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, 2000, Belgium

[2]Advanced Database Research and Modelling (ADReM), Department of Mathematics and Computer Science, University of Antwerp, Antwerp, 2020, Belgium

[3]Biomedical Informatics Research Network Antwerp (biomina), University of Antwerp, Antwerp, 2020, Belgium

[4]German Center for Infection Research, Partner Site Hamburg-Lübeck-Borstel-Riems, D-23845 Borstel, Germany

[5]Molecular and Experimental Mycobacteriology, Priority Area Infections, Research Center Borstel, D-23845 Borstel, Germany

[6]Department of Biosystems Science and Engineering, ETH Zürich, 4058 Basel, Switzerland

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

[8]Bernhard Nocht Institute for Tropical Medicine, 20359 Hamburg, Germany

[9]University Hospital Zurich, 8091 Zurich, Switzerland

[10]National Tuberculosis Program, B.P 1197, Kinshasa, Democratic Republic of Congo

[&]equal contribution; *corresponding author: cmeehan@itg.be; +32-32476318

Keywords

Mycobacterium tuberculosis; MDR-TB molecular epidemiology; transmission; spoligotyping; MIRU-VNTR; MLST; whole genome sequencing; outbreak detection

Abstract

Background: Tracking recent transmission is a vital part of controlling widespread pathogens such as *Mycobacterium tuberculosis*. Multiple methods with specific performance characteristics exist for detecting recent transmission chains, usually by clustering strains based on genotype similarities. With such a large variety of methods available, informed selection of an appropriate approach for determining transmissions within a given setting/time period is difficult.

Methods: This study combines whole genome sequence (WGS) data derived from 324 isolates collected 2005-2010 in Kinshasa, Democratic Republic of Congo (DRC), a high endemic setting, with phylodynamics to unveil the timing of transmission events posited by a variety of standard genotyping methods. Clustering data based on Spoligotyping, 24-loci MIRU-VNTR typing, WGS based SNP (Single Nucleotide Polymorphism) and core genome multi locus sequence typing (cgMLST) typing were evaluated.

Findings: Our results suggest that clusters based on Spoligotyping could encompass transmission events that occurred over 70 years prior to sampling while 24-loci-MIRU-VNTR often represented two or more decades of transmission. Instead, WGS based genotyping applying low SNP or cgMLST allele thresholds allows for determination of recent transmission events in timespans of up to 10 years e.g. for a 5 SNP/allele cut-off.

Interpretation: With the rapid uptake of WGS methods in surveillance and outbreak tracking, the findings obtained in this study can guide the selection of appropriate clustering methods for uncovering relevant transmission chains within a given time-period. For high resolution cluster analyses, WGS-SNP and cgMLST based analyses have similar clustering/timing characteristics even for data obtained from a high incidence setting.

53

Research in context

Evidence before this study

For nearly 30 years, molecular genotyping tools have been used to define transmission chains/clusters of *Mycobacterium tuberculosis* strains. A variety of tools are used for such analysis e.g. the presence/absence of spacers sequences (Spoligotyping), the length of tandem repeat patterns (24-loci-MIRU-VNTR) or, more recently, nearly the complete genome by whole genome sequencing (WGS). Each method has been proposed as the gold standard genotyping technique for detecting transmission events in a certain timeframe and selection of the optimal method for a given question is difficult as important parameters (e.g. the time span a particular outbreak can encompass) are not well defined. Based on inferred mutation rates, there have been some time scales proposed for clusters based on WGS SNP-based methods, supported by contact tracing data to confirm epidemiological links. However, there is uncertainty around these timing estimates for SNP-based techniques, limited timing estimates available for classical genotyping techniques and no such estimates for cgMLST approaches. This makes it very difficult for researchers, public health workers and clinicians to correctly interpret reported clustering data. This is especially the case as WGS based methods are becoming rapidly ingrained in surveillance and clinical workflows.

Added value of this study

This study is the first to perform a comparative evaluation of cluster data defined by both classical and WGS-based *M. tuberculosis* genotyping approaches, especially with regard to transmission timing. While many studies have put forward various methods as the gold standard for *M. tuberculosis* transmission detection, we have tested clustering data

generated by the different methods in a Bayesian statistical framework to elucidate the true fraction of recent transmission each approach is detecting. When specifically looking at recent transmission (e.g. less than 10 years previous), our results indicate that classical genotyping methods vastly over estimate recent transmission events. This solidifies the need for WGS-based methods when searching for recent outbreaks of *M. tuberculosis*.

Implications of all the available evidence

Our study allows researchers and public health officials to select the appropriate genotyping method for assessing transmission with respect to the epidemiological setting and a given time-period. We also suggest the incorporation of particular genotyping methods in a cascade system with increasing resolution for various levels of surveillance e.g. from multi-country surveillance down to recent transmission and outbreak analyses. This is particularly important as each method comes with specific costs, infrastructure and computational requirements, human resources, and, last but not least interpretation complexities – all of which might not be feasible at any site or at any scale. Accordingly, our study can aid a cost/benefit analysis for selection of genotyping techniques in a cascade system, that might especially be used in high incidence, low resource settings.

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 (MDR) Mtbc strains 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]. Thus, interrupting transmission is key for the control of MDR-TB[3,4]. 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)[5]. Contact tracing is the primary non-molecular epidemiological method for investigating transmission networks of TB, mainly based on patient interviews[6]. 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[7] and misses many connections[8]. The implementation of molecular genotyping and epidemiological approaches has overcome these limitations and is often used as the main approach for transmission analyses. Classical genotyping has involved IS*6110* DNA fingerprinting[9], Spoligotyping[10], and variable-number tandem repeats of mycobacterial interspersed repetitive units (MIRU-VNTR)[11] which is the most common method at the moment[5]. The latter method is based on copy numbers of a sequence in tandem repeat patterns derived from 24 distinct loci within the genome[12]. If two patients have the same classical genotyping pattern such as a 24-loci MIRU-VNTR pattern (or up to one locus difference[12]) 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[12]. However,

125 examples of unlinked patients with identical patterns have been observed, suggesting that
126 this threshold covers too broad a genetic diversity and timespan between infections[7].
127 The application of (whole genome) sequence (WGS)-based approaches for similarity analysis
128 of Mtbc isolates and cluster determination is known to have high discriminatory power
129 when assessing transmission dynamics[7,13–16], either using core genome multi-locus
130 sequence typing (cgMLST)[17,18] or SNP distances[7,14,15,19]. WGS-based approaches compare the
131 genetic relatedness of the genomes of the clinical strains under consideration, albeit usually
132 excluding large repetitive portions of the genome, with the assumption that highly similar
133 strains are linked by a recent transmission event[7,14]. Although many SNP cut-offs for linking
134 isolates have been proposed[20], the most commonly employed is based on the finding that a
135 5 SNP cut-off will cluster the genomes of strains from the majority of epidemiologically
136 linked TB patients, with an upper bound of 12 SNPs between any two linked isolates[14]. The
137 emerging widespread use of WGS has quickly pushed these cut-offs to be considered the
138 new molecular gold standard of recent transmission linking, although SNP distances may
139 vary for technical reasons (e.g. assembly pipelines or filter criteria[21]) and between study
140 populations e.g. high and low incidence settings[19].
141 In addition to cluster detection, uncovering the timing of transmission events within a given
142 cluster is highly useful information for TB control e.g. for assessing the impact of
143 interventions on the spread of an outbreak or uncovering when MDR-TB transmission first
144 emerged in a particular setting. Accordingly, knowledge of the rate change associated with
145 different genotyping methods is essential for correct timing. The whole genome mutation
146 rate of Mtbc strains has been estimated by several studies as between 10^{-7} and 10^{-8}
147 substitutions per site per year or ~0·3-0·5 SNPs per genome per year[7,14,22–24], while the rate
148 of change in the MIRU-VNTR loci specifically is known to be quicker (~10^{-3})[25]. Since these

149 mutation rates have been shown to also vary by lineage[23,26] and over short periods of
150 time[22], such variation needs to be accounted for when estimating transmission times, e.g.
151 by using Bayesian phylogenetic dating techniques[3,22,25].
152 Considering the multiple genotyping methods currently available, many of them proposed
153 as a "gold standard", there is an urgent need to precisely define the individual capacity of
154 each method to accurately detect recent transmission events and perform timing of
155 outbreaks. To provide this essential information, this study harnesses the power of WGS-
156 based phylogenetic dating methods to assign timespans onto Mtbc transmission chains
157 encompassed by the different genotypic clustering methods commonly used in TB
158 transmission studies.
159

160 **Materials and Methods**

161 Dataset, ethical approval and sequencing

162 A set of 324 isolates from Kinshasa, Democratic Republic of Congo were collected from
163 consecutive retreatment TB patients between 2005 and 2010 at TB clinics, servicing an
164 estimated 30% of the population of Kinshasa. All isolates were phenotypically resistant to
165 rifampicin (RR-TB) and the majority are also isoniazid resistant (i.e. MDR-TB). Use of the
166 stored isolates without any linked personal information was approved by the health
167 authorities of the DRC and the Institutional Review Board of the ITM in Antwerp (ref no
168 945/14). Libraries for whole genome sequencing were prepared from extracted genomic
169 DNA with the Illumina Nextera XT kit, and run on the Illumina NextSeq platform in a
170 2x151bp run according to manufacturer's instructions. Illumina read sets are available on
171 the ENA (https://www.ebi.ac.uk/ena) under the accession number PRJEB27847.
172

173 Genome reconstruction

174 The MTBseq pipeline[27] was used to detect the SNPs for each isolate using the H37Rv
175 reference genome (NCBI accession number NC000962.3)[28]. Sites known to be involved in
176 drug resistance (as outlined in the PhyResSE list of drug mutations v27[29]) were excluded
177 from the alignment and additional filtering of sites with ambiguous calls in >5% of isolates
178 and those SNPs within a 12bp window of each other was also applied.

179

180 Transmission cluster estimation methods

181 Five standard transmission clustering approaches were chosen for comparison and analysis
182 (four additional methods are outlined in the supplementary methods). Clustering due to
183 convergent genotypic patterns were removed as outlined in the supplementary methods.
184 For each method, the total SNP distances were calculated to investigate the range of
185 variability encompassed within each cluster. Maximum SNP distances were derived from
186 pairwise comparisons of isolates within the SNP alignment using custom python scripts. A
187 clustering rate was calculated for each method using the formula $(n_c - c)/n$, where n_c is the
188 total number of isolates clustered by a given method, c is the number of clusters, and n is
189 the total number of isolates in the dataset (n=324).

190

191 Spoligotyping and MIRU-VNTR

192 Spoligotype patterns were obtained from membranes following the previously published
193 protocol[10]. Isolates were said to be clustered if all 43 spacers matched. Genotyping by
194 MIRU-VNTR was undertaken as previously described[12]. 2µl of DNA was extracted from
195 cultures and amplified using the 24 loci MIRU-VNTR typing kit (Genoscreen, Lille, France).
196 Analysis of patterns was undertaken using the ABI 3500 automatic sequencer (Applied

197 Biosystems, California, USA) and Genemapper software (Applied Biosystems). Isolates were
198 said to be clustered if all 24 loci matched. Mixed MIRU-VNTR patterns were observed in 18
199 isolates although this mixing was not observed in the WGS data, likely due to subculturing
200 for sequencing.
201
202 SNP cut-off clustering
203 In this study, we employed the widely used 5 SNP (proposed by Walker *et al.* [14] as the likely
204 boundary for linked transmission) and 12 SNP cut-offs (proposed maximum boundary) for
205 cluster definition. Additionally, we employed a lower cut-off of 1 SNP to look for clusters of
206 very highly related isolates. Pairwise SNP distances were calculated between all isolates. A
207 loose cluster definition was used, where every isolate in a cluster at most the SNP cut-off
208 from at least 1 other isolate in the cluster.
209
210 Estimation of transmission times
211 To estimate the age and timespan of potential transmission clusters, SNP alignments were
212 created from the convergence-free version of the three primary clustering types:
213 Spoligotyping, MIRU-VNTR, and 12 SNP clusters.
214 A Bayesian approach to transmission time estimation was then undertaken. The SNP
215 alignments were created as above for the three high-level clustering types. Each cluster
216 method alignment was separately input to BEAST-2 v2.4.7[30] to create a time tree for those
217 isolates. These phylogenies were built using the following priors: GTR+GAMMA substitution
218 model, a log-normal relaxed molecular clock model to account for variation in mutation
219 rates[31] and coalescent constant size demographic model[32], both of which have been found
220 to be suitable for lineage 4 isolates in a previous study[19]. The MCMC chain was run six times

221 independently per alignment with a length of at least 400 million, sampled every 40,000th
222 step (MIRU & Spoligotyping: 600 million; 12 SNP: 500 million;). A log normal prior (mean
223 $1.5x10^{-7}$; variance 1·0) was used for the clock model to reflect the previously estimated
224 mutation rate of *M. tuberculosis* lineage 4[7,14,22–24], while allowing for variation as previously
225 suggested[22]. A 1/X non-informative prior was selected for the population size parameter of
226 the demographic model. Isolation dates were used as informative heterochronous tip dates
227 and the SNP alignment was augmented with a count of invariant sites for each of the four
228 nucleotide bases to avoid ascertainment bias[33]. Tracer v1.6 was used to determine
229 adequate mixing and convergence of chains (ESS >150) after a 25% burn-in. The chains were
230 combined via LogCombiner v2.4.8[30] to obtain a single chain for each clustering type with
231 high (>1000) effective sample sizes. The tree samples were combined in the same manner
232 and resampled at a lower frequency to create thinned samples of (minimum) 20,000 trees.
233 The algorithm for estimating the timespan of transmission events encompassed by each
234 method is outlined in Supplemental Figure 1. Briefly, for each cluster created by the given
235 method, we defined the MRCA node as the internal node that connects all taxa in that
236 cluster. The youngest node was then defined as the tip that is furthest from this MRCA
237 within the clade (i.e. the tip descendant from that node that was sampled closest to the
238 present time). To better account for changes in the mutation rate over short periods[22], all
239 trees estimated and sampled during the Bayesian MCMC process were used instead of only
240 a single summary phylogeny. For each retained tree in the MCMC process, the difference in
241 age between the MRCA node and youngest node was calculated. This gave a distribution of
242 likely maximum transmission event times within that cluster. For each method, these per-
243 cluster aggregated ages were then combined across all clusters to give a per-method
244 distribution of transmission event times represented by the clusters. The 95% Highest

245 Posterior Density (HPD) interval of these distributions was calculated with the
246 LaplacesDemon p.interval function in R v3.4.0 and the distribution within this interval for
247 each method along with the mean based upon this interval were then visualized in violin
248 plots per clustering method using ggplot2 in R.
249
250 **Results**
251 In this study, we assessed five different approaches for generating putative *M. tuberculosis*
252 transmission clusters: Spoligotyping, MIRU-VNTR, and SNP-based clustering using a 12, 5
253 and 1 SNP cut-off, using a dataset of 324 isolates collected 2005-2010 in Kinshasa,
254 Democratic Republic of Congo (DRC). The dataset contained 309 L4 and 15 L5 isolates, with
255 a maximum of 1671 SNPs between any two isolates. Bayesian phylodynamic dating
256 approaches implemented in BEAST-2[30] were then utilised to assign timespans to the
257 transmission events estimated by each genotyping method.
258 As expected, classical genotyping methods clustered many strains, with the lowest
259 resolution (i.e. highest clustering rate) (Figure 1, Table 1). Convergent evolution (defined as
260 the same pattern observed in unrelated strains; see supplemental methods) was found to
261 affect 39% (12) of Spoligotyping-based clusters. Although MIRU-VNTR performed far better
262 than spoligotyping, 16% (6) of clustering patterns were influenced by convergence in this
263 study (Table 1, Figure 1). Spoligotyping and MIRU-VNTR patterns that were affected by
264 convergence were removed from further timespan analyses. WGS-based methods had by
265 far the highest discriminatory power and low SNP cut-offs grouped isolates into smaller
266 clusters (e.g. 2-10 isolates per cluster for a 5 SNP cut-off) (Table 1, Figure 1). The high
267 percentage of strains in a 12 SNP cluster (75%) suggests high levels of transmission in this
268 population, making is suitable for further transmission analyses.

Bayesian phylogenetic dating of the timeframe associated with particular transmission chains showed large differences in estimated cluster ages between the different 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 (see methods and supplementary figure 1). 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 median ages of clusters derived from Spoligotyping were found to often be several decades old (median 77 years (95% HPD: 1-823)) (Table 1, Figure 2). MIRU-VNTR clustering encompassed more recent transmission events than Spoligotyping, but were still found to be often over 2 decades old (median 26 years (95% HPD: 0-162)). The combination of MIRU-VNTR and Spoligotyping resulted in cluster ages similar to MIRU-VNTR alone (supplemental table 1). Clusters based on SNP cut-offs correlated to 23 years using a 12 SNP cut-off (95% HPD: 0-103), 11 years using a 5 SNP cut-off (95% HPD: 0-47), and 4 years using a 1 SNP cut-off (95% HPD: 0-24) (Table 1, Figure 2). For comparison, the cgMLST method was also performed and discriminatory power, cluster sizes and ages based on cgMLST alleles were similar to the SNP-based clusters (see supplemental methods, supplemental table 1).

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/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 several decades old. This low discriminatory power coupled with the high rate of convergent evolution (the same spoligotype pattern found in phylogenetically distant isolates) adds weight to the previous suggestion that this technique is not suitable for recent transmission studies[34], although may be of use as a low-cost method of sorting Mtbc strains into the seven primary lineages[35,36].

In line with previous findings[34,37], 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 often spanned over two decades (Table 1, Figure 2), confirming previous studies showing over-estimation of recent transmission with this method[7,13,19,38].

For defining transmission events that occurred in more recent time frames before sampling, WGS-based methods 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

317 clustering isolates, often defines transmission events that occurred on average two decades
318 prior to sampling, similar in age to clusters estimated by MIRU-VNTR. This suggests that the
319 12 SNP cluster method may be a good replacement for MIRU-VNTR as it detects larger
320 transmission networks spanning similar transmission time periods but is less affected by
321 convergent evolution. Isolates clustered at a low (5 SNP) or nearly identical (1 SNP) cut-off
322 were found to represent transmission events occurring over a time span of up to ten years.
323 These findings correlate well with previous studies where confirmed contact tracing-based
324 epidemiological links were found between patients that were two[15,39] and three[7] SNPs
325 apart. The original paper that proposed the 5 and 12 SNP cut-offs found that serial isolates
326 that were 10 years apart differed by, on average, 6 SNPs, also agreeing with the findings
327 presented here [14]. Comparisons between the SNP-based (using almost all genomic
328 differences) and the cgMLST-based (using a defined core set of genes) methods
329 demonstrated that the latter approach gives similar estimates to full SNP approaches. This
330 supports the use of low SNP or cgMLST differences for detection or exclusion of very recent
331 transmission, although this low variability between isolates makes robust identification of
332 transmission direction difficult, especially during short timespans.
333 The mutation rate of *M. tuberculosis* has been estimated to be between 10^{-7} and 10^{-8}
334 substitutions per site per year[3,7,23]. Within the Bayesian analysis employed here, the
335 mutation rate was free to vary between these values but was found to strongly favour
336 ~$3x10^{-8}$ (ESS > 1000 for all runs; 95% HDP: $4x10^{-9}$ - $8x10^{-8}$), translating to approximately 0·3
337 SNPs per genome per year. While the mutation rate used here is in line with previous
338 estimates for lineage 4 [23] (which most of this dataset is comprised of), it may be similar in
339 other lineages, although this has only been shown for lineage 2[3,23]. Thus, per-lineage

340 estimates are required for all seven lineages to ensure similar transmission times are linked
341 to genotyping methods across the whole diversity of the Mtbc.
342 While this study has many advantages due to its five year population based design in an
343 endemic setting coupled with the application of three different genotyping methods
344 (Spoligotyping, 24-locus MIRU-VNTR and WGS), future confirmatory studies could address
345 the following drawbacks that are inherent to genomic epidemiology[16,21]: 1) studies
346 employing contact tracing and/or digital epidemiology[40] in conjunction with these
347 genotyping methods can help confirm transmission times associated with different clusters
348 (although these methods also have many limitations); 2) as outlined above, strains of other
349 lineages of the Mtbc should be analysed in a similar fashion to ensure transferability of
350 findings across the entire complex; 3) a broad range of drug resistance profiles should be
351 included to fully assess the impact of such mutations on transmission estimates; 4)
352 improved WGS methods, such as directly from clinical samples to help reduce culture
353 biases[41] and longer reads (e.g. PacBio SMRT or Nanopore MinION) to capture the entire
354 genome, including repetitive regions such as PE/PPE genes known to impact genome
355 remodelling[42,43], will ensure that the maximum diversity between isolates is captured; 5)
356 extensive panels of Spoligotyping and MIRU-VNTR results paired with WGS data will help
357 assess the extent of convergence in these methods and better correlate their clusters with
358 those of low SNP thresholds and 6) standardised SNP calling pipelines appropriate across all
359 lineages, with high true positive/low false negative rates, will ensure that Mtbc molecular
360 epidemiology can be uniformly implemented and comparable across studies.
361 Since each method was found to represent different timespans and clustering definitions,
362 they can be used in a stratified manner in an integrated epidemiological and public health
363 investigation addressing the transmission of Mtbc 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.

These potential transmission hotspots could be further investigated with contact tracing and/or WGS. Employment of different cut-offs and clustering approaches to WGS data can then address several questions. The 12 SNP/cgMLST allele cluster approaches serve well for high level surveillance targeting larger (older) transmission networks, akin to what is currently often done using MIRU-VNTR (e.g.[15,44]). Recent transmission events can then be detected through employment of low SNP cut-offs (e.g. 5 SNPs for transmission in the past 10 years or 1 SNPs for transmission in the past 5 years). In high incidence/low diversity settings where amalgamation of clusters may inadvertently obscure distinct hotspots of transmission at different time points, subdivision into distinct time-dependant clusters can be undertaken using the algorithm presented in such a study in East Greenland[19].

Overall, phylodynamic approaches applied to whole genome sequences, as undertaken here, are recommended to fully investigate the specific transmission dynamics within a study population to account for setting-specific conditions, such as low/high TB incidence, low/high pathogen population diversity, sampling fractions and social factors influencing transmission. As WGS methods become more commonplace and easier to implement in a variety of settings, 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.

Acknowledgements

The authors would like to thank Armand Van Deun and Koen Vandelannoote for valuable discussion and input and Cecile Uwizeye for aid with spoligotyping.

Funding sources

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, CU, PB 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.

Declaration of interests

The authors declare there are no conflicts of interest attached to this work.

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.

412 **References**

413 1. WHO. Global tuberculosis report 2017 [Internet]. 2018. Available from:
414 http://www.who.int/tb/publications/global_report/archive/en/
415 2. Kendall EA, Fofana MO, Dowdy DDW, WHO, Dye C, Garnett G, et al. Burden of
416 transmitted multidrug resistance in epidemics of tuberculosis: a transmission
417 modelling analysis. Lancet Respir Med [Internet]. 2015 Nov 12 [cited 2015 Nov
418 23];3(12):963–72. Available from:
419 http://linkinghub.elsevier.com/retrieve/pii/S2213260015004580
420 3. Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, Willery E, et al. Evolutionary
421 history and global spread of the Mycobacterium tuberculosis Beijing lineage. Nat
422 Genet [Internet]. 2015 Jan 19 [cited 2017 Sep 22];47(3):242–9. Available from:
423 http://www.nature.com/doifinder/10.1038/ng.3195
424 4. Shah NS, Auld SC, Brust JCM, Mathema B, Ismail N, Moodley P, et al. Transmission of
425 Extensively Drug-Resistant Tuberculosis in South Africa. N Engl J Med [Internet]. 2017
426 Jan 19 [cited 2017 Jul 24];376(3):243–53. Available from:
427 http://www.nejm.org/doi/10.1056/NEJMoa1604544
428 5. Merker M, Kohl TA, Niemann S, Supply P. The Evolution of Strain Typing in the
429 Mycobacterium tuberculosis Complex. In: Advances in experimental medicine and
430 biology [Internet]. 2017 [cited 2018 Apr 10]. p. 43–78. Available from:
431 http://www.ncbi.nlm.nih.gov/pubmed/29116629
432 6. Fox GJ, Barry SE, Britton WJ, Marks GB. Contact investigation for tuberculosis: a
433 systematic review and meta-analysis. Eur Respir J [Internet]. 2012 [cited 2017 Sep
434 6];41(1). Available from: http://erj.ersjournals.com/content/41/1/140.short
435 7. Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, et al. Whole genome

436 sequencing versus traditional genotyping for investigation of a Mycobacterium
437 tuberculosis outbreak: a longitudinal molecular epidemiological study. Neyrolles O,
438 editor. PLoS Med [Internet]. 2013 Jan 12 [cited 2014 Mar 31];10(2):e1001387.
439 Available from: http://dx.plos.org/10.1371/journal.pmed.1001387
440 8. Bjorn-Mortensen K, Lillebaek T, Koch A, Soborg B, Ladefoged K, Sørensen HCF, et al.
441 Extent of transmission captured by contact tracing in a tuberculosis high endemic
442 setting. Eur Respir J [Internet]. 2017 [cited 2017 Sep 13];49(3). Available from:
443 http://erj.ersjournals.com/content/49/3/1601851
444 9. Thierry D, Cave MD, Eisenach KD, Crawford JT, Bates JH, Gicquel B, et al. IS6110, an
445 IS-like element of Mycobacterium tuberculosis complex. Nucleic Acids Res [Internet].
446 1990 Jan 11 [cited 2017 Sep 19];18(1):188. Available from:
447 http://www.ncbi.nlm.nih.gov/pubmed/2155396
448 10. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, et al.
449 Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for
450 diagnosis and epidemiology. J Clin Microbiol [Internet]. 1997 Apr [cited 2017 May
451 30];35(4):907–14. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9157152
452 11. Supply P, Magdalena J, Himpens S, Locht C. Identification of novel intergenic
453 repetitive units in a mycobacterial two-component system operon. Mol Microbiol
454 [Internet]. 1997 Dec [cited 2017 Jul 7];26(5):991–1003. Available from:
455 http://www.ncbi.nlm.nih.gov/pubmed/9426136
456 12. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rusch-Gerdes S, Willery E, et al.
457 Proposal for Standardization of Optimized Mycobacterial Interspersed Repetitive
458 Unit-Variable-Number Tandem Repeat Typing of Mycobacterium tuberculosis. J Clin
459 Microbiol [Internet]. 2006 Dec 1 [cited 2017 May 3];44(12):4498–510. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/17005759

13. Wyllie D, Davidson J, Walker T, Rathod P, Peto T, Robinson E, et al. A quantitative evaluation of MIRU-VNTR typing against whole-genome sequencing for identifying Mycobacterium tuberculosis transmission: A prospective observational cohort study. bioRxiv [Internet]. 2018 Jan 24 [cited 2018 Mar 19];252734. Available from: https://www.biorxiv.org/content/early/2018/01/24/252734

14. Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis [Internet]. 2013 Feb [cited 2014 Feb 23];13(2):137–46. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3556524&tool=pmcentrez&rendertype=abstract

15. Walker TM, Merker M, Knoblauch AM, Helbling P, Schoch OD, van der Werf MJ, et al. A cluster of multidrug-resistant Mycobacterium tuberculosis among patients arriving in Europe from the Horn of Africa: a molecular epidemiological study. Lancet Infect Dis [Internet]. 2018 Jan 8 [cited 2018 Mar 20]; Available from: https://www.sciencedirect.com/science/article/pii/S1473309918300045

16. Comas I. Genomic Epidemiology of Tuberculosis. In Springer, Cham; 2017 [cited 2018 Mar 20]. p. 79–93. Available from: http://link.springer.com/10.1007/978-3-319-64371-7_4

17. Kohl TA, Diel R, Harmsen D, Rothgänger J, Walter KM, Merker M, et al. Whole-genome-based Mycobacterium tuberculosis surveillance: a standardized, portable, and expandable approach. J Clin Microbiol [Internet]. 2014 Jul 1 [cited 2017 Sep 18];52(7):2479–86. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24789177

484 18. Kohl TA, Harmsen D, Rothgänger J, Walker T, Diel R, Niemann S. Harmonized Genome
485 Wide Typing of Tubercle Bacilli Using a Web-Based Gene-By-Gene Nomenclature
486 System. EBioMedicine. 2018;
487 19. Bjorn-Mortensen K, Soborg B, Koch A, Ladefoged K, Merker M, Lillebaek T, et al.
488 Tracing Mycobacterium tuberculosis transmission by whole genome sequencing in a
489 high incidence setting: a retrospective population-based study in East Greenland. Sci
490 Rep [Internet]. 2016 Sep 12 [cited 2017 May 3];6:33180. Available from:
491 http://www.ncbi.nlm.nih.gov/pubmed/27615360
492 20. Hatherell H-A, Colijn C, Stagg HR, Jackson C, Winter JR, Abubakar I. Interpreting whole
493 genome sequencing for investigating tuberculosis transmission: a systematic review.
494 BMC Med [Internet]. 2016 Dec 23 [cited 2017 Sep 13];14(1):21. Available from:
495 http://www.biomedcentral.com/1741-7015/14/21
496 21. Guthrie JL, Gardy JL. A brief primer on genomic epidemiology: lessons learned from
497 Mycobacterium tuberculosis. Ann N Y Acad Sci [Internet]. 2017 Jan 1 [cited 2018 Apr
498 16];1388(1):59–77. Available from: http://doi.wiley.com/10.1111/nyas.13273
499 22. Bryant JM, Schürch AC, van Deutekom H, Harris SR, de Beer JL, de Jager V, et al.
500 Inferring patient to patient transmission of Mycobacterium tuberculosis from whole
501 genome sequencing data. BMC Infect Dis [Internet]. 2013 Feb 27 [cited 2014 Feb
502 13];13(1):110. Available from:
503 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3599118&tool=pmcentr
504 ez&rendertype=abstract
505 23. Duchêne S, Holt KE, Weill F-X, Le Hello S, Hawkey J, Edwards DJ, et al. Genome-scale
506 rates of evolutionary change in bacteria. Microb genomics [Internet]. 2016 Nov [cited
507 2017 Jul 7];2(11):e000094. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/28348834

24. Eldholm V, Monteserin J, Rieux A, Lopez B, Sobkowiak B, Ritacco V, et al. Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. Nat Commun [Internet]. 2015 May 11 [cited 2016 Aug 17];6:7119. Available from: http://www.nature.com/doifinder/10.1038/ncomms8119

25. Wirth T, Hildebrand F, Allix-Béguec C, Wölbeling F, Kubica T, Kremer K, et al. Origin, Spread and Demography of the Mycobacterium tuberculosis Complex. Achtman M, editor. PLoS Pathog [Internet]. 2008 Sep 26 [cited 2018 Apr 10];4(9):e1000160. Available from: http://dx.plos.org/10.1371/journal.ppat.1000160

26. Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T, et al. Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. Nat Genet [Internet]. 2013 Jul [cited 2017 Aug 30];45(7):784–90. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23749189

27. Kohl TA, Utpatel C, Schleusener V, De Filippo MR, Beckert P, Cirillo DM, et al. MTBseq: A comprehensive pipeline for whole genome sequence analysis of Mycobacterium tuberculosis complex isolates. Prep. 2018;

28. Lew JM, Kapopoulou A, Jones LM, Cole ST. TubercuList – 10 years after. Tuberculosis [Internet]. 2011 [cited 2017 May 3];91(1):1–7. Available from: http://www.sciencedirect.com/science/article/pii/S1472979210001113

29. Feuerriegel S, Schleusener V, Beckert P, Kohl TA, Miotto P, Cirillo DM, et al. PhyResSE: a Web Tool Delineating Mycobacterium tuberculosis Antibiotic Resistance and Lineage from Whole-Genome Sequencing Data. Carroll KC, editor. J Clin Microbiol [Internet]. 2015 Jun [cited 2017 May 3];53(6):1908–14. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/25854485

30. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, et al. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. Prlic A, editor. PLoS Comput Biol [Internet]. 2014 Apr 10 [cited 2017 May 3];10(4):e1003537. Available from: http://dx.plos.org/10.1371/journal.pcbi.1003537

31. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A, Rambaut A. Relaxed Phylogenetics and Dating with Confidence. Penny D, editor. PLoS Biol [Internet]. 2006 Mar 14 [cited 2017 May 3];4(5):e88. Available from: http://dx.plos.org/10.1371/journal.pbio.0040088

32. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. Mol Biol Evol [Internet]. 2005 Feb 9 [cited 2017 May 3];22(5):1185–92. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msi103

33. Leaché AD, Banbury BL, Felsenstein J, de Oca A nieto-M, Stamatakis A, K. S, et al. Short Tree, Long Tree, Right Tree, Wrong Tree: New Acquisition Bias Corrections for Inferring SNP Phylogenies. Syst Biol [Internet]. 2015 Nov [cited 2017 May 5];64(6):1032–47. Available from: https://academic.oup.com/sysbio/article-lookup/doi/10.1093/sysbio/syv053

34. Comas I, Homolka S, Niemann S, Gagneux S. Genotyping of Genetically Monomorphic Bacteria: DNA Sequencing in Mycobacterium tuberculosis Highlights the Limitations of Current Methodologies. PLoS One [Internet]. 2009 Nov 12 [cited 2018 Apr 10];4(11):e7815. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19915672

35. Kato-Maeda M, Gagneux S, Flores LL, Kim EY, Small PM, Desmond EP, et al. Strain classification of Mycobacterium tuberculosis: congruence between large sequence

556 polymorphisms and spoligotypes. Int J Tuberc Lung Dis [Internet]. 2011 Jan [cited
557 2017 Sep 6];15(1):131–3. Available from:
558 http://www.ncbi.nlm.nih.gov/pubmed/21276309
559 36. Filliol I, Motiwala AS, Cavatore M, Qi W, Hazbón MH, Bobadilla del Valle M, et al.
560 Global phylogeny of Mycobacterium tuberculosis based on single nucleotide
561 polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic
562 accuracy of other DNA fingerprinting systems, and recommendations for a minimal
563 standard SNP set. J Bacteriol [Internet]. 2006 Jan 15 [cited 2017 Sep 6];188(2):759–
564 72. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16385065
565 37. Scott AN, Menzies D, Tannenbaum T-N, Thibert L, Kozak R, Joseph L, et al. Sensitivities
566 and specificities of spoligotyping and mycobacterial interspersed repetitive unit-
567 variable-number tandem repeat typing methods for studying molecular epidemiology
568 of tuberculosis. J Clin Microbiol [Internet]. 2005 Jan [cited 2017 Jul 7];43(1):89–94.
569 Available from: http://www.ncbi.nlm.nih.gov/pubmed/15634955
570 38. Stucki D, Ballif M, Egger M, Furrer H, Altpeter E, Battegay M, et al. Standard
571 Genotyping Overestimates Transmission of Mycobacterium tuberculosis among
572 Immigrants in a Low-Incidence Country. J Clin Microbiol [Internet]. 2016 Jul 1 [cited
573 2017 Jul 21];54(7):1862–70. Available from:
574 http://www.ncbi.nlm.nih.gov/pubmed/27194683
575 39. Walker TM, Lalor MK, Broda A, Ortega LS, Morgan M, Parker L, et al. Assessment of
576 Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007-12, with whole
577 pathogen genome sequences: an observational study. Lancet Respir Med [Internet].
578 2014 Apr [cited 2018 Apr 10];2(4):285–92. Available from:
579 http://www.ncbi.nlm.nih.gov/pubmed/24717625

580 40. Salathé M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, et al. Digital
581 Epidemiology. PLoS Comput Biol [Internet]. 2012 Jul 26 [cited 2018 Apr
582 10];8(7):e1002616. Available from: http://dx.plos.org/10.1371/journal.pcbi.1002616
583 41. Sanoussi CN, Affolabi D, Rigouts L, Anagonou S, de Jong B. Genotypic characterization
584 directly applied to sputum improves the detection of Mycobacterium africanum West
585 African 1, under-represented in positive cultures. Picardeau M, editor. PLoS Negl Trop
586 Dis [Internet]. 2017 Sep 1 [cited 2018 Mar 22];11(9):e0005900. Available from:
587 http://dx.plos.org/10.1371/journal.pntd.0005900
588 42. Phelan JE, Coll F, Bergval I, Anthony RM, Warren R, Sampson SL, et al. Recombination
589 in pe/ppe genes contributes to genetic variation in Mycobacterium tuberculosis
590 lineages. BMC Genomics [Internet]. 2016 Feb 29 [cited 2018 Mar 16];17:151.
591 Available from: http://www.ncbi.nlm.nih.gov/pubmed/26923687
592 43. Ates LS, Dippenaar A, Ummels R, Piersma SR, van der Woude AD, van der Kuij K, et al.
593 Mutations in ppe38 block PE_PGRS secretion and increase virulence of
594 Mycobacterium tuberculosis. Nat Microbiol [Internet]. 2018 Feb 15 [cited 2018 Mar
595 16];3(2):181–8. Available from: http://www.nature.com/articles/s41564-017-0090-6
596 44. Guthrie JL, Kong C, Roth D, Jorgensen D, Rodrigues M, Hoang L, et al. Molecular
597 Epidemiology of Tuberculosis in British Columbia, Canada: A 10-Year Retrospective
598 Study. Clin Infect Dis [Internet]. 2018 Mar 5 [cited 2018 Mar 20];66(6):849–56.
599 Available from: http://www.ncbi.nlm.nih.gov/pubmed/29069284
600 45. Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. Compact graphical
601 representation of phylogenetic data and metadata with GraPhlAn. PeerJ [Internet].
602 2015 Jun 18 [cited 2017 Aug 23];3:e1029. Available from:
603 http://www.ncbi.nlm.nih.gov/pubmed/26157614

604

Table legends

Table 1: Clustering method overview.

For each clustering method, the general features are outlined in the table. Median ages and 95% HPD ranges are based upon the BEAST2 estimates of clade heights (see methods).

Method	Strains in clusters	Number of clusters	Percent of strains in clusters	Cluster sizes	Maximum SNP distances	Clustering rate	Median Timespan	Timespan 95% HPD
Spoligotype	118	21	36.42	2-28	0-189	0.2994	76.51	0.81 - 823.21
MIRU-VNTR	121	32	37.35	2-11	0-48	0.2747	26.08	0 - 162.27
12 SNP cluster	242	47	74.69	2-34	0-23	0.6019	23.63	0 - 102.58
5 SNP cluster	147	40	45.37	2-27	0-10	0.3302	10.86	0 - 47.07
1 SNP cluster	74	29	22.84	2-6	0-2	0.1389	3.91	0 - 23.54

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

For each approach the inclusion of an isolate into a cluster is outlined in the surrounding circles using GraPhlAn[45]. 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 ML phylogenetic tree is rooted between L4 and L5 isolates.

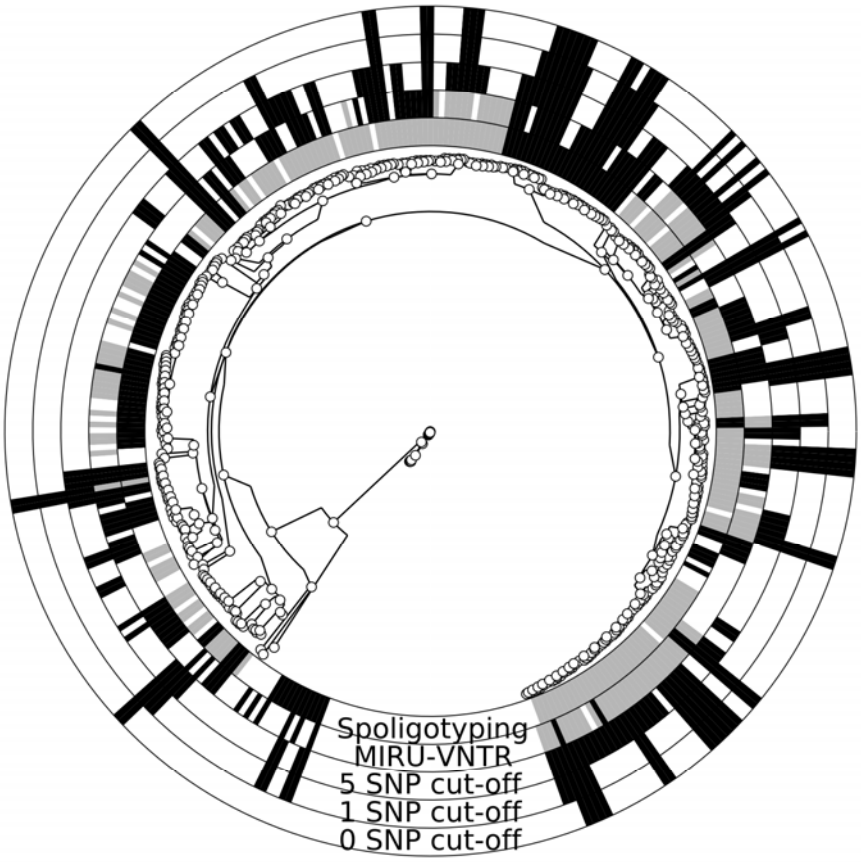
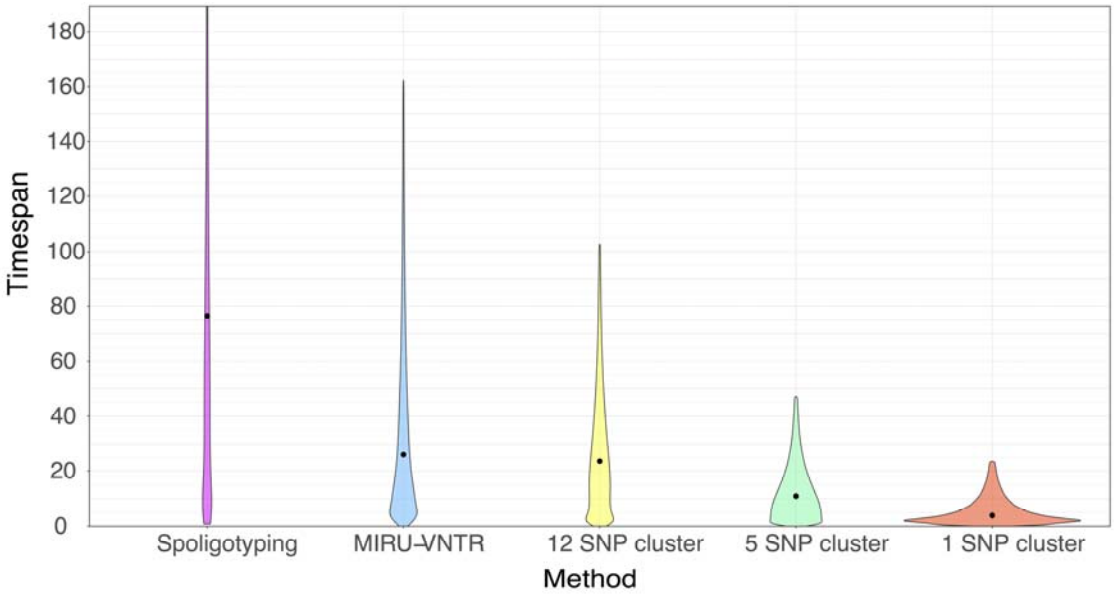


Figure 2: Timespans associated with transmission clusters

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). Note the y-axis is cut at 200 years. See Table 1 for full 95% HPD for each method.