.2.532-539.2002>

Driscoll, J. R. (2009). Spoligotyping for Molecular Epidemiology of the Mycobacterium tuberculosis Complex. In *Methods in molecular biology (Clifton, N.J.)* (Vol. 551, pp. 117–128). https://doi.org/10.1007/978-1-60327-999-4_10

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., Rambaut, A., & Rambaut, A. (2006). Relaxed Phylogenetics and Dating with Confidence. *PLoS Biology*, 4(5), e88. <https://doi.org/10.1371/journal.pbio.0040088>

Drummond, A. J., Rambaut, A., Shapiro, B., & Pybus, O. G. (2005). Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. *Molecular Biology and Evolution*, 22(5), 1185–1192. <https://doi.org/10.1093/molbev/msi103>

Duchêne, S., Holt, K. E., Weill, F.-X., Le Hello, S., Hawkey, J., Edwards, D. J., ... Holmes, E. C. (2016). Genome-scale rates of evolutionary change in bacteria. *Microbial Genomics*, 2(11), e000094. <https://doi.org/10.1099/mgen.0.000094>

Eldholm, V., & Balloux, F. (2016). Antimicrobial Resistance in Mycobacterium tuberculosis: The Odd One Out. *Trends in Microbiology*. <https://doi.org/10.1016/j.tim.2016.03.007>

Eldholm, V., Monteserin, J., Rieux, A., Lopez, B., Sobkowiak, B., Ritacco, V., ... Lemey, P. (2015). Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. *Nature Communications*, 6, 7119. <https://doi.org/10.1038/ncomms8119>

Feuerriegel, S., Schleusener, V., Beckert, P., Kohl, T. A., Miotto, P., Cirillo, D. M., ... Fellenberg, K. (2015). PhyResSE: a Web Tool Delineating Mycobacterium tuberculosis

Antibiotic Resistance and Lineage from Whole-Genome Sequencing Data. *Journal of Clinical Microbiology*, 53(6), 1908–1914. <https://doi.org/10.1128/JCM.00025-15>

Filliol, I., Motiwala, A. S., Cavatore, M., Qi, W., Hazbón, M. H., Bobadilla del Valle, M., ... Alland, D. (2006). Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. *Journal of Bacteriology*, 188(2), 759–72. <https://doi.org/10.1128/JB.188.2.759-772.2006>

Ford, C. B., Shah, R. R., Maeda, M. K., Gagneux, S., Murray, M. B., Cohen, T., ... Fortune, S. M. (2013). *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nature Genetics*, 45(7), 784–90. <https://doi.org/10.1038/ng.2656>

Fox, G. J., Barry, S. E., Britton, W. J., & Marks, G. B. (2012). Contact investigation for tuberculosis: a systematic review and meta-analysis. *European Respiratory Journal*, 41(1). Retrieved from <http://erj.ersjournals.com/content/41/1/140.short>

Gardy, J. L., Johnston, J. C., Sui, S. J. H., Cook, V. J., Shah, L., Brodtkin, E., ... Tang, P. (2011). Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. *New England Journal of Medicine*, 364(8), 730–739. <https://doi.org/10.1056/NEJMoa1003176>

Goguet de la Salmonière, Y. O., Li, H. M., Torrea, G., Bunschoten, A., van Embden, J., & Gicquel, B. (1997). Evaluation of spoligotyping in a study of the transmission of *Mycobacterium tuberculosis*. *Journal of Clinical Microbiology*, 35(9), 2210–4. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/9276389>

Guernier, V., Sola, C., Brudey, K., Guégan, J.-F., & Rastogi, N. (2008). Use of cluster-graphs

from spoligotyping data to study genotype similarities and a comparison of three indices to quantify recent tuberculosis transmission among culture positive cases in French Guiana during a eight year period. *BMC Infectious Diseases*, 8(1), 46.
<https://doi.org/10.1186/1471-2334-8-46>

Guthrie, J. L., & Gardy, J. L. (2017). A brief primer on genomic epidemiology: lessons learned from *Mycobacterium tuberculosis*. *Annals of the New York Academy of Sciences*, 1388(1), 59–77. <https://doi.org/10.1111/nyas.13273>

Guthrie, J. L., Kong, C., Roth, D., Jorgensen, D., Rodrigues, M., Hoang, L., ... Gardy, J. L. (2018). Molecular Epidemiology of Tuberculosis in British Columbia, Canada: A 10-Year Retrospective Study. *Clinical Infectious Diseases*, 66(6), 849–856.
<https://doi.org/10.1093/cid/cix906>

Hatherell, H.-A., Colijn, C., Stagg, H. R., Jackson, C., Winter, J. R., & Abubakar, I. (2016). Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review. *BMC Medicine*, 14(1), 21. <https://doi.org/10.1186/s12916-016-0566-x>

Jonsson, J., Hoffner, S., Berggren, I., Bruchfeld, J., Ghebremichael, S., Pennhag, A., & Groenheit, R. (2014). Comparison between RFLP and MIRU-VNTR Genotyping of *Mycobacterium tuberculosis* Strains Isolated in Stockholm 2009 to 2011. *PLoS ONE*, 9(4), e95159. <https://doi.org/10.1371/journal.pone.0095159>

Kamerbeek, J., Schouls, L., Kolk, A., van Agterveld, M., van Soolingen, D., Kuijper, S., ... van Embden, J. (1997). Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *Journal of Clinical Microbiology*, 35(4), 907–14. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/9157152>

Kato-Maeda, M., Gagneux, S., Flores, L. L., Kim, E. Y., Small, P. M., Desmond, E. P., &

656 Hopewell, P. C. (2011). Strain classification of Mycobacterium tuberculosis: congruence
657 between large sequence polymorphisms and spoligotypes. *The International Journal of*
658 *Tuberculosis and Lung Disease : The Official Journal of the International Union against*
659 *Tuberculosis and Lung Disease*, 15(1), 131–3. Retrieved from
660 <http://www.ncbi.nlm.nih.gov/pubmed/21276309>

661 Kendall, E. A., Fofana, M. O., Dowdy, D. D. W., WHO, Dye, C., Garnett, G., ... al., et. (2015).
662 Burden of transmitted multidrug resistance in epidemics of tuberculosis: a transmission
663 modelling analysis. *The Lancet Respiratory Medicine*, 3(12), 963–972.
664 [https://doi.org/10.1016/S2213-2600\(15\)00458-0](https://doi.org/10.1016/S2213-2600(15)00458-0)

665 Klopper, M., Warren, R. M., Hayes, C., Gey van Pittius, N. C., Streicher, E. M., Müller, B., ...
666 Trollip, A. P. (2013). Emergence and Spread of Extensively and Totally Drug-Resistant
667 Tuberculosis, South Africa. *Emerging Infectious Diseases*, 19(3), 449–455.
668 <https://doi.org/10.3201/eid1903.120246>

669 Kobert, K., Stamatakis, A., & Flouri, T. (2016). Efficient Detection of Repeating Sites to
670 Accelerate Phylogenetic Likelihood Calculations. *Systematic Biology*, 66(2), syw075.
671 <https://doi.org/10.1093/sysbio/syw075>

672 Kohl, T. A., Diel, R., Harmsen, D., Rothgänger, J., Walter, K. M., Merker, M., ... Niemann, S.
673 (2014). Whole-genome-based Mycobacterium tuberculosis surveillance: a
674 standardized, portable, and expandable approach. *Journal of Clinical Microbiology*,
675 52(7), 2479–86. <https://doi.org/10.1128/JCM.00567-14>

676 Kohl, T. A., Utpatel, C., Schleusener, V., De Filippo, M. R., Beckert, P., Cirillo, D. M., &
677 Niemann, S. (2018). MTBseq: A comprehensive pipeline for whole genome sequence
678 analysis of Mycobacterium tuberculosis complex isolates. *In Preparation*.

679 Kozlov, A. (2017). RAXML-NG.

680 <https://doi.org/https://zenodo.org/record/888146#.Wm8r2Uso9TY>

681 Kurbatova, E. V., Cavanaugh, J. S., Dalton, T., S. Click, E., & Cegielski, J. P. (2013).
682 Epidemiology of Pyrazinamide-Resistant Tuberculosis in the United States, 1999–2009.
683 *Clinical Infectious Diseases*, 57(8), 1081–1093. <https://doi.org/10.1093/cid/cit452>

684 Leaché, A. D., Banbury, B. L., Felsenstein, J., de Oca, A. nieta-M., Stamatakis, A., K., S., ... J.,
685 W. (2015). Short Tree, Long Tree, Right Tree, Wrong Tree: New Acquisition Bias
686 Corrections for Inferring SNP Phylogenies. *Systematic Biology*, 64(6), 1032–1047.
687 <https://doi.org/10.1093/sysbio/syv053>

688 Lew, J. M., Kapopoulou, A., Jones, L. M., & Cole, S. T. (2011). TubercuList – 10 years after.
689 *Tuberculosis*, 91(1), 1–7. <https://doi.org/10.1016/j.tube.2010.09.008>

690 Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete
691 morphological character data. *Systematic Biology*, 50(6), 913–25. Retrieved from
692 <http://www.ncbi.nlm.nih.gov/pubmed/12116640>

693 Médigue, C., Cole, S. T., Camus, J.-C., & Pryor, M. J. (2002). Re-annotation of the genome
694 sequence of Mycobacterium tuberculosis H37Rv. *Microbiology*, 148(10), 2967–2973.
695 <https://doi.org/10.1099/00221287-148-10-2967>

696 Merker, M., Blin, C., Mona, S., Duforet-Frebourg, N., Lecher, S., Willery, E., ... Wirth, T.
697 (2015). Evolutionary history and global spread of the Mycobacterium tuberculosis
698 Beijing lineage. *Nature Genetics*, 47(3), 242–249. <https://doi.org/10.1038/ng.3195>

699 Merker, M., Kohl, T. A., Niemann, S., & Supply, P. (2017). The Evolution of Strain Typing in
700 the Mycobacterium tuberculosis Complex. In *Advances in experimental medicine and*
701 *biology* (Vol. 1019, pp. 43–78). https://doi.org/10.1007/978-3-319-64371-7_3

702 Miotto, P., Cabibbe, A. M., Feuerriegel, S., Casali, N., Drobniewski, F., Rodionova, Y., ...
703 Cirillo, D. M. (2014). Mycobacterium tuberculosis pyrazinamide resistance

determinants: a multicenter study. *MBio*, 5(5), e01819-14.

<https://doi.org/10.1128/mBio.01819-14>

Mokrousov, I., Chernyaeva, E., Vyazovaya, A., Sinkov, V., Zhuravlev, V., & Narvskaya, O. (2016). Next-Generation Sequencing of Mycobacterium tuberculosis. *Emerging Infectious Diseases*, 22(6), 1127–9. <https://doi.org/10.3201/eid2206.152051>

Ngabonziza, J. C. S., Diallo, A. B., Tagliani, E., Diarra, B., Kadanga, A. E., Togo, A. C. G., ... group", the "Union short M.-T. regimen study. (2017). Half of rifampicin-resistant Mycobacterium tuberculosis complex isolated from tuberculosis patients in Sub-Saharan Africa have concomitant resistance to pyrazinamide. *PLOS ONE*, 12(10), e0187211. <https://doi.org/10.1371/journal.pone.0187211>

Phelan, J. E., Coll, F., Bergval, I., Anthony, R. M., Warren, R., Sampson, S. L., ... Clark, T. G. (2016). Recombination in pe/ppe genes contributes to genetic variation in Mycobacterium tuberculosis lineages. *BMC Genomics*, 17, 151. <https://doi.org/10.1186/s12864-016-2467-y>

R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from <http://www.r-project.org/>

Ragheb, M. N., Ford, C. B., Chase, M. R., Lin, P. L., Flynn, J. L., & Fortune, S. M. (2013). The mutation rate of mycobacterial repetitive unit loci in strains of M. tuberculosis from cynomolgus macaque infection. *BMC Genomics*, 14(1), 145. <https://doi.org/10.1186/1471-2164-14-145>

Rambaut, A. (2016). Figtree. Retrieved May 3, 2017, from <http://tree.bio.ed.ac.uk/software/figtree/>

Rambaut, A., Suchard, M. A., Xie, D., & Drummond, A. (2013). Tracer. Retrieved December 11, 2013, from <http://beast.bio.ed.ac.uk/Tracer>

728 Roetzer, A., Diel, R., Kohl, T. A., Rückert, C., Nübel, U., Blom, J., ... Niemann, S. (2013). Whole
729 genome sequencing versus traditional genotyping for investigation of a *Mycobacterium*
730 tuberculosis outbreak: a longitudinal molecular epidemiological study. *PLoS Medicine*,
731 10(2), e1001387. <https://doi.org/10.1371/journal.pmed.1001387>

732 Roetzer, A., Schuback, S., Diel, R., Gasau, F., Ubben, T., di Nauta, A., ... Niemann, S. (2011).
733 Evaluation of *Mycobacterium tuberculosis* typing methods in a 4-year study in
734 Schleswig-Holstein, Northern Germany. *Journal of Clinical Microbiology*, 49(12), 4173–
735 8. <https://doi.org/10.1128/JCM.05293-11>

736 Said, H. M., Kushner, N., Omar, S. V., Dreyer, A. W., Koornhof, H., Erasmus, L., ... Ismail, N. A.
737 (2016). A Novel Molecular Strategy for Surveillance of Multidrug Resistant Tuberculosis
738 in High Burden Settings. *PloS One*, 11(1), e0146106.
739 <https://doi.org/10.1371/journal.pone.0146106>

740 Salathé, M., Bengtsson, L., Bodnar, T. J., Brewer, D. D., Brownstein, J. S., Buckee, C., ...
741 Vespignani, A. (2012). Digital Epidemiology. *PLoS Computational Biology*, 8(7),
742 e1002616. <https://doi.org/10.1371/journal.pcbi.1002616>

743 Sanoussi, C. N., Affolabi, D., Rigouts, L., Anagonou, S., & de Jong, B. (2017). Genotypic
744 characterization directly applied to sputum improves the detection of *Mycobacterium*
745 africanum West African 1, under-represented in positive cultures. *PLOS Neglected*
746 *Tropical Diseases*, 11(9), e0005900. <https://doi.org/10.1371/journal.pntd.0005900>

747 Scott, A. N., Menzies, D., Tannenbaum, T.-N., Thibert, L., Kozak, R., Joseph, L., ... Behr, M. A.
748 (2005). Sensitivities and specificities of spoligotyping and mycobacterial interspersed
749 repetitive unit-variable-number tandem repeat typing methods for studying molecular
750 epidemiology of tuberculosis. *Journal of Clinical Microbiology*, 43(1), 89–94.
751 <https://doi.org/10.1128/JCM.43.1.89-94.2005>

752 Sengstake, S., Bergval, I. L., Schuitema, A. R., Beer, J. L., Phelan, J., Zwaan, R., ... Anthony, R.
753 M. (2017). Pyrazinamide resistance-conferring mutations in *pncA* and the transmission
754 of multidrug resistant TB in Georgia. *BMC Infectious Diseases* 2017 17:1, 17(1), 491.
755 <https://doi.org/10.1186/s12879-017-2594-3>

756 Shah, N. S., Auld, S. C., Brust, J. C. M., Mathema, B., Ismail, N., Moodley, P., ... Gandhi, N. R.
757 (2017). Transmission of Extensively Drug-Resistant Tuberculosis in South Africa. *New*
758 *England Journal of Medicine*, 376(3), 243–253.
759 <https://doi.org/10.1056/NEJMoa1604544>

760 Small, P. M., Hopewell, P. C., Singh, S. P., Paz, A., Parsonnet, J., Ruston, D. C., ... Schoolnik, G.
761 K. (1994). The Epidemiology of Tuberculosis in San Francisco -- A Population-Based
762 Study Using Conventional and Molecular Methods. *New England Journal of Medicine*,
763 330(24), 1703–1709. <https://doi.org/10.1056/NEJM199406163302402>

764 Statisticat. (2016). LaplacesDemon: Complete Environment for Bayesian Inference.
765 Bayesian-Inference.com. R package version 16.0.1. Retrieved from
766 [https://web.archive.org/web/20150206004624/http://www.bayesian-](https://web.archive.org/web/20150206004624/http://www.bayesian-inference.com/software)
767 [inference.com/software](https://web.archive.org/web/20150206004624/http://www.bayesian-inference.com/software)

768 Stucki, D., Ballif, M., Egger, M., Furrer, H., Altpeter, E., Battegay, M., ... Fenner, L. (2016).
769 Standard Genotyping Overestimates Transmission of *Mycobacterium tuberculosis*
770 among Immigrants in a Low-Incidence Country. *Journal of Clinical Microbiology*, 54(7),
771 1862–70. <https://doi.org/10.1128/JCM.00126-16>

772 Sukumaran, J., & Holder, M. T. (2010). DendroPy: a Python library for phylogenetic
773 computing. *Bioinformatics*, 26(12), 1569–1571.
774 <https://doi.org/10.1093/bioinformatics/btq228>

775 Supply, P., Allix, C., Lesjean, S., Cardoso-Oelemann, M., Rusch-Gerdes, S., Willery, E., ... van

776 Soolingen, D. (2006). Proposal for Standardization of Optimized Mycobacterial
777 Interspersed Repetitive Unit-Variable-Number Tandem Repeat Typing of
778 Mycobacterium tuberculosis. *Journal of Clinical Microbiology*, 44(12), 4498–4510.
779 <https://doi.org/10.1128/JCM.01392-06>

780 Supply, P., Magdalena, J., Himpens, S., & Locht, C. (1997). Identification of novel intergenic
781 repetitive units in a mycobacterial two-component system operon. *Molecular*
782 *Microbiology*, 26(5), 991–1003. Retrieved from
783 <http://www.ncbi.nlm.nih.gov/pubmed/9426136>

784 Tessema, B., Nabeta, P., Valli, E., Albertini, A., Collantes, J., Lan, N. H., ... Dolinger, D. L.
785 (2017). FIND Tuberculosis Strain Bank: a Resource for Researchers and Developers
786 Working on Tests To Detect Mycobacterium tuberculosis and Related Drug Resistance.
787 *Journal of Clinical Microbiology*, 55(4), 1066–1073.
788 <https://doi.org/10.1128/JCM.01662-16>

789 Thierry, D., Cave, M. D., Eisenach, K. D., Crawford, J. T., Bates, J. H., Gicquel, B., & Guesdon,
790 J. L. (1990). IS6110, an IS-like element of Mycobacterium tuberculosis complex. *Nucleic*
791 *Acids Research*, 18(1), 188. Retrieved from
792 <http://www.ncbi.nlm.nih.gov/pubmed/2155396>

793 Trauer, J. M., Denholm, J. T., & McBryde, E. S. (2014). Construction of a mathematical model
794 for tuberculosis transmission in highly endemic regions of the Asia-pacific. *Journal of*
795 *Theoretical Biology*, 358, 74–84. <https://doi.org/10.1016/j.jtbi.2014.05.023>

796 van Embden, J. D., Cave, M. D., Crawford, J. T., Dale, J. W., Eisenach, K. D., Gicquel, B., ... al.,
797 et. (1993). Strain identification of Mycobacterium tuberculosis by DNA fingerprinting:
798 recommendations for a standardized methodology. *Journal of Clinical Microbiology*,
799 31(2), 406–9. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/8381814>

Vluggen, C., Soetaert, K., Groenen, G., Wanlin, M., Spitaels, M., Arrazola de Oñate, W., ...
Mathys, V. (2017). Molecular epidemiology of Mycobacterium tuberculosis complex in
Brussels, 2010–2013. *PLOS ONE*, 12(2), e0172554.
<https://doi.org/10.1371/journal.pone.0172554>

Walker, T. M., Ip, C. L. C., Harrell, R. H., Evans, J. T., Kapatai, G., Dedicoat, M. J., ... Peto, T. E.
A. (2013). Whole-genome sequencing to delineate Mycobacterium tuberculosis
outbreaks: a retrospective observational study. *The Lancet Infectious Diseases*, 13(2),
137–46. [https://doi.org/10.1016/S1473-3099\(12\)70277-3](https://doi.org/10.1016/S1473-3099(12)70277-3)

Walker, T. M., Lalor, M. K., Broda, A., Ortega, L. S., Morgan, M., Parker, L., ... Conlon, C. P.
(2014). Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK,
2007-12, with whole pathogen genome sequences: an observational study. *The Lancet.
Respiratory Medicine*, 2(4), 285–292. [https://doi.org/10.1016/S2213-2600\(14\)70027-X](https://doi.org/10.1016/S2213-2600(14)70027-X)

Walker, T. M., Merker, M., Knoblauch, A. M., Helbling, P., Schoch, O. D., van der Werf, M. J.,
... Witschi, M. (2018). A cluster of multidrug-resistant Mycobacterium tuberculosis
among patients arriving in Europe from the Horn of Africa: a molecular epidemiological
study. *The Lancet Infectious Diseases*. [https://doi.org/10.1016/S1473-3099\(18\)30004-5](https://doi.org/10.1016/S1473-3099(18)30004-5)

Warren, R. M., Streicher, E. M., Sampson, S. L., van der Spuy, G. D., Richardson, M., Nguyen,
D., ... van Helden, P. D. (2002). Microevolution of the direct repeat region of
Mycobacterium tuberculosis: implications for interpretation of spoligotyping data.
Journal of Clinical Microbiology, 40(12), 4457–65. Retrieved from
<http://www.ncbi.nlm.nih.gov/pubmed/12454136>

WHO. (2015). *Guidelines for surveillance of drug resistance in tuberculosis* (5th ed.). Geneva.

WHO. (2018). *Global tuberculosis report 2017*. Retrieved from
http://www.who.int/tb/publications/global_report/archive/en/

824 Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

825 Retrieved from <http://ggplot2.org>

826 Wirth, T., Hildebrand, F., Allix-Béguet, C., Wölbeling, F., Kubica, T., Kremer, K., ... Niemann,

827 S. (2008). Origin, Spread and Demography of the Mycobacterium tuberculosis Complex.

828 *PLoS Pathogens*, 4(9), e1000160. <https://doi.org/10.1371/journal.ppat.1000160>

829 Witney, A. A., Cosgrove, C. A., Arnold, A., Hinds, J., Stoker, N. G., & Butcher, P. D. (2016).

830 Clinical use of whole genome sequencing for Mycobacterium tuberculosis. *BMC*

831 *Medicine*, 14(1), 46. <https://doi.org/10.1186/s12916-016-0598-2>

832 Wyllie, D., Davidson, J., Walker, T., Rathod, P., Peto, T., Robinson, E., ... Campbell, C. (2018).

833 A quantitative evaluation of MIRU-VNTR typing against whole-genome sequencing for

834 identifying Mycobacterium tuberculosis transmission: A prospective observational

835 cohort study. *BioRxiv*, 252734. <https://doi.org/10.1101/252734>

836 Xia, E., Teo, Y.-Y., & Ong, R. T.-H. (2016). SpoTyping: fast and accurate in silico

837 Mycobacterium spoligotyping from sequence reads. *Genome Medicine*, 8(1), 19.

838 <https://doi.org/10.1186/s13073-016-0270-7>

839 Xu, P., Wu, J., Yang, C., Luo, T., Shen, X., Zhang, Y., ... Gao, Q. (2016). Prevalence and

840 transmission of pyrazinamide resistant Mycobacterium tuberculosis in China.

841 *Tuberculosis*, 98, 56–61. <https://doi.org/10.1016/j.tube.2016.02.008>

842 Zignol, M., Cabibbe, A. M., Dean, A. S., Glaziou, P., Alikhanova, N., Ama, C., ... Raviglione, M.

843 C. (2018). Genetic sequencing for surveillance of drug resistance in tuberculosis in

844 highly endemic countries: a multi-country population-based surveillance study. *The*

845 *Lancet Infectious Diseases*. [https://doi.org/10.1016/S1473-3099\(18\)30073-2](https://doi.org/10.1016/S1473-3099(18)30073-2)

846 Zignol, M., Dean, A. S., Alikhanova, N., Andres, S., Cabibbe, A. M. A., Cirillo, D. M., ... Valafar,

847 F. (2016). Population-based resistance of Mycobacterium tuberculosis isolates to

848 pyrazinamide and fluoroquinolones: results from a multicountry surveillance project.

849 *The Lancet Infectious Diseases*. [https://doi.org/10.1016/S1473-3099\(16\)30190-6](https://doi.org/10.1016/S1473-3099(16)30190-6)

850

851 Supplementary legends

852 Supplementary Table 1: Spoligotyping patterns derived from membrane- and genome-based

853 methods

854 Spoligotyping patterns were estimated from both membrane (Mem-Spo) and genome (Gen-

855 Spo) approaches. Patterns are shown from both methods per isolates with the number of

856 mismatches between patterns recorded.

857

Isolate	Mem-Spoligotype	Gen-Spoligotype	Mismatches
DRC-052577	11111110000011111111101000000000010001111	111111100000111111111101000000000110001111	1
DRC-052578	1111111000000000000000000011111110000111111	1111111000000000000000000011111110000111111	0
DRC-052654	1111111111111111111110000011111110000111111	1101111111111111111110111111101110000110111	7
DRC-052667	11111111000111111110000000000000000001100	111111111110000111000000000000000000011100	9
DRC-052750	11111110000011111111101100000000001000111	111111100000111111111101000000000110001111	3
DRC-052956	1111111111111111111111111111111110000111011	1111111111111111111111111111111110000111011	0
DRC-052958	111111111111111111111111111111111011110000111011	111111111111111111111111111111111011110000111011	0
DRC-052959	1111111111111111111111111111111110000111111	11111111111111111111111111111111100000111111	1
DRC-052964	11111111000111111110000000000000000001111	11111111000111111110000000000000000001111	0
DRC-060116	1111111111111111111111111111111110000111011	1111111111111111111111111111111110000111011	0
DRC-060118	1111111111111111111111111111111110000111011	1111111111111111111111111111111110000111011	0
DRC-060120	1111111111111111111111111111111110000111011	1111111111111111111111111111111110000111011	0
DRC-060131	1101111111111111111110000111111110000111111	1101111111111111111110000111111110000111111	0
DRC-060237	1111111111111111111111111111111110000111111	1111111111111111111111111111111110000111111	0
DRC-060244	000000000000000000000000001111110000111111	0000000000000000000000000011111110000111011	3
DRC-060248	1111111111111111111111111111111110000111011	1111111111111111111111111111111110000111011	0
DRC-060253	1111111111111111111111111111111110000111011	1111111111111111111111111111111110000111011	0
DRC-060260	1111110000000000000000000011111110000101111	1111110000000000000000000011111110010111111	2
DRC-060360	11111010000011111111111000000000010000011	111111100000111111111110000000000110000011	2
DRC-060361	1101111111111111111110000111111110000111111	1101111111111111111110000111111110000111111	0
DRC-060362	11111110000011111111101100000000010001111	111111100000111111111010000000000110001111	2
DRC-060366	1111110000000000000000000011111110000101111	1111110000000000000000000011111110000111111	1
DRC-060375	11000000111111101111000011111110000111111	11000000111111101111000011111110000111111	2
DRC-060376	11011110001111111110000111111110000011111	11011111111111111110000111111110000111111	5

DRC-060381	11111111000111111110000000000000000011111	11111111000111111110000000000000000011111	0
DRC-060382	11111100000000000000000000111111100001111111	11111100000000000000000000111111100001111111	0
DRC-060514	1100000011111111011100001111111100001111111	1100000011111111011100001111111100001111111	0
DRC-060526	11111111000111111110000000000000000011111	11111111000111111110000000000000000011111	0
DRC-060527	1111111111110000111100001111111100001011111	1111111111110000111100001111111100001011111	0
DRC-060528	11111100000000000000000000111111100001011111	11111100000000000000000000111111100001111111	1
DRC-060549	10000011111111111101111111111100001110111	100000111111111111111111111111111100001111111	2
DRC-060562	111111111111111111111111111111100001000011	111111111111111111111111111111111100001000011	0
DRC-060565	111111111111111111111111111111100001110111	111111111111111111111111111111111100001110111	0
DRC-060756	11111100000000000000000000111111100001111111	11111100000000000000000000111111100001111111	0
DRC-060757	111111111111111111111111111111011100001110111	111111111111111111111111111111011100001110111	0
DRC-060772	111111100000111111111101000000000010001111	111111100000111111111101000000000010001111	1
DRC-060778	11111111000111111110000000000000000011111	11111111000111111110000000000000000011111	0
DRC-060856	11111111111100001100000000000000000011100	11111111111100001100000000000000000011100	0
DRC-060858	111111111111111111111000011111110000101111	1111111111111111111111111111011100001110111	7
DRC-060859	111111111111111111111000011111110000111111	111111111111111111111000011111110000111111	0
DRC-061069	111111100000111111111110000000000010001111	111111100000111111111110000000000010001111	1
DRC-061123	111111111111111111111000011111110000111111	111111111111111111111000011111110000111111	0
DRC-061124	110111111111111111111000011111110000111111	111111111111111111111000011111110000111111	1
DRC-061127	1111111000000000000000000011111110000111111	1111111000000000000000000011111110000111111	0
DRC-061136	1111110000000000000000000011111110000101111	1111110000000000000000000011111110000111111	1
DRC-061137	1111110000000000000000000011111110000101111	1111110000000000000000000011111110000111111	1
DRC-061138	1111111100011111111000000000000000001111	1111111100011111111000000000000000001111	0
DRC-061142	111111111111111111111000011111110000111111	111111111111111111111000011111110000111111	0
DRC-061431	1111111100011111111000000000000000001111	1111111100011111111000000000000000001111	0
DRC-061493	111111111111111111111111111111100001110111	1111111111111111111111111111111100001110111	1
DRC-062295	111111111111100000111111111111100001110111	1111111111111000001111111111111100001110111	0
DRC-062298	11111111111000011100000000000000000011100	11111111111000011100000000000000000011100	0
DRC-062303	110000001111110111100001111111100001111111	110000001111110111100001111111100001111111	2
DRC-062444	1111111100011111110000000000000000001111	1111111100011111110000000000000000001111	0
DRC-062446	111111111111111111111000011101110000111111	111111111111111111111000011101110000111111	0
DRC-062447	11000000111111101110000111111110000111111	11000000111111101110000111111110000111111	0
DRC-062495	11111111111111111111111111111110000111111	111111111111111111111111111111110000111111	0
DRC-062502	11111111111111111110000111111100001011111	11111111111111111110000111111100001011111	0
DRC-062503	11111111111111111111111111111110000111111	1111111111111111111111111111111100001110111	1

bioRxiv preprint doi: <https://doi.org/10.1101/302232>; this version posted April 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

DRC-100033	1111111111111111011111111111110000111111	1111111111111111101111111111110000111111	0
DRC-100035	1111111111110111111111111111110000111011	1111111111110111111111111111110000111011	0
DRC-100037	11011111111101111111000011111110000111111	11011111111101111111000011111110000111111	0
DRC-100099	11111100000000000000000000001111110000111111	11111100000000000000000000001111110000111111	0
DRC-100122	111111111111111111111111111100110000111011	111111111111111111111111111110000111011	2
DRC-100299	11111111111111111111000011111110000101111	11111111111111111111000011111110000101111	0
DRC-100300	111111100011111111000000000000000000001111	111111100011111111000000000000000000001111	0
DRC-100301	11111111111111111111000011111110000101111	11111111111111111111000011111110000101111	0
DRC-100307	111111111111111111110000110000110000111111	111111111111111111110000110000110000111111	0
DRC-100310	1111111111111100001111111111110000111011	1111111111111100001111111111110000111011	0
DRC-100503	1111111111111111111111111111110000111011	1111111111111111111111111111110000111011	0
DRC-100740	11111111111111111111000011111110000111111	11111111111111111111000011111110000111111	0
DRC-100753	11111111111111111111111111111101110000111111	1111111111111111111111111111110000111011	2
DRC-100759	11011111111111111111000011111110000111111	11011111111111111111000011111110000111111	0
DRC-100923	11111111111111111111000011101110000111111	11111111111111111111000011111110000101111	2
DRC-100929	111111100011111111000000000000000000001111	111111100011111111000000000000000000001111	0
DRC-100931	11011111111111111111000011111110000111111	11011111111111111111000011111110000111111	0
DRC-101072	11111111111111111111000011111110000111111	11111111111111111111000011111110000111111	0
DRC-101306	11111111111111111111000011101110000111111	11111111111111111111000011101110000111111	0
DRC-101308	11011111111111111111000011111110000111111	11011111111111111111000011111110000111111	0

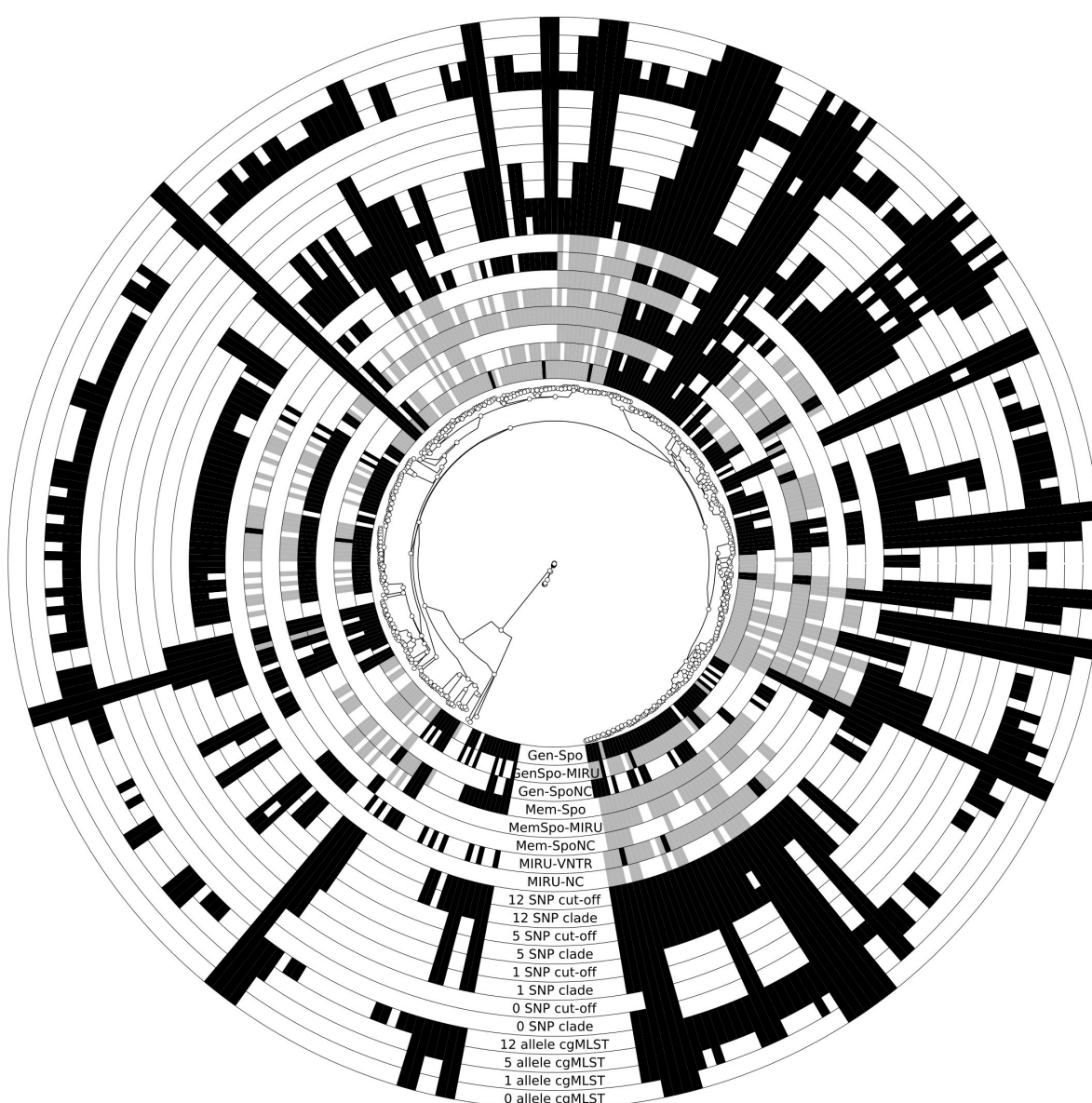
858

859

860

Supplementary Figure 1: Clustering of *M. tuberculosis* isolates based on 20 methods.

For each method, the inclusion of an isolate into a cluster is outlined in the surrounding circles using GraPhlAn (Asnicar et al., 2015). If an isolate is in a cluster not affected by convergence, it is highlighted in black for the given method. If an isolate is in a cluster affected by convergence, it is shown in grey.



Supplementary Figure 2: Algorithm of the phylogenetic inclusion method

An extension of the SNP-based clustering methods is used to detect complete chains of transmission that may be missed by pairwise SNP comparisons. In this example: A) shows the inferred phylogenetic sub-tree for five taxa and associated SNPs for each branch; B) a SNP cut-off is applied, which places 2 taxa (Taxon 1 and Taxon 4) together into a cluster; the most recent common ancestor (MRCA) of all taxa in this cluster (i.e. Taxon 1 & Taxon 4) is found; D) all descendants of this MRCA are added to the putative transmission chain.

