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 Pečerska[6], Suriya Akter[1], Matthias Merker[4,5], Christian Utpatel[4,5], Patrick Beckert[4,5], Florian Gehre[1,7,8], Pauline Lempens[1], Tanja Stadler[6], Michel K. Kaswa[1,10], Denise Kühnert[9], Stefan Niemann[4,5,&], Bouke C de Jong[1,&]

Affiliations

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

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

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

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

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

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

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

[8]Bernhard Nocht Institute for Tropical Medicine, Hamburg, Germany

[9]University Hospital Zurich, Zurich, Switzerland

[10]National 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 (Mtbc) strains, 10.4 million new patients develop tuberculosis (TB) every year[1]. In addition, the prevalence of multidrug resistant Mtbc strains (MDR-TB) is increasing[1], predominantly through ongoing transmission within large populations[2,3]. 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[2,4]. Thus, interrupting transmission is key

for the control of MDR-TB[3,5,6]. 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)[7]. Contact tracing is a primary epidemiological method for investigating transmission networks of TB, mainly based on patient interviews[8]. 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[9–13] and misses many connections[14,15]. 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 IS*6110* DNA fingerprinting[16,17], spoligotyping[18–20], and variable-number tandem repeats of mycobacterial interspersed repetitive units (MIRU-VNTR)[21] which is the most common method at the moment[7]. The latter method is based on copy numbers of a sequence in tandem repeat patterns derived from 24 distinct loci within the genome[22]. If two patients have the same classical genotyping pattern such as a 24-loci MIRU-VNTR pattern (or up to one locus difference[23,22]) 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[22]. 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[12,24].

The application of (whole genome) sequence-based approaches for similarity analysis of Mtbc isolates and cluster determination is known to have high discriminatory power when assessing transmission dynamics[12,25–28]. 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[29]. However, this is limited by the low occurrence of PZA resistance, even in MDR-TB isolates[30–33]. The advent of widespread whole genome sequencing (WGS) capabilities has allowed for highly discriminatory analyses of Mtbc strains either using core genome multi-locus sequence typing (cgMLST)[34] or SNP distances[12,24,26,27,35]. 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[12,26]. Although many SNP cut-offs for linking isolates have been proposed[36], 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[26]. 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[37]) and between study populations e.g. high and low incidence settings[35].

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 Mtbc 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[12,26,38–41] while the rate of change in the MIRU-VNTR loci specifically is known to be quicker (~10^{-3}) [42,43]. Since these mutation rates have been shown to also vary by lineage[39,44] and over short periods of time[38], such variation needs to be accounted for, e.g. in Bayesian phylogenetic dating techniques[3,38,42].

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[45] 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).

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)

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 Mtbc 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[46], although Mem-Spo may be of use as a low-cost method of sorting Mtbc strains into the seven primary lineages[47,48]. 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

192 patterns, requiring more investigation into which method is closer to the 'true'
193 spoligotyping pattern within a genome[49–52].
194
195 In line with previous findings[46,53], convergent evolution of 24-loci MIRU-VNTR patterns was
196 rarer than observed for spoligotyping, but did occur in 16% of MIRU-VNTR-based clusters.
197 Additionally, the transmission times encompassed by MIRU-VNTR clusters spanned several
198 decades (Table 1b, Figure 2a), confirming previous studies showing over-estimation of
199 recent transmission with this method[12,25,35,54].
200
201 The combination of MIRU-VNTR or spoligotyping with *pncA* mutations (MIRU-NC and Gen-
202 SpoNC/Mem-SpoNC) appeared to reflect true clusters of PZA resistance transmission based
203 on the relatively young ages of such transmission clusters (Table 1b). Thus, as discussed
204 before[55,56], although transmission of *pncA* mutations seems to occur, further investigation is
205 needed to find out whether *pncA* mutants are less transmissible than those with a wildtype
206 gene.
207
208 For defining transmission events that occurred in more recent time frames before sampling,
209 WGS-based methods (SNP or cgMLST) were found to be better suited than classical
210 genotyping methods (Table 1, Figure 2). The 12 SNP cut-off, currently the recommended
211 upper bound for clustering isolates, likely defines transmission events that occurred on
212 average three decades prior to sampling, similar in age to clusters estimated by MIRU-VNTR.
213 This suggests that the 12 SNP cluster method may be a good replacement for MIRU-VNTR as
214 it detects larger transmission networks spanning similar transmission time periods but is
215 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[57], three[12] or five[26] 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[27]. 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.[34]) 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.[57]). 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[3,12,39]. Within the Bayesian analysis employed here, the mutation rate was free to vary between these values but was found to strongly favour ~$3x10^{-8}$ (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[39], it may be similar in other lineages, although this has only been shown for lineage 2[3,39]. 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 Mtbc.

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[28,37]: 1) studies employing contact tracing and/or digital epidemiology[58] 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 Mtbc should be

264 analysed in a similar fashion to ensure transferability of findings across the entire complex;
265 3) a broad range of drug resistance profiles should be included to fully assess the impact of
266 such mutations on transmission estimates; 4) improved WGS methods, such as directly from
267 clinical samples to help reduce culture biases[59] and longer reads (e.g. PacBio SMRT or
268 Nanopore MinION) to capture the entire genome, including repetitive regions such as
269 PE/PPE genes known to impact genome remodelling[60,61], will ensure that the maximum
270 diversity between isolates is captured and 5) standardised SNP calling pipelines appropriate
271 across all lineages, with high true positive/low false negative rates, will ensure that Mtbc
272 molecular epidemiology can be uniformly implemented and comparable across studies.
273
274 In conclusion, since each method was found to represent different timespans and clustering
275 definitions, they can be used in a stratified manner in an integrated epidemiological and
276 public health investigation addressing the transmission of Mtbc strains. For instance,
277 although spoligotyping clusters represented potentially very old transmission events, the
278 low associated cost and its ability to be applied directly on sputum helps reduce culture bias
279 and thus robustly assign lineages. Thus, spoligotyping and/or MIRU-VNTR would serve well
280 as first-line surveillance of potential transmission events in the population, guiding further
281 investigations and resource allocations.
282
283 These potential transmission hotspots could be further investigated with contact tracing
284 and/or WGS. Employment of different cut-offs and clustering approaches to WGS data can
285 then address several questions. The 12 SNP cluster/clade or 12 allele cgMLST approaches
286 serve well for high level surveillance targeting larger (older) transmission networks, akin to
287 what is currently often done using MIRU-VNTR (e.g.[27,62]). Recent transmission events can

288 then be detected through employment of low SNP or cgMLST-based cut-offs (e.g. 5 SNPs for
289 transmission in the past 15 years or 0-1 SNPs for transmission in the past 5 years). These
290 clusters can then be linked to historical isolates or other clusters through employment of
291 the phylogenetic inclusion method to resolve the local circulating genotypes. This is
292 especially useful if bursts of sampling are undertaken such as in drug resistance surveys[63],
293 which are increasingly employing WGS approaches[32,64,65]. Alternatively, in high
294 incidence/low diversity settings where amalgamation of clusters may inadvertently obscure
295 distinct hotspots of transmission at different time points, subdivision into distinct time-
296 dependant clusters can be undertaken using the algorithm presented in such a study in East
297 Greenland[35].
298
299 Overall, phylodynamic approaches applied to whole genome sequences, as undertaken
300 here, are recommended to fully investigate the specific transmission dynamics within a
301 study population to account for setting-specific conditions, such as low/high TB incidence,
302 low/high pathogen population diversity, sampling fractions and social factors influencing
303 transmission. Thus, each genotyping method can be employed as part of an overall evidence
304 gathering program for transmission, placing molecular epidemiological approaches as an
305 integral part in tracking and stopping the spread of TB.
306
307 **Materials and Methods**
308 Dataset and sequencing
309 A set of 324 isolates from Kinshasa, Democratic Republic of Congo were collected from
310 consecutive retreatment TB patients between 2005 and 2010 at TB clinics, servicing an
311 estimated 30% of the population of Kinshasa. All isolates were phenotypically resistant to

312 rifampicin (RR-TB) and the majority are also isoniazid resistant (i.e. MDR-TB). Use of the
313 stored isolates without any linked personal information was approved by the health
314 authorities of the DRC and the Institutional Review Board of the ITM in Antwerp (ref no
315 945/14). Libraries for whole genome sequencing were prepared from extracted genomic
316 DNA with the Illumina Nextera XT kit, and run on the Illumina NextSeq platform in a
317 2x151bp run according to manufacturer's instructions. Illumina read sets will be available at
318 ReSeqTB (platform.reseqtb.org) upon publication.
319
320 Genome reconstruction and maximum likelihood phylogeny estimation
321 The MTBseq pipeline[66] was used to detect the SNPs for each isolate using the H37Rv
322 reference genome (NCBI accession number NC000962.3)[67,68]. Sites known to be involved in
323 drug resistance (as outlined in the PhyResSE list of drug mutations v27[69]) were excluded
324 from the alignment and additional filtering of sites with ambiguous calls in >5% of isolates
325 and those SNPs within a 12bp window of each other was also applied.
326
327 The SNP alignment of all isolates was used as the basis for creating a maximum likelihood
328 (ML) phylogeny. RAxML-NG version 0.5.1b[70] was used to reconstruct the phylogeny from
329 this alignment using a GTR+GAMMA model of evolution, accounting for ascertainment
330 bias[71] with the Stamatakis reconstituted DNA approach[72] and site repeat optimisation[73] with
331 20 different starting trees and 100 bootstraps. All subsequent topology visualisation was
332 undertaken using FigTree version 1.4.3[74] and GraPhlAn[75].
333
334 Transmission cluster estimation methods

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

Spoligotyping

Spoligotype patterns were estimated by 2 methods: membrane-based and genome-based. Membrane-based patterns were obtained following the previously published protocol[20]. This method is referred to as Mem-Spo. Genome-based spoligotyping was derived from the Illumina reads of each isolate using SpoTyping v2.1[49]. 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[22]. 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

359 clustered if both the spoligotyping pattern and the 24 loci MIRU-VNTR pattern matched.
360 These clustering methods are referred to as MemSpo-MIRU and GenSpo-MIRU.

361 SpoNC

362 Transmission estimation using spoligotyping has been shown to be improved if combined
363 with *pncA* mutations[29]. This method, referred to as SpoNC, was applied to both Mem-Spo
364 (Mem-SpoNC) and Gen-Spo (Gen-SpoNC). Mutations in *pncA* were extracted from the
365 MTBseq tabular output for each isolate. All mutations were selected, regardless of drug
366 resistance association, as is done in the SpoNC approach. The upstream promoter region of
367 *pncA* did not reveal any mutations in this dataset. Isolates were said to be clustered if all 43
368 spacers matched and the *pncA* mutation was the same in both isolates. MIRU-VNTR
369 patterns were combined with *pncA* mutations in a similar manner. This is referred to as
370 MIRU-NC.
371

372 SNP cut-off clustering

373 The advent of whole genome reconstruction has allowed for genome-based comparisons
374 for transmission clustering. Previous work has suggested that linked and recent transmission
375 can be estimated by comparison of SNP differences between isolates. The cut-offs proposed
376 by Walker *et al.*[26] are the most widely used and have been employed in multiple studies[76–
377 [78]. In this study, we employed both the 5 SNP (proposed by Walker *et al*. as the likely
378 boundary for linked transmission) and 12 SNP cut-offs (proposed maximum boundary) for
379 cluster definition. Additionally, we employed lower cut-offs of 0 and 1 SNPs to look for
380 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[79]. 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[34]. 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 Mtbc 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[51,53,80,81]. Convergence was detected and cross-checked with two methods. Firstly, Mtbc lineage and sub-lineage numbering[82] was applied to all isolates based on the PhyResSE lineage-defining SNP list v27[69]. 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[45] 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[83] and coalescent constant size demographic model[84], both of which have been found to be suitable for lineage 4 isolates in a previous study[35]. 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 $1.5x10^{-7}$; variance 1.0) was used for the clock model to reflect the previously estimated mutation rate of *M. tuberculosis* lineage 4[12,26,38–41], while allowing for variation as previously suggested[38]. 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[72]. Tracer v1.6[85] 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[45] to obtain a single chain

453 for each clustering type with high (>1000) effective sample sizes. The tree samples were
454 combined in the same manner and resampled at a lower frequency to create thinned
455 samples of (minimum) 20,000 trees.
456
457 The timespan of transmission events estimated by each method was then calculated as
458 follows: for each cluster created by the given method, we defined the MRCA node as the
459 internal node that connects all taxa in that cluster. The youngest node was then defined as
460 the tip that is furthest from this MRCA within the clade (i.e. the tip descendant from that
461 node that was sampled closest to the present time). For each retained tree in the MCMC
462 process, the difference in age between the MRCA node and youngest node was calculated.
463 This gave a distribution of likely maximum transmission event times within that cluster. For
464 each method, these per-cluster aggregated ages were then combined across all clusters to
465 give a per-method distribution of transmission event times represented by the clusters. The
466 95% HPD interval of these distributions was calculated with the LaplacesDemon p.interval
467 function[86] in R v3.4.0[87] and the distribution within this interval for each method along with
468 the mean based upon this interval were then visualized in violin plots per clustering method
469 using ggplot2[88] in R.
470
471 **References**

472 1. WHO. *Global tuberculosis report 2017*. (2018).
473 2. Kendall, E. A. *et al.* Burden of transmitted multidrug resistance in epidemics of
474 tuberculosis: a transmission modelling analysis. *Lancet Respir. Med.* **3,** 963–972
475 (2015).
476 3. Merker, M. *et al.* Evolutionary history and global spread of the Mycobacterium

tuberculosis Beijing lineage. *Nat. Genet.* **47,** 242–249 (2015).

4. Trauer, J. M., Denholm, J. T. & McBryde, E. S. Construction of a mathematical model for tuberculosis transmission in highly endemic regions of the Asia-pacific. *J. Theor. Biol.* **358,** 74–84 (2014).

5. Shah, N. S. *et al.* Transmission of Extensively Drug-Resistant Tuberculosis in South Africa. *N. Engl. J. Med.* **376,** 243–253 (2017).

6. Klopper, M. *et al.* Emergence and Spread of Extensively and Totally Drug-Resistant Tuberculosis, South Africa. *Emerg. Infect. Dis.* **19,** 449–455 (2013).

7. Merker, M., Kohl, T. A., Niemann, S. & Supply, P. in *Advances in experimental medicine and biology* **1019,** 43–78 (2017).

8. Fox, G. J., Barry, S. E., Britton, W. J. & Marks, G. B. Contact investigation for tuberculosis: a systematic review and meta-analysis. *Eur. Respir. J.* **41,** (2012).

9. Small, P. M. *et al.* The Epidemiology of Tuberculosis in San Francisco -- A Population-Based Study Using Conventional and Molecular Methods. *N. Engl. J. Med.* **330,** 1703–1709 (1994).

10. Behr, M. A. *et al.* Predictive Value of Contact Investigation for Identifying Recent Transmission of Mycobacterium tuberculosis. *Am. J. Respir. Crit. Care Med.* **158,** 465–469 (1998).

11. Diel, R. *et al.* Epidemiology of Tuberculosis in Hamburg, Germany: Long-Term Population-Based Analysis Applying Classical and Molecular Epidemiological Techniques. *J. Clin. Microbiol.* **40,** 532–539 (2002).

12. Roetzer, A. *et al.* Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. *PLoS Med.* **10,** e1001387 (2013).

13. Roetzer, A. *et al.* Evaluation of Mycobacterium tuberculosis typing methods in a 4-year study in Schleswig-Holstein, Northern Germany. *J. Clin. Microbiol.* **49,** 4173–8 (2011).

14. Bjorn-Mortensen, K. *et al.* Extent of transmission captured by contact tracing in a tuberculosis high endemic setting. *Eur. Respir. J.* **49,** (2017).

15. Vluggen, C. *et al.* Molecular epidemiology of Mycobacterium tuberculosis complex in Brussels, 2010–2013. *PLoS One* **12,** e0172554 (2017).

16. van Embden, J. D. *et al.* Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology. *J. Clin. Microbiol.* **31,** 406–9 (1993).

17. Thierry, D. *et al.* IS6110, an IS-like element of Mycobacterium tuberculosis complex. *Nucleic Acids Res.* **18,** 188 (1990).

18. Guernier, V., Sola, C., Brudey, K., Guégan, J.-F. & Rastogi, N. 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 Infect. Dis.* **8,** 46 (2008).

19. Goguet de la Salmonière, Y. O. *et al.* Evaluation of spoligotyping in a study of the transmission of Mycobacterium tuberculosis. *J. Clin. Microbiol.* **35,** 2210–4 (1997).

20. Kamerbeek, J. *et al.* Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. *J. Clin. Microbiol.* **35,** 907–14 (1997).

21. Supply, P., Magdalena, J., Himpens, S. & Locht, C. Identification of novel intergenic repetitive units in a mycobacterial two-component system operon. *Mol. Microbiol.* **26,** 991–1003 (1997).

525 22. Supply, P. *et al.* Proposal for Standardization of Optimized Mycobacterial Interspersed
526 Repetitive Unit-Variable-Number Tandem Repeat Typing of Mycobacterium
527 tuberculosis. *J. Clin. Microbiol.* **44,** 4498–4510 (2006).
528 23. Jonsson, J. *et al.* Comparison between RFLP and MIRU-VNTR Genotyping of
529 Mycobacterium tuberculosis Strains Isolated in Stockholm 2009 to 2011. *PLoS One* **9,**
530 e95159 (2014).
531 24. Gardy, J. L. *et al.* Whole-Genome Sequencing and Social-Network Analysis of a
532 Tuberculosis Outbreak. *N. Engl. J. Med.* **364,** 730–739 (2011).
533 25. Wyllie, D. *et al.* A quantitative evaluation of MIRU-VNTR typing against whole-
534 genome sequencing for identifying Mycobacterium tuberculosis transmission: A
535 prospective observational cohort study. *bioRxiv* 252734 (2018). doi:10.1101/252734
536 26. Walker, T. M. *et al.* Whole-genome sequencing to delineate Mycobacterium
537 tuberculosis outbreaks: a retrospective observational study. *Lancet Infect. Dis.* **13,**
538 137–46 (2013).
539 27. Walker, T. M. *et al.* A cluster of multidrug-resistant Mycobacterium tuberculosis
540 among patients arriving in Europe from the Horn of Africa: a molecular
541 epidemiological study. *Lancet Infect. Dis.* (2018). doi:10.1016/S1473-3099(18)30004-5
542 28. Comas, I. in 79–93 (Springer, Cham, 2017). doi:10.1007/978-3-319-64371-7_4
543 29. Said, H. M. *et al.* A Novel Molecular Strategy for Surveillance of Multidrug Resistant
544 Tuberculosis in High Burden Settings. *PLoS One* **11,** e0146106 (2016).
545 30. Kurbatova, E. V., Cavanaugh, J. S., Dalton, T., S. Click, E. & Cegielski, J. P. Epidemiology
546 of Pyrazinamide-Resistant Tuberculosis in the United States, 1999–2009. *Clin. Infect.*
547 *Dis.* **57,** 1081–1093 (2013).
548 31. Xu, P. *et al.* Prevalence and transmission of pyrazinamide resistant Mycobacterium

549 tuberculosis in China. *Tuberculosis* **98,** 56–61 (2016).

550 32. Zignol, M. *et al.* Population-based resistance of Mycobacterium tuberculosis isolates
551 to pyrazinamide and fluoroquinolones: results from a multicountry surveillance
552 project. *Lancet Infect. Dis.* (2016). doi:10.1016/S1473-3099(16)30190-6

553 33. Ngabonziza, J. C. S. *et al.* Half of rifampicin-resistant Mycobacterium tuberculosis
554 complex isolated from tuberculosis patients in Sub-Saharan Africa have concomitant
555 resistance to pyrazinamide. *PLoS One* **12,** e0187211 (2017).

556 34. Kohl, T. A. *et al.* Whole-genome-based Mycobacterium tuberculosis surveillance: a
557 standardized, portable, and expandable approach. *J. Clin. Microbiol.* **52,** 2479–86
558 (2014).

559 35. Bjorn-Mortensen, K. *et al.* Tracing Mycobacterium tuberculosis transmission by whole
560 genome sequencing in a high incidence setting: a retrospective population-based
561 study in East Greenland. *Sci. Rep.* **6,** 33180 (2016).

562 36. Hatherell, H.-A. *et al.* Interpreting whole genome sequencing for investigating
563 tuberculosis transmission: a systematic review. *BMC Med.* **14,** 21 (2016).

564 37. Guthrie, J. L. & Gardy, J. L. A brief primer on genomic epidemiology: lessons learned
565 from *Mycobacterium tuberculosis*. *Ann. N. Y. Acad. Sci.* **1388,** 59–77 (2017).

566 38. Bryant, J. M. *et al.* Inferring patient to patient transmission of Mycobacterium
567 tuberculosis from whole genome sequencing data. *BMC Infect. Dis.* **13,** 110 (2013).

568 39. Duchêne, S. *et al.* Genome-scale rates of evolutionary change in bacteria. *Microb.*
569 *genomics* **2,** e000094 (2016).

570 40. Eldholm, V. & Balloux, F. Antimicrobial Resistance in Mycobacterium tuberculosis:
571 The Odd One Out. *Trends Microbiol.* (2016). doi:10.1016/j.tim.2016.03.007

572 41. Eldholm, V. *et al.* Four decades of transmission of a multidrug-resistant

573 Mycobacterium tuberculosis outbreak strain. *Nat. Commun.* **6,** 7119 (2015).

574 42. Wirth, T. *et al.* Origin, Spread and Demography of the Mycobacterium tuberculosis
575 Complex. *PLoS Pathog.* **4,** e1000160 (2008).

576 43. Ragheb, M. N. *et al.* The mutation rate of mycobacterial repetitive unit loci in strains
577 of M. tuberculosis from cynomolgus macaque infection. *BMC Genomics* **14,** 145
578 (2013).

579 44. Ford, C. B. *et al.* Mycobacterium tuberculosis mutation rate estimates from different
580 lineages predict substantial differences in the emergence of drug-resistant
581 tuberculosis. *Nat. Genet.* **45,** 784–90 (2013).

582 45. Bouckaert, R. *et al.* BEAST 2: A Software Platform for Bayesian Evolutionary Analysis.
583 *PLoS Comput. Biol.* **10,** e1003537 (2014).

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

587 47. Kato-Maeda, M. *et al.* Strain classification of Mycobacterium tuberculosis:
588 congruence between large sequence polymorphisms and spoligotypes. *Int. J. Tuberc.*
589 *Lung Dis.* **15,** 131–3 (2011).

590 48. Filliol, I. *et al.* Global phylogeny of Mycobacterium tuberculosis based on single
591 nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution,
592 phylogenetic accuracy of other DNA fingerprinting systems, and recommendations
593 for a minimal standard SNP set. *J. Bacteriol.* **188,** 759–72 (2006).

594 49. Xia, E., Teo, Y.-Y. & Ong, R. T.-H. SpoTyping: fast and accurate in silico Mycobacterium
595 spoligotyping from sequence reads. *Genome Med.* **8,** 19 (2016).

596 50. Coll, F. *et al.* SpolPred: rapid and accurate prediction of Mycobacterium tuberculosis

spoligotypes from short genomic sequences. *Bioinformatics* **28,** 2991–3 (2012).

51. Warren, R. M. *et al.* Microevolution of the direct repeat region of Mycobacterium tuberculosis: implications for interpretation of spoligotyping data. *J. Clin. Microbiol.* **40,** 4457–65 (2002).

52. Mokrousov, I. *et al.* Next-Generation Sequencing of Mycobacterium tuberculosis. *Emerg. Infect. Dis.* **22,** 1127–9 (2016).

53. Scott, A. N. *et al.* Sensitivities and specificities of spoligotyping and mycobacterial interspersed repetitive unit-variable-number tandem repeat typing methods for studying molecular epidemiology of tuberculosis. *J. Clin. Microbiol.* **43,** 89–94 (2005).

54. Stucki, D. *et al.* Standard Genotyping Overestimates Transmission of Mycobacterium tuberculosis among Immigrants in a Low-Incidence Country. *J. Clin. Microbiol.* **54,** 1862–70 (2016).

55. den Hertog, A. L., Sengstake, S. & Anthony, R. M. Pyrazinamide resistance in Mycobacterium tuberculosis fails to bite? *Pathog. Dis.* **73,** ftv037 (2015).

56. Sengstake, S. *et al.* Pyrazinamide resistance-conferring mutations in pncA and the transmission of multidrug resistant TB in Georgia. *BMC Infect. Dis. 2017 171* **17,** 491 (2017).

57. Walker, T. M. *et al.* Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an observational study. *Lancet. Respir. Med.* **2,** 285–292 (2014).

58. Salathé, M. *et al.* Digital Epidemiology. *PLoS Comput. Biol.* **8,** e1002616 (2012).

59. Sanoussi, C. N., Affolabi, D., Rigouts, L., Anagonou, S. & de Jong, B. Genotypic characterization directly applied to sputum improves the detection of Mycobacterium africanum West African 1, under-represented in positive cultures. *PLoS Negl. Trop.*

Dis. **11,** e0005900 (2017).

60. Phelan, J. E. *et al.* Recombination in pe/ppe genes contributes to genetic variation in Mycobacterium tuberculosis lineages. *BMC Genomics* **17,** 151 (2016).

61. Ates, L. S. *et al.* Mutations in ppe38 block PE_PGRS secretion and increase virulence of Mycobacterium tuberculosis. *Nat. Microbiol.* **3,** 181–188 (2018).

62. Guthrie, J. L. *et al.* Molecular Epidemiology of Tuberculosis in British Columbia, Canada: A 10-Year Retrospective Study. *Clin. Infect. Dis.* **66,** 849–856 (2018).

63. WHO. *Guidelines for surveillance of drug resistance in tuberculosis*. (2015).

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

65. Zignol, M. *et al.* Genetic sequencing for surveillance of drug resistance in tuberculosis in highly endemic countries: a multi-country population-based surveillance study. *Lancet Infect. Dis.* (2018). doi:10.1016/S1473-3099(18)30073-2

66. Kohl, T. A. *et al.* MTBseq: A comprehensive pipeline for whole genome sequence analysis of Mycobacterium tuberculosis complex isolates. *Prep.* (2018).

67. Lew, J. M., Kapopoulou, A., Jones, L. M. & Cole, S. T. TubercuList – 10 years after. *Tuberculosis* **91,** 1–7 (2011).

68. Médigue, C., Cole, S. T., Camus, J.-C. & Pryor, M. J. Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv. *Microbiology* **148,** 2967–2973 (2002).

69. Feuerriegel, S. *et al.* PhyResSE: a Web Tool Delineating Mycobacterium tuberculosis Antibiotic Resistance and Lineage from Whole-Genome Sequencing Data. *J. Clin. Microbiol.* **53,** 1908–1914 (2015).

70. Kozlov, A. RAxML-NG. (2017). doi:https://zenodo.org/record/888146#.Wm8r2Uso9TY

645 71. Lewis, P. O. A likelihood approach to estimating phylogeny from discrete
646 morphological character data. *Syst. Biol.* **50,** 913–25 (2001).
647 72. Leaché, A. D. *et al.* Short Tree, Long Tree, Right Tree, Wrong Tree: New Acquisition
648 Bias Corrections for Inferring SNP Phylogenies. *Syst. Biol.* **64,** 1032–1047 (2015).
649 73. Kobert, K., Stamatakis, A. & Flouri, T. Efficient Detection of Repeating Sites to
650 Accelerate Phylogenetic Likelihood Calculations. *Syst. Biol.* **66,** syw075 (2016).
651 74. Rambaut, A. Figtree. (2016). Available at: http://tree.bio.ed.ac.uk/software/figtree/.
652 (Accessed: 3rd May 2017)
653 75. Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C. & Segata, N. Compact graphical
654 representation of phylogenetic data and metadata with GraPhlAn. *PeerJ* **3,** e1029
655 (2015).
656 76. Tessema, B. *et al.* FIND Tuberculosis Strain Bank: a Resource for Researchers and
657 Developers Working on Tests To Detect Mycobacterium tuberculosis and Related
658 Drug Resistance. *J. Clin. Microbiol.* **55,** 1066–1073 (2017).
659 77. Casali, N. *et al.* Whole Genome Sequence Analysis of a Large Isoniazid-Resistant
660 Tuberculosis Outbreak in London: A Retrospective Observational Study. *PLOS Med.*
661 **13,** e1002137 (2016).
662 78. Witney, A. A. *et al.* Clinical use of whole genome sequencing for Mycobacterium
663 tuberculosis. *BMC Med.* **14,** 46 (2016).
664 79. Sukumaran, J. & Holder, M. T. DendroPy: a Python library for phylogenetic
665 computing. *Bioinformatics* **26,** 1569–1571 (2010).
666 80. Driscoll, J. R. in *Methods in molecular biology (Clifton, N.J.)* **551,** 117–128 (2009).
667 81. Miotto, P. *et al.* Mycobacterium tuberculosis pyrazinamide resistance determinants: a
668 multicenter study. *MBio* **5,** e01819-14 (2014).

669 82. Coll, F. *et al.* A robust SNP barcode for typing Mycobacterium tuberculosis complex
670 strains. *Nat. Commun.* **5,** 4812 (2014).
671 83. Drummond, A. J., Ho, S. Y. W., Phillips, M. J., Rambaut, A. & Rambaut, A. Relaxed
672 Phylogenetics and Dating with Confidence. *PLoS Biol.* **4,** e88 (2006).
673 84. Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. Bayesian Coalescent
674 Inference of Past Population Dynamics from Molecular Sequences. *Mol. Biol. Evol.* **22,**
675 1185–1192 (2005).
676 85. Rambaut, A., Suchard, M. A., Xie, D. & Drummond, A. Tracer. (2013). Available at:
677 http://beast.bio.ed.ac.uk/Tracer. (Accessed: 11th December 2013)
678 86. Statisticat. LaplacesDemon: Complete Environment for Bayesian Inference. Bayesian-
679 Inference.com. R package version 16.0.1. (2016). Available at:
680 https://web.archive.org/web/20150206004624/http://www.bayesian-
681 inference.com/software.
682 87. R Core Team. R: A language and environment for statistical computing. (2017).
683 88. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York,
684 2009).
685

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;

Author contributions

CJM, FG and BCdJ conceived the study. MKK and BCdJ oversaw collection of isolates and ethical approval. TAK, SA, MM, PB and SN undertook classic genotyping and sequencing of isolates. CJM, PM, TA, CU and PL undertook WGS assembly and data preparation. CJM undertook all convergence and clustering analyses. CJM, PM, JP, MM, TS and DK undertook all phylodynamics. CJM, PM, SN and BCdJ wrote the manuscript. All authors read and revised the manuscript and approved its final form.

Competing interests

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

701 **Table legends**

702 Table 1: Clustering method overview.

703 For each clustering method, the general features are outlined in the tables. a) All clusters for each method affected by convergence. b) Clusters

704 derived only from non-convergent patterns. c) SNP- and cgMLST-based methods Mean ages and 95% HPD ranges are based upon the BEAST2

705 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

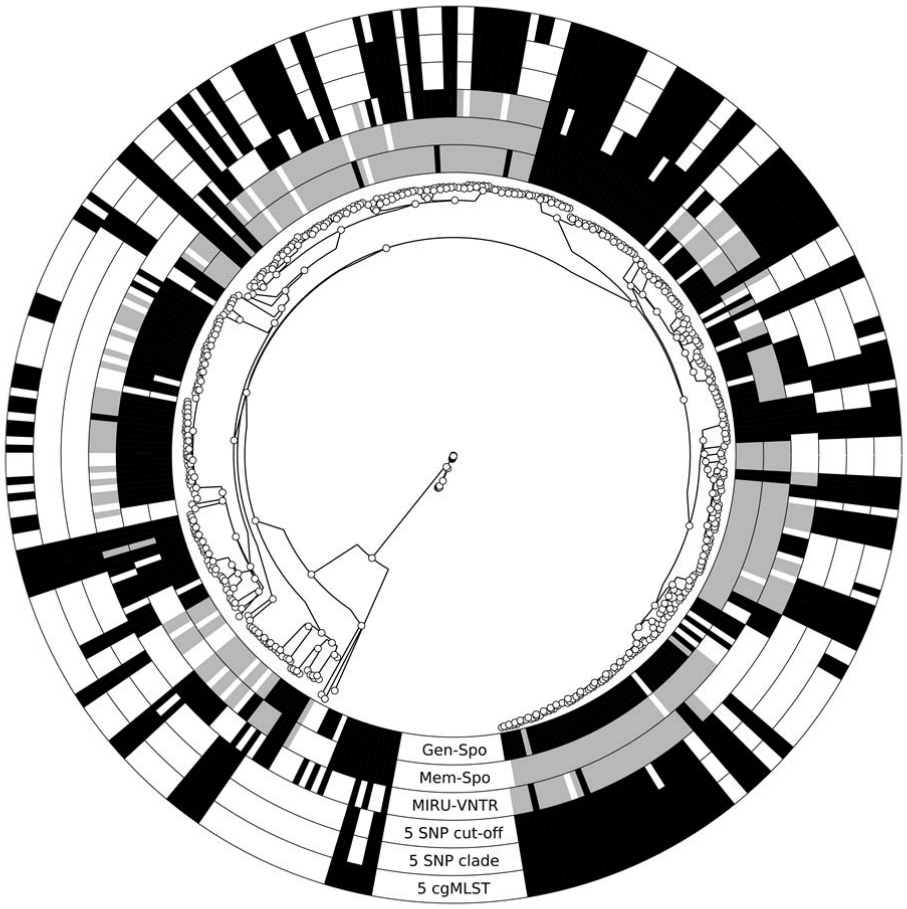
c)

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

706

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

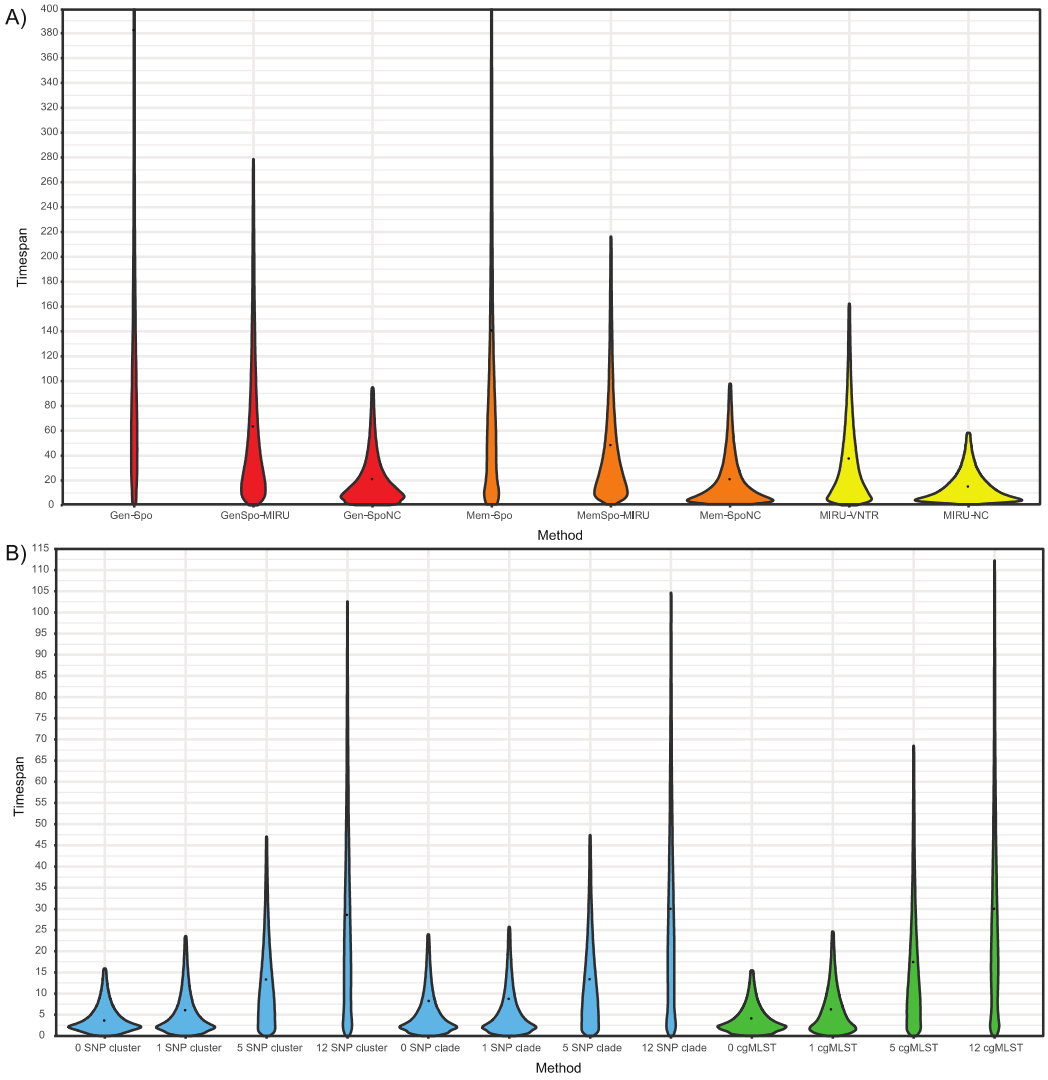
708 For a representative approach of each of the main methods (Mem-Spo, Gen-Spo, MIRU-
709 VNTR, 5 SNP cut-off, 5 SNP clade and 5 cgMLST) the inclusion of an isolate into a cluster is
710 outlined in the surrounding circles using GraPhlAn[75]. If an isolate is in a cluster not affected
711 by convergence, it is highlighted in black for the given method. If an isolate is in a cluster
712 affected by convergence, it is shown in grey. The clustering based on all 20 approaches is
713 shown in Supplementary Figure 1.



714

715 Figure 2: Timespans associated with transmission clusters

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



723