Genome-based transmission modeling separates imported tuberculosis from recent transmission within an immigrant population

November 29, 2017

Diepreye Ayabina[1] Janne O Rønning[2] Kristian Alfsnes[2] Nadia Debech[2] Ola B Brynildsrud[2] Trude Arnesen[2] Gunnstein Norheim[2] Anne-Torunn Mengshoel[2] Rikard Rykkvin[2] Ulf R Dahle[2] Caroline Colijn[1]* Vegard Eldholm[2]*

[1] *Department of Mathematics, Imperial College London, South Kensington Campus, London SW7 2AZ, UK*

[2] *Infection Control and Environmental Health, Norwegian Institute of Public Health, Lovisengerggata 8, 0456 Oslo, Norway*

Abstract

In many Western countries tuberculosis (TB) incidence is low and largely shaped by immigrant populations originating from high-burden countries. A variable latent period, low rates of evolution and structured social networks, makes separating repeated import from within-border transmission a major conundrum to TB-control efforts in many low-incidence countries. This is the case in Norway, where TB incidence dropped to very low levels during the second half of the 20th century (6 per 100,000 in 2016) and more than 80 per cent of TB cases are now found among immigrants from high-incidence countries. Immigrants from the Horn of Africa constitute the largest group of TB patients in Norway, making up a third of all TB cases in the country over the last decade. One particular genotype-cluster strongly associated with people originating in this region has been identified regularly over a 20-year period. Here we apply transmission modeling methods to whole-genome sequence data to estimate the times at which individual patients were originally infected. By contrasting these estimates with time of arrival in Norway, we estimate on a case-by-case basis whether individual patients were likely to have been infected before or after arrival. Although import was responsible for the majority of cases, we find that transmission is also occurring in Norway. Our approach is very broadly applicable and relevant to many settings where TB control programs can benefit from an understanding of when (and consequently where) individuals have acquired a tuberculosis infection.

Introduction

The 'End TB strategy' of the World Health Organization aims to reduce the global incidence of tuberculosis (TB) to 100 cases per million by 2035. Reaching this goal entails reducing the global TB incidence to the rates currently seen in countries with the lowest incidence, whereas low-incidence countries must aim for further reduction in incidence levels.

As the TB epidemic is fading out of the local population, the TB incidence in Norway now largely reflects the level of immigration from high-TB-incidence countries [3, 14, 5]. With effective case finding and case management, TB transmission from immigrant populations to the Norwegian-born population has been found to be very limited [4], but it does occasionally occur [13]. In low-incidence countries, preventing transmission within immigrant groups originating from high-incidence countries is of utmost importance if the overall TB incidence is to be reduced further. However, detecting transmission within immigrant populations remains a complicated task as it requires the ability to distinguish between import and transmission post arrival. Social networks are not randomly formed, and are among other things shaped by shared cultural and ethnic backgrounds [22]. An immigrant might thus have been exposed to the same circulating TB strains along the whole spatio-temporal trajectory beginning in the country of origin and ending in a low-incidence country, often via a lengthy and complex journey.

Genome-level analyses have rapidly become important tools for molecular epidemiological studies as they deliver massively improved resolution and detail, relative to traditional genotyping methods. Fast and large-scale sequencing of pathogen genomes can provide stronger and more accurate evidence to exlude and sometimes confirm transmission, and is increasingly applied for disease outbreak management [15]. However, even when genome sequences are available for analysis, the reconstruction of detailed transmission histories is far from straightforward. As a result of a low evolutionary rate and highly variable and stochastic latency periods, this is especially true for TB [13, 7].

Several approaches have been developed to make use of genomic data to infer infector/infectee relationships. They differ in their statistical approaches, the complexity of the underlying epidemiological models and the assumptions they make about unsampled cases, transmission bottlenecks, pathogen evolution, diversity inside hosts and the likelihood of transmission events. Key parameters that must be specified or estimated include the generation time (time from an individual becoming infected until infecting others) and patient + health system delay (time from symptom onset to diagnosis). In addition, inference requires timing information and a model of the pathogen's evolution to connect the acquisition of polymorphisms to the (variable and uncertain) time that has elapsed.[8, 18, 24, 19, 6].

Didelot et al. developed a method to infer transmission networks from time-labelled phylogenies assuming a Susceptible-Infectious-Removed epidemiological model [8] . Building on the same approach, Eldholm et al. implemented a latent state ("Exposed" category) in a similar framework to estimate the probability of pairs of patients being linked by a transmission event [13]. TransPhylo [7] is a Bayesian method for inference of transmission trees that also accounts for unsampled cases and infers transmission events and their timing given mutational events captured in a time-labelled phylogeny.

Here, we apply a novel approach to tackle a major public health conundrum, namely whether immigrants diagnosed with TB were infected before or after arrival in a low-incidence country (imported cases or local transmission, respectively). We focused on a large genotypic *M. tuberculosis* cluster (Norwegian-African large Lineage 3 cluster; NAL3C), strongly associated with immigrants from the Horn of Africa, that has been identified in Norway consistently for almost 20 years [16]. Building on a temporal phylogeny built on genome-wide SNPs, we infer the posterior distribution of infection times T_{inf} using TransPhylo. We then compare these with the time of arrival of individual patients in Norway, available through the Norwegian Surveillance System for Communicable Diseases (MSIS) to ascertain probabilistically whether

they became infected before or after their arrival in Norway. We confirm that most TB patients were indeed infected prior to arrival, but show that about 25% of the patients likely contracted TB after arrival in Norway.

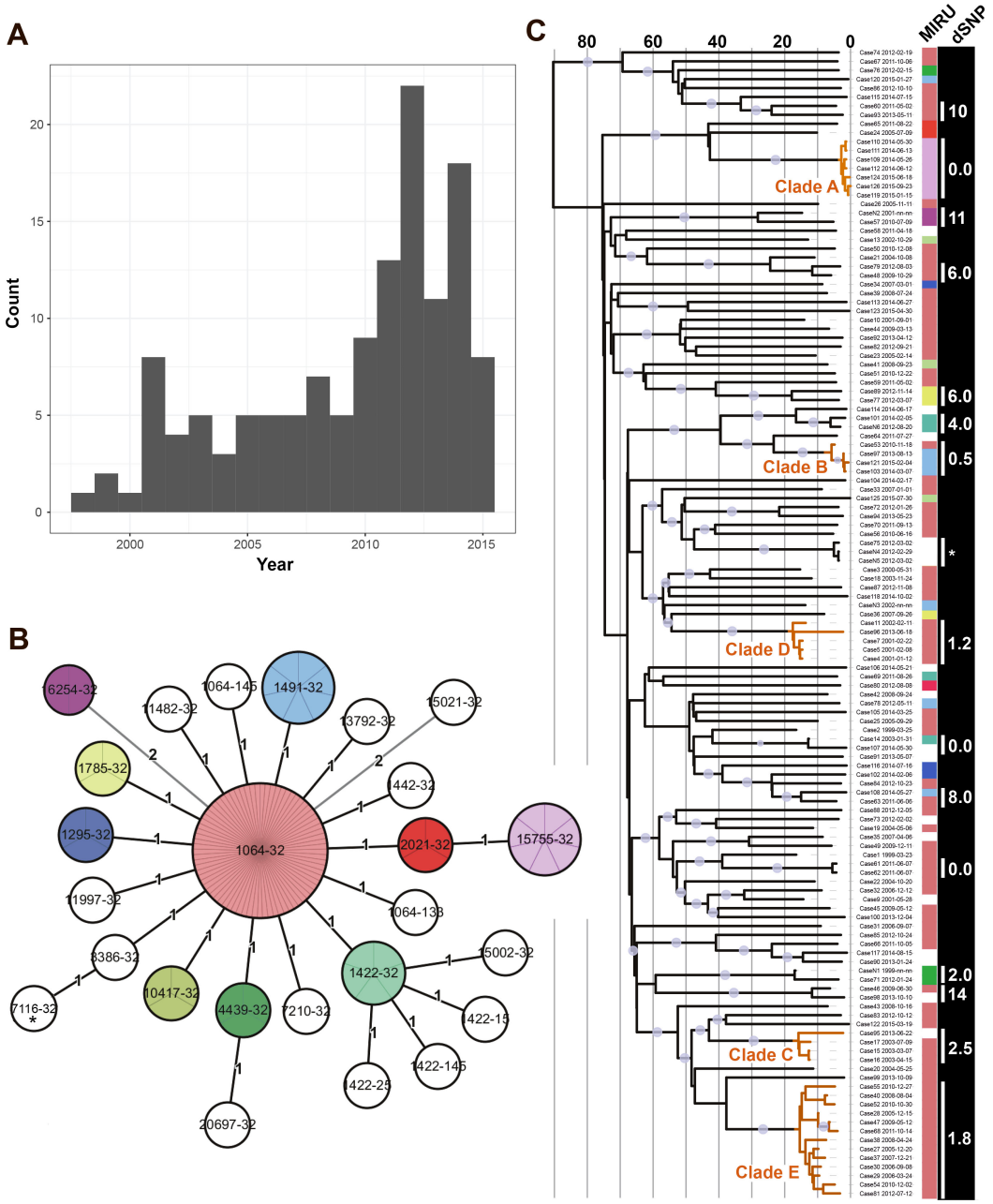

Figure 1: *Clinical M. tuberculosis isolates and phylogenetic reconstruction. (A) Histogram illustrating sampling times for NAL3C isolates. (B) NAL3C minimum-spanning tree based on 24-loci MIRU genotypes. MIRU profiles identified in more than one isolate were assigned individual colors. (C) Temporal phylogeny constructed from genome-wide SNPs. The color strip next to the phylogeny indicates MIRU genotype, whereas the dSNP column denotes mean pairwise-SNP distances within each cluster. Clades amenable for TransPhylo transmission reconstruction are highlighted in orange. The time-axis on the phylogeny corresponds to years before 2015. An asterisk denotes three samples isolated from the same patient. Grey dots on branches indicate posterior probability of* > 0.8

Results

Based on Mycobacterial Interspersed Repetitive Unit (MIRU) genotyping, 129 clinical *M.tuberculosis* NAL3C isolates from 127 patients, collected between 1997 and 2015 (Fig. 1A and 1B) were sequenced at the National Reference Laboratory for Mycobacteria (NRLM). A total of 1418 variable sites were identified, resulting in a mean pairwise-SNP distance of 43.22, which is high relative to what we would expect from an outbreak (see e.g. [11]. A temporal phylogeny was estimated in Beast 1.8.4 [9] utilizing sampling dates for temporal calibration (Fig. 1C).

Comparing the MIRU-based minimum-spanning network and the whole-genome phylogeny of our samples, it was clear that the true genomic diversity of NAL3C was not at all captured by MIRU typing. Strikingly, we also find that the micro-evolution of MIRU loci within the cluster evolved in a way that was not informative for molecular epidemiological purposes. In fact, a number of homoplasic events led to the repeated evolution of identical MIRU types across the NAL3C (Fig. 1). Based on these analyses it is clear that MIRU typing worked rather well for crude grouping of isolates, but that micro-evolutionary events, such as the mutation of a single MIRU locus, is not necessarily informative for molecular epidemiological inference. NAL3C belongs to lineage 3, an understudied *M. tuberculosis* lineage. Whether the mode and rate of MIRU evolution differs between lineages is a question that deserves attention, as it would clearly affect the interpretation of MIRU data.

The high genetic diversity within the NAL3C cluster, combined with an overall phylogenetic structure characterized by multiple long terminal branches interspersed by a handful of tight clusters, suggested that the clinical TB cases in Norway represented samples drawn from a larger population of mainly unsampled cases presumably circulating in the Horn of Africa. We reasoned that the tight sub-clades could correspond to clusters of transmission in Norway. As the vast majority of immigrants from the Horn of Africa came to Norway after 1995, following the withdrawal of UN from Somalia, we only included clades with an inferred most recent ancestor younger than 20 years (corresponding to 1995). As a tree must include at least four cases for meaningful modeling in TransPhylo, this inclusion criterion was also applied. This resulted in a total of five clades (clades A,B,C,D and E shown in Figure 1, meeting the criteria for detailed transmission modelling. Most of the cases in these clades come from countries in the horn of Africa, two from Sudan and one case each from Ghana, Gambia, Iran, Thailand and Norway.

Figure 2 shows the arrival times of all cases from these clades for whom arrival times were retrievable (all in clades A, B and E), plotted on top of the posterior infection time distribution for each individual (see methods section for details). It is quite clear that some of these patients (cases 30,37,40,47,54,68 and 126) arrived in Norway before the estimated time T_{inf}. For all other cases, the estimated range for the time of infection has at least some overlap with the time of arrival. Using the posterior densities of infection times alongside the arrival times of the cases, we obtain probabilities of infection post arrival in Norway ($P(t_{inf}$ after $t_{arrv})$; see the Methods section for details), and these are listed in Table S3 (supplementary document). The cumulative frequency plot of these probabilities (Fig. 3) shows that there are 16 cases with $P(t_{inf}$ after $t_{arrv}) > 0.5$ and 12 cases with $P(t_{inf}.$ after $t_{arrv}) > 0.9$ where t_{arrv} is the time of arrival.

In clades A,B and E, cases 28 and 29 lacked arrival time information. For case 29 (Somali) we could conclude that the patient likely contracted TB in Norway, as the inferred infector was also infected in Norway. Case 28 was an immigrant from Ghana and the isolate was genomically identical to the isolate from case 29, so case 28 was also probably infected in Norway. Next, we looked into clades C and D for which

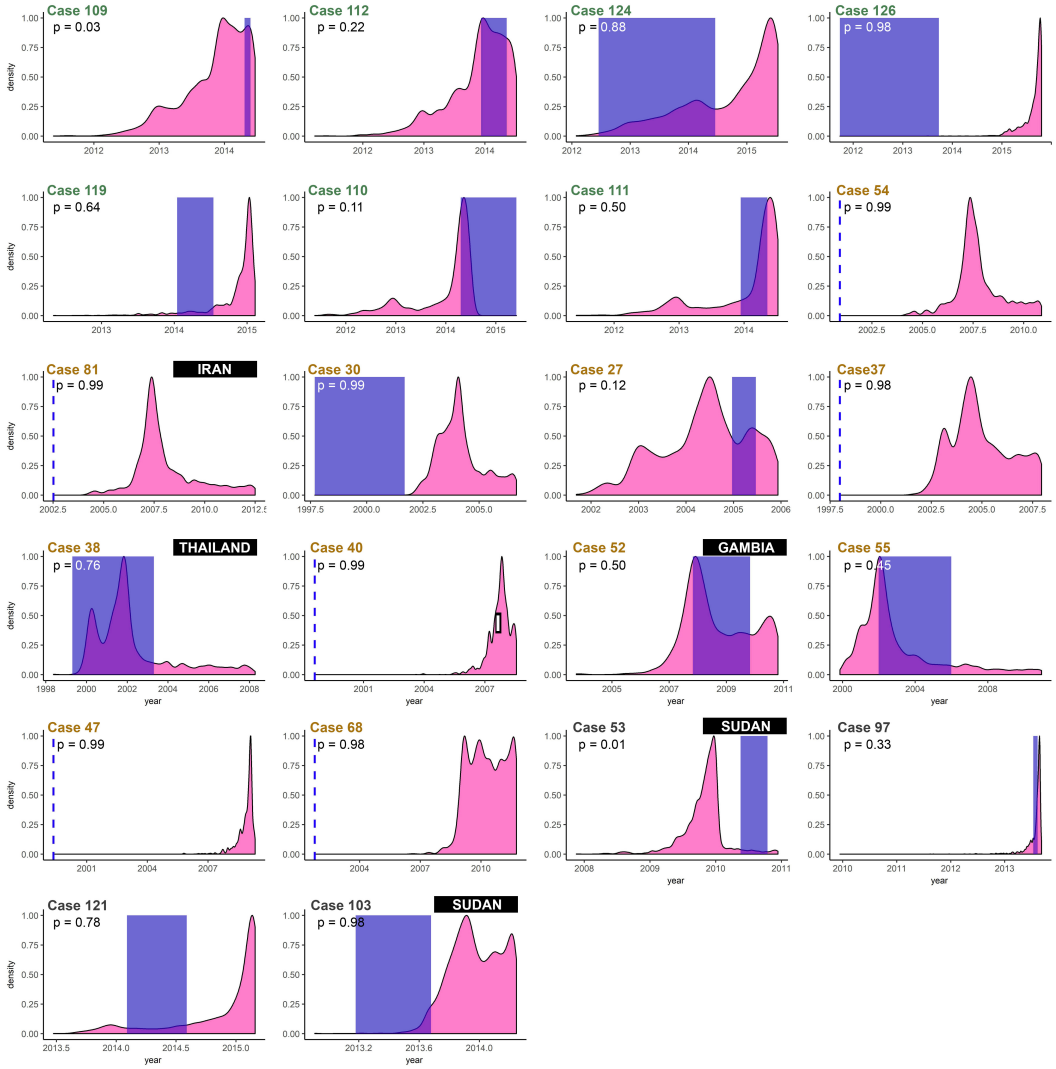

Figure 2: *Arrival times (in blue) plotted on top of estimated infection times for all cases of interest with available data. The case numbers are colored by clade assignment (clade A in green, clade E in orange and clade B in grey). The blue shaded area covers the time from earliest and latest possible arrival times, whereas a dotted single line indicates the latest possible arrival time. P-values indicate probability of infection after arrival in Norway, averaged over 10 different TransPhylo inference procedures. The country of origin of patients not originating from the Horn of Africa is annotated in black boxes.*

arrival information was lacking for all isolates. For clade C, TransPhylo inferred that the same unsampled case had infected both a Norwegian, Ethiopian and two Somali patients, as well as a final Somali patient via another unsampled intermediate (see Figure S7 in supplementary document), though we note that with long and variable infectious periods, as is the case for TB, considerable uncertainty remains in the details of reconstructed transmission events. However, our results suggest that all five patients contracted TB in Norway. In clade D, all patients were Somali. This, combined with a lack of arrival information for these patients, makes it impossible to distinguish between transmission before or after arrival. Finally, in order to obtain a more complete picture of transmission in Norway, beyond clades that were amenable to transmission inference using TransPhylo, we manually investigated the temporal

	Yes	Undetermined	No	Clades
TransPhylo inference (with arrival info)	16	0	6	A,B,C
TransPhylo inference (without arrival info)	7	4	0	C,D,E
Smaller clusters	3	6	5	–
Other cases	0	0	80	-
Total (%)	26(20)	10(8)	91(72)	-

Table 1: *Summary of transmission inference.*

phylogeny for pairs and triplets of closely related isolates for evidence of transmission in Norway. Following the inclusion criteria applied for TransPhylo inference, we only included pairs and triplets with an estimated most recent common ancestor after 1995. Based on a combination of arrival times, disease manifestation and country of origin, we were able to identify another three instances of very probable transmission in Norway. For five cases we could conclude that transmission in Norway was highly unlikely, whereas no conclusion could be drawn for six of the cases (see Tables S04 and S05 for a summary of the evidence).

Altogether, using our main workflow of contrasting TransPhylo-inferred time of infection with time of arrival in Norway, we identified 16 cases in clades A, B and E that had a higher than 0.5 probability of having contracted TB in Norway. Two additional patients in these clades lacking arrival information were also determined to have likely contracted TB in Norway, as were all the five patients in clade C despite a lack of arrival times for these patients. Finally, three additional instances of probable transmission in Norway, represented by two pairs of closely related isolates were identified. In total, we conclude that 26 out of 129 NAL3C cases were probably infected in Norway. For 10 cases we were unable to conclude, whereas the remaining 91 probably represented instances of imported TB (see Table 1).

Retrospectively, we retrieved available contact tracing information and extended epidemiological data for cases belonging to clades A, B and E. The data were incomplete, but useful information was available for six of the cases. For five of these patients the extended data supported our inference, in that cases we estimated were likely infected in Norway had known TB contacts in Norway (cases 40, 47 and 81) or in one case our estimated transmission time concurred with an earlier episode of tuberculosis (case 55). Case 103 had a negative TB screen upon arrival in Norway, consistent with our inference that case 103 was infected after arrival. However, in once case (119) we estimated infection in Norway but the patient reported to have had an episode of symptomatic, untreated TB before arrival; the recent isolate could reflect a re-infection, or our estimation could be incorrect.

Discussion

This work represents a conceptually novel approach to tackle an important public health issue in low TB incidence countries. Here we combine whole genome sequencing with epidemiological modeling to estimate the time of infection for individual patients. We have applied a Bayesian method of transmission inference, TransPhylo, to do this analysis for clusters of tuberculosis cases in Norway. We used individual-level data on arrival time, sputum smear status, time of sampling and whether the patients' tuberculosis disease was pulmonary or extra-pulmonary to refine our inference of the transmission process. We compared the posterior estimates of case infection times to the times when patients arrived in Norway; where arrival time was uncertain (a range rather than a date) we integrated over this time-range in order

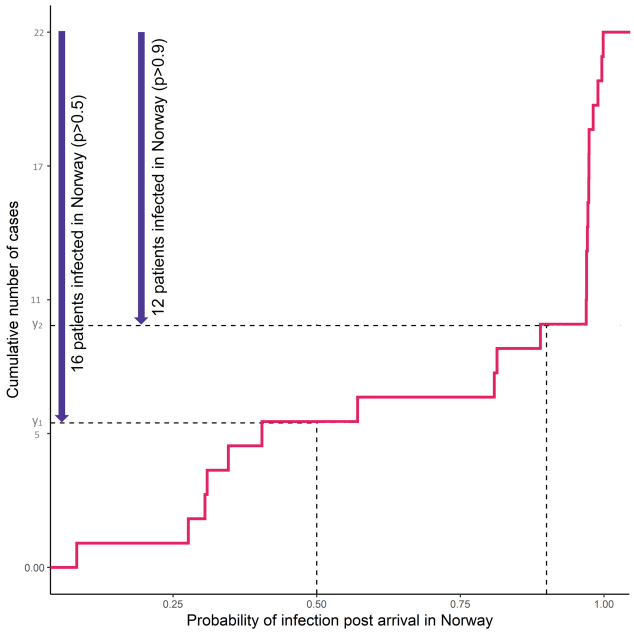

Figure 3: *Probability of infection after arrival in Norway for 22 cases included in TransPhylo analyses. The lines annotate the number of cases with probabilities equal to* 0.5 *and* 0.9 *in the cumulative distribution plot.*

to estimate the probability that cases were infected after arrival. For many patients, the genetic structure of the isolates and/or a lack of epidemiological information left considerable uncertainty in their time of infection (see Results section). However, for some patients, by taking all the available evidence into account we were able to infer with relative certainty whether they had contracted TB before or after arrival in Norway. Overall, we found that there is substantial evidence of ongoing transmission in Norway. Our approach and findings highlight the importance of collecting and keeping good epidemiological records on individual patients, as well as the importance of active contact-tracing in high-risk groups in low-incidence countries..

Even without reconstructing transmission events explicitly, some conclusions about the timing of transmission are evident in the timed phylogenetic tree. Each tip of the tree corresponds to a different host, which implies that there must be at least one transmission event on the path between every pair of tips [8, 7]. Recent branching in the timed phylogeny therefore indicates recent infection times. The analysis we have done goes further, by producing probability distributions for infection times for individual patients.

Five of the patients belonging to potential transmission clusters and for whom arrival times were known originated from countries other than those in the Horn of Africa (Fig. 2). Four of these, originating from Thailand, Iran, Sudan and the Gambia, were inferred to have been infected in Norway ($p \geq 0.5$). The fifth patient was from Sudan, and was inferred to have been infected prior to arrival. The *M. tuberculosis* isolates from two additional Sudanese patients (Case 24 and 83) were found alone on long terminal branches of the phylogeny (Fig. 1), also suggesting that they were independently imported cases. Two Tanzanian patients also represented relatively clear import events with clinical isolates situated on long terminal branches (Case 33 and Case 57). Apart from patients from the Horn of Africa, the only foreign patients we could conclude had been infected prior to arrival originated from Tanzania and Sudan. Sudan shares a border with Eritrea and Ethiopia whereas

Tanzania lies to the south of Somalia and Ethiopia, with Kenya in between. Taken together, these findings suggest that the original geographic range of the NAL3 cluster stretches beyond the Horn of Africa into neighboring countries including Sudan, Kenya and Tanzania. However we did not have access to patient travel records and migration throughout the region is a possible route for exposure.

We find that 26 patients belonging to the NAL3C cluster were most probably infected in Norway, whereas 91 patients had probably contracted TB prior to arrival in Norway. For 10 patients, our analyses were inconclusive. In addition, we were unable to retrieve samples for DNA extraction for three patients, originating from Norway, the former Yugoslavia and Namibia respectively. Based on their country of origin alone, these were all probably infected in Norway. Only considering patients for which a conclusion could be drawn, about a quarter were most likely to have contracted TB in Norway, a far from trivial proportion. It should be noted that some patients might also have been infected during travels to the country of origin after arrival in Norway, but we were not in a position to investigate this possible order of events. A very recent study from the Netherlands and Denmark identified the 1064-32 MIRU-type among refugees and immigrants with a similar country-profile as observed in Norway [17]. Whole genome sequencing (WGS) of 40 isolates revealed a pairwise SNP-distance of 80, almost twice as high as observed here. Although formal transmission reconstruction was not performed, the high diversity supports their conclusion that transmission in the Netherlands and Denmark is very limited. The lower diversity in Norway also suggests that recent transmission in the country has affected the observed population structure of NAL3C.

Immigrants from high-incidence countries typically make up a significant portion of TB cases in low-incidence countries such as Norway, but it is difficult to elucidate whether TB disease among them is a result of import or recent transmission in the receiving country. This difficulty has multiple roots, including the fact that cultural and ethnic identity play a role in forming social connections and low institutional trust in some immigrant groups, which can lead to an unwillingness to share information necessary for contact tracing. In Norway, contact tracing is initiated around all pulmonary TB cases, but the intensity of the effort is higher when recent transmission is suspected. In practical terms, the bar for initiating broad contact tracing efforts will thus often be higher for TB-cases belonging to high-risk groups such as immigrants from high-incidence countries. Routine use of WGS for molecular epidemiology is expected to provide a more solid evidence-base to inform the intensity of contact tracing efforts, but even with WGS data, the detailed reconstruction of transmission events is not trivial [23]. The approach we have presented here, applying transmission modelling to clinical, epidemiological and genomic data, can assist public health authorities in understanding where and when patients are infected, and can aid in the design of appropriate TB control measures.

Methods

Sample collection and inclusion criteria

The NRLM maintains a national culture collection consisting of all culture-positive TB cases in Norway and is responsible for susceptibility testing and genotyping. From 1997 till 2010 *IS6110* restriction fragment length polymorphism (RFLP) was the routine method for molecular epidemiological surveillance. In this period, a large *IS6110*-RFLP-cluster associated with patients from the Horn of Africa was identified [16]. Following the replacement of *IS6110*-RFLP typing with 24-loci Mycobacterial Interspered Repetitive Unit (MIRU) typing at the NRLM, these isolates were re-

typed and the majority of isolates belonged to the MIRU type 1064-32 following on the MTBC 15-9 nomenclature [2]. All subsequent *M. tuberculosis* isolates have been MIRU typed. In order to study the transmission dynamic of this cluster, we included all isolates sampled between 1997 and 2015 that differed at zero to two loci relative to the 1064-32 genotype. In total, 133 isolates matched the inclusion criteria, of which 130 could be retrieved and were submitted to whole-genome sequencing on the Illumina platform. Initial analyses revealed one of these to be a clear outlier only distantly related to the other 129 isolates; it was thus excluded. The 129 isolates represented 127 patients (three isolates were from the same patient). The cluster, as defined by *IS6110*-RFLP genotyping, was recently termed "Cluster X" [16]. However, we coined the more informative term Norwegian-African large lineage 3 cluster (NAL3C) for the current study. In addition, we extracted data from the Norwegian Surveillance System for Communicable Diseases (MSIS), which stores clinical and epidemiological data on all TB cases notified by clinicians. The study protocol was approved by the Regional Ethics Committee (reference 2015/2127).

Variant calling

DNA was extracted from *M.tuberculosis* grown on Lowenstein-Jensen slants as described previously [12]. Paired-end sequences were generated on the Illumina MiSeq and NextSeq platforms (250 and 150 bp read length respectively). High quality single nucleotide polymorphisms were identified following the same procedures as described in [12]. After removal of the single outlier isolate, this resulted in 1418 variable sites that were used for evolutionary analyses as outlined below. Median sequencing depth of the 129 genomes ranged from 20x to 161x. All sequence reads are available under ENA study accession PRJEB23495. Individual run accessions, sampling years and MIRU data for all NAL3C isolates are listed in dataset S01.

Bayesian Evolutionary analyses

Marginal likelihood estimates in Beast 1.8.4 [9] were performed to identify the optimal substitution, clock and demographic models for Bayesian evolutionary analyses. We tested the HKY and GTR substitution models combined with either a strict or uncorrelated relaxed clock and a constant, logistic, exponential or Skyride demographic model. A GTR model with relaxed clock combined with a Skyride demographic model was favored (see supplementary material). Three independent Markov chain Monte Carlo (MCMC) chains consisting of 200 million steps were performed and the output combined after inspection of convergence within and between chains. These analyses resulted in an estimated substitution rate of 8.99E-8 (95 per cent HPD: 5.07E-8, 1.31E-7) substitutions per genome per year. To verify the presence of sufficient temporal signal in the data, tip-randomization was performed utilizing the 'tipdatingbeast' R package [21]. Of 20 tip-randomized runs, the 95 % HPD interval of a single run overlapped with the tree height 95% HPD interval generated in the combined non-randomized data (see supplementary material), indicating that the strength of the temporal signal was acceptable [10, 20].

Transmission reconstruction

The maximum credibility tree of the 129 isolates is characterized by long branch lengths with a few clades that have relatively short branch lengths. As the branch lengths of a timed phylogenetic tree represent duration of evolution [1], it is intuitive to assume that these clades represent densely sampled clusters of cases whereas the

long branches represent cases with unsampled infectors. We thus reasoned that putative transmission clusters in Norway would be represented by sub-clades of closely related isolates within the larger NAL3C cluster. Based on a previous comprehensive study [23] and the overall structure of the phylogeny, we selected clades with a minimum of four cases, and with a maximal mean pairwise SNP-distance of five or fewer SNPs within the clade for transmission inference. Five clades matched this criteria (clades A,B,C,D and E shown in Figure 1). There are a total of 33 cases in the selected clades, with times of arrival into Norway available for 22 of them.

We used the R package TransPhylo [7] to reconstruct the outbreaks. TransPhylo allows for in-host diversity; individuals may harbour more than one pathogen variant (though they may not). TransPhylo uses a branching process model to compute the likelihood of a set of transmission events based on the likelihoods of the times between individuals becoming infected and infecting others (generation times) and times between infection and sampling (becoming known to public health authorities/TB sample taken), alongside a negative binomial distribution for the number of secondary cases an individual will cause. The epidemiological model requires the user to specify a gamma distribution for the generation time, and similarly, a gamma distribution for the sampling time. Estimating these parameters is challenging, especially when the sampling density is unknown. However, for the five sub-clades identified above, we can assume a high sampling density based on the extremely limited observed diversity within each clade. We estimated the parameters of the sampling and generation time distributions from the subtree of the least diverse clade (clade A; mean pairwise SNP-distance $= 0$), assuming 95% sampling. A gamma distribution is used for the prior generation time distribution in order to reflect the variable progression of tuberculosis, which could either be rapid with short time interval from the time of infection to the onset of infectiousness, or very long with infection leading to long latent periods before the onset of infectiousness. We chose shape and scale parameters of the gamma distribution that give a mean of $4-5$ years for the generation time distribution. Case finding and management of tuberculosis is quite effective in Norway, and as such, we chose a gamma sampling distribution with mean between $2.5-3$ years. Altogether, we chose ten different shape and scale parameters that meet these criteria (Table S1a, supplementary document). We start off with clade A, run the inference procedure using these different choices for the priors of the sampling and generation times, whilst assuming a very high sampling proportion. We thus obtain ten different posterior sampling and generation time distributions whose shape and scale parameters (obtained by fitting a gamma distribution to the posterior generation and sampling times respectively using the function $fitdistr$ in the MASS package in R; these are shown in Table S1 of the supplementary document) are then used as inputs for the inference of transmission events on the other clades. The results reported here are the output from the first run except where otherwise stated.

Smear negative patients are less infectious than smear positive patients, and it is reasonable to assume that they transmit tuberculosis with less efficiency. We applied a penalty to transmission events that have smear negative infectors by multiplying the probability of the transmission tree by 0.75. As such the inferred transmission trees are pulled away from maximum likelihood estimates (in the baseline model without smear status), and pulled towards estimates that have fewer transmission events from smear-negative patients. We assume that patients with extra-pulmonary tuberculosis are only 1% as likely to transmit the disease as pulmonary tuberculosis patients, and also apply this penalization to the likelihood of a transmission tree.

A key feature of TransPhylo is that it infers transmission events using a two-step

procedure: obtaining a timed phylogenetic tree and inferring transmission events given this phylogenetic tree. This approach therefore makes it difficult to pass the uncertainty in the phylogenetic reconstruction to transmission inference especially when a posterior distribution of phylogenetic trees are produced in the first step. In order to account for this, the transmission inference was applied to a random sample of phylogenetic trees obtained in the tree reconstruction step. Also there is uncertainty in the choice of prior parameters for the generation and sampling time distributions. These were therefore chosen over a wide range of combinations that depict a tuberculosis outbreak.

Probability of infection prior to arrival

We can quantify the probability that an individual was exposed to their TB strain prior to their arrival in Norway, using the posterior times of infection. If we know the arrival time t^i_{arr} for case i, and we let the posterior time of infection density be called $L^i(\tau)$, then the probability that i was infected after arrival is just the portion of the posterior that lies above t^i_{arr}:

$$P(t^i_{inf} \text{ after } t^i_{arr}) = \int_{t^i_{arr}}^{t_{max}} L^i(\tau) d\tau. \tag{1}$$

If the arrival is uncertain, and we only know that case j arrived between minimum time m_j and maximum time M_j, then we can integrate out the unknown time of arrival to find the marginal probability that j was infected after arrival in Norway:

$$P(t^j_{inf} \text{ after } t^j_{arr}) = \int_{m_j}^{M_j} P(t_{arr} = s) P(t^i_{inf} \text{ after } s) ds$$

and we use (1) to obtain $P(t^i_{inf}$ after $s)$, and a uniform distribution (blue rectangles in the Figures) for $P(t_{arr} = s)$. These probabilities are averaged over 10 inference procedures using different prior distribution parameters.

Acknowledgments

We would like to acknowledge the Norwegian Sequencing Center for sequencing a subset of the samples on the NextSeq platform and the staff at the National Reference Laboratory for Mycobacteria, Norway, for maintenance and genotyping of a national TB culture collection. CC is funded by the Engineering and Physical Sciences Research Council of the United Kingdom (EPSRC EP/N014529/1 and EPSRC EP/K026003/1) and DA receives funding from Nigerian Universities Commission (under the presidential special scholarship scheme for innovation and development (PRESSID)).

1 References

[1] D. Alexei and B. Remco. *Bayesian evolutionary analysis with BEAST*. Cambridge University Press, 2015.

[2] C. Allix-Béguec, M. Fauville-Dufaux, and P. Supply. Three-year population-based evaluation of standardized mycobacterial interspersed repetitive-unit-variable-number tandem-repeat typing of Mycobacterium tuberculosis. *Journal of Clinical Microbiology*, 46(4):1398–406, Apr 2008.

[3] M. Carballo and M. Mboup. International migration and health: A paper prepared for the policy analysis and research programme of the global commission on international migration. 2005.

[4] U. R. Dahle, V. Eldholm, B. A. Winje, T. Mannsåker, and E. Heldal. Impact of immigration on the molecular epidemiology of *Mycobacterium tuberculosis* in a low-incidence country. *American Journal of Respiratory and Critical Care Medicine*, 176(9):930–935, nov 2007.

[5] U. R. Dahle, P. Sandven, E. Heldal, and D. A. Caugant. Continued low rates of transmission of *Mycobacterium tuberculosis* in Norway. *Journal of Clinical Microbiology*, 41(7):2968–73, Jul 2003.

[6] N. De Maio, C.-H. Wu, D. J. Wilson, S. Gaudieri, S. Pham, and A. Chopra. SCOTTI: Efficient reconstruction of transmission within outbreaks with the structured coalescent. *PLOS Computational Biology*, 12(9):e1005130, Sep 2016.

[7] X. Didelot, C. Fraser, J. Gardy, and C. Colijn. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular Biology and Evolution*, 195(4):msw075, Jan 2017.

[8] X. Didelot, J. Gardy, and C. Colijn. Bayesian inference of infectious disease transmission from whole-genome sequence data. *Molecular Biology and Evolution*, 31(7):1869–79, Jul 2014.

[9] A. J. Drummond, M. A. Suchard, D. Xie, and A. Rambaut. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29(8):1969–1973, Aug 2012.

[10] S. Duchêne, K. E. Holt, F.-X. Weill, S. Le Hello, J. Hawkey, D. J. Edwards, M. Fourment, and E. C. Holmes. Genome-scale rates of evolutionary change in bacteria. *Microbial Genomics*, 2(11):e000094, Nov 2016.

[11] V. Eldholm, J. Monteserin, A. Rieux, B. Lopez, B. Sobkowiak, V. Ritacco, and F. Balloux. Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. *Nature Communications*, 6:7119, May 2015.

[12] V. Eldholm, J. H.-O. Pettersson, O. B. Brynildsrud, A. Kitchen, E. M. Rasmussen, T. Lillebaek, J. O. Rønning, V. Crudu, A. T. Mengshoel, N. Debech, K. Alfsnes, J. Bohlin, C. S. Pepperell, and F. Balloux. Armed conflict and population displacement as drivers of the evolution and dispersal of *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences of the United States of America*, 113(48):13881–13886, Nov 2016.

[13] V. Eldholm, A. Rieux, J. Monteserin, J. M. Lopez, D. Palmero, B. Lopez, V. Ritacco, X. Didelot, and F. Balloux. Impact of HIV co-infection on the

evolution and transmission of multidrug-resistant tuberculosis. *eLife*, 5:e16644, Aug 2016.

[14] M. Farah, H. Meyer, R. Selmer, E. Heldal, and G. Bjune. Long-term risk of tuberculosis among immigrants in Norway. *International Journal of Epidemiology*, 34(5):1005–1011, Jul 2005.

[15] M. Gilmour, M. Graham, A. Reimer, and G. Van Domselaar. Public health genomics and the new molecular epidemiology of bacterial pathogens. *Public Health Genomics*, 16(1-2):25–30, 2013.

[16] B. R. Guzman Herrador, K. Rønning, K. Borgen, T. Mannsåker, and U. R. Dahle. Description of the largest cluster of tuberculosis notified in Norway 1997–2011: is the Norwegian tuberculosis control programme serving its purpose for high risk groups? *BMC Public Health*, 15(1):367, Dec 2015.

[17] R. Jajou, A. de Neeling, E. M. Rasmussen, A. Norman, A. Mulder, R. van Hunen, G. de Vries, W. Haddad, R. Anthony, T. Lillebaek, W. van der Hoek, and D. van Soolingen. A predominant VNTR cluster of Mycobacterium tuberculosis isolates among asylum seekers in the Netherlands and Denmark deciphered by whole genome sequencing. *Journal of Clinical Microbiology*, Nov. 2017.

[18] M. Kendall, D. Ayabina, and C. Colijn. Estimating transmission from genetic and epidemiological data: a metric to compare transmission trees. sep 2016.

[19] D. Klinkenberg, J. A. Backer, X. Didelot, C. Colijn, J. Wallinga, and D. Haydon. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLOS Computational Biology*, 13(5):e1005495, May 2017.

[20] A. Rieux and F. Balloux. Inferences from tip-calibrated phylogenies: a review and a practical guide. *Molecular Ecology*, 25(9):1911–1924, May 2016.

[21] A. Rieux and C. E. Khatchikian. tipdatingbeast: an r package to assist the implementation of phylogenetic tip-dating tests using beast. *Molecular Ecology Resources*, 17(4):608–613, Jul 2017.

[22] A. Sandgren, M. S. Schepisi, G. Sotgiu, E. Huitric, G. B. Migliori, D. Manissero, M. J. van der Werf, and E. Girardi. Tuberculosis transmission between foreign- and native-born populations in the EU/EEA: a systematic review. *The European Respiratory Journal*, 43(4):1159–71, Apr 2014.

[23] T. M. Walker, C. L. C. Ip, R. H. Harrell, J. T. Evans, G. Kapatai, M. J. Dedicoat, D. W. Eyre, D. J. Wilson, P. M. Hawkey, D. W. Crook, J. Parkhill, D. Harris, A. S. Walker, R. Bowden, P. Monk, E. G. Smith, and T. E. A. Peto. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *The Lancet. Infectious Diseases*, 13(2):137–46, Feb 2013.

[24] C. J. Worby, M. Lipsitch, and W. P. Hanage. Shared genomic variants: identification of transmission routes using pathogen deep sequence data. *American Journal of Epidemiology*, Jun 2017.