



OPEN

# Phylogenetic assessment of SNP distances from whole genome sequencing for determining *Mycobacterium tuberculosis* transmission

Bastiaan R. Van der Roest<sup>1</sup>✉, Martin C. J. Bootsma<sup>1,2,6</sup>, Egil A. J. Fischer<sup>3</sup>,  
Matthias I. Gröschel<sup>4</sup>, Richard M. Anthony<sup>5</sup>, Rina de Zwaan<sup>5</sup>,  
Mirjam E. E. Kretzschmar<sup>1,6,8</sup> & Don Klinkenberg<sup>7</sup>

The global tuberculosis (TB) epidemic is driven by primary transmission. Pathogen genome sequencing is increasingly used in molecular epidemiology and outbreak investigations. Based on contact tracing and epidemiological links, Single Nucleotide Polymorphism (SNP) cut-offs, ranging from 3 to 12 SNPs, identify probable transmission clusters or exclude direct transmission. However, contact tracing can be limited by recall bias and inconsistent methodologies across TB settings. We propose phylogenetic models, i.e. methods to infer transmission processes from pathogen genomes and associated epidemiological data, as an alternative reference to infer transmission events. We analyzed 2,008 whole-genome sequences from Dutch TB patients collected from 2015 to 2019. Genetic clusters were defined within a 20-SNP range, and the phylogenetic model phylbreak was employed to infer transmission. Probable transmission SNP cut-offs were assessed by the proportion of inferred transmission events with a SNP distance below these cut-offs. A total of 79 clusters were identified, with a median size of 4 isolates (IQR = 3–8). A SNP cut-off of 4 captured 98% of inferred transmission events while reducing pairs without transmission links. A cut-off beyond 12 SNPs effectively excluded transmission. Phylogenetic approaches provide a valuable alternative to contact tracing for defining SNP cut-offs, allowing for a more precise assessment of transmission events.

**Keywords** Phylogenetic modeling, Whole Genome Sequencing, SNP cut-off assessment, Tuberculosis, Transmission

Tuberculosis (TB), caused by members of the *Mycobacterium tuberculosis* complex (Mtb), remains a global health concern. Primary transmission of active TB is the main contributor to the current global epidemic. To reduce TB incidence in a population, it is essential to understand transmission patterns. Upon infection, only a small fraction (between 2 to 10%) develop active TB in the first 1–2 years<sup>1</sup>. The remaining infected individuals may carry a (latent) TB infection for decades, without transmitting to others or having any symptoms unless the infection becomes active. Cases are often classified as due to recent transmission, i.e. having been infected or having infected someone else within 2 years, or not<sup>2</sup>. This is often done by bacterial (geno)typing to rule out recent transmission, and to identify probable transmission clusters.

<sup>1</sup>Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, P.O.Box 85500 Utrecht, The Netherlands. <sup>2</sup>Department of Mathematics, Faculty of Science, University Utrecht, Utrecht, The Netherlands. <sup>3</sup>Population Health Sciences, Faculty of Veterinary Medicine, University Utrecht, Utrecht, The Netherlands. <sup>4</sup>Department of Infectious Diseases, Respiratory and Critical Care Medicine, Charité-Universitätsmedizin Berlin, Berlin, Germany. <sup>5</sup>Tuberculosis Reference Laboratory, Centre for Infectious Disease Control, National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands. <sup>6</sup>Centre for Complex System Studies (CCSS), University Utrecht, Utrecht, The Netherlands. <sup>7</sup>National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands. <sup>8</sup>Institute of Epidemiology and Social Medicine, University of Münster, Münster, Germany. ✉email: b.r.vanderroest-2@umcutrecht.nl

Potential transmission events are inferred when the genotyping method applied finds the isolates sufficiently similar: When typing with the Mycobacterial Identification Repetitive Unit - Variable Number of Tandem Repeat (MIRU-VNTR) method, transmission is not ruled out unless the two cases have identical MIRU-VNTR profiles. Transmission is considered confirmed if interviews with the two cases suggest that they may have been in contact<sup>3</sup>.

Until recently, MIRU-VNTR typing was the method of choice for bacterial genotyping and hence genetic transmission inference. However, several studies based on Whole Genome Sequences (WGS) show that the proportion of isolates where transmission can be ruled out is greater if Single Nucleotide Polymorphisms (SNPs) are used<sup>4,5</sup>, i.e., isolates are added to a cluster if the number of SNPs with at least one other isolate in the cluster does not exceed a SNP cut-off. This cut-off can be as low as a few SNPs, as the mutation rate is only 0.3 per genome per year for *Mtbc*.<sup>6</sup>

SNP cut-offs are determined for the goals of TB control using contact tracing data as the determinant for transmission. However, contact tracing is not always reliable, i.e., in high TB incidence settings where contacts can be missed or not reported<sup>7</sup>, and can be subject to recall bias. Several SNP cut-offs have been reported in the literature. To rule out potential transmission, a concordance between studies was found for a SNP cut-off of 12<sup>3,8,9</sup>. However, cases that were in contact with each other, leading to potential transmission, were included in the same transmission cluster with a SNP cut-off varying from 3 to 12<sup>10–17</sup>.

Where in a low-incidence setting, a SNP cut-off of 5 was shown to cluster 99.3% of the cases with confirmed contact. In 61.8% of case pairs, whose isolates had a distance below 5 SNPs, no contact between the paired members could be identified<sup>18</sup>. In other low-incidence settings, the 5 SNP cut-off positively ruled out 17% to 95% of the probable links<sup>19</sup>. Although these differences could partly be explained by the SNP calling pipelines used and different mutation rates among the lineages, the performance of a SNP cut-off also depends on the collection, interpretation, and quality of the contact tracing data. Therefore, it is hard to select the optimal SNP cut-off.

Given the practical value of a SNP threshold to assess potential transmission events in understanding local TB transmission, it is useful to put the SNP cut-offs derived by contact tracing as a reference into perspective by using a different reference. In the field of phylodynamics, genomic data is used to study the transmission of infectious diseases, e.g., to infer transmission events by combining WGS data, epidemiological data, and information on disease dynamics such as the serial interval<sup>20,21</sup>. For example, using the TransPhylo model TB outbreaks have been studied, showing that transmission events could be revealed that were otherwise overlooked<sup>22</sup>. Furthermore, using the same method it was revealed that one-quarter of patients contracted TB outside Norway, which is a setting with primarily TB cases among migrants<sup>23</sup>. However, identification of transmission events comes with great uncertainty<sup>22</sup>. TransPhylo links all observed cases into a single transmission tree, i.e. the sequence of transmission between cases, and it may impute unobserved cases. The method assumes that all observed and unobserved cases have the same time-to-detection and generation time distributions. However, the transmission dynamics and case detection may differ between countries, so this assumption may be problematic for datasets with many immigrants who may have been infected in their country of origin. This is often the case in low-incidence countries in Europe and North America.

An alternative method to infer transmission events is *phybreak*, which does not impute single unobserved cases but assumes that there is a source population of unobserved cases that may have resulted in multiple index cases of smaller transmission trees with observed cases<sup>24,25</sup>. Furthermore, it only assumes the same time-to-detection and generation time distributions for transmission events between observed cases. This makes *phybreak* more suitable for studying low-incidence settings where a significant proportion of distant source cases will not be known. Although *phybreak* is employed in studies to risk factor of TB transmission<sup>26,27</sup>, inference results were not used before to assess SNP cut-offs used to make genetic clusters.

Here, we conducted a retrospective genomic epidemiology study using pathogen WGS data routinely collected in The Netherlands to assess widely used SNP distance cut-offs using *phybreak* to infer transmission events. The events are used as a reference to assess SNP cut-offs for the two objectives of epidemiological investigation: 1) ruling out potential recent transmission events or 2) identification of TB transmission clusters.

## Methods

All code used for the methods described below can be found at [https://github.com/bastiaanvdroest/TB\\_SNP\\_cutoff\\_assessment](https://github.com/bastiaanvdroest/TB_SNP_cutoff_assessment).

## Data

Our data contains 2008 whole genome sequences from *Mtbc* culture-positive cases sampled by the National Institute for Public Health and the Environment (RIVM) in the Netherlands between December 2015 and December 2019 with only strains typed as *M. tuberculosis sensu stricto* (lineages 1–4) or *M. africanum* (lineages 5 and 6). Isolates from the Netherlands were sequenced at the RIVM Bilthoven as follows: genomic DNA was extracted from positive MGIT culture tubes (Becton Dickinson, NJ, USA) using the QIAamp DNA mini kit (QIAGEN GmbH, Hilden, Germany) and sequencing was performed on Illumina HiSeq2500 with 2 x 125bp reads following Nextera XT library preparation. Sequences are stored under BioProjects PRJNA882748, PRJNA1079737, PRJEB32037 and PRJEB25592, and SRA accessions found in Supplementary Table 2. Reads were first trimmed using *fastp* v0.20.1<sup>28</sup> and aligned to the reference genome H37Rv (accession NC\_000962.3) using *BWA mem* v0.7.17 (parameters: default)<sup>29</sup>. Duplicate reads were removed with *Picard* v2.25.1 (<http://broadinstitute.github.io/picard/>). SNPs were called using *Pilon* v1.24 (parameters: default) requiring confidence base calls (pass filter)<sup>30</sup>. To analyze the genetic variation in our sample we created a genotypes matrix based on the variant call format files. We excluded all sites that had an Empirical Base-level Recall (EBR) score of <0.9<sup>31</sup>, and all sites located in mobile genetic elements (e.g., transposases, integrases, phages, or insertion sequences). We

also dropped SNP sites that are missing in  $> 10\%$  of our strains. We cannot infer any information on these sites, as they are missing because of real deletions or low sequence quality. If a base call at a specific reference position for an isolate did not meet the filter criteria that allele was coded as missing. Our filtered genotypes matrix had dimensions of 2,008x325275 SNPs representing 325275 SNP sites across 2,008 strains. The phylogenetic lineage was determined using a refined lineage-calling scheme based on 96-SNPs with fast-lineage-caller<sup>32</sup>. For all isolates, the phylogenetic lineage type was known (Table 3). Besides sequences, we had clinical and demographic data, comorbidity information, and data on social risk factors for TB for all cases<sup>33</sup>. Gröschel et al.<sup>33</sup> described data of close contacts. However, we could not use these data, because close contacts were mostly latent TB cases, of which no sequence was available, and we did not have information to link active TB cases via contact. Furthermore, we used both pulmonary TB and extra pulmonary TB (EPTB) cases in our study, although cases with EPTB have a significant lower infectiousness. EPTB are unlikely to be infectors but could be present in transmission events.

### Assessment pipeline

The reconstruction of transmission events and assessment of the SNP cut-offs involves multiple steps (Figure 1). First, we split the set of isolates into genetic clusters with a genetic distance between clusters of more than 20 SNPs, so that recent transmission can be ruled out for isolates in different clusters<sup>34</sup>. Next, we used the sequences to infer phylogenetic trees of Mtb lineages, to obtain mutation rates of these lineages. Then, we inferred for each genetic cluster potential transmission events. Finally, we use the inferred transmission events as a reference for the assessment of the SNP cut-offs.

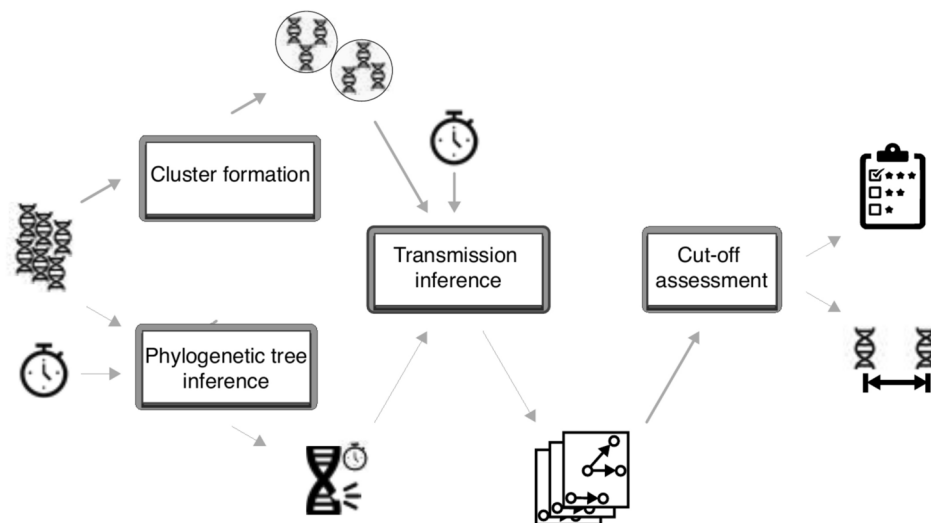
### Cluster formation

We used the R package *adegenet*<sup>35</sup> to perform genetic transitive clustering with a distance cut-off of 20 SNPs. This ensures that each sequence in a cluster of size two or larger has a SNP distance of 20 or less to at least one other sequence in the same cluster. A distance of 20 SNPs is assumed to be large enough to ensure recent transmission events between clusters can be ruled out. Clusters of size 2 are dropped from the analysis because they contain too little information for reliable inference. Once genetic clusters were defined, SNPs that were found in subsets of isolates in different clusters were removed. These are believed to be due to miscalling of nucleotides.

### Phylogenetic tree inference

Since the genetic clusters are too small to estimate a reliable mutation rate, we estimated the mutation rate from sequences from all isolates with the same Mtb lineage in a phylogenetic analysis with BEAST 1.7<sup>36</sup>. We used the GTR mutation model with a relaxed clock<sup>23,37</sup> combined with a Skyride demographic model<sup>38</sup>. We ran two MCMC chains of length 100,000,000 and derived the mean mutation rate and standard deviation over all branches. The mean mutation rate per lineage was used as a fixed parameter in the transmission tree inference.

TreeAnnotator, included in the BEAST software package<sup>36</sup>, was used to get the consensus phylogenetic tree to find the posterior support of the genetic clusters made by the 20 SNP cut-off in the phylogenetic tree. If the



**Fig. 1.** Flow diagram of the SNP cut-off assessment pipeline. Sequences, and their sampling times, are used to infer phylogenetic trees per Mtb lineage and to estimate mutation rates for these lineages with BEAST. Independently of the phylogenetic tree inference, the sequences are grouped in genetic clusters based on a SNP cut-off. These clusters are used, together with the sampling times of sequences, and the mutation rate of the corresponding Mtb lineage, to infer transmission trees with *phybreak*. Finally, inferred transmission events and unconnected cases serve as a reference to assess the performance of SNP cut-offs in determining transmission events.

posterior support of the most recent common ancestor of a genetic cluster is 1, we conclude that this cluster is complete, i.e. the cluster is fully distinctive from all other isolates.

Transmission tree inference

For the inference of transmission, we used the `phybreak` package<sup>24</sup> in the R programming language<sup>39</sup>, which contains a Bayesian inference method to reconstruct transmission trees. Using genetic sequences and their sampling times, `phybreak` infers transmission chains by simultaneously sampling phylogenetic and transmission trees from the posterior distribution. In particular, `phybreak` allows for multiple transmission trees in a genetic cluster representing multiple index cases coming from an unsampled source population. In `phybreak`, a complete bottleneck at transmission is assumed, which is in line with observation of genetic bottlenecks at transmission in case of *Mtbc*<sup>33</sup>. On the other hand, the model does allow for within-host diversity, resulting genetic variation within and between hosts as observed in experiments<sup>40</sup>. In a sensitivity analysis we used the package `TransPhylo`<sup>22</sup>, which is also a Bayesian inference method for transmission trees. `TransPhylo` uses a phylogenetic tree as input on which it places transmission events including potential unobserved cases between the observed cases. The input phylogenetic trees are inferred from sequences and sampling times in BEAST.

We ran the transmission tree inference with `phybreak` on each cluster defined above, using the prior distributions for the generation and sampling time interval shown as in<sup>23</sup> (Table 1), assuming that all cases other than the index cases developed active TB soon after infection, without a phase of latent TB. Furthermore, we chose the substitution rate to be constant at the mean substitution rate of the cluster’s lineage, estimated in the phylogenetic analysis with BEAST: 0.355 and 0.3 SNPs/genome/year for lineages 3 and 4, which corresponds to a substitution rate of 0.2–0.5 SNPs per genome per year found in literature<sup>41</sup>. We ran for each cluster two MCMC chains of 10,000 cycles and combine the results of both chains. The time of the transmission cluster inference strongly depends on the genetic cluster size, but was less than 10 minutes on a Apple M1 chip for the largest cluster of 23 isolates and two MCMC chains of 10,000 cycles each. The sensitivity analyses with `TransPhylo` were done with the settings as in<sup>23</sup> (Table 1).

We decided not to estimate the mean generation time and mean sampling time from the data because most clusters were small and did therefore not contain sufficient information for estimation, and to ensure that the means were the same across clusters. Instead, we carried out sensitivity analyses for the means. We varied means for the generation time, i.e. 1, 1.3, and 2 years, and the sampling time, i.e. 0.3, 0.8, and 1.3 years.

Assessment of SNP cut-offs

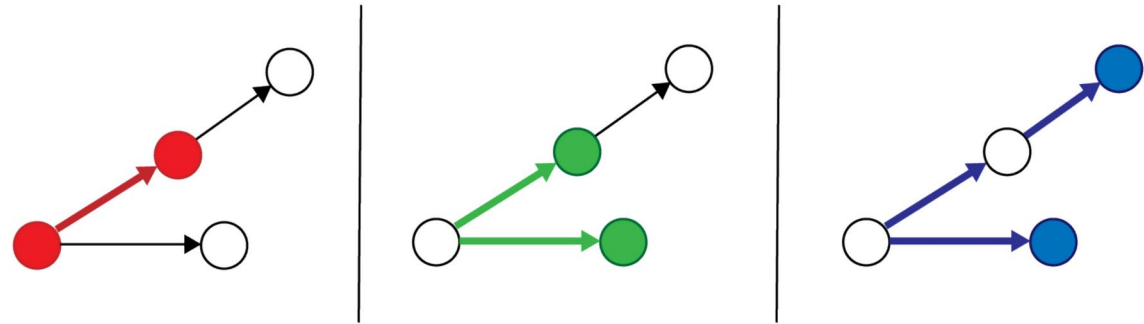
We assessed SNP cut-offs from 0 to 20 to either rule out recent transmission or identify transmission clusters. The SNP distance between two isolates is calculated using all SNPs in the dataset, including the non-cluster-unique SNPs, to make a better comparison with previous SNP cut-offs. We used three definitions of pairs being linked by recent transmission (Figure 2): direct linkage, i.e. one having infected the other (in red); infector linkage, i.e. both cases having the same infector (in green); transmission tree linkage, i.e. both cases being part of the same transmission tree (in blue). We reserve the words linked and linkage to describe the relation of a pair in the inferred transmission trees. Two isolates are truly identified as having a transmission event between them, i.e. True Positive, if the event is inferred and the distance between the isolates is lower than or equal to the SNP cut-off (Table 2). We calculated the sensitivity, i.e. the fraction of isolate pairs with a distance below or equal to the SNP cut-off among all linked pairs:  $\frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$ . In addition, we calculated the precision, i.e. the fraction of pairs with an inferred transmission event among all pairs with a distance below or equal to the SNP cut-off:  $\frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$ .

To account for the uncertainty in the transmission trees, we built the confusion matrix with 1000 samples from the `phybreak` posterior distributions of trees instead of a single consensus tree. As a result, the linkage relations between all pairs were appropriately weighed by its support in the posterior. Confidence intervals for sensitivity and precision were obtained by bootstrapping the `phybreak` posterior distributions of trees and thus creating 1000 bootstrap samples of the confusion matrix.

A sensitivity analysis on the effect of transmission inference with `phybreak` on the sensitivity and precision scores was done by inferring transmission events with `TransPhylo`.

	Phybreak	TransPhylo
Generation time interval	$\Gamma(2.77, 0.47)$	
Sampling time interval	$\Gamma(1.45, 0.55)$	
Substitution rate	0.355 SNPs/genome/year (Lineage 3) 0.3 SNPs/genome/year (Lineage 4)	
Sampling probability	NA	0.95
Effective population size	NA	1.48
Offspring distribution	NA	1

Table 1. Parameter values for main analysis with `phybreak` and sensitivity analysis with `TransPhylo`.



**Fig. 2.** Three possible types of linkage between a pair of isolates. The linkage types are given between two cases (colored dots) in a transmission tree, with the transmission direction given by arrows. In red; direct linkage, i.e., infector-infectee pair, in green; infector linkage, i.e., linkage via the same infector, and in blue; transmission tree linkage by the same transmission tree.

	True	False
Positive	Pair with inferred event and distance $\leq$ cut-off	Pair without inferred event and distance $\leq$ cut-off
Negative	Pair without inferred event and distance $>$ cut-off	Pair with inferred event and distance $>$ cut-off

**Table 2.** Definition of terms in the confusion matrix. The confusion matrix used for the calculation of precision and sensitivity contains for each term the sum of the support of all pairs in that term

	Number of isolates	Number of unclustered isolates	Number of clusters with more than 2 isolates
Lineage 1	197	188	1
Lineage 2	177	152	2
Lineage 3	422	250	21
Lineage 4	1198	892	58
Lineage 5	6	6	0
Lineage 6	8	8	0
Total	2008	1496	82

**Table 3.** Lineage typing of isolates and number of clusters in the lineages.

**Results**  
**Data overview**

We used 2,008 sequences from isolates sampled from active TB cases in the Netherlands between December 2015 and December 2019, where each case was sampled once. The main Mtb lineages were L3 (N=422) and L4 (N=1196), followed by L1 (N=197) and L2 (N=177) (Table 3). Of all sequences, 78% were sampled from first- or second-generation migrants, with half of them born in Sudan, Eritrea, Ethiopia, and Somalia (Supplementary Table 1).

**Cluster formation**

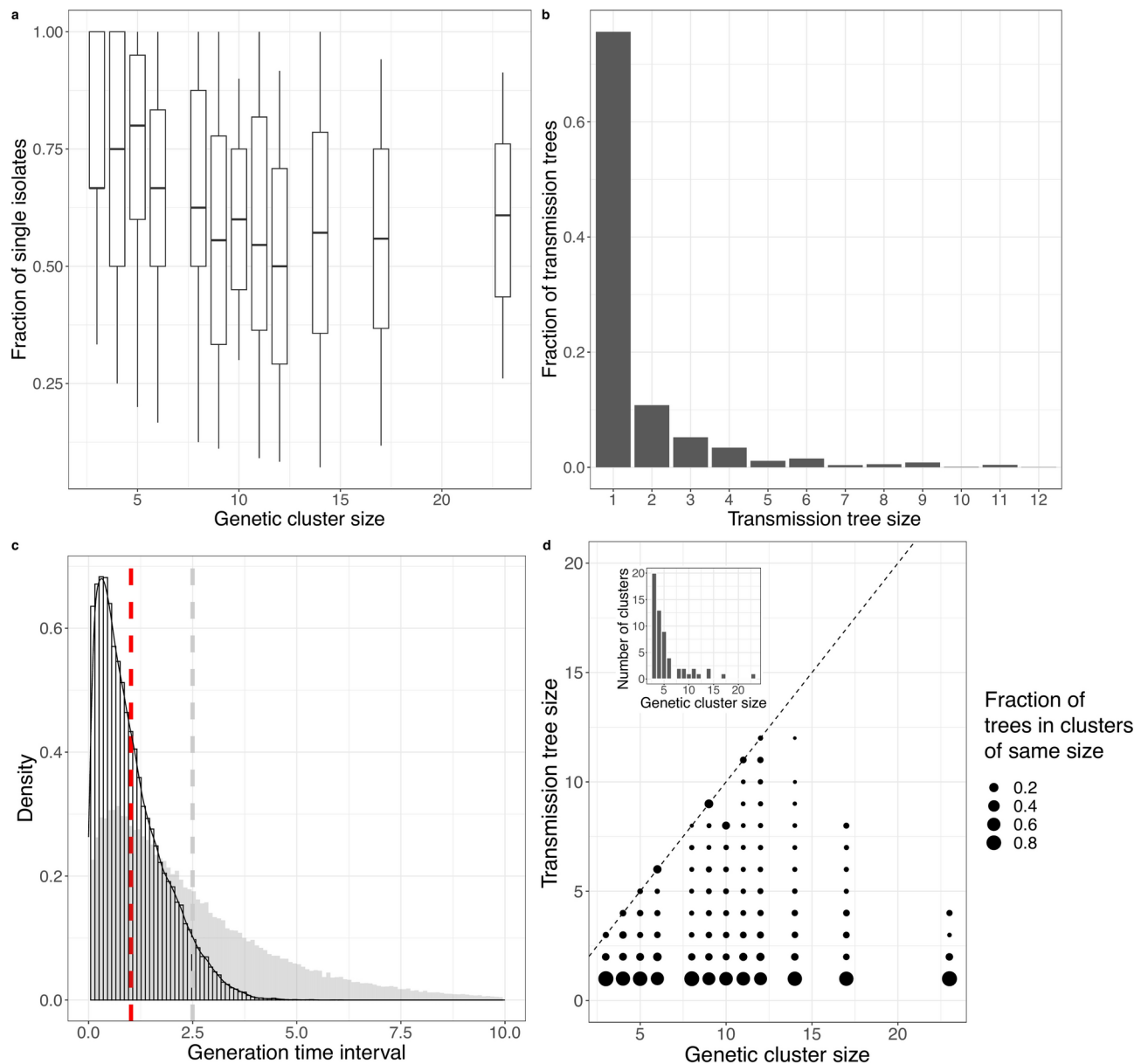
For the 14 strains of L5 and L6 each pair of isolates had a SNP distance above 200 (Supplementary Figure 1), so these isolates were excluded from further analyses. We applied a SNP cut-off of 20 SNPs to create genetic clusters of cases with possible transmission events. We found only three genetic clusters with more than 2 isolates for lineages 1 and 2 together (Table 3). Each of these 3 genetic clusters contained no more than 4 isolates. Hence, we excluded also lineages 1 and 2 from our analyses and focused on isolates from lineages 3 and 4 (1,620 isolates). We defined 21 and 58 genetic clusters of size 3 or larger for lineage 3 and 4, with in total 478/1620 of the isolates. The median size of the genetic clusters was 4 isolates (IQR: 3-8, median of L3 was 7 isolates and median of L4 was 4) and the mean time between the first and last diagnosis in the clusters was 2.2 years (SD, 1.0 years). Most clustered isolates came from people born in the Netherlands, Morocco, Sudan, Eritrea, or Ethiopia (72.5% of all clustered isolates). Isolates from migrants from Southeast Asia and the Western Pacific were underrepresented in genetic clusters of size 3 or larger, i.e., the occurrence reduced from 11.9% and 5.7% of lineages 3 and 4 isolates to 1.5% and 0.4% of clustered isolates, respectively (Supplementary Table 1).

### Transmission inference

We found a probability of 66% (95% CI: 61% - 70%) that an isolate belonging to a genetic cluster of size 3 or larger is not linked to any other case. This probability does not strongly depend on the cluster size (Figure 3A). Furthermore, 95% of all inferred transmission trees, including unlinked cases, had a size of 4 or less (Figure 3B), although larger trees were found for large clusters (Figure 3D). The mean sampled generation time within the inferred transmission trees was 1.02 years (Figure 3C).

### Pairwise SNP distances of transmission links

To obtain a distribution of the pairwise SNP distance, we used all transmission trees found in the inference. On average, we found a total number of 231 pairs of isolates linked by a transmission event. A direct transmission event linked 44.6% of these pairs, of which 95% had a SNP distance between 0 and 4, with a median of 0 (Supplementary Figure 3). Furthermore, 22.9% of the linked pairs were infected by the same parent, of which 95% had a SNP distance between 0 and 5, with a median of 0. The other 32.9% of the linked pairs were isolates in the same transmission tree, but not directly linked or linked via the same parent. Of these pairs, 95% had



**Fig. 3.** Transmission inference results. **(a)** Fraction of unlinked TB cases in clusters. Fractions are calculated per genetic cluster size. **(b)** Distribution of the transmission tree size for all clusters. **(c)** Distribution of the sampled generation time interval, in years. The red dashed line is the mean generation time of 1.02 years. In grey: the prior generation time distribution with mean = 2.5 years as a dashed line. **(d)** Distribution of transmission tree size per genetic cluster size. The size of points represents the fraction of transmission trees in the genetic clusters of a specific size. Inlay gives the distribution of genetic cluster sizes.



a SNP distance between 0 and 4, with a median of 0. The maximal within-tree distance for 95% of all trees is between 0 and 4, with a median of 1, whereas the minimal between-tree distance is between 0 and 18 for 95% of all trees, with a median of 4 (Figure 4). Distances of 0 SNPs between trees were found because we looked at the transmission trees in all MCMC cycles.

### Assessment of SNP cut-offs

The sensitivity, i.e. the fraction of all linked pairs classified as having a link, rapidly increases for higher cut-offs and reaches 1 for a SNP cut-off of 12. A SNP cut-off of 4 yields a sensitivity of 98% while minimizing the number of falsely classified links. At a SNP cut-off of 7, the sensitivity is 100%, so we state that transmission is excluded for any isolate pair with a distance of more than 7 SNPs. No differences are found in the sensitivities calculated with direct links, infector links, or transmission tree links. For the precision, i.e. the fraction of pairs classified as linked, which are inferred to be linked, we see an effect of the different linkage types, where the classification of transmission tree links is more precise than the classification of the more stringent transmission linkage types (Figure 5).

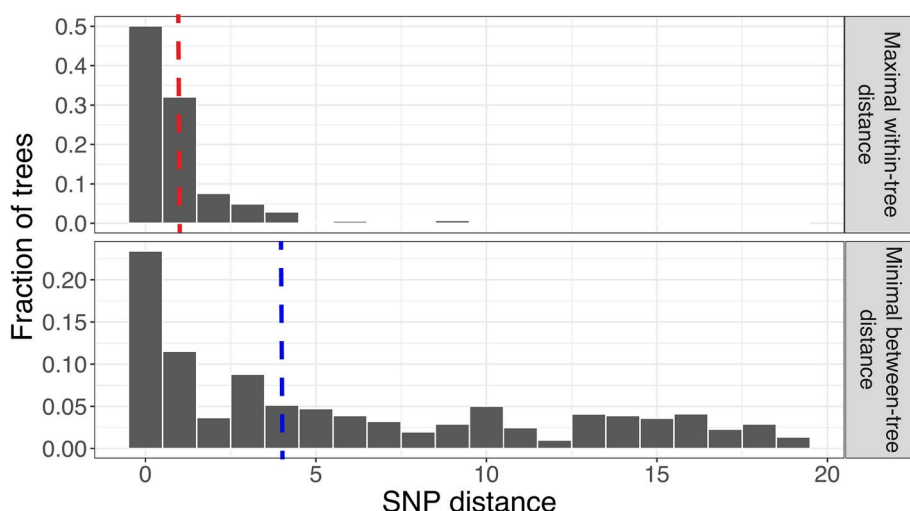
### Sensitivity analyses

Although sensitivity and precision do change when using an alternative generation time interval distribution, no other conclusions on SNP cut-off values would be drawn (Supplementary Figure 4). Furthermore, sensitivity depends on the SNPs used to calculate pair-wise distances. Using only cluster-unique SNPs yields higher sensitivity of the SNP cut-offs, reaching 98% sensitivity with a 2 SNP cut-off instead of 4 when using all SNPs in the data (Supplementary Figure 5). As an alternative for inferred transmission events from phybreak, we used the events inferred by TransPhylo to compute the sensitivity and precision (Supplementary Figure 6). The assessment results of the TransPhylo inference give higher sensitivity scores, being approximately 1 from a cut-off of 1 SNP onwards. However, the precision is an order of magnitude lower than the precision assessed with the phybreak inference results.

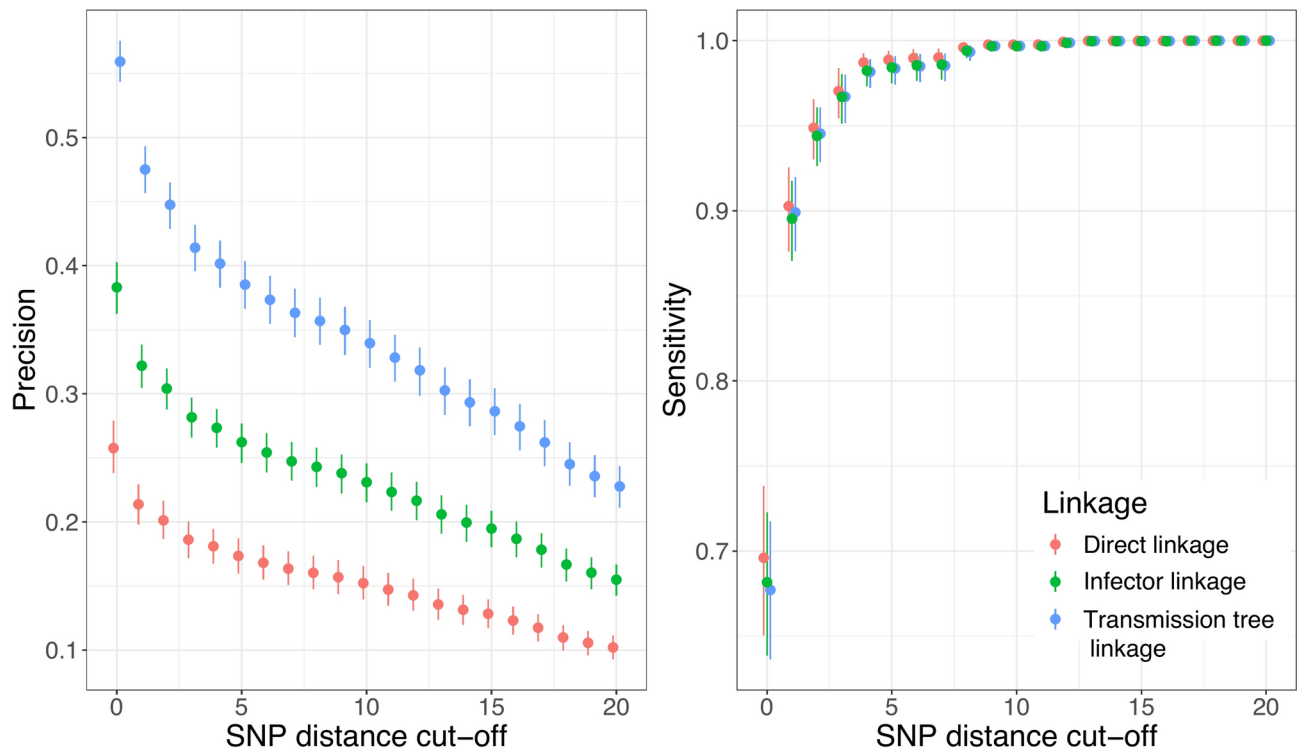
### Discussion

In the molecular surveillance of human-adapted *Mycobacterium tuberculosis* complex, clusters of possible links between cases are identified and monitored based on the number of SNPs between their sequences. However, the SNP cut-off values used to define these clusters vary and depend for some extent on the collection, interpretation, and quality of contact tracing results. To put the cut-off values derived with contact tracing data in perspective, we used inferred transmission events from phylodynamic models as an alternative reference set to estimate how well certain SNP cut-off values perform in ruling out potential transmission events or in identifying transmission clusters. Cut-offs of  $\geq 12$  SNPs yielded a sensitivity of 100%, ruling in all types of recent transmission links, i.e. transmission happened within 2 years before detection, which is consistent with literature<sup>8,9,17</sup>. In fact, our analysis suggests that a lower cut-off can be used: we can confidently rule out a transmission tree link when the SNP distance between two isolates exceeds 7. Furthermore, 98% of all transmission links will be correctly classified with a SNP cut-off of 4.

To validate our choice of phybreak as a suitable inference method, we also reconstructed transmission events with TransPhylo (Supplementary Figure 6). Despite using the same prior distributions for the generation time and sampling time intervals, the precision curve derived with TransPhylo inference results



**Fig. 4.** Distributions of maximal within-tree and minimal between-tree SNP distances. The maximal within-tree distance was calculated for each tree in all MCMC cycles per cluster. Unconnected cases are not taken into account. The minimal between-tree distance was calculated for each tree in all MCMC cycles per cluster, including unconnected cases. Medians of the within-tree distance (red) and between-tree distance (blue) are given in dashed lines



**Fig. 5.** Precision of classification of transmission links for different SNP distance cut-offs. Red: Classification of direct linkage, green: classification of direct linkage and infector linkage, blue: classification of transmission tree linkage.

was an order of magnitude lower than the curve computed from the *phybreak* inference. This discrepancy is because *TransPhylo* infers fewer transmission events. Because the sequence data do not contain many SNPs, the placement of the branching points on the time axis by *BEAST* is highly variable. When *TransPhylo* uses these phylogenetic trees, actual transmission events may not be identified if the placement of branching points prevents sampling of well-supported generation and sampling intervals. *phybreak* uses the sequence data directly so that the phylogenetic tree is also informed by the generation and sampling time intervals.

In 2020, the WHO proposed a novel conceptualization of the progression from inhalation of *Mtb* bacilli to possible clinically manifested TB disease as a continuous spectrum, traversing diverse stages based on the metabolic activity of *Mtb* and the host's innate and acquired immune responses<sup>42</sup>. The initial stage is TB infection, previous known as latent TB, where bacilli are dormant and the host is noninfectious. In the subsequent stage, incipient TB, bacilli show periods of slow metabolic activity, where hosts may be infectious<sup>43</sup>. Infectiousness emerges in the subsequent stage of subclinical TB. Here, bacilli are replicating and metabolic active, though patients do not exhibit clinical TB-related symptoms. Consequently, patients in this stage are frequently overlooked, leading to the continuous transmission of *Mtb* infection. In our study, we assumed that all sequences were sampled from active TB cases, with non-index cases traversed the previous TB stages in a short period. However, cases could have been infected decades earlier and entered the active TB stage in our study period, thus appearing unlinked in our results. Moreover, we may have missed infectors of various other reasons: they were sampled before Dec 2015, they were outside the Dutch population, or they had subclinical TB and were not sampled. As a result, genetic clusters made with a low SNP cut-off may still include unlinked cases or multiple transmission chains.

Our assessment was performed on the specific setting of a low-prevalence country with only L3 and L4 lineages of the *Mtb* complex, as our data did not contain enough samples from other lineages to assess SNP cut-offs. Lineages L1 and L2 are shown to have respectively higher and lower mutation rates than L3 and L4<sup>13</sup>, the suggested SNP cut-offs are not generalizable to these lineages. For precise cut-off assessment, a lineage specific study must be performed, in which the method described in this study can be used.

Generalization of the suggested SNP cut-offs to a high-prevalence setting may lead to very large genetic clusters with probable transmissions, because the number of cases with a similar bacterial strain may be high. As a result, the uncertainty of transmission becomes high. However, assessing smaller SNP cut-offs with *phybreak* is not possible. A limitation of the method is the need for all cases to be sampled. In a high-prevalence setting, often not all cases are sampled, if all are detected.

Migrants were shown to be the main group of TB cases in the Netherlands, a substantial proportion of which come from countries with high TB prevalence. Migrants with genetically similar TB strains and assigned to the same transmission cluster because of the suggested SNP distance cut-off, could have been infected in the Netherlands, in the country of origin, or during migration. To distinguish between these three options more



information is needed on the patients, i.e. information on migration data, contact information and detailed origin information, which is not available in our data. As a result, the number of introductions in the Netherlands may be slightly underestimated.

Besides the quality of contact tracing data, the variability of SNP cut-offs between publications can also be explained by the variety of sequencing techniques and quality control used for WGS. By following a strict quality control pipeline and only selecting cluster-unique SNPs, we aimed to get strong evidence for our transmission events. Since in practice, an isolate will be assigned to a cluster based on its SNPs, we assessed the identification of transmission clusters based on all SNPs. However, using only cluster-unique SNPs, SNP cut-offs became more sensitive for transmission events. Therefore, we believe that a second step of filtering isolates on cluster-unique SNPs will give additional information to identify likely links.

Although we assessed the SNP cut-off values independently of contact data, these data remain a valuable source of information. Campbell et al. have shown that combining contact data and genetic data leads to higher support of the inferred transmission events<sup>44</sup>. Evaluating the SNP cut-off values based on a reference set inferred with contact and genetic data could result in higher precision of the classifier. Besides contact data, the presence of more epidemiological data, such as the residency of cases and the geographical movement of migrants in the Netherlands, can serve as external controls of the transmission events, elevating the support even more. As tracking of such data is sensitive to privacy concerns, making utilization in research challenging, our assessment of SNP cut-offs without this data is valuable.

In conclusion, the optimal cut-off values found in this paper align with those found in previous research based on contact tracing. Therefore, phylodynamics is shown to be a valuable alternative for the identification of TB transmission events and the found SNP cut-offs can be used for efficient targeted contact tracing, which needs to be the next step to find true transmission events.

## Data availability

The datasets supporting the conclusions of this article are available in the BioProject repository [PRJNA882748, PRJNA1079737, PRJEB32037 and PRJEB25592, <https://www.ncbi.nlm.nih.gov/bioproject/>] and with SRA accessions found in Supplementary Table 2.

Received: 25 September 2024; Accepted: 17 March 2025

Published online: 28 March 2025

## References

- Behr, M. A., Edelstein, P. H. & Ramakrishnan, L. Revisiting the timetable of tuberculosis. *BMJ* k2738, <https://doi.org/10.1136/bmj.k2738> (2018).
- Smith, R. A., Ionides, E. L. & King, A. A. Infectious disease dynamics inferred from genetic data via sequential Monte Carlo. *Molecular Biology and Evolution* **34**, 2065–2084. <https://doi.org/10.1093/molbev/msx124> (2017).
- Bryant, J. M. et al. Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data. *BMC Infectious Diseases* **13**, 110. <https://doi.org/10.1186/1471-2334-13-110> (2013).
- Jajou, R. et al. Epidemiological links between tuberculosis cases identified twice as efficiently by whole genome sequencing than conventional molecular typing: A population-based study. *PLOS ONE* **13**, e0195413. <https://doi.org/10.1371/journal.pone.0195413> (2018).
- Nikolayevskyy, V. et al. Role and value of whole genome sequencing in studying tuberculosis transmission. *Clinical Microbiology and Infection* **25**, 1377–1382. <https://doi.org/10.1016/j.cmi.2019.03.022> (2019).
- Eldholm, V. & Balloux, F. Antimicrobial Resistance in Mycobacterium tuberculosis: The Odd One Out. *Trends in Microbiology* **24**, 637–648. <https://doi.org/10.1016/j.tim.2016.03.007> (2016).
- Fox, G. J., Barry, S. E., Britton, W. J. & Marks, G. B. Contact investigation for tuberculosis: a systematic review and meta-analysis. *European Respiratory Journal* **41**, 140–156. <https://doi.org/10.1183/09031936.00070812> (2013).
- Fiebig, L. et al. A joint cross-border investigation of a cluster of multidrug-resistant tuberculosis in Austria, Romania and Germany in 2014 using classic, genotyping and whole genome sequencing methods: lessons learnt. *Eurosurveillance* **22**, <https://doi.org/10.2807/1560-7917.ES.2017.22.2.30439> (2017).
- Walker, T. M. et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *The Lancet Infectious Diseases* **13**, 137–146. [https://doi.org/10.1016/S1473-3099\(12\)70277-3](https://doi.org/10.1016/S1473-3099(12)70277-3) (2013).
- Alvarez, G. G. et al. Molecular Epidemiology of *Mycobacterium tuberculosis* To Describe the Transmission Dynamics Among Inuit Residing in Iqaluit Nunavut Using Whole-Genome Sequencing. *Clinical Infectious Diseases* **72**, 2187–2195. <https://doi.org/10.1093/cid/ciaa420> (2021).
- Asare, P. et al. Whole Genome Sequencing and Spatial Analysis Identifies Recent Tuberculosis Transmission Hotspots in Ghana. *Frontiers in Medicine* **7**, <https://doi.org/10.3389/fmed.2020.00161> (2020).
- Chee, C. B. E. et al. Whole genome sequencing analysis of multidrug-resistant tuberculosis in Singapore, 2006–2018. *European Journal of Clinical Microbiology & Infectious Diseases* **40**, 1079–1083. <https://doi.org/10.1007/s10096-020-04100-6> (2021).
- Guerra-Assunção, J. et al. Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. *eLife* **4**, <https://doi.org/10.7554/eLife.05166> (2015).
- Jiang, Q. et al. Citywide Transmission of Multidrug-resistant Tuberculosis Under China's Rapid Urbanization: A Retrospective Population-based Genomic Spatial Epidemiological Study. *Clinical Infectious Diseases* **71**, 142–151. <https://doi.org/10.1093/cid/ci790> (2020).
- Macedo, R. et al. Evaluation of a gene-by-gene approach for prospective whole-genome sequencing-based surveillance of multidrug resistant Mycobacterium tuberculosis. *Tuberculosis* **115**, 81–88. <https://doi.org/10.1016/j.tube.2019.02.006> (2019).
- Verza, M. et al. Genomic epidemiology of Mycobacterium tuberculosis in Santa Catarina, Southern Brazil. *Scientific Reports* **10**, 12891. <https://doi.org/10.1038/s41598-020-69755-9> (2020).
- Walker, T. M. et al. Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007–12, with whole pathogen genome sequences: an observational study. *The Lancet Respiratory Medicine* **2**, 285–292. [https://doi.org/10.1016/S2213-2600\(14\)70027-X](https://doi.org/10.1016/S2213-2600(14)70027-X) (2014).
- Diel, R. et al. Accuracy of whole-genome sequencing to determine recent tuberculosis transmission: an 11-year population-based study in Hamburg, Germany. *European Respiratory Journal* **54**, 1901154. <https://doi.org/10.1183/13993003.01154-2019> (2019).

19. Nikolayevskiy, V., Kranzer, K., Niemann, S. & Drobniewski, F. Whole genome sequencing of *Mycobacterium tuberculosis* for detection of recent transmission and tracing outbreaks: A systematic review. *Tuberculosis* **98**, 77–85. <https://doi.org/10.1016/j.tube.2016.02.009> (2016).
20. Jombart, T. et al. Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. *PLoS Computational Biology* **10**, <https://doi.org/10.1371/journal.pcbi.1003457> (2014).
21. Ypma, R. J., van Ballegooijen, W. M. & Wallinga, J. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics* **195**, 1055–1062. <https://doi.org/10.1534/genetics.113.154856> (2013).
22. Didelot, X., Fraser, C., Gardy, J., Colijn, C. & Malik, H. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular Biology and Evolution* **34**, 997–1007. <https://doi.org/10.1093/molbev/msw275> (2017).
23. Ayabina, D. et al. Genome-based transmission modelling separates imported tuberculosis from recent transmission within an immigrant population. *Microbial Genomics* **4**, <https://doi.org/10.1099/mgen.0.000219> (2018).
24. Klinkenberg, D., Backer, J. A., Didelot, X., Colijn, C. & Wallinga, J. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Computational Biology* **13**, <https://doi.org/10.1371/journal.pcbi.1005495> (2017).
25. Van der Roest, B. R., Bootsma, M. C. J., Fischer, E. A. J., Klinkenberg, D. & Kretzschmar, M. E. A Bayesian inference method to estimate transmission trees with multiple introductions; applied to SARS-CoV-2 in Dutch mink farms. *PLOS Computational Biology* **19**, e1010928. <https://doi.org/10.1371/journal.pcbi.1010928> (2023).
26. Li, M. et al. Genotypic and spatial analysis of transmission dynamics of tuberculosis in Shanghai, China: a 10-year prospective population-based surveillance study. *The Lancet Regional Health - Western Pacific* **38**, 100833. <https://doi.org/10.1016/j.lanwpc.2023.100833> (2023).
27. Quan, Z. et al. Transmission of tuberculosis in rural Henan, China: a prospective population-based genomic spatial epidemiological study. *Emerging microbes & infections* **13**, 2399273. <https://doi.org/10.1080/22221751.2024.2399273> (2024).
28. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890. <https://doi.org/10.1093/bioinformatics/bty560> (2018).
29. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM (2013).
30. Walker, B. J. et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS ONE* **9**, e112963. <https://doi.org/10.1371/journal.pone.0112963> (2014).
31. Marin, M. et al. Genomic sequence characteristics and the empiric accuracy of short-read sequencing. *bioRxiv* 2021.04.08.438862, <https://doi.org/10.1101/2021.04.08.438862> (2021).
32. Freschi, L. et al. Population structure, biogeography and transmissibility of *Mycobacterium tuberculosis*. *Nature Communications* **12**, 6099. <https://doi.org/10.1038/s41467-021-26248-1> (2021).
33. Gröschel, M. I. et al. Differential rates of *Mycobacterium tuberculosis* transmission associate with host-pathogen sympatry. *Nature Microbiology* **9**, 2113–2127. <https://doi.org/10.1038/s41564-024-01758-y> (2024).
34. de Vries, G., van den Boogaard, J. & Schenk, C. Nationaal plan tuberculosebestrijding, Update 2021–2025. Tuberculosebestrijding en infectieziektebestrijding: samen sterker (2021).
35. Jombart, T. <adeget>: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405. <https://doi.org/10.1093/bioinformatics/btn129> (2008).
36. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* **29**, 1969–1973. <https://doi.org/10.1093/molbev/mss075> (2012).
37. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed Phylogenetics and Dating with Confidence. *PLoS Biology* **4**, e88. <https://doi.org/10.1371/journal.pbio.0040088> (2006).
38. Minin, V. N., Bloomquist, E. W. & Suchard, M. A. Smooth Skyride through a Rough Skyline: Bayesian Coalescent-Based Inference of Population Dynamics. *Molecular Biology and Evolution* **25**, 1459–1471. <https://doi.org/10.1093/molbev/msn090> (2008).
39. R Core Team. R: A Language and Environment for Statistical Computing (2022).
40. Pérez-Lago, L. et al. Whole Genome Sequencing Analysis of Inpatient Microevolution in *Mycobacterium tuberculosis*: Potential Impact on the Inference of Tuberculosis Transmission. *The Journal of Infectious Diseases* **209**, 98–108. <https://doi.org/10.1093/infdis/jit439> (2014).
41. Nelson, K. N. et al. Mutation of *Mycobacterium tuberculosis* and Implications for Using Whole-Genome Sequencing for Investigating Recent Tuberculosis Transmission. *Frontiers in Public Health* **9**, <https://doi.org/10.3389/fpubh.2021.790544> (2022).
42. World Health Organization. WHO consolidated guidelines on tuberculosis. Module 1: prevention - tuberculosis preventive treatment. Tech. Rep., Geneva (2020).
43. Drain, P. K. et al. Incipient and Subclinical Tuberculosis: a Clinical Review of Early Stages and Progression of Infection. *Clinical Microbiology Reviews* **31**, <https://doi.org/10.1128/CMR.00021-18> (2018).
44. Campbell, F., Cori, A., Ferguson, N. & Jombart, T. Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. *PLoS Computational Biology* **15**, 1–20. <https://doi.org/10.1371/journal.pcbi.1006930> (2019).

## Acknowledgements

This work was performed as part of the research program of the Netherlands Centre for One Health (www.ncoh.nl). BR, EF, and MK acknowledge support from the Netherlands Organization for Health Research and Development for the project "NCOH Pandemic Preparedness Research Kickstarter" (ZonMw; Grant no. 10710022210003). We thank Hester Korthals Altes for the discussions about TB transmission in the Netherlands.

## Author contributions

BR, MB, EF, MK, DK - concept and study design. BR - analyses. BR - manuscript text. MB, EF, MG, RA, RZ, MK, DK - manuscript editing. All authors read and approved the final manuscript.

## Declaration

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-94646-2>.

**Correspondence** and requests for materials should be addressed to B.R.R.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025