# 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 (WHO, 2018). In addition, the prevalence of multidrug resistant Mtbc strains (MDR-TB) is increasing (WHO, 2018), predominantly through ongoing transmission within large populations (Kendall et al., 2015; Merker et al., 2015). The tracking and timing of recent transmission chains allows TB control programs to effectively pinpoint transmission hotspots and employ targeted intervention measures. This is especially important for the transmission of drug resistant strains as it appears that drug resistance may be transmitted more frequently than acquired

48 (Kendall et al., 2015; Trauer, Denholm, & McBryde, 2014). Thus, interrupting transmission is
49 key for the control of MDR-TB (Klopper et al., 2013; Merker et al., 2015; Shah et al., 2017).
50 For the development of the most effective control strategies, there is a strong need for (i)
51 appropriate identification of relevant transmission chains, risk factors and hotspots and (ii)
52 robust timing of when outbreaks first arose.
53
54 Epidemiological TB studies often apply genotyping methods to Mtbc strains to determine
55 whether two or more patients are linked within a transmission chain (molecular
56 epidemiology) (Merker, Kohl, Niemann, & Supply, 2017). Contact tracing is a primary
57 epidemiological method for investigating transmission networks of TB, mainly based on
58 patient interviews (Fox, Barry, Britton, & Marks, 2012). Although this method is often seen as
59 a gold standard of transmission linking, it does not always match the true transmission
60 patterns, even in low incidence settings (Behr et al., 1998; Diel et al., 2002; Roetzer et al.,
61 2011, 2013; Small et al., 1994) and misses many connections (Bjorn-Mortensen et al., 2017;
62 Vluggen et al., 2017). The implementation of molecular epidemiological approaches has
63 overcome these limitations and is often used as the main approach for cluster analysis.
64 Classical genotyping has involved IS*6110* DNA fingerprinting (Thierry et al., 1990; van Embden
65 et al., 1993), spoligotyping (Goguet de la Salmonière et al., 1997; Guernier, Sola, Brudey,
66 Guégan, & Rastogi, 2008; Kamerbeek et al., 1997), and variable-number tandem repeats of
67 mycobacterial interspersed repetitive units (MIRU-VNTR) (Supply, Magdalena, Himpens, &
68 Locht, 1997) which is the most common method at the moment (Merker et al., 2017). The
69 latter method is based on copy numbers of a sequence in tandem repeat patterns derived
70 from 24 distinct loci within the genome (Supply et al., 2006). If two patients have the same
71 classical genotyping pattern such as a 24-loci MIRU-VNTR pattern (or up to one locus

72 difference (Jonsson et al., 2014; Supply et al., 2006)) they are considered to be within a local
73 transmission chain. The combination of spoligotyping and MIRU-VNTR-typing, where patterns
74 must match in both methods to be considered a transmission link, is often considered the
75 molecular gold standard for transmission linking and genotyping (Supply et al., 2006).
76 However, examples of unlinked patients with identical patterns have been observed,
77 suggesting that this threshold covers too broad a genetic diversity and timespan between
78 infections (Gardy et al., 2011; Roetzer et al., 2013).
79
80 The application of (whole genome) sequence-based approaches for similarity analysis of Mtbc
81 isolates and cluster determination is known to have high discriminatory power when
82 assessing transmission dynamics (Comas, 2017; Roetzer et al., 2013; Walker et al., 2013, 2018;
83 Wyllie et al., 2018). Single nucleotide polymorphisms (SNPs) in the *pncA* gene are associated
84 with resistance to pyrazinamide (PZA) and can be used to improve the discriminatory power
85 of spoligotyping in a method referred to as SpoNC (Said et al., 2016). However, this is limited
86 by the low occurrence of PZA resistance, even in MDR-TB isolates (Kurbatova, Cavanaugh,
87 Dalton, S. Click, & Cegielski, 2013; Ngabonziza et al., 2017; Xu et al., 2016; Zignol et al., 2016).
88 The advent of widespread whole genome sequencing (WGS) capabilities has allowed for
89 highly discriminatory analyses of Mtbc strains either using core genome multi-locus sequence
90 typing (cgMLST) (Thomas A. Kohl et al., 2014) or SNP distances (Bjorn-Mortensen et al., 2016;
91 Gardy et al., 2011; Roetzer et al., 2013; Walker et al., 2013, 2018). WGS-based approaches
92 compare the genetic relatedness of the genomes of the clinical strains under consideration,
93 albeit usually excluding large repetitive portions of the genome, with the assumption that
94 highly similar strains are linked by a recent transmission event (Roetzer et al., 2013; Walker
95 et al., 2013). Although many SNP cut-offs for linking isolates have been proposed (Hatherell

96 et al., 2016), the most commonly employed is based on the finding that a 5 SNP cut-off will
97 cluster the genomes of strains from the majority of epidemiologically linked TB patients, with
98 an upper bound of 12 SNPs between any two linked isolates (Walker et al., 2013). The
99 widespread use of WGS has quickly pushed these cut-offs to be considered the new molecular
100 gold standard of recent transmission linking, although SNP distances may vary for technical
101 reasons (e.g. assembly pipelines or filter criteria (Guthrie & Gardy, 2017)) and between study
102 populations e.g. high and low incidence settings (Bjorn-Mortensen et al., 2016).
103
104 In addition to cluster detection, uncovering the timing of transmission events within a given
105 cluster is highly useful information for TB control e.g. for assessing the impact of interventions
106 on the spread of an outbreak. Accordingly, knowledge of the rate change associated with
107 different genotyping methods is essential for correct timing. The whole genome mutation
108 rate of Mtbc strains has been estimated by several studies as between $10^{-7}$ and $10^{-8}$
109 substitutions per site per year or ~0.3-0.5 SNPs per genome per year (Bryant et al., 2013;
110 Duchêne et al., 2016; Eldholm et al., 2015; Eldholm & Balloux, 2016; Roetzer et al., 2013;
111 Walker et al., 2013) while the rate of change in the MIRU-VNTR loci specifically is known to
112 be quicker (~$10^{-3}$) (Ragheb et al., 2013; Wirth et al., 2008). Since these mutation rates have
113 been shown to also vary by lineage (Duchêne et al., 2016; Ford et al., 2013) and over short
114 periods of time (Bryant et al., 2013), such variation needs to be accounted for, e.g. in Bayesian
115 phylogenetic dating techniques (Bryant et al., 2013; Merker et al., 2015; Wirth et al., 2008).
116
117 Considering the multiple genotyping methods currently available, many of them proposed as
118 a “gold standard”, there is an urgent need to precisely define the individual capacity of each
119 method to accurately detect recent transmission events and perform timing of outbreaks. To

120 provide this essential information, this study harnesses the power of WGS-based
121 phylogenetic dating methods to assign timespans onto Mtbc transmission chains
122 encompassed by the different genotypic clustering methods commonly used in TB
123 transmission studies.
124
125 **Results**
126 In this study, we assessed 20 different approaches for generating putative *M. tuberculosis*
127 transmission clusters (see methods for approaches and naming schemes) using a dataset of
128 324 phenotypically rifampicin resistant isolates collected 2005-2010 from retreatment cases
129 in Kinshasa, Democratic Republic of Congo (DRC). These 20 sets of clustering patterns were
130 then characterised using whole genome sequence data and the propensity for convergence
131 of clustering patterns was estimated (see methods). Bayesian phylodynamic approaches
132 implemented in BEAST-2 (Bouckaert et al., 2014) were then utilised to assign timespans to
133 the transmission events estimated by each genotyping method.
134
135 As expected, both the genome- and membrane-based spoligotyping approaches (named Gen-
136 Spo and Mem-Spo respectively), clustered the most strains, with the lowest resolution (i.e.
137 highest clustering rate) (Figure 1, Table 1). Convergent evolution (defined as the same pattern
138 observed in unrelated strains; see methods) was found to affect 39% (12) of Mem-Spo
139 clusters and 25% (7) of Gen-Spo clusters. Additionally, some discrepancies between the Mem-
140 Spo and Gen-Spo patterns of each isolate were observed, with 291 isolates (90%) having the
141 same pattern in both Mem-Spo and Gen-Spo approaches with 1 mismatch allowed
142 (Supplementary table 1). The remaining 33 isolates mismatched with 2 to 17 spacers (average
143 of 5 spacers). Although MIRU-VNTR performed far better than spoligotyping, 16% (6) of

144 clustering patterns were influenced by convergence in this study (see methods) (Table 1,
145 Figure 1). Mixed MIRU-VNTR patterns were observed in 18 isolates although this mixing was
146 not observed in the WGS data.
147
148 WGS-based methods had by far the highest discriminatory power and low SNP cut-offs
149 grouped isolates into smaller clusters (e.g. 2-10 isolates per cluster for a 5 SNP cut-off) (Table
150 1, Figure 1). When the clusters were expanded to better represent transmission chains using
151 the novel phylogenetic inclusion method implemented here (see methods), the resulting SNP
152 clusters often did not increase dramatically in size (Table 1). Discriminatory power and cluster
153 sizes based on cgMLST alleles were similar to the SNP-based clusters (Table 1, Figure 1).
154

155 Table 1: Clustering method overview.

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

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

158 estimates of clade mean heights.

**a)**

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

**b)**

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

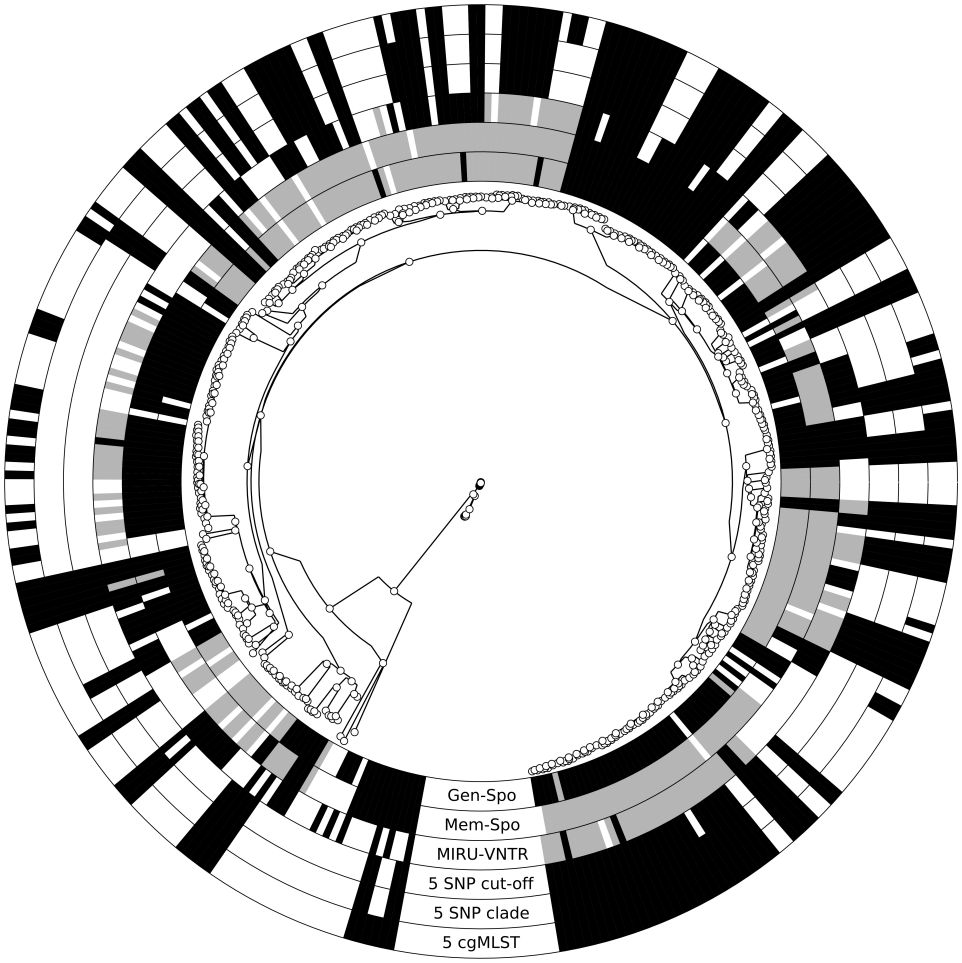
**c)**

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

159

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

161 For a representative approach of each of the main methods (Mem-Spo, Gen-Spo, MIRU-VNTR,
162 5 SNP cut-off, 5 SNP clade and 5 cgMLST) the inclusion of an isolate into a cluster is outlined
163 in the surrounding circles using GraPhlAn(Asnicar, Weingart, Tickle, Huttenhower, & Segata,
164 2015). If an isolate is in a cluster not affected by convergence, it is highlighted in black for the
165 given method. If an isolate is in a cluster affected by convergence, it is shown in grey. The
166 clustering based on all 20 approaches is shown in Supplementary Figure 1.




167

168 Statistical estimation of the timeframe associated with particular transmission chains showed
169 large differences in estimated cluster ages between the genotyping approaches used (Table
170 1, Figure 2), correlating well with the difference in discriminatory power. Cluster ages are
171 defined here as the most ancient transmission event that links any two isolates within a
172 specific cluster. Thus, in phylogenetic terms, the cluster age is the difference in time between
173 when the most recent common ancestor (MCRA) of the entire cluster existed and the date of
174 isolation of the furthest isolate from this ancestor. The aggregate mean ages of clusters
175 derived from spoligotyping approaches were found to often be several hundreds of years old
176 (Gen-Spo: 383 years ago (95% HPD: 1-1893); Mem-Spo: 141 years ago (95% HPD: 1-823))
177 (Table 1b, Figure 2a). The addition of MIRU-VNTR or *pncA* mutation data to spoligotyping
178 resulted in clusters that, on average, originated less than 100 years ago (Table 1b, Figure 2a).
179 MIRU-VNTR alone gave similar cluster ages as to when combined with spoligotyping (MIRU-
180 VNTR: 38 (0-162); GenSpo-MIRU: 64 (0-279); MemSpo-MIRU: 49 (1-216)) (Table 1b, Figure
181 2a).
182
183 Clusters based on SNP cut-offs correlated to 4 years of transmission using a 0 SNP cut-off (95%
184 HPD: 0-16), 6 years using a 1 SNP cut-off (95% HPD: 0-24), 13 years using a 5 SNP cut-off (95%
185 HPD: 0-47), and 29 years using a 12 SNP cut-off (95% HPD: 0-103) (Table 1c, Figure 2b).
186 Extension on the tree using the phylogenetic inclusion approach to form SNP clades did not
187 greatly increase the lengths of transmissions encompassed by clusters (one year increase, on
188 average) (Table 1c). Similar findings were obtained when clusters were based on allele
189 differences in the cgMLST method: 4 years of transmission using a 0 cgMLST cut-off (95% HPD:
190 0-15), 6 years using a 1 cgMLST cut-off (95% HPD: 0-25), 18 years using a 5 cgMLST cut-off

191 (95% HPD: 0-68), and 30 years using a 12 cgMLST cut-off (95% HPD: 0-112) (Table 1c, Figure

192 2b)

193 Figure 2: Timespans associated with transmission clusters

194 For each clustering method, the timespan associated with a cluster was estimated using

195 BEAST-2. The ages of each cluster (Y-axis) was aggregated per clustering method (X-axis).

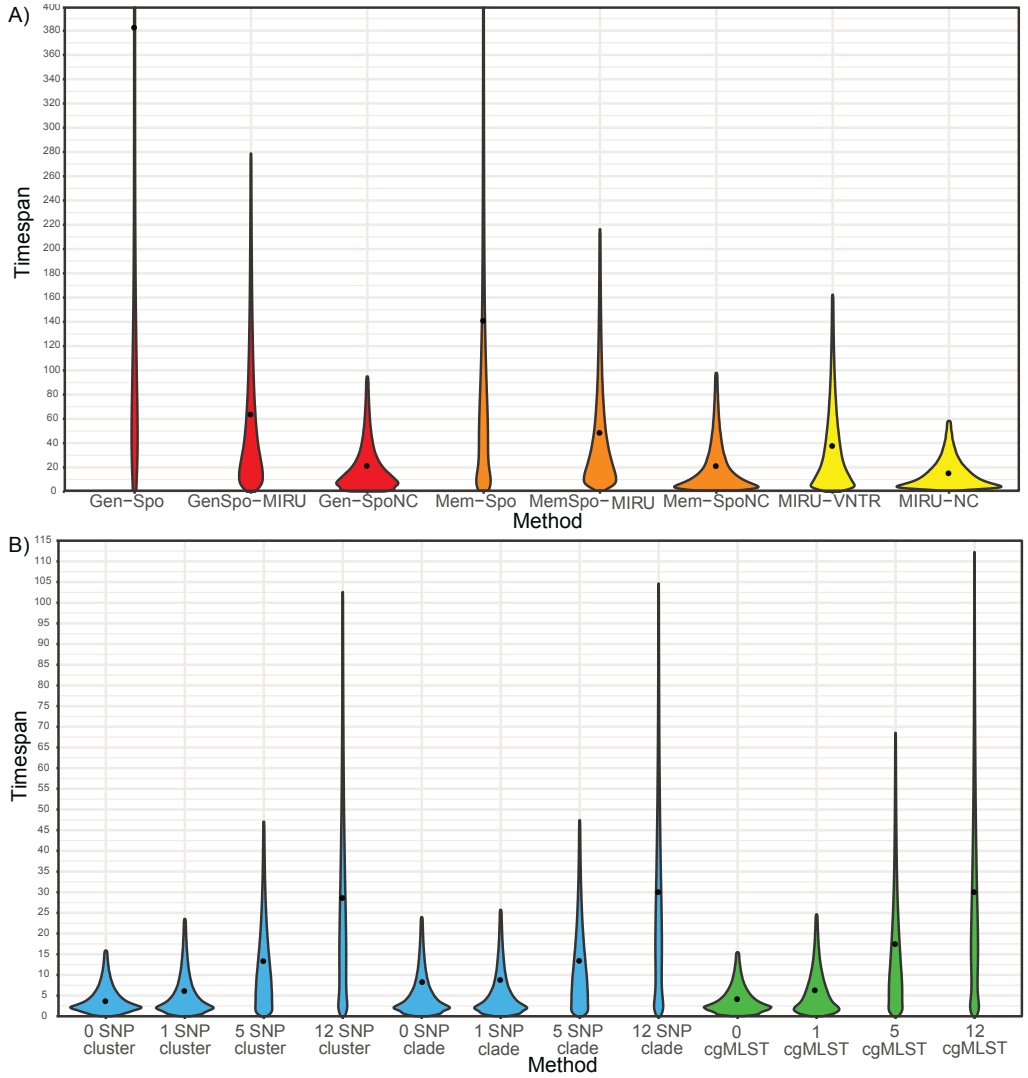
196 Violin plots show the mean (black dot) for timespans along with the proportion of clusters

197 with a given age (coloured kernel plots). Methods are split as follows: A) Spoligotype-based

198 (Gen-Spo-based (red), Mem-Spo-based (orange)) and MIRU-VNTR-based (yellow), B) SNP-

199 based (blue) and cgMLST-based (green). Note the y-axis is different for each and panel A) is

200 cut at 400 years.




201

202 **Discussion**

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

216

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

226 requiring more investigation into which method is closer to the 'true' spoligotyping pattern
227 within a genome (Coll et al., 2012; Mokrousov et al., 2016; Warren et al., 2002; Xia, Teo, &
228 Ong, 2016).
229
230 In line with previous findings (Comas et al., 2009; Scott et al., 2005), convergent evolution of
231 24-loci MIRU-VNTR patterns was rarer than observed for spoligotyping, but did occur in 16%
232 of MIRU-VNTR-based clusters. Additionally, the transmission times encompassed by MIRU-
233 VNTR clusters spanned several decades (Table 1b, Figure 2a), confirming previous studies
234 showing over-estimation of recent transmission with this method (Bjorn-Mortensen et al.,
235 2016; Roetzer et al., 2013; Stucki et al., 2016; Wyllie et al., 2018).
236
237 The combination of MIRU-VNTR or spoligotyping with *pncA* mutations (MIRU-NC and Gen-
238 SpoNC/Mem-SpoNC) appeared to reflect true clusters of PZA resistance transmission based
239 on the relatively young ages of such transmission clusters (Table 1b). Thus, as discussed
240 before (den Hertog, Sengstake, & Anthony, 2015; Sengstake et al., 2017), although
241 transmission of *pncA* mutations seems to occur, further investigation is needed to find out
242 whether *pncA* mutants are less transmissible than those with a wildtype gene.
243
244 For defining transmission events that occurred in more recent time frames before sampling,
245 WGS-based methods (SNP or cgMLST) were found to be better suited than classical
246 genotyping methods (Table 1, Figure 2). The 12 SNP cut-off, currently the recommended
247 upper bound for clustering isolates, likely defines transmission events that occurred on
248 average three decades prior to sampling, similar in age to clusters estimated by MIRU-VNTR.
249 This suggests that the 12 SNP cluster method may be a good replacement for MIRU-VNTR as

250 it detects larger transmission networks spanning similar transmission time periods but is less
251 affected by convergent evolution. Isolates clustered at identical (0 SNP) or nearly identical (1
252 SNP) cut-offs were found to represent transmission events occurring four to six years
253 previous. These findings correlate well with previous studies where confirmed contact
254 tracing-based epidemiological links were found between patients that were two (Walker et
255 al., 2014), three (Roetzer et al., 2013) or five (Walker et al., 2013) SNPs apart. Indeed, a recent
256 study of a cross-country MDR-TB outbreak found only a maximum of two SNP differences
257 between all 29 isolates involved in the origin of the outbreak (Walker et al., 2018). Although
258 this supports their use for detection or exclusion of very recent transmission, this low
259 variability between isolates makes robust identification of transmission direction impossible,
260 especially during short timespans.
261
262 Comparisons between the SNP-based (using almost all genomic differences) and the cgMLST-
263 based cluster detection (using a defined core set of genes) demonstrated that the latter
264 approach gives similar estimations to full SNP approaches. However, as current SNP assembly
265 pipelines for Illumina data exclude repetitive region such as PE/PPE genes, larger differences
266 between cgMLST and full SNP estimation may be seen once all aspects of the genome can be
267 utilised.
268
269 Different clustering approaches can be applied when grouping isolates by SNP distance. Two
270 partitional clustering methods are primarily utilised: either the creation of tight clusters
271 (where the maximum pairwise distance between isolates in a cluster is less than the SNP cut-
272 off; e.g. (Thomas A. Kohl et al., 2014)) or loose clusters (where each isolate is less than the
273 SNP cut-off distance from at least one other isolate in the cluster; e.g. (Walker et al., 2014)).

274 Tight clusters ensure high connectivity within clusters, but may result in isolates belonging to
275 multiple groups, making interpretation and delineation of transmission events difficult. Loose
276 clusters (the definition used in this study), separate isolates into non-overlapping clusters, but
277 may result in low connectivity within clusters. Here we present an extension of the loose
278 cluster, termed the phylogenetic inclusion method, which adds all other isolates with the
279 same phylogenetically defined common ancestor to the cluster, potentially identifying larger
280 circulating genotypes. Tight, loose and phylogenetic inclusion clusters each aim to define
281 different levels of connectivity through time, an aspect that should be considered when
282 selecting the appropriate clustering approach.
283
284 The mutation rate of *M. tuberculosis* has been estimated to be between $10^{-7}$ and $10^{-8}$
285 substitutions per site per year (Duchêne et al., 2016; Merker et al., 2015; Roetzer et al., 2013).
286 Within the Bayesian analysis employed here, the mutation rate was free to vary between
287 these values but was found to strongly favour ~$3x10^{-8}$ (ESS > 1000 for all runs), translating to
288 approximately 0.3 SNPs per genome per year. While the mutation rate used here is primarily
289 applicable for lineage 4 (which most of this dataset is comprised of) and in line with previous
290 estimates for this lineage (Duchêne et al., 2016), it may be similar in other lineages, although
291 this has only been shown for lineage 2 (Duchêne et al., 2016; Merker et al., 2015). Thus, per-
292 lineage estimates are required for all seven lineages to ensure similar transmission times are
293 linked to genotyping methods across the whole population diversity of the Mtbc.
294
295 While this study has many advantages due to its five year population based design in an
296 endemic setting coupled with the application of three different genotyping methods
297 (membrane based spoligotyping analysis, 24-locus MIRU-VNTR and WGS), future

298 confirmatory studies could address the following drawbacks that are inherent to genomic
299 epidemiology (Comas, 2017; Guthrie & Gardy, 2017): 1) studies employing contact tracing
300 and/or digital epidemiology (Salathé et al., 2012) in conjunction with these genotyping
301 methods can help confirm transmission times associated with different clusters; 2) as outlined
302 above, strains of other lineages of the Mtbc should be analysed in a similar fashion to ensure
303 transferability of findings across the entire complex; 3) a broad range of drug resistance
304 profiles should be included to fully assess the impact of such mutations on transmission
305 estimates; 4) improved WGS methods, such as directly from clinical samples to help reduce
306 culture biases (Sanoussi, Affolabi, Rigouts, Anagonou, & de Jong, 2017) and longer reads (e.g.
307 PacBio SMRT or Nanopore MinION) to capture the entire genome, including repetitive regions
308 such as PE/PPE genes known to impact genome remodelling (Ates et al., 2018; Phelan et al.,
309 2016), will ensure that the maximum diversity between isolates is captured and 5)
310 standardised SNP calling pipelines appropriate across all lineages, with high true positive/low
311 false negative rates, will ensure that Mtbc molecular epidemiology can be uniformly
312 implemented and comparable across studies.
313
314 In conclusion, since each method was found to represent different timespans and clustering
315 definitions, they can be used in a stratified manner in an integrated epidemiological and
316 public health investigation addressing the transmission of Mtbc strains. For instance,
317 although spoligotyping clusters represented potentially very old transmission events, the low
318 associated cost and its ability to be applied directly on sputum helps reduce culture bias and
319 thus robustly assign lineages. Thus, spoligotyping and/or MIRU-VNTR would serve well as
320 first-line surveillance of potential transmission events in the population, guiding further
321 investigations and resource allocations.

322

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

339

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

345 program for transmission, placing molecular epidemiological approaches as an integral part
346 in tracking and stopping the spread of TB.
347
348 **Materials and Methods**
349 Dataset and sequencing
350 A set of 324 isolates from Kinshasa, Democratic Republic of Congo were collected from
351 consecutive retreatment TB patients between 2005 and 2010 at TB clinics, servicing an
352 estimated 30% of the population of Kinshasa. All isolates were phenotypically resistant to
353 rifampicin (RR-TB) and the majority are also isoniazid resistant (i.e. MDR-TB). Use of the
354 stored isolates without any linked personal information was approved by the health
355 authorities of the DRC and the Institutional Review Board of the ITM in Antwerp (ref no
356 945/14). Libraries for whole genome sequencing were prepared from extracted genomic DNA
357 with the Illumina Nextera XT kit, and run on the Illumina NextSeq platform in a 2x151bp run
358 according to manufacturer's instructions. Illumina read sets will be available at ReSeqTB
359 (platform.reseqtb.org) upon publication.
360
361 Genome reconstruction and maximum likelihood phylogeny estimation
362 The MTBseq pipeline(Thomas Andreas Kohl et al., 2018) was used to detect the SNPs for each
363 isolate using the H37Rv reference genome (NCBI accession number NC000962.3) (Lew,
364 Kapopoulou, Jones, & Cole, 2011; Médigue, Cole, Camus, & Pryor, 2002). Sites known to be
365 involved in drug resistance (as outlined in the PhyResSE list of drug mutations v27 (Feuerriegel
366 et al., 2015)) were excluded from the alignment and additional filtering of sites with
367 ambiguous calls in >5% of isolates and those SNPs within a 12bp window of each other was
368 also applied.

369

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

377

378 Transmission cluster estimation methods

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

386

387 Spoligotyping

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

393 parameters and the 43 spacer values were extracted from the output. This method is referred
394 to as Gen-Spo. For both methods, isolates were said to be clustered if all 43 spacers matched.
395
396 MIRU-VNTR
397 Genotyping by MIRU-VNTR was undertaken as previously described (Supply et al., 2006). 2 µl
398 of DNA was extracted from cultures and amplified using the 24 loci MIRU-VNTR typing kit
399 (Genoscreen, Lille, France). Analysis of patterns was undertaken using the ABI 3500 automatic
400 sequencer (Applied Biosystems, California, USA) and Genemapper software (Applied
401 Biosystems). Isolates were said to be clustered if all 24 loci matched. MIRU-VNTR patterns
402 were also combined with spoligotyping patterns for additional refinement of clusters. Isolates
403 were clustered if both the spoligotyping pattern and the 24 loci MIRU-VNTR pattern matched.
404 These clustering methods are referred to as MemSpo-MIRU and GenSpo-MIRU.

405 SpoNC
406 Transmission estimation using spoligotyping has been shown to be improved if combined
407 with *pncA* mutations (Said et al., 2016). This method, referred to as SpoNC, was applied to
408 both Mem-Spo (Mem-SpoNC) and Gen-Spo (Gen-SpoNC). Mutations in *pncA* were extracted
409 from the MTBseq tabular output for each isolate. All mutations were selected, regardless of
410 drug resistance association, as is done in the SpoNC approach. The upstream promoter region
411 of *pncA* did not reveal any mutations in this dataset. Isolates were said to be clustered if all
412 43 spacers matched and the *pncA* mutation was the same in both isolates. MIRU-VNTR
413 patterns were combined with *pncA* mutations in a similar manner. This is referred to as MIRU-
414 NC.
415

416 SNP cut-off clustering

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

428

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

440 referred to as extended SNP clades to distinguish them from SNP clusters as created by the
441 standard non-phylogenetic method above. The python script that implements this method
442 can be found at https://github.com/conmeehan/pathophy.
443
444 cgMLST
445 An alternative approach to clustering using WGS data is the concept of core genome MLST
446 (cgMLST) patterns (Thomas A. Kohl et al., 2014). Since SNP detection can be variable between
447 assembly pipelines, SNP clusters between studies may be difficult to compare. The cgMLST
448 approach standardises comparisons by ensuring the same core genes are always compared.
449 BAM files for all isolates are input into Ridom SeqSphere$^{+}$ software (Ridom GmbH, Münster,
450 Germany) to compile an allelic distance matrix based on the cgMLST v2 scheme consisting of
451 2,891 core Mtbc genes. Loose clusters were then defined as above using allelic differences of
452 0, 1, 5 and 12 as cut-offs. These methods are referred to as 0/1/5/12 cgMLST respectively.
453
454 Detection of convergent evolution
455 Convergent evolution towards identical patterns may occur for Spoligotyping, MIRU-VNTR
456 and *pncA* mutations (Driscoll, 2009; Miotto et al., 2014; Scott et al., 2005; Warren et al., 2002).
457 Convergence was detected and cross-checked with two methods. Firstly, Mtbc lineage and
458 sub-lineage numbering (Coll et al., 2014) was applied to all isolates based on the PhyResSE
459 lineage-defining SNP list v27 (Feuerriegel et al., 2015). If the same clustering pattern was
460 observed in two different sub-lineages, with other patterns seen in-between, this was flagged
461 as potential convergence. Additional convergence confirmation was also undertaken using
462 phylogenetic distances, as estimated by DendroPy. If the phylogenetic distance (combined
463 branch lengths that separate 2 isolates) between two isolates with identical clustering

464 patterns was greater than 0.0005, this was flagged as potential convergence. Any isolates
465 flagged by both methods (lineage-based and distance-based) were marked as clustered by
466 convergence. For example, if isolates with the same spoligotyping pattern appeared in lineage
467 4,1 and 4,6 with different patterns in-between and these isolates were distant on the tree
468 (distance greater than 0.0005), this was confirmed as a convergent pattern. Convergence was
469 checked for all approaches except the SNP cut-off clusters/clades, which, by definition, could
470 not be convergent. Clustering methods that combined two other methods (e.g. Gen-SpoNC)
471 were first checked separately for convergence and then combined to create the final clusters.
472
473 Estimation of transmission times
474 To estimate the age and timespan of potential transmission clusters, SNP alignments were
475 created from the convergence-free version of the five primary clustering types: Gen-Spo,
476 Mem-Spo, MIRU-VNTR, extended 12 SNP clades and 12 allele cgMLST. All other methods are
477 sub-clustering methods of at least one of these five methods (e.g. Mem-SpoNC clusters are
478 inherently included in any Mem-Spo clusters, and all SNP-based clusters are sub-clusters of
479 the 12 SNP clades).
480
481 A Bayesian approach to transmission time estimation was then undertaken. The SNP
482 alignments were created as above for the five high-level clustering types. Each cluster method
483 alignment was separately input to BEAST-2 v2.4.7 (Bouckaert et al., 2014) to create a time
484 tree for those isolates. These phylogenies were built using the following priors: GTR+GAMMA
485 substitution model, a log-normal relaxed molecular clock model to account for variation in
486 mutation rates (Alexei J Drummond, Ho, Phillips, Rambaut, & Rambaut, 2006) and coalescent
487 constant size demographic model (A. J. Drummond, Rambaut, Shapiro, & Pybus, 2005), both

488 of which have been found to be suitable for lineage 4 isolates in a previous study (Bjorn-
489 Mortensen et al., 2016). The MCMC chain was run six times independently per alignment with
490 a length of at least 400 million, sampled every 40,000$^{th}$ step (Gen-Spo: 400 million; extended
491 12 SNP & cgMLST: 500 million; MIRU & MemSpo: 600 million). A log normal prior (mean
492 $1.5x10^{-7}$; variance 1.0) was used for the clock model to reflect the previously estimated
493 mutation rate of *M. tuberculosis* lineage 4 (Bryant et al., 2013; Duchêne et al., 2016; Eldholm
494 et al., 2015; Eldholm & Balloux, 2016; Roetzer et al., 2013; Walker et al., 2013), while allowing
495 for variation as previously suggested (Bryant et al., 2013). A 1/X non-informative prior was
496 selected for the population size parameter of the demographic model. Isolation dates were
497 used as informative heterochronous tip dates and the SNP alignment was augmented with a
498 count of invariant sites for each of the four nucleotide bases to avoid ascertainment
499 bias(Leaché et al., 2015). Tracer v1.6 (Rambaut, Suchard, Xie, & Drummond, 2013) was used
500 to determine adequate mixing and convergence of chains (ESS >150) after a 25% burn-in. The
501 chains were combined via LogCombiner v2.4.8 (Bouckaert et al., 2014) to obtain a single chain
502 for each clustering type with high (>1000) effective sample sizes. The tree samples were
503 combined in the same manner and resampled at a lower frequency to create thinned samples
504 of (minimum) 20,000 trees.
505
506 The timespan of transmission events estimated by each method was then calculated as
507 follows: for each cluster created by the given method, we defined the MRCA node as the
508 internal node that connects all taxa in that cluster. The youngest node was then defined as
509 the tip that is furthest from this MRCA within the clade (i.e. the tip descendant from that node
510 that was sampled closest to the present time). For each retained tree in the MCMC process,
511 the difference in age between the MRCA node and youngest node was calculated. This gave

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

519 **Acknowledgements/Funding**

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

529 **Competing interests**

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

531

532 **References**

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

536 Ates, L. S., Dippenaar, A., Ummels, R., Piersma, S. R., van der Woude, A. D., van der Kuij, K.,
537 … Bitter, W. (2018). Mutations in ppe38 block PE_PGRS secretion and increase
538 virulence of Mycobacterium tuberculosis. *Nature Microbiology*, *3*(2), 181–188.
539 https://doi.org/10.1038/s41564-017-0090-6
540 Behr, M. A., Hopewell, P. C., Antonio Paz, E., Masae Kawamura, L., Schecter, G. F., & Small,
541 P. M. (1998). Predictive Value of Contact Investigation for Identifying Recent
542 Transmission of Mycobacterium tuberculosis. *American Journal of Respiratory and*
543 *Critical Care Medicine*, *158*(2), 465–469.
544 https://doi.org/10.1164/ajrccm.158.2.9801062
545 Bjorn-Mortensen, K., Lillebaek, T., Koch, A., Soborg, B., Ladefoged, K., Sørensen, H. C. F., …
546 Andersen, A. B. (2017). Extent of transmission captured by contact tracing in a
547 tuberculosis high endemic setting. *European Respiratory Journal*, *49*(3). Retrieved from
548 http://erj.ersjournals.com/content/49/3/1601851
549 Bjorn-Mortensen, K., Soborg, B., Koch, A., Ladefoged, K., Merker, M., Lillebaek, T., … Kohl, T.
550 A. (2016). Tracing Mycobacterium tuberculosis transmission by whole genome
551 sequencing in a high incidence setting: a retrospective population-based study in East
552 Greenland. *Scientific Reports*, *6*, 33180. https://doi.org/10.1038/srep33180
553 Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., … Drummond, A. J.
554 (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS*
555 *Computational Biology*, *10*(4), e1003537.
556 https://doi.org/10.1371/journal.pcbi.1003537
557 Bryant, J. M., Schürch, A. C., van Deutekom, H., Harris, S. R., de Beer, J. L., de Jager, V., … van
558 Soolingen, D. (2013). Inferring patient to patient transmission of Mycobacterium
559 tuberculosis from whole genome sequencing data. *BMC Infectious Diseases*, *13*(1), 110.

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

Cabibbe, A. M., & Cirillo, D. M. (2016). Best approaches to drug-resistance surveillance at the country level. *International Journal of Mycobacteriology*, *5*, S40–S41. https://doi.org/10.1016/j.ijmyco.2016.09.010

Casali, N., Broda, A., Harris, S. R., Parkhill, J., Brown, T., & Drobniewski, F. (2016). Whole Genome Sequence Analysis of a Large Isoniazid-Resistant Tuberculosis Outbreak in London: A Retrospective Observational Study. *PLOS Medicine*, *13*(10), e1002137. https://doi.org/10.1371/journal.pmed.1002137

Coll, F., Mallard, K., Preston, M. D., Bentley, S., Parkhill, J., McNerney, R., … Clark, T. G. (2012). SpolPred: rapid and accurate prediction of Mycobacterium tuberculosis spoligotypes from short genomic sequences. *Bioinformatics (Oxford, England)*, *28*(22), 2991–3. https://doi.org/10.1093/bioinformatics/bts544

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

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

Comas, I., Homolka, S., Niemann, S., & Gagneux, S. (2009). Genotyping of Genetically Monomorphic Bacteria: DNA Sequencing in Mycobacterium tuberculosis Highlights the Limitations of Current Methodologies. *PLoS ONE*, *4*(11), e7815. https://doi.org/10.1371/journal.pone.0007815

den Hertog, A. L., Sengstake, S., & Anthony, R. M. (2015). Pyrazinamide resistance in Mycobacterium tuberculosis fails to bite? *Pathogens and Disease*, *73*(6), ftv037. https://doi.org/10.1093/femspd/ftv037

584 Diel, R., Schneider, S., Meywald-Walter, K., Ruf, C.-M., Rusch-Gerdes, S., & Niemann, S.
585 (2002). Epidemiology of Tuberculosis in Hamburg, Germany: Long-Term Population-
586 Based Analysis Applying Classical and Molecular Epidemiological Techniques. *Journal of*
587 *Clinical Microbiology*, *40*(2), 532–539. https://doi.org/10.1128/JCM.40.2.532-539.2002
588 Driscoll, J. R. (2009). Spoligotyping for Molecular Epidemiology of the Mycobacterium
589 tuberculosis Complex. In *Methods in molecular biology (Clifton, N.J.)* (Vol. 551, pp. 117–
590 128). https://doi.org/10.1007/978-1-60327-999-4_10
591 Drummond, A. J., Ho, S. Y. W., Phillips, M. J., Rambaut, A., & Rambaut, A. (2006). Relaxed
592 Phylogenetics and Dating with Confidence. *PLoS Biology*, *4*(5), e88.
593 https://doi.org/10.1371/journal.pbio.0040088
594 Drummond, A. J., Rambaut, A., Shapiro, B., & Pybus, O. G. (2005). Bayesian Coalescent
595 Inference of Past Population Dynamics from Molecular Sequences. *Molecular Biology*
596 *and Evolution*, *22*(5), 1185–1192. https://doi.org/10.1093/molbev/msi103
597 Duchêne, S., Holt, K. E., Weill, F.-X., Le Hello, S., Hawkey, J., Edwards, D. J., … Holmes, E. C.
598 (2016). Genome-scale rates of evolutionary change in bacteria. *Microbial Genomics*,
599 *2*(11), e000094. https://doi.org/10.1099/mgen.0.000094
600 Eldholm, V., & Balloux, F. (2016). Antimicrobial Resistance in Mycobacterium tuberculosis:
601 The Odd One Out. *Trends in Microbiology*. https://doi.org/10.1016/j.tim.2016.03.007
602 Eldholm, V., Monteserin, J., Rieux, A., Lopez, B., Sobkowiak, B., Ritacco, V., … Lemey, P.
603 (2015). Four decades of transmission of a multidrug-resistant Mycobacterium
604 tuberculosis outbreak strain. *Nature Communications*, *6*, 7119.
605 https://doi.org/10.1038/ncomms8119
606 Feuerriegel, S., Schleusener, V., Beckert, P., Kohl, T. A., Miotto, P., Cirillo, D. M., …
607 Fellenberg, K. (2015). PhyResSE: a Web Tool Delineating Mycobacterium tuberculosis

608 Antibiotic Resistance and Lineage from Whole-Genome Sequencing Data. *Journal of*
609 *Clinical Microbiology*, *53*(6), 1908–1914. https://doi.org/10.1128/JCM.00025-15
610 Filliol, I., Motiwala, A. S., Cavatore, M., Qi, W., Hazbón, M. H., Bobadilla del Valle, M., …
611 Alland, D. (2006). Global phylogeny of Mycobacterium tuberculosis based on single
612 nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution,
613 phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for
614 a minimal standard SNP set. *Journal of Bacteriology*, *188*(2), 759–72.
615 https://doi.org/10.1128/JB.188.2.759-772.2006
616 Ford, C. B., Shah, R. R., Maeda, M. K., Gagneux, S., Murray, M. B., Cohen, T., … Fortune, S. M.
617 (2013). Mycobacterium tuberculosis mutation rate estimates from different lineages
618 predict substantial differences in the emergence of drug-resistant tuberculosis. *Nature*
619 *Genetics*, *45*(7), 784–90. https://doi.org/10.1038/ng.2656
620 Fox, G. J., Barry, S. E., Britton, W. J., & Marks, G. B. (2012). Contact investigation for
621 tuberculosis: a systematic review and meta-analysis. *European Respiratory Journal*,
622 *41*(1). Retrieved from http://erj.ersjournals.com/content/41/1/140.short
623 Gardy, J. L., Johnston, J. C., Sui, S. J. H., Cook, V. J., Shah, L., Brodkin, E., … Tang, P. (2011).
624 Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak.
625 *New England Journal of Medicine*, *364*(8), 730–739.
626 https://doi.org/10.1056/NEJMoa1003176
627 Goguet de la Salmonière, Y. O., Li, H. M., Torrea, G., Bunschoten, A., van Embden, J., &
628 Gicquel, B. (1997). Evaluation of spoligotyping in a study of the transmission of
629 Mycobacterium tuberculosis. *Journal of Clinical Microbiology*, *35*(9), 2210–4. Retrieved
630 from http://www.ncbi.nlm.nih.gov/pubmed/9276389
631 Guernier, V., Sola, C., Brudey, K., Guégan, J.-F., & Rastogi, N. (2008). Use of cluster-graphs

632 from spoligotyping data to study genotype similarities and a comparison of three
633 indices to quantify recent tuberculosis transmission among culture positive cases in
634 French Guiana during a eight year period. *BMC Infectious Diseases*, *8*(1), 46.
635 https://doi.org/10.1186/1471-2334-8-46
636 Guthrie, J. L., & Gardy, J. L. (2017). A brief primer on genomic epidemiology: lessons learned
637 from *Mycobacterium tuberculosis*. *Annals of the New York Academy of Sciences*,
638 *1388*(1), 59–77. https://doi.org/10.1111/nyas.13273
639 Guthrie, J. L., Kong, C., Roth, D., Jorgensen, D., Rodrigues, M., Hoang, L., … Gardy, J. L.
640 (2018). Molecular Epidemiology of Tuberculosis in British Columbia, Canada: A 10-Year
641 Retrospective Study. *Clinical Infectious Diseases*, *66*(6), 849–856.
642 https://doi.org/10.1093/cid/cix906
643 Hatherell, H.-A., Colijn, C., Stagg, H. R., Jackson, C., Winter, J. R., & Abubakar, I. (2016).
644 Interpreting whole genome sequencing for investigating tuberculosis transmission: a
645 systematic review. *BMC Medicine*, *14*(1), 21. https://doi.org/10.1186/s12916-016-
646 0566-x
647 Jonsson, J., Hoffner, S., Berggren, I., Bruchfeld, J., Ghebremichael, S., Pennhag, A., &
648 Groenheit, R. (2014). Comparison between RFLP and MIRU-VNTR Genotyping of
649 Mycobacterium tuberculosis Strains Isolated in Stockholm 2009 to 2011. *PLoS ONE*,
650 *9*(4), e95159. https://doi.org/10.1371/journal.pone.0095159
651 Kamerbeek, J., Schouls, L., Kolk, A., van Agterveld, M., van Soolingen, D., Kuijper, S., … van
652 Embden, J. (1997). Simultaneous detection and strain differentiation of Mycobacterium
653 tuberculosis for diagnosis and epidemiology. *Journal of Clinical Microbiology*, *35*(4),
654 907–14. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9157152
655 Kato-Maeda, M., Gagneux, S., Flores, L. L., Kim, E. Y., Small, P. M., Desmond, E. P., &

656 Hopewell, P. C. (2011). Strain classification of Mycobacterium tuberculosis: congruence
657 between large sequence polymorphisms and spoligotypes. *The International Journal of*
658 *Tuberculosis and Lung Disease : The Official Journal of the International Union against*
659 *Tuberculosis and Lung Disease*, *15*(1), 131–3. Retrieved from
660 http://www.ncbi.nlm.nih.gov/pubmed/21276309
661 Kendall, E. A., Fofana, M. O., Dowdy, D. D. W., WHO, Dye, C., Garnett, G., … al., et. (2015).
662 Burden of transmitted multidrug resistance in epidemics of tuberculosis: a transmission
663 modelling analysis. *The Lancet Respiratory Medicine*, *3*(12), 963–972.
664 https://doi.org/10.1016/S2213-2600(15)00458-0
665 Klopper, M., Warren, R. M., Hayes, C., Gey van Pittius, N. C., Streicher, E. M., Müller, B., …
666 Trollip, A. P. (2013). Emergence and Spread of Extensively and Totally Drug-Resistant
667 Tuberculosis, South Africa. *Emerging Infectious Diseases*, *19*(3), 449–455.
668 https://doi.org/10.3201/eid1903.120246
669 Kobert, K., Stamatakis, A., & Flouri, T. (2016). Efficient Detection of Repeating Sites to
670 Accelerate Phylogenetic Likelihood Calculations. *Systematic Biology*, *66*(2), syw075.
671 https://doi.org/10.1093/sysbio/syw075
672 Kohl, T. A., Diel, R., Harmsen, D., Rothgänger, J., Walter, K. M., Merker, M., … Niemann, S.
673 (2014). Whole-genome-based Mycobacterium tuberculosis surveillance: a
674 standardized, portable, and expandable approach. *Journal of Clinical Microbiology*,
675 *52*(7), 2479–86. https://doi.org/10.1128/JCM.00567-14
676 Kohl, T. A., Utpatel, C., Schleusener, V., De Filippo, M. R., Beckert, P., Cirillo, D. M., &
677 Niemann, S. (2018). MTBseq: A comprehensive pipeline for whole genome sequence
678 analysis of Mycobacterium tuberculosis complex isolates. *In Preparation*.
679 Kozlov, A. (2017). RAxML-NG.

https://doi.org/https://zenodo.org/record/888146#.Wm8r2Uso9TY

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

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

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

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

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

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

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

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

704 determinants: a multicenter study. *MBio*, *5*(5), e01819-14.
705 https://doi.org/10.1128/mBio.01819-14
706 Mokrousov, I., Chernyaeva, E., Vyazovaya, A., Sinkov, V., Zhuravlev, V., & Narvskaya, O.
707 (2016). Next-Generation Sequencing of Mycobacterium tuberculosis. *Emerging*
708 *Infectious Diseases*, *22*(6), 1127–9. https://doi.org/10.3201/eid2206.152051
709 Ngabonziza, J. C. S., Diallo, A. B., Tagliani, E., Diarra, B., Kadanga, A. E., Togo, A. C. G., …
710 group", the "Union short M.-T. regimen study. (2017). Half of rifampicin-resistant
711 Mycobacterium tuberculosis complex isolated from tuberculosis patients in Sub-
712 Saharan Africa have concomitant resistance to pyrazinamide. *PLOS ONE*, *12*(10),
713 e0187211. https://doi.org/10.1371/journal.pone.0187211
714 Phelan, J. E., Coll, F., Bergval, I., Anthony, R. M., Warren, R., Sampson, S. L., … Clark, T. G.
715 (2016). Recombination in pe/ppe genes contributes to genetic variation in
716 Mycobacterium tuberculosis lineages. *BMC Genomics*, *17*, 151.
717 https://doi.org/10.1186/s12864-016-2467-y
718 R Core Team. (2017). R: A language and environment for statistical computing. R Foundation
719 for Statistical Computing, Vienna, Austria. Retrieved from http://www.r-project.org/
720 Ragheb, M. N., Ford, C. B., Chase, M. R., Lin, P. L., Flynn, J. L., & Fortune, S. M. (2013). The
721 mutation rate of mycobacterial repetitive unit loci in strains of M. tuberculosis from
722 cynomolgus macaque infection. *BMC Genomics*, *14*(1), 145.
723 https://doi.org/10.1186/1471-2164-14-145
724 Rambaut, A. (2016). Figtree. Retrieved May 3, 2017, from
725 http://tree.bio.ed.ac.uk/software/figtree/
726 Rambaut, A., Suchard, M. A., Xie, D., & Drummond, A. (2013). Tracer. Retrieved December
727 11, 2013, from http://beast.bio.ed.ac.uk/Tracer

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

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

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

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

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

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

752 Sengstake, S., Bergval, I. L., Schuitema, A. R., Beer, J. L., Phelan, J., Zwaan, R., … Anthony, R.
753 M. (2017). Pyrazinamide resistance-conferring mutations in pncA and the transmission
754 of multidrug resistant TB in Georgia. *BMC Infectious Diseases 2017 17:1*, *17*(1), 491.
755 https://doi.org/10.1186/s12879-017-2594-3
756 Shah, N. S., Auld, S. C., Brust, J. C. M., Mathema, B., Ismail, N., Moodley, P., … Gandhi, N. R.
757 (2017). Transmission of Extensively Drug-Resistant Tuberculosis in South Africa. *New*
758 *England Journal of Medicine*, *376*(3), 243–253.
759 https://doi.org/10.1056/NEJMoa1604544
760 Small, P. M., Hopewell, P. C., Singh, S. P., Paz, A., Parsonnet, J., Ruston, D. C., … Schoolnik, G.
761 K. (1994). The Epidemiology of Tuberculosis in San Francisco -- A Population-Based
762 Study Using Conventional and Molecular Methods. *New England Journal of Medicine*,
763 *330*(24), 1703–1709. https://doi.org/10.1056/NEJM199406163302402
764 Statisticat. (2016). LaplacesDemon: Complete Environment for Bayesian Inference.
765 Bayesian-Inference.com. R package version 16.0.1. Retrieved from
766 https://web.archive.org/web/20150206004624/http://www.bayesian-
767 inference.com/software
768 Stucki, D., Ballif, M., Egger, M., Furrer, H., Altpeter, E., Battegay, M., … Fenner, L. (2016).
769 Standard Genotyping Overestimates Transmission of Mycobacterium tuberculosis
770 among Immigrants in a Low-Incidence Country. *Journal of Clinical Microbiology*, *54*(7),
771 1862–70. https://doi.org/10.1128/JCM.00126-16
772 Sukumaran, J., & Holder, M. T. (2010). DendroPy: a Python library for phylogenetic
773 computing. *Bioinformatics*, *26*(12), 1569–1571.
774 https://doi.org/10.1093/bioinformatics/btq228
775 Supply, P., Allix, C., Lesjean, S., Cardoso-Oelemann, M., Rusch-Gerdes, S., Willery, E., … van

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

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

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

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

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

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

800 Vluggen, C., Soetaert, K., Groenen, G., Wanlin, M., Spitaels, M., Arrazola de Oñate, W., …
801 Mathys, V. (2017). Molecular epidemiology of Mycobacterium tuberculosis complex in
802 Brussels, 2010–2013. *PLOS ONE*, *12*(2), e0172554.
803 https://doi.org/10.1371/journal.pone.0172554
804 Walker, T. M., Ip, C. L. C., Harrell, R. H., Evans, J. T., Kapatai, G., Dedicoat, M. J., … Peto, T. E.
805 A. (2013). Whole-genome sequencing to delineate Mycobacterium tuberculosis
806 outbreaks: a retrospective observational study. *The Lancet Infectious Diseases*, *13*(2),
807 137–46. https://doi.org/10.1016/S1473-3099(12)70277-3
808 Walker, T. M., Lalor, M. K., Broda, A., Ortega, L. S., Morgan, M., Parker, L., … Conlon, C. P.
809 (2014). Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK,
810 2007-12, with whole pathogen genome sequences: an observational study. *The Lancet.*
811 *Respiratory Medicine*, *2*(4), 285–292. https://doi.org/10.1016/S2213-2600(14)70027-X
812 Walker, T. M., Merker, M., Knoblauch, A. M., Helbling, P., Schoch, O. D., van der Werf, M. J.,
813 … Witschi, M. (2018). A cluster of multidrug-resistant Mycobacterium tuberculosis
814 among patients arriving in Europe from the Horn of Africa: a molecular epidemiological
815 study. *The Lancet Infectious Diseases*. https://doi.org/10.1016/S1473-3099(18)30004-5
816 Warren, R. M., Streicher, E. M., Sampson, S. L., van der Spuy, G. D., Richardson, M., Nguyen,
817 D., … van Helden, P. D. (2002). Microevolution of the direct repeat region of
818 Mycobacterium tuberculosis: implications for interpretation of spoligotyping data.
819 *Journal of Clinical Microbiology*, *40*(12), 4457–65. Retrieved from
820 http://www.ncbi.nlm.nih.gov/pubmed/12454136
821 WHO. (2015). *Guidelines for surveillance of drug resistance in tuberculosis* (5th ed.). Geneva.
822 WHO. (2018). *Global tuberculosis report 2017*. Retrieved from
823 http://www.who.int/tb/publications/global_report/archive/en/

824 Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
825 Retrieved from http://ggplot2.org
826 Wirth, T., Hildebrand, F., Allix-Béguec, C., Wölbeling, F., Kubica, T., Kremer, K., … Niemann,
827 S. (2008). Origin, Spread and Demography of the Mycobacterium tuberculosis Complex.
828 *PLoS Pathogens*, *4*(9), e1000160. https://doi.org/10.1371/journal.ppat.1000160
829 Witney, A. A., Cosgrove, C. A., Arnold, A., Hinds, J., Stoker, N. G., & Butcher, P. D. (2016).
830 Clinical use of whole genome sequencing for Mycobacterium tuberculosis. *BMC*
831 *Medicine*, *14*(1), 46. https://doi.org/10.1186/s12916-016-0598-2
832 Wyllie, D., Davidson, J., Walker, T., Rathod, P., Peto, T., Robinson, E., … Campbell, C. (2018).
833 A quantitative evaluation of MIRU-VNTR typing against whole-genome sequencing for
834 identifying Mycobacterium tuberculosis transmission: A prospective observational
835 cohort study. *BioRxiv*, 252734. https://doi.org/10.1101/252734
836 Xia, E., Teo, Y.-Y., & Ong, R. T.-H. (2016). SpoTyping: fast and accurate in silico
837 Mycobacterium spoligotyping from sequence reads. *Genome Medicine*, *8*(1), 19.
838 https://doi.org/10.1186/s13073-016-0270-7
839 Xu, P., Wu, J., Yang, C., Luo, T., Shen, X., Zhang, Y., … Gao, Q. (2016). Prevalence and
840 transmission of pyrazinamide resistant Mycobacterium tuberculosis in China.
841 *Tuberculosis*, *98*, 56–61. https://doi.org/10.1016/j.tube.2016.02.008
842 Zignol, M., Cabibbe, A. M., Dean, A. S., Glaziou, P., Alikhanova, N., Ama, C., … Raviglione, M.
843 C. (2018). Genetic sequencing for surveillance of drug resistance in tuberculosis in
844 highly endemic countries: a multi-country population-based surveillance study. *The*
845 *Lancet Infectious Diseases*. https://doi.org/10.1016/S1473-3099(18)30073-2
846 Zignol, M., Dean, A. S., Alikhanova, N., Andres, S., Cabibbe, A. M. A., Cirillo, D. M., … Valafar,
847 F. (2016). Population-based resistance of Mycobacterium tuberculosis isolates to

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

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

850

851 Supplementary legends

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

853 methods

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

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

856 mismatches between patterns recorded.

857

| Isolate | Mem-Spoligotype | Gen-Spoligotype | Mismatches |
|---|---|---|---|
| DRC-052577 | 111111100000111111111101000000000000010001111 | 111111100000111111111101000000000000110001111 | 1 |
| DRC-052578 | 111111100000000000000000011111100001111111 | 111111000000000000000000011111100001111111 | 0 |
| DRC-052654 | 111111111111111111100000111111100001111111 | 110111111111111111110111111101100001101111 | 7 |
| DRC-052667 | 111111110001111111100000000000000000000001100 | 111111111110000111000000000000000000111100 | 9 |
| DRC-052750 | 111111100000111111111101100000000000010001111 | 111111100000111111111101000000000000110001111 | 3 |
| DRC-052956 | 111111111111111111111111111111111100001110111 | 111111111111111111111111111111111100001110111 | 0 |
| DRC-052958 | 111111111111111111111111111101111100001110111 | 111111111111111111111111111101111100001110111 | 0 |
| DRC-052959 | 111111111111111111111111111111111100001111111 | 111111111111111111111111111111111000001111111 | 1 |
| DRC-052964 | 111111110001111111100000000000000000000011111 | 111111110001111111100000000000000000000011111 | 0 |
| DRC-060116 | 111111111111111111111111111111111100001110111 | 111111111111111111111111111111111100001110111 | 0 |
| DRC-060118 | 111111111111111111111111111111111100001110111 | 111111111111111111111111111111111100001110111 | 0 |
| DRC-060120 | 111111111111111111111111111111111100001110111 | 111111111111111111111111111111111100001110111 | 0 |
| DRC-060131 | 110111111111111111110000111111100001111111 | 110111111111111111110000111111100001111111 | 0 |
| DRC-060237 | 111111111111111111111111111111100001111111 | 111111111111111111111111111111100001111111 | 0 |
| DRC-060244 | 000000000000000000000000011111100001111111 | 000000000000000000000000111111100001110111 | 3 |
| DRC-060248 | 111111111111111111111111111111111100001110111 | 111111111111111111111111111111111100001110111 | 0 |
| DRC-060253 | 111111111111111111111111111111111100001110111 | 111111111111111111111111111111111100001110111 | 0 |
| DRC-060260 | 111111000000000000000000011111100001011111 | 111111000000000000000000011111100101111111 | 2 |
| DRC-060360 | 111110100000111111111111000000000000010000011 | 111111100000111111111111000000000000110000011 | 2 |
| DRC-060361 | 110111111111111111110000111111100001111111 | 110111111111111111110000111111100001111111 | 0 |
| DRC-060362 | 111111100000111111111101100000000000010001111 | 111111100000111111111101000000000000110001111 | 2 |
| DRC-060366 | 111111000000000000000000011111100001011111 | 111111000000000000000000011111100001111111 | 1 |
| DRC-060375 | 110000001111111011110000111111100001111111 | 110000001111111011110000111111100001111111 | 2 |
| DRC-060376 | 110111100011111111110000111111100000011111 | 110111111111111111110000111111100001111111 | 5 |

| | | | |
|---|---|---|---|
| DRC-060381 | 1111111100011111111000000000000000000011111 | 1111111100011111111000000000000000000011111 | 0 |
| DRC-060382 | 1111110000000000000000000111111100001111111 | 1111110000000000000000000111111100001111111 | 0 |
| DRC-060514 | 1100000011111111011100001111111100001111111 | 1100000011111111011100001111111100001111111 | 0 |
| DRC-060526 | 1111111100011111111000000000000000000011111 | 1111111100011111111000000000000000000011111 | 0 |
| DRC-060527 | 1111111111110000111100001111111100001011111 | 1111111111110000111100001111111100001011111 | 0 |
| DRC-060528 | 1111110000000000000000000111111100001011111 | 1111110000000000000000000111111100001111111 | 1 |
| DRC-060549 | 1000001111111111111011111111111100001110111 | 1000001111111111111111111111111100001111111 | 2 |
| DRC-060562 | 1111111111111111111111111111111100001000011 | 1111111111111111111111111111111100001000011 | 0 |
| DRC-060565 | 1111111111111111111111111111111100001110111 | 1111111111111111111111111111111100001110111 | 0 |
| DRC-060756 | 1111110000000000000000000111111100001111111 | 1111110000000000000000000111111100001111111 | 0 |
| DRC-060757 | 1111111111111111111111111110111100001110111 | 1111111111111111111111111110111100001110111 | 0 |
| DRC-060772 | 1111111000001111111111010000000000010001111 | 1111111000001111111111010000000000110001111 | 1 |
| DRC-060778 | 1111111100011111111000000000000000000011111 | 1111111100011111111000000000000000000011111 | 0 |
| DRC-060856 | 1111111111110000111000000000000000000111100 | 1111111111110000111000000000000000000111100 | 0 |
| DRC-060858 | 1111111111111111111000011111111100001011111 | 1111111111111111111111111110111100001110111 | 7 |
| DRC-060859 | 1111111111111111111000011111111100001111111 | 1111111111111111111000011111111100001111111 | 0 |
| DRC-061069 | 1111111000001111111111100000000000010001111 | 1111111000001111111111100000000000110001111 | 1 |
| DRC-061123 | 1111111111111111111000011111111100001111111 | 1111111111111111111000011111111100001111111 | 0 |
| DRC-061124 | 1101111111111111111000011111111100001111111 | 1111111111111111111000011111111100001111111 | 1 |
| DRC-061127 | 1111110000000000000000000111111100001111111 | 1111110000000000000000000111111100001111111 | 0 |
| DRC-061136 | 1111110000000000000000000111111100001011111 | 1111110000000000000000000111111100001111111 | 1 |
| DRC-061137 | 1111110000000000000000000111111100001011111 | 1111110000000000000000000111111100001111111 | 1 |
| DRC-061138 | 1111111100011111111000000000000000000011111 | 1111111100011111111000000000000000000011111 | 0 |
| DRC-061142 | 1111111111111111111000011111111100001111111 | 1111111111111111111000011111111100001111111 | 0 |
| DRC-061431 | 1111111100011111111000000000000000000011111 | 1111111100011111111000000000000000000011111 | 0 |
| DRC-061493 | 1111111111111111111111111111111100001110111 | 1111111110111111111111111111111100001110111 | 1 |
| DRC-062295 | 1111111111111100000111111111111100001110111 | 1111111111111100000111111111111100001110111 | 0 |
| DRC-062298 | 1111111111110000111000000000000000000111100 | 1111111111110000111000000000000000000111100 | 0 |
| DRC-062303 | 1100000011111110111100001111111100001111111 | 1100000011111110111100001111111100001111111 | 2 |
| DRC-062444 | 1111111100011111111000000000000000000011111 | 1111111100011111111000000000000000000011111 | 0 |
| DRC-062446 | 1111111111111111111000011110111100001111111 | 1111111111111111111000011110111100001111111 | 0 |
| DRC-062447 | 1100000011111110111100001111111100001111111 | 1100000011111110111100001111111100001111111 | 0 |
| DRC-062495 | 1111111111111111111111111111111100001111111 | 1111111111111111111111111111111100001111111 | 0 |
| DRC-062502 | 1111111111111111111000011111111100001011111 | 1111111111111111111000011111111100001011111 | 0 |
| DRC-062503 | 1111111111111111111111111111111100001111111 | 1111111111111111111111111111111100001110111 | 1 |

| | | | |
|---|---|---|---|
| DRC-062822 | 1101111111111111111100001111111100001111111 | 1101111111111111111100001111111100001111111 | 0 |
| DRC-062827 | 1101111111111111111100001111111100001111111 | 1101111111111111111100001111111100001111111 | 0 |
| DRC-062840 | 1111111111111111111111111111111100001111111 | 1111111111111111111111111111111100001111111 | 0 |
| DRC-062841 | 1101111111111111111100000011011100001111111 | 1111111111111111111100000011011100001111111 | 2 |
| DRC-062844 | 1101111111111111111100001111111100001111111 | 1101111111111111111100001111111100001111111 | 0 |
| DRC-062849 | 1111111111011111111100001111111100001111111 | 1111111111011111111100001111111100001111111 | 0 |
| DRC-063522 | 1111111111111111111111111111111100001110111 | 1111111111111111111111111111111100001110111 | 0 |
| DRC-063540 | 1111111100011111111000000000000000000011111 | 1111111100011111111000000000000000000011111 | 0 |
| DRC-063541 | 1111111000000000000000000111111100001111111 | 1111111000000000000000000111111100001111111 | 0 |
| DRC-063545 | 1111111000001111111111010000000000010001111 | 1111111000001111111111010000000000110001111 | 1 |
| DRC-063556 | 1111111100011111111000000000000000000011111 | 1111111100011111111000000000000000000011111 | 0 |
| DRC-063557 | 1111111000000000000000000111111100001111111 | 1111111000000000000000000111111100001111111 | 0 |
| DRC-063645 | 1111111111111111111100001111111100001111111 | 1111111111110111111100000011011100001111111 | 3 |
| DRC-063667 | 1101111111111111111100001111111100001111111 | 1101111111111111111100001111111100001111111 | 0 |
| DRC-063803 | 1111111111111111111100000011011100001111111 | 1111111111111111111100000011011100001111111 | 1 |
| DRC-063887 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-063888 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-063889 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-063896 | 1111111111111111111111111111111100001111111 | 1111111111111111111111111111111100001111111 | 0 |
| DRC-063898 | 1111111111111111111100001110111100001111111 | 1111111111111111111100001110111100001111111 | 0 |
| DRC-064081 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 1 |
| DRC-064085 | 1111111111111111111111111111111100001110111 | 1111111111111111111111111111111100001110111 | 0 |
| DRC-064090 | 1111110000000000000000001111111100001111111 | 1111110000000000000000001111111100001111111 | 0 |
| DRC-064101 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-064208 | 1111111111111111111111111111111100001110111 | 1111111111111111111111111111111100001110111 | 0 |
| DRC-070012 | 1011111111010111111111111111111100001110111 | 1011111111110111111111111111111100001110111 | 1 |
| DRC-070038 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-070043 | 1110000111111111111110000000000000111111111 | 1110000111111111111110000000000000111111111 | 0 |
| DRC-070049 | 1111111100011111111000000000000000000011111 | 1111111100011111111000000000000000000011111 | 0 |
| DRC-070777 | 1111111111111111111100001110111100001111111 | 1111111111111111111100001110111100001111111 | 0 |
| DRC-070779 | 1111110000000000000000001111111100001111111 | 1111110000000000000000001111111100001111111 | 0 |
| DRC-070781 | 1111111111110111111111111111111100001110111 | 1111111111110111111111111111111100001110111 | 0 |
| DRC-070783 | 1111111111111111111100000000000000001111111 | 1111111111011111111100000000000000001111111 | 1 |
| DRC-070786 | 1111111000001111111111010000000000010001111 | 1111111000001111111111010000000000110001111 | 1 |
| DRC-070788 | 1111000011110000111111111111111100001111111 | 1111000011110000111111111111111100001111111 | 0 |

| | | | |
|---|---|---|---|
| DRC-070790 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-070792 | 1111111111111111111111100000001100001111111 | 1111111111111111111111100000001100001111111 | 0 |
| DRC-070794 | 1000001111111111111111111111111100001111111 | 1000001111111111111111111111111100001111111 | 0 |
| DRC-070804 | 1111111111111111111100001111111100001011111 | 1111111111111111111100001111111100001011111 | 0 |
| DRC-070807 | 1111111111111111111111111101111100001110111 | 1111111111111111111111111101111100001110111 | 0 |
| DRC-070808 | 1111111111111111111111111111111100001110111 | 1111111111111111111111111111111100001110111 | 0 |
| DRC-070809 | 1101111111111111111100001111111100001111111 | 1101111111111111111100001111111100001111111 | 0 |
| DRC-071052 | 1111110000000000000000001111111100001111111 | 1111110000000000000000001111111100001111111 | 0 |
| DRC-071057 | 1111111111111111111111111111111100001110111 | 1111111111111111111111111111111100001110111 | 0 |
| DRC-071074 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-071075 | 1111110000000000000000001111111100001111111 | 1111110000000000000000001111111100001111111 | 0 |
| DRC-071076 | 1111111000001111111110100000000000010001111 | 1111111000001111111110100000000000110001111 | 1 |
| DRC-071458 | 1111111111111111111111111111111100001110111 | 1111111111111111111111111111111100001110111 | 0 |
| DRC-071460 | 1111111111111111111100011111111100001110111 | 1111111111111111111100011111111100001110111 | 0 |
| DRC-071472 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-071479 | 1111111111111111111111111111110100001111111 | 1111111111111111111111111111111100001111111 | 1 |
| DRC-071686 | 1111111111111111111111111111111100001110111 | 1111111111111111111111111111111100001110111 | 0 |
| DRC-071691 | 1101111111110111111100001111111100001111111 | 1111111111101111111100001111111100001111111 | 3 |
| DRC-071696 | 1111111111101111111111111111111100001110111 | 1111111111101111111111111111111100001110111 | 0 |
| DRC-071697 | 1111111111111111111111000111110100001111111 | 1111111111111111111111000111111100001111111 | 1 |
| DRC-071712 | 1111111111111111111111111111111100001110111 | 1111111111111111111111111111111100001110111 | 0 |
| DRC-071960 | 1101111111111111111100001111111100001111111 | 1101111111111111111100001111111100001111111 | 0 |
| DRC-071968 | 1111111111111111111111111111111100001110111 | 1111111111111111111111111111111100001110111 | 0 |
| DRC-071972 | 1111111111100001110000000000000000000111100 | 1111111111100001110000000000000000000111100 | 0 |
| DRC-071973 | 1111111100011111110000000000000000000011111 | 1111111100011111110000000000000000000011111 | 0 |
| DRC-071975 | 1111111111101111111111111111111100001110111 | 1111111111101111111111111111111100001110111 | 0 |
| DRC-071988 | 1111111111111111111100000011011100000011111 | 1111111111111111111100000011011100001111111 | 3 |
| DRC-071990 | 1111110000000000000000001111111100001111111 | 1111110000000000000000001111111100001111111 | 0 |
| DRC-071998 | 1111111100011111110000000000000000000011111 | 1111111100011111110000000000000000000011111 | 0 |
| DRC-072061 | 1111110000000000000000001111111100001111111 | 1111110000000000000000001111111100001111111 | 0 |
| DRC-072379 | 1111111100011111110000000000000000000011111 | 1111111100011111110000000000000000000011111 | 0 |
| DRC-072384 | 1111101000001111111111101000000000010000011 | 1111111000001111111111100000000000110000011 | 3 |
| DRC-072390 | 1101111111111111111100001111111100001111111 | 1101111111111111111100001111111100001111111 | 0 |
| DRC-072603 | 1111111111111111101111111111111100001111111 | 1111111111111111101111111111111100001111111 | 0 |
| DRC-072604 | 1111000000000000000000001111111100001110111 | 1111000000000000000000001111111100001110111 | 0 |

| | | | |
|---|---|---|---|
| DRC-072605 | 1111110000000000000000000111111100001111111 | 1111110000000000000000000111111100001111111 | 0 |
| DRC-072606 | 1111110000000000000000000111111100001111111 | 1111110000000000000000000111111100001111111 | 0 |
| DRC-072608 | 1111111111110111111111111111111111100001110111 | 1111111111110111111111111111111111100001110111 | 0 |
| DRC-072609 | 1111111111111111111111111111111111100001110111 | 1101111111110111111100001111111100001111111 | 7 |
| DRC-072613 | 1101111111011111111100001111111100001111111 | 1111111111011111111100001111111100001111111 | 1 |
| DRC-072615 | 1111110000000000000000000111111100001111111 | 1111110000000000000000000111111100001111111 | 0 |
| DRC-072618 | 1111111111110000111000000000000000000111100 | 1111111111110000111000000000000000000111100 | 0 |
| DRC-072625 | 1111111100011111111000000000000000000011111 | 1111111100011111111000000000000000000011111 | 0 |
| DRC-072818 | 1101111111111111111100001111111100001111111 | 1111111111110000111000000000000000000111100 | 17 |
| DRC-072819 | 1111110000000000000000000111111100001111111 | 1111110000000000000000000111111100001111111 | 0 |
| DRC-072820 | 1111110000000000000000000111111100001111111 | 1111110000000000000000000111111100001111111 | 0 |
| DRC-072824 | 1111110000000000000000000111111100001111111 | 1111110000000000000000000111111100001111111 | 0 |
| DRC-072825 | 1111111111111111111111111111111111100001110111 | 1111111111111111111111111111111111100001110111 | 0 |
| DRC-072827 | 1111111111111111111111111111111111100001110111 | 1111111111111111111111111111111111100001110111 | 0 |
| DRC-072829 | 1111111111111111111111111111111111100001110111 | 1111111111111111111111111111111111100001110111 | 0 |
| DRC-072830 | 1111111111110000111000000000000000000111100 | 1111111111110000111000000000000000000111100 | 0 |
| DRC-072834 | 1111111111101111111111111111111111100001110111 | 1111111111101111111111111111111111100001110111 | 0 |
| DRC-072836 | 1111111111111111111101100111111100001010111 | 1111111111111111111101100111111100001110111 | 1 |
| DRC-072837 | 1111111111111111111111111111111111100001110111 | 1111111111111111111111111111111111100001110111 | 0 |
| DRC-072943 | 1111111111110000111000000000000000000111100 | 1111111111110000111000000000000000000111100 | 0 |
| DRC-072945 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-072957 | 1101111111101111111100001111111100001111111 | 1111111111111111111111111111111100001110111 | 7 |
| DRC-072959 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-072963 | 1111111111111111101111111110111100001111111 | 1111111111111111111111111110111100001111111 | 1 |
| DRC-072974 | 1111111111111111111100001111111100001011111 | 1111111111111111111100001111111100001011111 | 0 |
| DRC-072986 | 1111111111111111111100001100001100001011111 | 1111111111111111111100001100001100001111111 | 1 |
| DRC-072988 | 1111110000000000000000000111111100001110111 | 1111110000000000000000000111111100001110111 | 0 |
| DRC-073093 | 1111111000001111111111011000000000010001111 | 1111111000001111111111010000000000110001111 | 2 |
| DRC-073099 | 1111101000001111111111110000000000010000011 | 1111111000001111111111110000000000110000011 | 3 |
| DRC-073168 | 1111111111110000111000000000000000000111100 | 1111111111110000111000000000000000000111100 | 0 |
| DRC-073169 | 1111111111111111111100001111011100001110111 | 1111111111111111111100001111011100001110111 | 0 |
| DRC-073170 | 1111111111111111111111111111111111100001110111 | 1111111111111111111111111111111111100001110111 | 0 |
| DRC-073173 | 1101111111101111111100001111111100001111111 | 1101111111101111111100001111111100001111111 | 0 |
| DRC-073175 | 1101111111111111111111111111111111100001110111 | 1101111111111111111111111111111111100001110111 | 0 |
| DRC-073183 | 1111111111111111101111111111111111100001111111 | 1111111111111111111111111111111111100001111111 | 1 |

| | | | |
|---|---|---|---|
| DRC-073186 | 1111111111111111111111111110111100001110111 | 1111111111111111111111111110111100001110111 | 0 |
| DRC-073413 | 1111111111111111111100001111011100001111111 | 1111111111111110111100001111011100001111111 | 1 |
| DRC-073414 | 0000000000000000000000001111111100001110111 | 0000000000000000000000001111111100001110111 | 0 |
| DRC-073415 | 0000000000000000000000001111111100001110111 | 0000000000000000000000001111111100001110111 | 0 |
| DRC-073420 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-073422 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-073423 | 1111111111111111101111111111111100001100001 | 1111111111111111101111111111111100001100001 | 0 |
| DRC-073424 | 1111111111111111111100001111111100001111111 | 1111111111011111111100001111111100001111111 | 1 |
| DRC-073429 | 1111110000000000000000001111111100001111111 | 1111110000000000000000001111111100001111111 | 0 |
| DRC-073430 | 1111111111111111111111111111111100001110111 | 1111111111111111111111111111111100001110111 | 0 |
| DRC-073432 | 1111111111111111111111111111111100001110111 | 1111111111011111111111111111111100001110111 | 1 |
| DRC-073434 | 1111111111101111101111111111111100001110111 | 1111111111101111101111111111111100001110111 | 0 |
| DRC-073443 | 1111111111100001110000000000000000000111100 | 1111111111100001110000000000000000000111100 | 0 |
| DRC-073515 | 1111110000000000000000001111111100001111111 | 1111110000000000000000001111111100001111111 | 0 |
| DRC-073712 | 1100000011111111011100001111111100001111111 | 1100000011111111011100001111111100001111111 | 0 |
| DRC-073716 | 1101111111111111111100001111111100001111111 | 1101111111111111111100001111111100001111111 | 0 |
| DRC-080008 | 1111111100011111110000000000000000000011111 | 1111111100011111110000000000000000000011111 | 0 |
| DRC-080016 | 1111111111111101111111111111111100001110111 | 1111111111111111111111111111111100001110111 | 1 |
| DRC-080021 | 1111110000000000000000001111111100001111111 | 1111110000000000000000001111111100001111111 | 0 |
| DRC-080028 | 1111111100011111110000000000000000000011111 | 1111111100011111110000000000000000000011111 | 0 |
| DRC-080040 | 1111110000000000000000001111111100001111111 | 1111110000000000000000001111111100001111111 | 0 |
| DRC-080046 | 0000000000000000000000001111111100001110111 | 0000000000000000000000001111111100001110111 | 0 |
| DRC-080545 | 1111111111111111111111001111111100001111111 | 1111111111111111111111001111111100001111111 | 0 |
| DRC-080548 | 1111111000001111111110110000000000010001111 | 1111111000001111111110100000000000110001111 | 3 |
| DRC-080552 | 1111111110111111111100001111111100001011111 | 1111111110111111111100001111111100001011111 | 0 |
| DRC-080553 | 1111111111111111111111111111111100001110111 | 1111111111111111111111111111111100001110111 | 0 |
| DRC-080561 | 1111111111111111111100001111111100001011111 | 1111111111111111111100001111111100001011111 | 0 |
| DRC-080567 | 1111111111111111111111111110111100001110111 | 1111111111111111111111111110111100001110111 | 0 |
| DRC-081290 | 1111111110111111111111111111111100001111111 | 1111111110111111111100001111111100001111111 | 4 |
| DRC-081527 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-081585 | 1111111111111111111111111111111100001110111 | 1111111111111111111111111111111100001110111 | 0 |
| DRC-081588 | 1100000011111111011100001111111100001111111 | 1100000011111111011100001111111100001111111 | 0 |
| DRC-081589 | 1110111111110000111000000000000000000111100 | 1110111111110000111000000000000000000111100 | 0 |
| DRC-081592 | 1111111111111101111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 1 |
| DRC-081593 | 1111111111111111111111111111111100001110111 | 1111111111111111111111111111111100001110111 | 0 |

| | | | |
|---|---|---|---|
| DRC-081595 | 1111111111111111111111111111111100001110111 | 1111111111111111111111111111111100001110111 | 0 |
| DRC-081600 | 1111110000000000000000000111111100001111111 | 1111110000000000000000000111111100001111111 | 0 |
| DRC-081602 | 1111111111110111111111111111111100001110111 | 1111111111110111111111111111111100001110111 | 0 |
| DRC-081610 | 1111111100111111111111111111111100001111111 | 1111111100111111111111111111111100001111111 | 0 |
| DRC-081622 | 1111111111111111111100001100001100001111111 | 1111111111111111111100001100001100001111111 | 0 |
| DRC-081722 | 1101111111111111111100001111111100001111111 | 1101111111111111111100001111111100001111111 | 0 |
| DRC-081873 | 1111111111111111111100001111111100001011111 | 1111111111111111111100001111111100001011111 | 0 |
| DRC-081879 | 1111111111111111101111111111111100001100001 | 1111111111111111101111111111111100001100001 | 0 |
| DRC-081881 | 1111110000000000000000000111111100001111111 | 1111110000000000000000000111111100001111111 | 0 |
| DRC-081888 | 1111111100011111110000000000000000000011111 | 1111111100011111110000000000000000000011111 | 0 |
| DRC-081985 | 1111111111111111101111111111111100001111111 | 1111111111111111101111111111111100001111111 | 0 |
| DRC-081986 | 1111111111101011111111111111111100001110111 | 1111111111101111111111111111111100001110111 | 1 |
| DRC-081989 | 1111110000000000000000000111111100001111111 | 1111110000000000000000000111111100001111111 | 0 |
| DRC-081990 | 0000000000000000000000000000000000111111111 | 0000000000000000000000000000000000111111111 | 0 |
| DRC-082120 | 1111111111111111111111111111111100001110111 | 1111111111111111111111111111111100001110111 | 0 |
| DRC-082121 | 1111111111111111111111111111111100001110111 | 1111111111111111111111111111111100001110111 | 0 |
| DRC-082130 | 1111111111111111111100001100001100001001111 | 0111111111111111111100001100001100001001111 | 1 |
| DRC-082131 | 1111111111111111101111111111111100001111111 | 1111111111111111101111111111111100001111111 | 0 |
| DRC-082317 | 1111111111111111111101100111111100001010111 | 1111111111111111111111111111111100001110111 | 4 |
| DRC-082318 | 1111110000000000000000000111111100001111111 | 1111110000000000000000000111111100001111111 | 0 |
| DRC-082327 | 1111111111111101111111111110111100001110111 | 1111111111111111111111111110111100001110111 | 1 |
| DRC-082328 | 1111111100011111110000000000000000000011111 | 1111111100011111110000000000000000000011111 | 0 |
| DRC-082330 | 1101111111111111111100001111111100001111111 | 1101111111111111111100001111111100001111111 | 0 |
| DRC-082584 | 1111111111111111111111111111111100001110111 | 1111111111111111111111111111111100001110111 | 0 |
| DRC-082592 | 1111111111111111111101100111111100001010111 | 1111111111111111111101100111111100001110111 | 1 |
| DRC-082639 | 1111111100011111110000000000000000000011111 | 1111111100011111110000000000000000000011111 | 0 |
| DRC-082656 | 1111111111111111111100001110111100001111111 | 1111111111111111111100001110111100001111111 | 0 |
| DRC-082700 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-082919 | 1100000011111110111000011111111100001111111 | 1100000011111110111000011111111100001111111 | 0 |
| DRC-082923 | 1111111000001111111110110000000000010001111 | 1111111000001111111110100000000000110001111 | 2 |
| DRC-082939 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-082940 | 1111111100011111110000000000000000000011111 | 1111111100011111110000000000000000000011111 | 0 |
| DRC-082945 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-082946 | 1111111111111111111111111111111100001110111 | 1111111111111111111111111111111100001110111 | 0 |
| DRC-082948 | 1111111111111111111100001110111100001110111 | 1111111111111111111100001110111100001110111 | 0 |

| | | | |
|---|---|---|---|
| DRC-082949 | 1111111111111101111111111111111111100001111111 | 1111111111111111111111111111111111100001111111 | 1 |
| DRC-083047 | 1111111111111111111100001111011100001111111 | 1111111111111111111100001111011100001111111 | 0 |
| DRC-083048 | 1111111111111111111100000011011100001111111 | 1111111111111111111100000011011100001111111 | 1 |
| DRC-083050 | 1111111111111111111100001111011100001001111 | 1111111111111111111100001111011100001001111 | 0 |
| DRC-083051 | 1111111111111111111111111111111100001110111 | 1111111111111111111111111111111100001110111 | 0 |
| DRC-083052 | 1111111111111111111111000111111100001111111 | 1111111111111111111111000111111100001111111 | 0 |
| DRC-083066 | 1111111111110000111000000000000000000111100 | 1111111111110000111000000000000000000111100 | 0 |
| DRC-083067 | 1111111111101111111111111111111100001110111 | 1111111111101111111111111111111100001110111 | 0 |
| DRC-083190 | 1111111111111111111100001111011100001111111 | 1111111111111111111100001111111100001111111 | 1 |
| DRC-083191 | 1111111111111111111100001111011100001111111 | 1111111111111111111100001111011100001111111 | 0 |
| DRC-083200 | 1111110000000000000000000111111100001111111 | 1111110000000000000000000111111100001111111 | 0 |
| DRC-083201 | 1000001111111111111000001111111100001111111 | 1000001111111111111111111111111100001111111 | 5 |
| DRC-083202 | 1111111111111111111100001111011100001001111 | 1111111111111111111100001111011100001001111 | 0 |
| DRC-083204 | 1111111100011111111000000000000000000011111 | 1111111100011111111000000000000000000011111 | 0 |
| DRC-083226 | 1111110101111111111111111000000000000000000 | 1111110101111111111111111000000000000000000 | 1 |
| DRC-083250 | 1111111111111111111111111111110100001111111 | 1111111111111111111111111111111100001111111 | 1 |
| DRC-083736 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-083801 | 1111111111111111111111111111111100001110111 | 1111111111111111111111111111111100001110111 | 0 |
| DRC-084109 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-084204 | 1101111101111111111100001111111100001110111 | 1101111101111111111100001111111100001110111 | 0 |
| DRC-084210 | 1111111100011111111000000000000000000011111 | 1111111100011111111000000000000000000011111 | 0 |
| DRC-090010 | 1111111111111111111100001111011100001111111 | 1111111111111111111100001111011100001111111 | 0 |
| DRC-090012 | 1101111111111111111100001111111100001111111 | 1101111111111111111100001111111100001111111 | 0 |
| DRC-090016 | 1111111111111111111100001111011100001111111 | 1111111111111111111100001111011100001111111 | 0 |
| DRC-090022 | 1111111111101111111111111111111100001110111 | 1111111111101111111111111111111100001110111 | 0 |
| DRC-090037 | 1111111111111111111100001100001100001111111 | 1111111111111111111100001100001100001111111 | 0 |
| DRC-090066 | 1111111000001111111111011000000000010001111 | 1111111000001111111111010000000000110001111 | 2 |
| DRC-090069 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-090077 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-090078 | 1111111111111111111100001111111100001011111 | 1111111111111111111100001111111100001011111 | 0 |
| DRC-090165 | 1111111111111111111111111111111100001110111 | 1111111111111111111111111111111100001110111 | 0 |
| DRC-090172 | 1111111111110000111000000000000000000111100 | 1111111111110000111000000000000000000111100 | 0 |
| DRC-090299 | 1101111111111111111100001111111100001111111 | 1101111111111111111100001111111100001111111 | 0 |
| DRC-090308 | 1111110000000000000000000111111100001111111 | 1111110000000000000000000111111100001111111 | 0 |
| DRC-090309 | 1111111111101111111100001100001100001111111 | 1111111111101111111100001100001100001111111 | 0 |

| | | | |
|---|---|---|---|
| DRC-090310 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-090432 | 1101111111111111111100001111111100001111111 | 1101111111111111111100001111111100001111111 | 0 |
| DRC-090451 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-090452 | 1111111100011111111000000000000000000011111 | 1111111100011111111000000000000000000011111 | 0 |
| DRC-090456 | 1111111111111111111111111101111100001110111 | 1111111111111111111111111101111100001110111 | 0 |
| DRC-090702 | 1111110000000000000000000111111100001111111 | 1111110000000000000000000111111100001111111 | 0 |
| DRC-090708 | 1111111111111111101111111111111100001100001 | 1111111111111111101111111111111100001100001 | 0 |
| DRC-090710 | 1111111111100001110000000000000000000111100 | 1111111111100001110000000000000000000111100 | 0 |
| DRC-090909 | 1101111111111111111100001111111100001111111 | 1101111111111111111100001111111100001111111 | 0 |
| DRC-090913 | 1111111111111111111111111111111100001110111 | 1111111111111111111111111111111100001110111 | 0 |
| DRC-090947 | 1111111100011111111000000000000000000011111 | 1111111100011111111000000000000000000011111 | 0 |
| DRC-090976 | 1111111111111111111100001100001100001111111 | 1111111111111111111100001100001100001111111 | 0 |
| DRC-090977 | 1101111111111111111100001111111100001111111 | 1101111111111111111100001111111100001111111 | 0 |
| DRC-090999 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-091003 | 1111111111111111111111111101111100001110111 | 1111111101111111111111111101111100001110111 | 1 |
| DRC-091004 | 1111110000000000000000000111111100001111111 | 1111110000000000000000000111111100001111111 | 0 |
| DRC-091005 | 1111111111101011111111111111111100001111111 | 1111110000000000000000000111111100001111111 | 16 |
| DRC-091168 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-091171 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111011100001111111 | 1 |
| DRC-091172 | 1111111111111111111111111101111100001110111 | 1111111111111111111111111101111100001110111 | 0 |
| DRC-091696 | 1111111110111111111100001111111100001111111 | 1111111110111111111100001111111100001111111 | 0 |
| DRC-091821 | 1111111111111111111111111101111100001110111 | 1111111111111111111111111101111100001110111 | 0 |
| DRC-092401 | 1111111111111111111111000111111100001111111 | 1111111111111111111111000111111100001111111 | 0 |
| DRC-092631 | 1111111111111111111100001100001100001110000 | 1111111111111111111100001100001100001110000 | 0 |
| DRC-092748 | 1101111111111111111100001111111100001111111 | 1101111111111111111100001111111100001111111 | 0 |
| DRC-092750 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-093215 | 1101111101111111111100001111111100001110111 | 1101111101111111111100001111111100001110111 | 0 |
| DRC-093480 | 1111111111111111111111111111111100001110111 | 1111111111111111111111111111111100001110111 | 0 |
| DRC-093482 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-093488 | 1101111111111111111100001111111100001111111 | 1101111111111111111100001111111100001111111 | 0 |
| DRC-093493 | 1111111111111111111100001111111100001111111 | 1111111111111111111100001111111100001111111 | 0 |
| DRC-093495 | 1111111111111111111111111101111100001110111 | 1111111111111111111100001111111100001111111 | 6 |
| DRC-093664 | 1111110000000000000000000111111100001111111 | 1111110000000000000000000111111100001111111 | 0 |
| DRC-093666 | 1111111111101111111100010111110000001111111 | 1111111111101111111100010111110000001111111 | 1 |
| DRC-093674 | 1111111111111111111111111111111100001111111 | 1111111111111111111111111111111100001111111 | 0 |

| | | | |
|---|---|---|---|
| DRC-100033 | 1111111111111111101111111111111100001111111 | 1111111111111111101111111111111100001111111 | 0 |
| DRC-100035 | 1111111111101111111111111111111100001110111 | 1111111111101111111111111111111100001110111 | 0 |
| DRC-100037 | 1101111111101111111000011111111100001111111 | 1101111111101111111000011111111100001111111 | 0 |
| DRC-100099 | 1111110000000000000000001111111100001111111 | 1111110000000000000000001111111100001111111 | 0 |
| DRC-100122 | 1111111111111111111111111111001100001110111 | 1111111111111111111111111111111100001110111 | 2 |
| DRC-100299 | 1111111111111111111000011111111100001011111 | 1111111111111111111000011111111100001011111 | 0 |
| DRC-100300 | 1111111100011111111000000000000000000011111 | 1111111100011111111000000000000000000011111 | 0 |
| DRC-100301 | 1111111111111111111000011111111100001011111 | 1111111111111111111000011111111100001011111 | 0 |
| DRC-100307 | 1111111111111111111000011000011100001111111 | 1111111111111111111000011000011100001111111 | 0 |
| DRC-100310 | 1111111111111100000111111111111100001110111 | 1111111111111100000111111111111100001110111 | 0 |
| DRC-100503 | 1111111111111111111111111111111100001110111 | 1111111111111111111111111111111100001110111 | 0 |
| DRC-100740 | 1111111111111111111000011111111100001111111 | 1111111111111111111000011111111100001111111 | 0 |
| DRC-100753 | 1111111111111111111111111111011100001111111 | 1111111111111111111111111111111100001110111 | 2 |
| DRC-100759 | 1101111111111111111000011111111100001111111 | 1101111111111111111000011111111100001111111 | 0 |
| DRC-100923 | 1111111111111111111000011110111100001111111 | 1111111111111111111000011111111100001011111 | 2 |
| DRC-100929 | 1111111100011111111000000000000000000011111 | 1111111100011111111000000000000000000011111 | 0 |
| DRC-100931 | 1101111111111111111000011111111100001111111 | 1101111111111111111000011111111100001111111 | 0 |
| DRC-101072 | 1111111111111111111000011111111100001111111 | 1111111111111111111000011111111100001111111 | 0 |
| DRC-101306 | 1111111111111111111000011110111100001111111 | 1111111111111111111000011110111100001111111 | 0 |
| DRC-101308 | 1101111111111111111000011111111100001111111 | 1101111111111111111000011111111100001111111 | 0 |

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

862 For each method, the inclusion of an isolate into a cluster is outlined in the surrounding

863 circles using GraPhlAn (Asnicar et al., 2015). If an isolate is in a cluster not affected by

864 convergence, it is highlighted in black for the given method. If an isolate is in a cluster

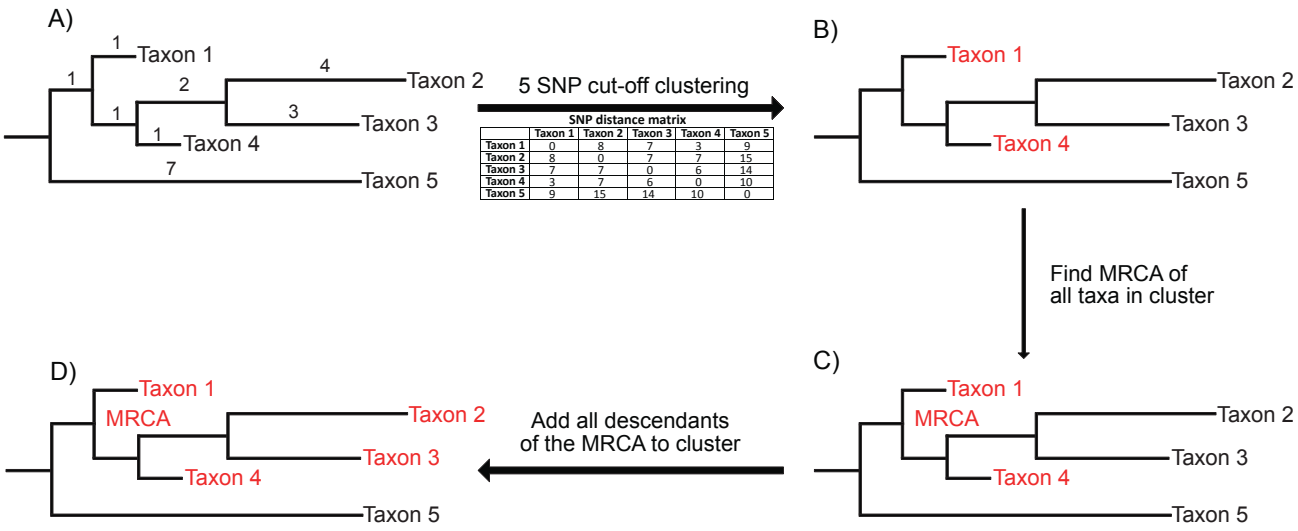
865 affected by convergence, it is shown in grey.

866

867

Supplementary Figure 2: Algorithm of the phylogenetic inclusion method

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




SNP distance matrix

| | Taxon 1 | Taxon 2 | Taxon 3 | Taxon 4 | Taxon 5 |
|---|---|---|---|---|---|
| Taxon 1 | 0 | 8 | 7 | 3 | 9 |
| Taxon 2 | 8 | 0 | 7 | 7 | 15 |
| Taxon 3 | 7 | 7 | 0 | 6 | 14 |
| Taxon 4 | 3 | 7 | 6 | 0 | 10 |
| Taxon 5 | 9 | 15 | 14 | 10 | 0 |