EBioMedicine 37 (2018) 410–416

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

EBioMedicine
Published by THE LANCET

Research paper

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

Conor J. Meehan [a,*], Pieter Moris [a,b,c], Thomas A. Kohl [d,e], Jūlija Pečerska [f], Suriya Akter [a], Matthias Merker [d,e], Christian Utpatel [d,e], Patrick Beckert [d,e], Florian Gehre [a,g,h], Pauline Lempens [a], Tanja Stadler [f], Michel K. Kaswa [a,j], Denise Kühnert [i], Stefan Niemann [d,e,1], Bouke C. de Jong [a,1]

[a] Unit of Mycobacteriology, Biomedical Sciences, Institute of Tropical Medicine, Antwerp 2000, Belgium
[b] Adrem Data Lab (Adrem), Department of Mathematics and Computer Science, University of Antwerp, Antwerp 2020, Belgium
[c] Biomedical Informatics Research Network Antwerp (biomina), University of Antwerp, Antwerp 2020, Belgium
[d] German Center for Infection Research, Partner Site Hamburg-Lübeck-Borstel-Riems, D-23845 Borstel, Germany
[e] Molecular and Experimental Mycobacteriology, Priority Area Infections, Research Center Borstel, D-23845 Borstel, Germany
[f] Swiss Institute of Bioinformatics (SIB), 1015 Lausanne, Switzerland
[g] Vaccines and Immunity Theme, Medical Research Council Unit The Gambia, Serekunda, Gambia
[h] Department Infectious Diseases Epidemiology, Bernhard Nocht Institute for Tropical Medicine, Hamburg 20359, Germany
[i] Max Planck Institute for the Science of Human History, 07745 JENA, Germany
[j] National Tuberculosis Program, Kinshasa, DR Congo

ARTICLE INFO

Article history:
Received 31 July 2018
Received in revised form 17 September 2018
Accepted 3 October 2018
Available online 16 October 2018

Keywords:
Mycobacterium tuberculosis
MDR-TB molecular epidemiology
Transmission
Spoligotyping
MIRU-VNTR
MLST
Whole genome sequencing
Outbreak detection

ABSTRACT

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

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

Findings: Our results suggest that clusters based on Spoligotyping could encompass transmission events that occurred almost 200 years prior to sampling while 24-loci-MIRU-VNTR often represented three decades of transmission. Instead, WGS based genotyping applying low SNP or cgMLST allele thresholds allows for determination of recent transmission events, e.g. in timespans of up to 10 years for a 5 SNP/allele cut-off.

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

1. Introduction

Despite the large global efforts at curbing the spread of *Mycobacterium tuberculosis* complex (Mtbc) strains, 10.4 million new patients develop tuberculosis (TB) every year [1]. In addition, the prevalence of multidrug resistant (MDR) Mtbc strains is increasing [1], predominantly through ongoing transmission within large populations [2,3]. The tracking and timing of recent transmission chains allows TB control programs to effectively pinpoint transmission hotspots and employ targeted intervention measures. This is especially important for the transmission of drug resistant strains as it appears that drug resistance may be transmitted more frequently than acquired [2]. Thus, interrupting transmission is key for the control of MDR-TB [3,4]. For the development of the most effective control strategies, there is a strong need for (i) appropriate identification of relevant transmission chains, risk factors and hotspots and (ii) robust timing of when outbreaks first arose.

* Corresponding author.
E-mail address: cmeehan@itg.be (C.J. Meehan).
[1] Equal contribution.

https://doi.org/10.1016/j.ebiom.2018.10.013

Research in context

Evidence before this study

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

Added value of this study

This study is the first to perform a comparative evaluation of cluster data defined by both classical and WGS-based *M. tuberculosis* genotyping approaches, especially with regard to transmission timing. While many studies have put forward various methods as the gold standard for *M. tuberculosis* transmission detection, we have tested clustering data generated by the different methods in a Bayesian statistical framework to elucidate the true fraction of recent transmission each approach is detecting. When specifically looking at recent transmission (e.g. <10 years previous), our results indicate that classical genotyping methods vastly over estimate recent transmission events. This solidifies the need for WGS-based methods when searching for recent outbreaks of *M. tuberculosis*.

Implications of all the available evidence

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

Epidemiological TB studies often apply genotyping methods to Mtbc strains to determine whether two or more patients are linked within a transmission chain (molecular epidemiology) [5]. Contact tracing is the primary non-molecular epidemiological method for investigating transmission networks of TB, mainly based on patient interviews [6]. Although this method is often seen as a gold standard of transmission linking, it does not always match the true transmission patterns, even in low incidence settings [7] and misses many connections [8]. The implementation of molecular genotyping and epidemiological approaches has overcome these limitations and is often used as the main approach for transmission analyses. Classical genotyping has involved IS*6110* DNA fingerprinting [9], Spoligotyping (CRISPR-based) [10], and variable-number tandem repeats of mycobacterial interspersed repetitive units (MIRU-VNTR) [11] which is the most common method at the moment [5]. The latter method is based on copy numbers of a sequence in tandem repeat patterns derived from 24 distinct loci within the genome [12]. If two patients have the same classical genotyping pattern such as a 24-loci MIRU-VNTR pattern (or up to one locus difference [12]) they are considered to be within a local transmission chain. The combination of Spoligotyping and MIRU-VNTR-typing, where patterns must match in both methods to be considered a transmission link, is often considered the molecular gold standard for transmission linking and genotyping [12]. However, examples of unlinked patients with identical patterns have been observed, suggesting that this threshold covers too broad a genetic diversity and timespan between infections [7].

The application of (whole genome) sequence (WGS)-based approaches for similarity analysis of Mtbc isolates and cluster determination is known to have high discriminatory power when assessing transmission dynamics [7,13–16], either using core genome multilocus sequence typing (cgMLST) [17,18] or SNP distances [7,14,15,19]. WGS-based approaches compare the genetic relatedness of the genomes of the clinical strains under consideration, albeit usually excluding large repetitive portions of the genome (>10% for the PE/PPE genes alone [20]), with the assumption that highly similar strains are linked by a recent transmission event [7,14]. Although many SNP cut-offs for linking isolates have been proposed [21], the most commonly employed is based on the finding that a 5 SNP cut-off will cluster the genomes of strains from the majority of epidemiologically linked TB patients, with an upper bound of 12 SNPs between any two linked isolates [14]. The emerging widespread use of WGS has quickly pushed these cut-offs to be considered the new molecular gold standard of recent transmission linking, although SNP distances may vary for technical reasons (e.g. assembly pipelines or filter criteria [22]) and between study populations e.g. high and low incidence settings [19].

In addition to cluster detection, uncovering the timing of transmission events within a given cluster is highly useful information for TB control e.g. for assessing the impact of interventions on the spread of an outbreak or uncovering when MDR-TB transmission first emerged in a particular setting. Accordingly, knowledge of the rate change associated with different genotyping methods is essential for correct timing. The whole genome mutation rate of Mtbc strains has been estimated by several studies as between 10^{-7} and 10^{-8} substitutions per site per year or ~0·3–0·5 SNPs per genome per year [7,14,23–25], while the rate of change in the MIRU-VNTR loci specifically is known to be quicker (~10^{-3}) [26,27]. Since these mutation rates have been shown to also vary by lineage [24,28] and over short periods of time [23], such variation needs to be accounted for when estimating transmission times, e.g. by using Bayesian phylogenetic dating techniques [3,23,26].

Considering the multiple genotyping methods currently available, many of them proposed as a “gold standard”, there is an urgent need to precisely define the individual capacity of each method to accurately detect recent transmission events and perform timing of outbreaks. To provide this essential information, this study harnesses the power of WGS-based phylogenetic dating methods to assign timespans onto Mtbc transmission chains encompassed by the different genotypic clustering methods commonly used in TB transmission studies.

2. Materials and methods

2.1. Dataset, ethical approval and sequencing

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

2.2. Genome reconstruction

The MTBseq pipeline [29] was used to detect the SNPs for each isolate using the H37Rv reference genome (NCBI accession number NC000962.3) [30]. Unambiguous allele calls were based on the following parameters: four forward and four reverse reads indicating the allele, four reads indicating the allele with a phred score of 20 and a 75% allele frequency. All samples had over 95% coverage of H37Rv (median of 98%) with genome depth ranging from 54× to 290× (median of 160.5×). For creation of the SNP alignments, genes known to be involved in drug resistance (as outlined in the PhyResSE list of drug mutations v27 [31]) were excluded from the alignment and additional filtering of sites with ambiguous calls in >5% of isolates and those SNPs within a 12 bp window of each other was also applied.

2.3. Transmission cluster estimation methods

Six standard transmission clustering approaches were chosen for comparison and analysis: Spoligotyping, MIRU-VNTR, Spoligotyping + MIRU-VNTR, SNP-based clustering and cgMLST-based clustering. The latter two approaches were undertaken at 3 different cut-offs (1, 5 and 12 SNPs/alleles). The total SNP distances were calculated, per method, to investigate the range of variability encompassed within each cluster. Maximum SNP distances were derived from pairwise comparisons of isolates within the SNP alignment using custom python scripts. A clustering rate was calculated for each method using the formula $(n_c - c)/n$, where n_c is the total number of isolates clustered by a given method, c is the number of clusters, and n is the total number of isolates in the dataset ($n = 324$).

2.4. Spoligotyping and MIRU-VNTR

Spoligotype patterns were obtained from membranes following the previously published protocol [10]. Isolates were said to be clustered if all 43 spacers matched. Genotyping by MIRU-VNTR was undertaken as previously described [12]. 2 μl of DNA was extracted from cultures and amplified using the 24 loci MIRU-VNTR typing kit (Genoscreen, Lille, France). Analysis of patterns was undertaken using the ABI 3500 automatic sequencer (Applied Biosystems, California, USA) and Genemapper software (Applied Biosystems). Isolates were said to be clustered if all 24 loci matched. Mixed MIRU-VNTR patterns were observed in 18 isolates although this mixing was not observed in the WGS data, likely due to subculturing for sequencing. MIRU-VNTR patterns were also combined with spoligotyping patterns for additional refinement of clusters. Isolates were clustered if both the spoligotyping pattern and the 24 loci MIRU-VNTR pattern matched. Spoligotyping and MIRU-VNTR patterns are available on figshare [32,33].

2.5. SNP and cgMLST cut-off clustering

In this study, we employed the widely used 5 SNP (proposed by Walker et al. [14] as the likely boundary for linked transmission) and 12 SNP cut-offs (proposed maximum boundary) for cluster definition. Additionally, we employed a lower cut-off of 1 SNP to look for clusters of very highly related isolates. Pairwise SNP distances were calculated between all isolates. A loose cluster definition was used, where every isolate in a cluster at most the SNP cut-off from at least 1 other isolate in the cluster.

An alternative approach to clustering using WGS data is the concept of core genome MLST (cgMLST) patterns [17,18]. BAM files for all isolates are input into Ridom SeqSphere+ software (Ridom GmbH, Münster, Germany) to compile an allelic distance matrix based on the cgMLST v2 scheme consisting of 2891 core Mtbc genes [18]. Loose clusters were then defined using allelic differences of 1, 5 and 12 as cut-offs. These methods are referred to as 1/5/12 cgMLST respectively.

2.6. Estimation of transmission times

To estimate the age and timespan of potential transmission clusters, SNP alignments were created for the four primary clustering types: Spoligotyping, MIRU-VNTR, 12 SNP clusters and 12 allele cgMLST clusters.

A Bayesian approach to transmission time estimation was then undertaken. Each cluster methods alignment was separately input to BEAST-2 v2.4.7 [34] to create a time tree for those isolates. These phylogenies were built using the following priors: GTR + GAMMA substitution model, a log-normal relaxed molecular clock model to account for variation in mutation rates [35] and coalescent constant size demographic model [36], which assumes a low sampling proportion, as observed here [37]. This combination of parameters has been tested previously within a Bayesian framework and been shown to be suitable for lineage 4 isolates [19,25,38,39], including in Brazzaville, the city neighbouring Kinshasa in the Republic of the Congo [40]. The MCMC chain was run six times independently per alignment with a length of at least 400 million, sampled every 40,000th step (Spoligotyping: 400 M; MIRU: 700 M; 12 SNP and cgMLST: 500 M). A log normal prior (mean 1.5×10^{-7}; variance $1 \cdot 0$) was used for the clock model to reflect the previously estimated mutation rate of *M. tuberculosis* lineage 4 [7,14,23–25], while allowing for variation as previously suggested [23]. A 1/X non-informative prior was selected for the population size parameter of the demographic model. Isolation dates were used as informative heterochronous tip dates and the SNP alignment was augmented with a count of invariant sites for each of the four nucleotide bases to avoid ascertainment bias [41]. Tracer v1.6 was used to determine adequate mixing and convergence of chains (effective sample sizes (ESS) >200 for all except Spoligotyping with ESS >100) after a 25% burn-in. The chains were combined via LogCombiner v2.4.8 [34] to obtain a single chain for each clustering type with high (>700) ESS. The tree samples were combined in the same manner and resampled at a lower frequency to create thinned samples of (minimum) 20,000 trees. Tip date randomisation was undertaken to check for temporal signal of the data. The R package 'TipDatingBeast' [42] was used to randomly reassign tip dates across the 12 SNP-based alignment. Ten repetitions were undertaken and BEAST-2 run as above. Rate mean and tree heights differed significantly between the random date and true dataset log files, suggesting a sufficient temporal signal was present in the data.

The algorithm for estimating the timespan of transmission events encompassed by each method is outlined in Supplemental Fig. 1. Briefly, for each cluster created by the given method, we defined the MRCA node as the internal node that connects all taxa in that cluster. The

youngest node was then defined as the tip that is furthest from this MRCA within the clade (i.e. the tip descendant from that node that was sampled closest to the present time). To better account for changes in the mutation rate over short periods [23], all trees estimated and sampled during the Bayesian MCMC process were used instead of only a single summary phylogeny. For each retained tree in the MCMC process, the difference in age between the MRCA node and youngest node was calculated. This gave a distribution of likely maximum transmission event times within that cluster. For each method, these per-cluster aggregated ages were then combined across all clusters to give a per-method distribution of transmission event times represented by the clusters. The 95% Highest Posterior Density (HPD) interval of these distributions was calculated with the LaplacesDemon [43] p.interval function in R v3.4.0 [44].

3. Results

In this study, we assessed five different approaches for generating putative *M. tuberculosis* transmission clusters: Spoligotyping, MIRU-VNTR, Spoligotyping and MIRU-VNTR, SNP-based clustering using a 12, 5 and 1 SNP cut-off, and cgMLST allele clustering with 12, 5 and 1 allele cut-offs, using a dataset of 324 isolates collected 2005–2010 in Kinshasa, Democratic Republic of Congo (DRC). The dataset contained 309 L4 and 15 L5 isolates, with a maximum of 1671 SNPs between any two isolates. Bayesian phylodynamic dating approaches implemented in BEAST-2 [34] were then utilised to assign timespans to the transmission events estimated by each genotyping method.

As expected, classical genotyping methods clustered the most strains, with the lowest resolution (i.e. highest clustering rate) (Fig. 1, Table 1). WGS-based methods had by far the highest discriminatory power and low SNP cut-offs grouped isolates into smaller clusters (e.g. 2–10 isolates per cluster for a 5 SNP cut-off) (Table 1, Fig. 1). The high percentage of strains in a 12 SNP cluster (75%) suggests high levels of transmission in this population, making is suitable for further transmission analyses, despite the estimated low sampling proportion (2% based on demographic data).

Bayesian phylogenetic dating of the timeframe associated with particular transmission chains showed large differences in estimated cluster ages between the different genotyping approaches used (Table 1), correlating well with the difference in discriminatory power. Cluster ages are defined here as the most ancient transmission event that links any two isolates within a specific cluster (see methods and supplemental Fig. 1). Thus, in phylogenetic terms, the cluster age is the difference in time between when the most recent common ancestor (MRCA) of the entire cluster existed and the date of isolation of the furthest isolate from this ancestor.

The aggregate median ages of clusters derived from Spoligotyping were found to often be several hundred years old (median 178 years

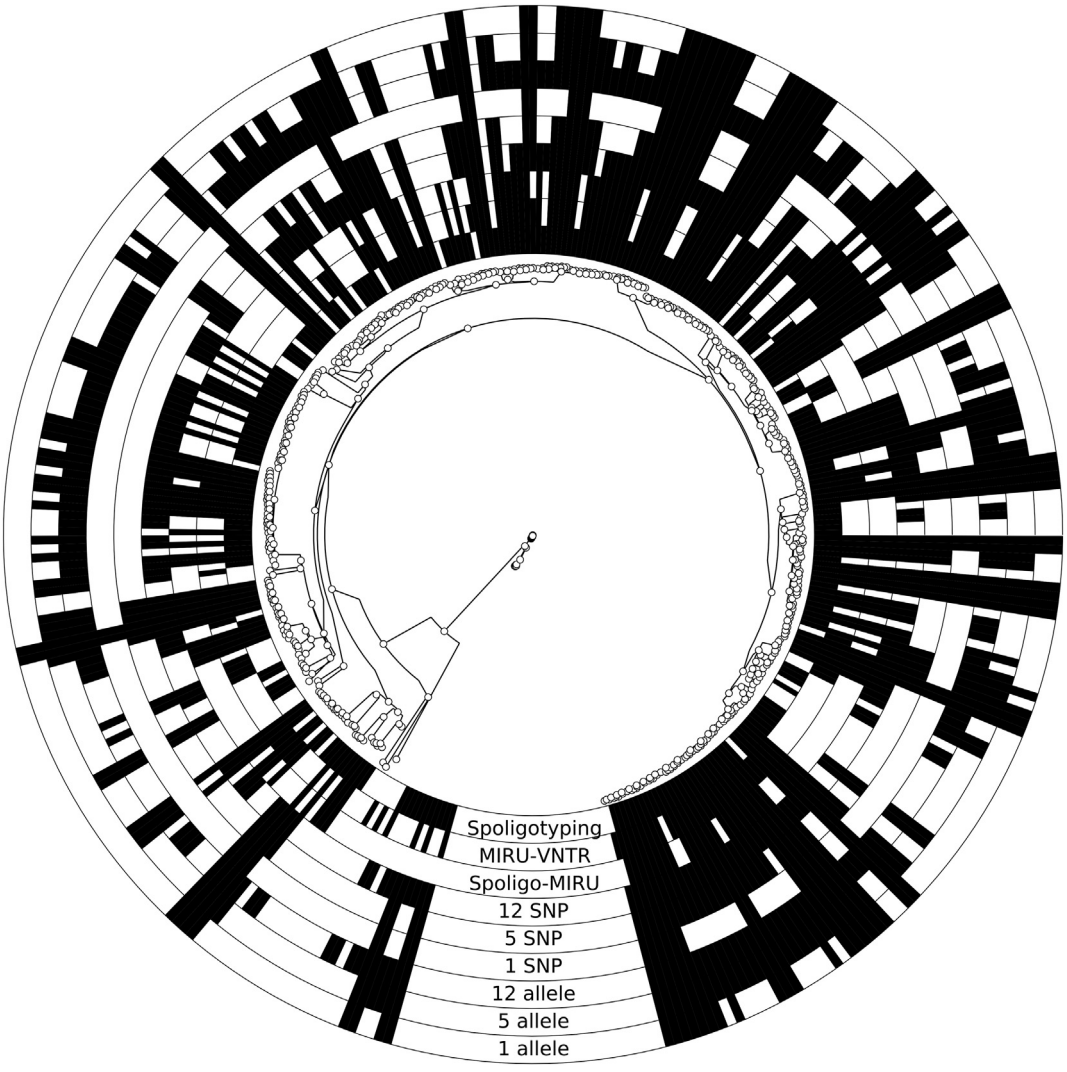

Fig. 1. Clustering of *M. tuberculosis* isolates. For each approach the inclusion of an isolate into a cluster is outlined in the surrounding circles using GraPhlAn [59]. The ML phylogenetic tree was created using RAxML-NG [60] (see supplemental material) and is rooted between L4 and L5 isolates.

Table 1
Clustering method overview for each clustering method, the general features are outlined in the table. Median ages and 95% HPD ranges are based upon the BEAST-2 estimates of clade heights (see methods).

Method	Strains in clusters	Number of clusters	Percent of strains in clusters	Cluster sizes	Maximum SNP distances	Clustering rate	Mean timespan	Timespan 95% HPD
Spoligotyping	276	33	85.19	2–39	1–685	0.75	178.35	0.34–7747
MIRU-VNTR	207	38	63.89	2–30	0–611	0.5216	35.58	0–1830
Spoligo-MIRU	174	36	53.7	2–25	0–611	0.4259	36.38	0–1969
12 SNP cluster	242	47	74.69	2–34	0–23	0.6019	23.63	0–102.58
5 SNP cluster	147	40	45.37	2–27	0–10	0.3302	10.86	0–47.07
1 SNP cluster	74	29	22.84	2–6	0–2	0.1389	3.91	0–23.54
12 allele cgMLST	254	45	78.4	2–39	0–51	0.6451	24.06	0–112.25
5 allele cgMLST	173	42	53.4	2–28	0–22	0.4043	13.4	0–68.53
1 allele cgMLST	80	31	24.69	2–6	0–4	0.1512	4.73	0–24.65

(95% HPD: 0–7747)) (Table 1). MIRU-VNTR clustering encompassed more recent transmission events than Spoligotyping, but were still found to be often over three decades old (median 36 years (95% HPD: 0–1830)). The combination of MIRU-VNTR and Spoligotyping resulted in cluster ages similar to MIRU-VNTR alone (Table 1). Clusters based on SNP cut-offs correlated to 23 years using a 12 SNP cut-off (95% HPD: 0–103), 11 years using a 5 SNP cut-off (95% HPD: 0–47), and 4 years using a 1 SNP cut-off (95% HPD: 0–24) (Table 1). Cluster sizes and ages based on cgMLST alleles were similar to the SNP-based clusters (Table 1).

4. Discussion

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

As the most extreme example, Spoligotyping-derived clusters were associated with transmission events that can be several hundred years old. This is due to the low discriminatory power coupled with the high rate of convergent evolution (the same spoligotype pattern found in phylogenetically distant isolates). When convergent patterns are removed, the median and maximum transmission ages drop dramatically (see Supplementary table 1). However, in practise, such pattern removal is impossible without WGS data. Thus, these findings add weight to the previous suggestion that this technique is not suitable for recent transmission studies [45], although may be of use as a low-cost method of sorting Mtbc strains into the seven primary lineages [46,47]. The transmission times encompassed by MIRU-VNTR clusters often spanned over three decades (Table 1), confirming previous studies showing over-estimation of recent transmission with this method [7,13,19,48]. In line with previous findings [45,49], convergent evolution of 24-loci MIRU-VNTR patterns was rarer than observed for Spoligotyping, but did occur in 16% of MIRU-VNTR-based clusters. Removal of such convergent patterns did not drastically change the median transmission ages for MIRU-VNTR (36 vs 26 years) but did affect the maximum ages (Supplementary table 1). As with Spoligotyping, such patterns cannot be easily detected and thus the impact of convergence in other datasets cannot be estimated. Combination of these two classical methods was similar to MIRU-VNTR alone, further limiting the use of Spoligotyping for molecular epidemiology.

For defining transmission events that occurred in more recent time frames before sampling, WGS-based methods were found to be better suited than classical genotyping methods (Table 1). The 12 SNP cut-off, currently the recommended upper bound for clustering isolates, often defines transmission events that occurred on average two decades prior to sampling, slightly younger in median age to clusters estimated by MIRU-VNTR, but also drastically more recent in maximum ages. This suggests that the 12 SNP cluster method may be a good replacement for MIRU-VNTR as it detects larger transmission networks spanning similar transmission time periods but is less affected by convergent evolution. Isolates clustered at a low (5 SNP) or nearly identical (1 SNP) cut-off were found to represent transmission events occurring over a time span of up to ten years. These findings correlate well with previous studies where confirmed contact tracing-based epidemiological links were found between patients that were two [15,50] and three [7] SNPs apart. The original paper that proposed the 5 and 12 SNP cut-offs found that serial isolates that were 10 years apart differed by, on average, 6 SNPs, also agreeing with the findings presented here [14]. Comparisons between the SNP-based (using almost all genomic differences) and the cgMLST-based (using a defined core set of genes) methods demonstrated that the latter approach gives similar estimates to full SNP approaches. This supports the use of low SNP or cgMLST differences for detection or exclusion of very recent transmission, although basing clustering on such low numbers of SNPs makes robust identification of transmission direction difficult.

The mutation rate of *M. tuberculosis* has been estimated to be between 10^{-7} and 10^{-8} substitutions per site per year [3,7,24]. Within the Bayesian analysis employed here, the mutation rate was free to vary between these values but was found to strongly favour ~3 $\times 10^{-8}$ (ESS > 1000 for all runs; 95% HDP: 4×10^{-9} - 8×10^{-8}), translating to approximately 0·13 SNPs per genome per year (95% HDP: 0.017 - 0.35). While the mutation rate used here is in line with previous estimates for lineage 4 [24] (which most of this dataset is comprised of), it may be similar in other lineages, although this has only been shown for lineage 2 [3,24]. Thus, per-lineage estimates are required for all seven lineages to ensure similar transmission times are linked to genotyping methods across the whole diversity of the Mtbc.

While this study has many advantages due to its five year population based design in an endemic setting coupled with the application of three different genotyping methods (Spoligotyping, 24-locus MIRU-VNTR and WGS), future confirmatory studies could address the following drawbacks that are inherent to genomic epidemiology [16,22]: 1) studies employing contact tracing and/or digital epidemiology [51] in conjunction with these genotyping methods can help confirm transmission times associated with different clusters and increase the sampling proportion (although these methods also have many limitations); 2) as outlined above, strains of other lineages of the Mtbc should be analysed in a similar fashion to ensure transferability of findings

across the entire complex; 3) a broad range of drug resistance profiles should be included to fully assess the impact of such mutations on transmission estimates; 4) improved WGS methods, such as directly from clinical samples to help reduce culture biases [52] and longer reads (e.g. PacBio SMRT or Nanopore MinION) to capture the entire genome, including repetitive regions such as PE/PPE genes known to impact genome remodelling [53,54], will ensure that the maximum diversity between isolates is captured; 6) extensive panels of Spoligotyping and MIRU-VNTR results paired with WGS data will help assess the extent of convergence in these methods and better correlate their clusters with those of low SNP thresholds and 7) standardised SNP calling pipelines appropriate across all lineages, with high true positive/low false negative rates, will ensure that Mtbc molecular epidemiology can be uniformly implemented and comparable across studies. Additionally, extensions of the current WGS-based strategies, such as including within-patient diversity [55,56] (may be missed by single colony picking for WGS) or counting inferred transmissions instead of SNPs [57] are required to truly understand the underlying dynamics of the *M. tuberculosis* transmission network.

Since each method was found to represent different timespans and clustering definitions, they can be used in a stratified manner in an integrated epidemiological and public health investigation addressing the transmission of Mtbc strains. For instance, although Spoligotyping clusters represented potentially very old transmission events, the low associated cost and its ability to be applied directly on sputum helps reduce culture bias and thus robustly assign lineages. This may aid public health officials in high burden settings understand (changes in) the population structure of the MTBc lineages, including ruling out instances of relapse or laboratory contamination in case patterns differ. However, due to the problems outlined above, the usefulness of this method in public health initiatives is limited. MIRU-VNTR may serve well as first-line surveillance of potential transmission events in the population, guiding further investigations and resource allocations. Although with the ever decreasing cost and increasing speed of WGS methods, the expense and workload of MIRU-VNTR makes it difficult to justify over the vast increase in data gained from genomics.

If classical genotyping methods are employed, any potential transmission hotspots should then be further investigated with contact tracing and/or WGS. Employment of different cut-offs and clustering approaches to WGS data can then address several questions. The 12 SNP/cgMLST allele cluster approaches serve well for high level surveillance targeting larger (older) transmission networks, akin to what is currently often done using MIRU-VNTR (e.g. [15,58]). Recent transmission events can then be detected through employment of low SNP cut-offs (e.g. 5 SNPs for transmission in the past 10 years or 1 SNPs for transmission in the past 5 years). In high incidence/low diversity settings where amalgamation of clusters may inadvertently obscure distinct hotspots of transmission at different time points, subdivision into distinct time-dependant clusters can be undertaken using the algorithm presented in such a study in East Greenland [19].

Overall, phylodynamic approaches applied to whole genome sequences, as undertaken here, are recommended to fully investigate the specific transmission dynamics within a study population to account for setting-specific conditions, such as low/high TB incidence, low/high pathogen population diversity, and sparse/dense sampling fractions. As WGS methods become more commonplace and easier to implement in a variety of settings, each genotyping method can be employed as part of an overall evidence gathering program for transmission, placing molecular epidemiological approaches as an integral part in tracking and stopping the spread of TB.

Acknowledgements

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

Funding sources

This work was supported by an ERC grant [INTERRUPTB; no. 311725] to BdJ, FG and CJM; an ERC grant to TS [PhyPD; no. 335529]; an FWO PhD fellowship to PM [grant number 1141217N]; the Leibniz Science Campus EvolLUNG for MM and SN; the German Centre for Infection Research (DZIF) for TAK, MM, CU, PB and SN; a SNF SystemsX grant (TBX) to JP and TS and a Marie Heim-Vögtlin fellowship granted to DK by the Swiss National Science Foundation. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government – department EWI.

Declaration of interests

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

Author contributions

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ebiom.2018.10.013.

References

[1] WHO. Global TB Rep 2017:2018.
[2] Kendall EA, Fofana MO, Dowdy DDW, Who, Dye C, Garnett G, et al. Burden of transmitted multidrug resistance in epidemics of tuberculosis: a transmission modelling analysis. Lancet Respir Med 2015 Nov 12;3(12):963–72.
[3] Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, Willery E, et al. Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage. Nat Genet 2015 Jan 19;47(3):242–9.
[4] Shah NS, Auld SC, Brust JCM, Mathema B, Ismail N, Moodley P, et al. Transmission of extensively drug-resistant tuberculosis in South Africa. N Engl J Med 2017 Jan 19;376(3):243–53.
[5] Merker M, Kohl TA, Niemann S, Supply P. The evolution of strain typing in the Mycobacterium tuberculosis complex. Advances in Experimental Medicine and Biology; 2017. p. 43–78.
[6] Fox GJ, Barry SE, Britton WJ, Marks GB. Contact investigation for tuberculosis: a systematic review and meta-analysis. Eur Respir J 2012;41(1).
[7] Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, et al. Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. (Neyrolles O, editor) PLoS Med 2013 Jan 12;10(2):e1001387.
[8] Bjorn-Mortensen K, Lillebaek T, Koch A, Soborg B, Ladefoged K, Sørensen HCF, et al. Extent of transmission captured by contact tracing in a tuberculosis high endemic setting. Eur Respir J 2017;49(3).
[9] Thierry D, Cave MD, Eisenach KD, Crawford JT, Bates JH, Gicquel B, et al. IS6110, an IS-like element of Mycobacterium tuberculosis complex. Nucleic Acids Res 1990 Jan 11;18(1):188.
[10] Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, et al. Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. J Clin Microbiol 1997 Apr;35(4):907–14.
[11] Supply P, Magdalena J, Himpens S, Locht C. Identification of novel intergenic repetitive units in a mycobacterial two-component system operon. Mol Microbiol 1997 Dec;26(5):991–1003.
[12] Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rusch-Gerdes S, Willery E, et al. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of mycobacterium tuberculosis. J Clin Microbiol 2006 Dec 1;44(12):4498–510.
[13] Wyllie DH, Davidson JA, Grace Smith E, Rathod P, Crook DW, Peto TEA, et al. A quantitative evaluation of MIRU-VNTR Typing against whole-genome sequencing for identifying mycobacterium tuberculosis transmission: a prospective observational cohort study. EBioMedicine 2018 Aug;34:122–30.

[14] Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis 2013 Feb;13(2):137–46.
[15] Walker TM, Merker M, Knoblauch AM, Helbling P, Schoch OD, van der Werf MJ, et al. A cluster of multidrug-resistant Mycobacterium tuberculosis among patients arriving in Europe from the Horn of Africa: a molecular epidemiological study. Lancet Infect Dis 2018 Jan;8.
[16] Comas I. Genomic Epidemiology of Tuberculosis. Cham: Springer; 2017; 79–93.
[17] Kohl TA, Diel R, Harmsen D, Rothgänger J, Walter KM, Merker M, et al. Whole-genome-based Mycobacterium tuberculosis surveillance: a standardized, portable, and expandable approach. J Clin Microbiol 2014 Jul 1;52(7):2479–86.
[18] Kohl TA, Harmsen D, Rothgänger J, Walker T, Diel R, Niemann S. Harmonized genome wide typing of tubercle bacilli using a web-based gene-by-gene nomenclature system. EBioMedicine 2018 Jul 30;0(0).
[19] Bjorn-Mortensen K, Soborg B, Koch A, Ladefoged K, Merker M, Lillebaek T, et al. Tracing mycobacterium tuberculosis transmission by whole genome sequencing in a high incidence setting: a retrospective population-based study in East Greenland. Sci Rep 2016 Sep 12;6:33180.
[20] Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of mycobacterium tuberculosis from the complete genome sequence. Nature 1998 Jun 11;393(6685):537–44.
[21] Hatherell H-A, Colijn C, Stagg HR, Jackson C, Winter JR, Abubakar I. Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review. BMC Med 2016 Dec 23;14(1):21.
[22] Guthrie JL, Gardy JL. A brief primer on genomic epidemiology: lessons learned from Mycobacterium tuberculosis. Ann N Y Acad Sci 2017 Jan 1;1388(1):59–77.
[23] Bryant JM, Schürch AC, van Deutekom H, Harris SR, de Beer JL, de Jager V, et al. Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data. BMC Infect Dis 2013 Feb 27;13(1):110.
[24] Duchêne S, Holt KE, Weill F-X, Le Hello S, Hawkey J, Edwards DJ, et al. Genome-scale rates of evolutionary change in bacteria. Microb Genomics 2016 Nov;2(11): e000094.
[25] Eldholm V, Monteserin J, Rieux A, Lopez B, Sobkowiak B, Ritacco V, et al. Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. Nat Commun 2015 May 11;6:7119.
[26] Wirth T, Hildebrand F, Allix-Béguec C, Wölbeling F, Kubica T, Kremer K, et al. Origin, spread and demography of the mycobacterium tuberculosis complex. Achtman M, editor PLoS Pathog 2008 Sep 26;4(9):e1000160.
[27] Ragheb MN, Ford CB, Chase MR, Lin PL, Flynn JL, Fortune SM. The mutation rate of mycobacterial repetitive unit loci in strains of M. tuberculosis from cynomolgus macaque infection. BMC Genomics 2013 Mar 5;14(1):145.
[28] Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T, et al. Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. Nat Genet 2013 Jul;45(7): 784–90.
[29] Kohl TA, Utpatel C, Schleusener V, De Filippo MR, Beckert P, Cirillo DM, et al. MTBseq: A Comprehensive Pipeline for Whole Genome Sequence Analysis of Mycobacterium Tuberculosis Complex Isolates. PeerJ; 2018 (in press).
[30] Lew JM, Kapopoulou A, Jones LM, Cole ST. TubercuList – 10 years after. Tuberculosis 2011;91(1):1–7.
[31] Feuerriegel S, Schleusener V, Beckert P, Kohl TA, Miotto P, Cirillo DM, et al. PhyResSE: a web tool delineating mycobacterium tuberculosis antibiotic resistance and lineage from whole-genome sequencing data. (Carroll KC, editor) J Clin Microbiol 2015 Jun; 53(6):1908–14.
[32] Meehan CJ, de Jong BC. Membrane Spoligotype Mtbc Kinshasa 2005–2010; 2018.
[33] Meehan CJ, de Jong BC. [dataset] MIRU-VNTR Mtbc Kinshasa 2005–2010; 2018.
[34] Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, et al. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. (Prlic A, editor) PLoS Comput Biol 2014 Apr 10;10(4):e1003537.
[35] Drummond AJ, Ho SYW, Phillips MJ, Rambaut A, Rambaut A. Relaxed phylogenetics and dating with confidence. (Penny D, editor) PLoS Biol 2006 Mar 14;4(5):e88.
[36] Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from Molecular Sequences. Mol Biol Evol 2005 Feb 9;22 (5):1185–92.
[37] Stadler T, Vaughan TG, Gavryushkin A, Guindon S, Kühnert D, Leventhal GE, et al. How well can the exponential-growth coalescent approximate constant-rate birth–death population dynamics? Proc R Soc Lond B Biol Sci 2015;282(1806).
[38] Lee RS, Radomski N, Proulx J-F, Levade I, Shapiro BJ, McIntosh F, et al. Population genomics of Mycobacterium tuberculosis in the Inuit. Proc Natl Acad Sci U S A 2015 Nov 3;112(44):13609–14.
[39] Didelot X, Fraser C, Gardy J, Colijn C. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. Mol Biol Evol 2017 Jan 18;34(4) (msw075).
[40] Malm S, Linguissi LSG, Tekwu EM, Vouvoungui JC, Kohl TA, Beckert P, et al. New *Mycobacterium tuberculosis* complex Sublineage, Brazzaville. Congo Emerg Infect Dis 2017 Mar;23(3):423–9.
[41] Leaché AD, Banbury BL, Felsenstein J, de Oca A Nieto-M, Stamatakis A, et al. Short tree, long tree, right tree, wrong tree: new acquisition bias corrections for inferring SNP phylogenies. Syst Biol 2015 Nov;64(6):1032–47.
[42] Rieux A, Khatchikian C. TipDatingBeast: Using Tip Dates with Phylogenetic Trees in BEAST (Software for Phylogenetic Analysis); 2018.
[43] Staticat LLC. LaplacesDemon: Complete Environment for Bayesian Inference. Bayesian-Inference.com. (R package version 16.0.1. [Internet]. Available from)2016: https://cran.r-project.org/web/packages/LaplacesDemon/citation.html
[44] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; 2017 (Vienna, Austria).
[45] Comas I, Homolka S, Niemann S, Gagneux S. Genotyping of genetically monomorphic bacteria: DNA sequencing in Mycobacterium tuberculosis highlights the limitations of current methodologies. PLoS One 2009 Nov 12;4(11):e7815.
[46] Kato-Maeda M, Gagneux S, Flores LL, Kim EY, Small PM, Desmond EP, et al. Strain classification of Mycobacterium tuberculosis: congruence between large sequence polymorphisms and spoligotypes. Int J Tuberc Lung Dis 2011 Jan;15 (1):131–3.
[47] Filliol I, Motiwala AS, Cavatore M, Qi W, Hazbón MH, Bobadilla del Valle M, et al. Global phylogeny of Mycobacterium tuberculosis based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. J Bacteriol 2006 Jan 15;188(2):759–72.
[48] Stucki D, Ballif M, Egger M, Furrer H, Altpeter E, Battegay M, et al. Standard genotyping overestimates transmission of Mycobacterium tuberculosis among immigrants in a low-incidence country. J Clin Microbiol 2016 Jul 1;54(7): 1862–70.
[49] Scott AN, Menzies D, Tannenbaum T-N, Thibert L, Kozak R, Joseph L, et al. Sensitivities and specificities of spoligotyping and mycobacterial interspersed repetitive unit-variable-number tandem repeat typing methods for studying molecular epidemiology of tuberculosis. J Clin Microbiol 2005 Jan;43(1):89–94.
[50] Walker TM, Lalor MK, Broda A, Ortega LS, Morgan M, Parker L, et al. Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007–12, with whole pathogen genome sequences: an observational study. Lancet Respir Med 2014 Apr;2(4):285–92.
[51] Salathé M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, et al. Digital Epidemiology. PLoS Comput Biol 2012 Jul 26;8(7):e1002616.
[52] Sanoussi CN, Affolabi D, Rigouts L, Anagonou S, de Jong B. Genotypic characterization directly applied to sputum improves the detection of Mycobacterium africanum West African 1, under-represented in positive cultures. (Picardeau M, editor) PLoS Negl Trop Dis 2017 Sep 1;11(9):e0005900.
[53] Phelan JE, Coll F, Bergval I, Anthony RM, Warren R, Sampson SL, et al. Recombination in pe/ppe genes contributes to genetic variation in Mycobacterium tuberculosis lineages. BMC Genomics 2016 Feb 29;17:151.
[54] Ates LS, Dippenaar A, Ummels R, Piersma SR, van der Woude AD, van der Kuij K, et al. Mutations in ppe38 block PE_PGRS secretion and increase virulence of Mycobacterium tuberculosis. Nat Microbiol 2018 Feb 15;3(2):181–8.
[55] Casali N, Broda A, Harris SR, Parkhill J, Brown T, Drobniewski F. Whole genome sequence analysis of a large isoniazid-resistant tuberculosis outbreak in London: A Retrospective Observational Study. (Metcalfe JZ, editor) PLoS Med 2016 Oct 4;13 (10):e1002137.
[56] Martin M, Lee RS, Cowley LA, Gardy JL, Hanage WP. Within-host diversity and its utility for Mycobacterium tuberculosis transmission inferences. Microb Genomics October 2018(10) (in press).
[57] Stimson J, Gardy JL, Mathema B, Crudu V, Cohen T, Colijn C. Beyond the SNP Threshold: Identifying Outbreak Clusters Using Inferred Transmissions bioRxiv, 319707; 2018 May 10.
[58] Guthrie JL, Kong C, Roth D, Jorgensen D, Rodrigues M, Hoang L, et al. Molecular epidemiology of tuberculosis in British Columbia, Canada: a 10-year retrospective study. Clin Infect Dis 2018 Mar 5;66(6):849–56.
[59] Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ 2015 Jun 18;3: e1029.
[60] Kozlov A. RAxML-NG; 2017. https://doi.org/10.5281/zenodo.593079.