




The population genomics of within-host *Mycobacterium tuberculosis*

Ana Y. Morales-Arce¹ · Susanna J. Sabin¹ · Anne C. Stone ^{1,2} · Jeffrey D. Jensen^{1,3}

Received: 29 June 2020 / Revised: 2 October 2020 / Accepted: 3 October 2020 / Published online: 15 October 2020

© The Author(s), under exclusive licence to The Genetics Society 2020

Abstract

Recent progress in genomic sequencing from patient samples has allowed for the first detailed insight into the within-host genetic diversity of *Mycobacterium tuberculosis* (*M.TB*), revealing remarkably low levels of variation. While this has often been attributed to low mutation rates, other factors have been described, including resistance evolution (i.e., selective sweeps), widespread purifying and background selection, and, more recently, progeny skew. Here we review recent findings pertaining to the processes governing the evolutionary dynamics of *M.TB*, discuss their implications for improving our understanding of this important human pathogen, and make recommendations for future work. Significantly, this emerging evolutionary framework involving the joint estimation of demographic, selective, and reproductive processes is forming a new paradigm for the study of within-host pathogen evolution that will be widely applicable across organisms.

Introduction

Mycobacterium tuberculosis (*M.TB*) is considered one of the most successful pathogens in human history (Galagan 2014). Infection caused by *M.TB* or other species in the *Mycobacterium tuberculosis* complex (MTBC), known as tuberculosis (TB), has persistently been among the top ten global causes of death per decade, with at least 8 million new cases every year (World Health Organization 2019). Two important factors have kept TB as the major focus in clinical research: (a) increasing antibiotic resistance in patients since the 1970s (Alanis 2005; Toungoussova et al. 2006; Eldholm et al. 2016), and (b) the rise of human immunodeficiency virus (HIV) infection since the 1980s, and the resulting immunosuppression leading to an

increased incidence of *M.TB* (Centers for Disease Control and Prevention CDC 1989; Müller et al. 2013). More recently, whole-genome sequencing (WGS) technologies have opened new avenues, particularly with regard to drug-resistance surveillance (e.g., Köser et al. 2014; Zignol et al. 2018), epidemiology for infection prevention and outbreak control (e.g., Dheda et al. 2017; Dicks and Stout 2019), and for the more general study of pathogen evolution and transmission dynamics (e.g., Trauner et al. 2017; Payne et al. 2019). Leveraging these data to study the evolutionary forces driving the evolution of *M.TB*, which will be the focus of this review, has required a population-genetic perspective. Specifically, we explore how *M.TB* evolves within human hosts, and summarize recent results of relevance related to population-genetic theory and statistical inference. We argue that newly developed null modeling has shed light on earlier discrepancies and paradoxes, and provides an appropriate framework for studying the evolution of drug resistance in a variety of pathogens.

Deep sequencing as a measure of variation

During the last decade, WGS methods have allowed for detailed insight into DNA sequence variation. For pathogen samples, the number of sequences in the alignment is taken to be representative of an entire within-host population, and commonly this diversity is summarized as a consensus sequence for subsequent between-host analyses. Such analyses in *M.TB* have revealed low genetic diversity, with few SNPs differing between patients in an outbreak. In some

Associate editor: Frank Hailer

✉ Ana Y. Morales-Arce
amoral70@asu.edu

✉ Jeffrey D. Jensen
Jeffrey.D.Jensen@asu.edu

¹ Center for Evolution and Medicine, Arizona State University, Tempe, AZ, USA

² School of Human Evolution and Social Change, Arizona State University, Tempe, AZ, USA

³ School of Life Sciences, Arizona State University, Tempe, AZ, USA

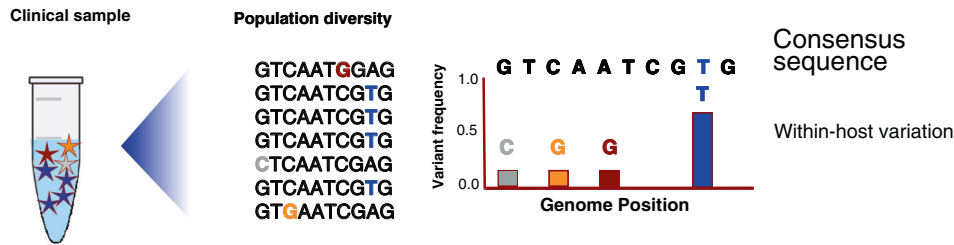


Fig. 1 *M.TB* population diversity from a host sample. An example of a clinical sample, demonstrating how a consensus sequence-based analysis neglects low-frequency variants (which are expected to

constitute the majority of within-host variation). Each star in the clinical sample represents a unique variant. In this example, only one of the four variants would be represented in the consensus.

cases, there have been fewer than 6 nucleotide differences across more than 100 patient samples (e.g., Walker et al. 2013; Séraphin et al. 2019). Similarly, low variation has been reported for thousands of global samples with only ~2200 SNPs separating any two MTBC genomes, and ~1000 on average differing between any two strains (e.g., Liu et al. 2018; Ruesen et al. 2018). Consistently, global phylogenetic levels of genome-wide nucleotide diversity have been observed to be on the order of 10^{-4} /site (O'Neill et al. 2015, 2019). While the limits of SNP detection naturally differ by sequencing technology (see reviews of Pfeifer 2017; Meehan et al. 2019), reducing the analysis to the most common alleles between hosts (i.e., a consensus sequence) as is common practice, has limited our progress toward understanding *M.TB* evolution, as the full spectrum of diversity remains underexplored (Fig. 1; see discussion of Renzette et al. 2017). For example, recent studies have revealed that *M.TB* nucleotide variation is considerably higher within host than previously thought—particularly if a sample is directly sequenced from the sputum rather than cultured before sequencing (Lee et al. 2020)—reaching up to 10^{-3} /site (Séraphin et al. 2019; Shockey et al. 2019; Nimmo et al. 2019). Additionally, WGS analysis has revealed the large number of rare alleles present in a host population, with a portion being associated with drug resistance (Operario et al. 2017; O'Donnell et al. 2019). Rare alleles are most often excluded in common bioinformatic practice in *M.TB*, as establishing a minimum allele frequency is necessary for determining unambiguous variants (which will differ by sequencing technology and processing pipeline, as noted above), and would by definition be neglected in a consensus. These factors, combined with a lack of standardization in workflow for WGS analysis of *M.TB* with different coverage requirements and SNP filtering criteria, appear to have led to consistent underestimation of within- and between-host genetic variation (Meehan et al. 2019; Ley et al. 2019).

We next briefly summarize the primary processes shaping within-patient levels and patterns of variation that have been described to date.

Determinants of *M.TB* diversity

Within-host spatial structure

Sequentially sampled sputum from individual patients analyzed with WGS has revealed that the genetic separation of bacteria can be sufficiently large to resemble sub-populations (Ford et al. 2012; Liu et al. 2015), a phenomenon also observed in autopsies across different anatomical compartments from HIV patients with *M.TB* infection (Lieberman et al. 2016). Relatedly, infection experiments performed on nonhuman primates show limited evidence of interlesion mixing of bacterial populations, indicating that granulomas (aggregations of immune cells—a host response for isolating infection) may only rarely exchange bacteria after dissemination (Martin et al. 2017). However, infection is associated with a variety of types of pathological lesions, and the ability of each to effectively contribute to pathogen migration remains an open question (and see review of Leanerts et al. 2015). Hence, interpretations pertaining to the entire population, based on a single sample per patient, must be made with caution—as different samples may be associated with different variants (Dheda et al. 2018; Cohen et al. 2019). Furthermore, such structuring may effectively serve as a reservoir of variation, an important consideration in designing treatment strategies (Navarro et al. 2017; Cadena et al. 2017).

Purifying selection

The selective removal of deleterious mutations (purifying selection), as well as the resulting background-selection effects (Charleworth et al. 1993), strongly impacts *M.TB* populations, given both a coding-dense genome and a lack of recombination (Dos Vultos et al. 2008; Morales-Arce et al. 2020). Furthermore, as a haploid, there is an expectation that selection may generally act more efficiently, as fitness-impacting mutations may not be masked by dominant alleles (Kondrashov and Crow 1991; Otto and Gerstein 2008). Other claims have been made of relaxed purifying

selection in *M.TB*, given what appears to be an excess of nonsynonymous variants attributed to a lack of selective constraint (Hershberg et al. 2008; Pepperell et al. 2013; Lee et al. 2015). More recently, WGS data have allowed for the observation that rare alleles are common in patient samples (Trauner et al. 2017; Shockey et al. 2019), and that most of these rare alleles are eventually lost (S  raphin et al. 2019), as would be expected. This excess of rare variation is represented by the strongly left-skewed site-frequency spectra (SFS) generally observed in individual samples (Trauner et al. 2017), and despite the loss of diversity when calling consensus sequences, similar patterns are observed when comparing consensus sequences across patients (Chiner-Oms et al. 2019). In terms of parameter fitting, infection histories combined with a mix of both deleterious and neutrally evolving sites produce the nearest fit to the observed SFS in *M.TB* populations (Pepperell et al. 2013; Morales-Arce et al. 2020)—consistent with the suggestion that purifying selection effects are widespread in the *M.TB* genome, serving to greatly reduce variation and skew the SFS (Pepperell et al. 2010; Namouchi et al. 2012; Liu et al. 2014; Minias et al. 2018). Further, when considering the relative roles of selective compared with stochastic (e.g., genetic drift) factors in dictating observed variation, it is important to appreciate that this is not simply a question of the underlying selection coefficients themselves, but also of the underlying effective population size (i.e., the relevant

quantity is the product of the two)—with selection acting more efficiently in large relative to small populations (Ohta 1973). Further, the effective population size itself is shaped by population history, progeny skew, and selective effects, resulting in heterogeneity in the efficacy of selection (see Charlesworth and Charlesworth 2010; Walsh and Lynch 2018).

Population bottlenecks

Experimental data and physiological studies in *M.TB* and other members of the MTBC have shown that pulmonary infection can be established by as few as 1–3 bacteria (Rich 1946; Sonin 1951; O’Grady and Riley 1963; Dean et al. 2005; see the review of Ryndak and Laal 2019). This infection bottleneck is one avenue through which genetic drift shapes *M.TB* evolution, the effect of which would be amplified at each transmission. *M.TB* populations also experience subsequent population bottlenecks after transmission, related to within-host dissemination (Martin et al. 2017) as well as drug treatment (Cohen et al. 2019). Unfortunately, with the severity of the infection bottleneck combined with a lack of recombination, distinguishing selective from demographic signals can be challenging (Thornton and Jensen 2007; Crisci et al. 2013), even under standard models. In sum, both population-size changes and purifying selection likely contribute to the observed low

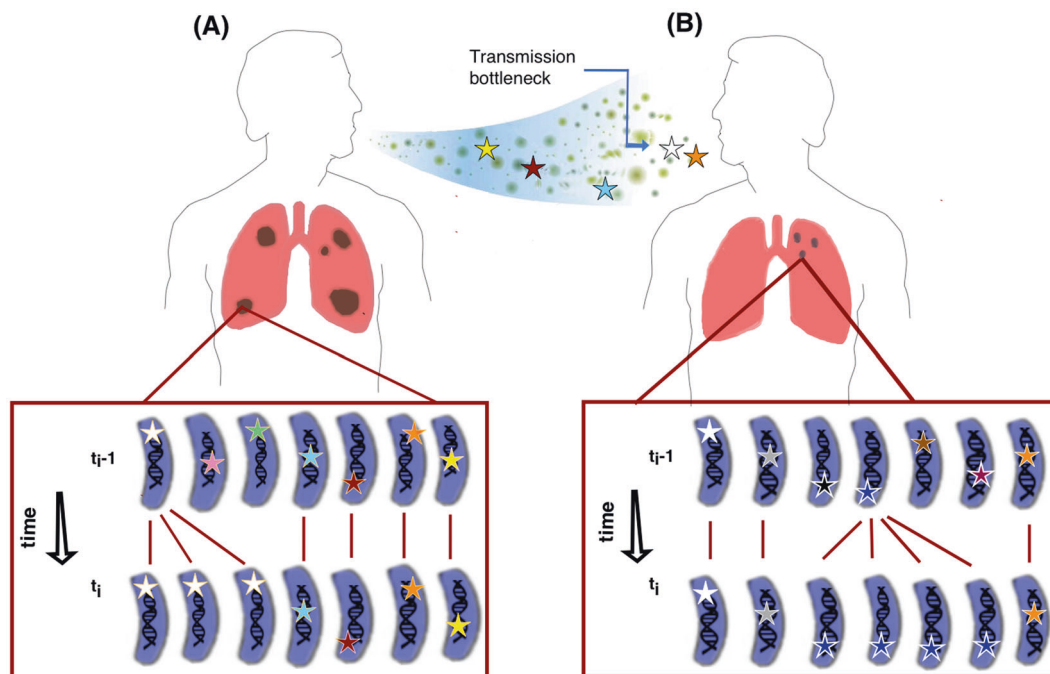


Fig. 2 Infection dynamics and population-genetic diversity. **A** The diversity of *M.TB* within a patient is reduced relative to Wright–Fisher expectations owing to stochastic progeny skew, as shown in the inset in which a single clone in generation t_{i-1} leaves a large proportion of progeny in the following generation t_i . Each colored star represents a

unique variant. **B** The transmission bottleneck associated with a new patient infection additionally acts to reduce diversity to a subset of that present in the infecting individual. Further, the cycle of clonality will continue to reduce variation and alter allele frequencies.

levels of genetic diversity, and to the excess of low-frequency variants.

Progeny skew and mutation rate

One underappreciated process in *M.TB* has been the large variance in progeny distributions, and recent multiple-merger coalescent (MMC) modeling has been proposed as the appropriate population-genetics framework to examine *M.TB* diversity (Morales-Arce et al. 2020; Menardo et al. 2020). Martin et al. (2017) provided experimental evidence for the appropriateness of MMC for *M.TB* through digital barcoding experiments, in which individual bacteria were tracked from the moment of infection in macaques. The experiments demonstrated that most diversity was produced inside the granuloma, where for each infection, a single dominant clone generated large population sizes (up to ~53,710 CFU in a 4-week period). As the common Wright–Fisher (WF) model assumes progeny distributions to be small, *M.TB* likely violates this assumption, necessitating the development of alternative inference approaches (Tellier and Lemaire 2014; Irwin et al. 2016; Sackman et al. 2019). Specifically considering MMC modeling of within-host *M.TB* populations sequenced from serial sputum samples, recent work has demonstrated that such progeny skew (Ψ) is likely an additional and important factor in reducing variation (Fig. 2), and further that a neglect of this parameter has resulted in a downward bias in the estimation of de novo mutation rates (Morales-Arce et al. 2020). In other words, under a null model not accounting for Ψ or background selection, lower mutation rates are inferred in order to fit the paucity of observed variation. However, once accounted for, inference suggested an underlying de novo mutation rate on the order of ~6e–8 per site per replication, and a mean progeny distribution strongly differing from WF assumptions (Morales-Arce et al. 2020). Thus, once these biologically relevant, diversity-reducing processes are considered, it is necessary to invoke higher mutation rates in order to fit observed within-patient levels of variation.

However, there are a number of challenges in directly relating these within-patient mutation-rate estimates with those previously made within a phylogenetic context, which are rather based on the comparison of between-patient consensus sequences. First, the latter estimates are generally given per year, whereas the population-genetic estimates are per generation. While there is some support for a generation time of 1 day (Cole et al. 1998), the uncertainty surrounding this conversion makes direct comparisons difficult. Second, the population-genetic estimates concern the total mutation rate, whereas phylogenetic estimates measure the neutral mutation rate (i.e., mutations with an appreciable

probability of fixation). As such, determining the fraction of the total distribution of fitness effects that is represented by neutral mutations represents another highly tenuous conversion factor. Finally, consensus sequences by their nature neglect the vast majority of sequence variation (Fig. 1); thus, within-patient genome sequencing measures variation on a much finer scale.

Positive selection

While the above processes are highly significant for understanding *M.TB* evolution, the identification of positively selected sites remains a major focus owing to the high incidence of drug resistance (as such, this search has been well-reviewed elsewhere—see Kurz et al. 2016; Dookie et al. 2018; Singh et al. 2020), and thus will not be extensively rediscussed here given length limitations. Briefly, a primary objective of many WGS efforts has been the characterization of de novo mutations and genes specifically involved in mechanisms associated with specific drug treatments (Gygli et al. 2017; Dookie et al. 2018; Ghajavand et al. 2019). This framework has led to the proposal of novel multidrug combinations that may reduce the possibility of adaptive escape (Moreno-Gamez et al. 2015; Trauner et al. 2017). Furthermore, the adaptive potential of *M.TB* has been partially attributed to the existence of pulmonary cavities and lesions that can effectively protect *M.TB* populations in the presence of drug treatment, harboring potential resistance variants (Moreno-Gamez et al. 2015).

Yet, differentiating the genomic effects of positive selection from population bottlenecks can be challenging as described, and even more so with the addition of Ψ , and a neglect of an appropriate null model in such genomic scans can often result in extreme false-positive rates (i.e., misidentifying large numbers of SNPs as being positively selected; Crisci et al. 2013; Harris et al. 2018; Jensen et al. 2019). Statistical analyses designed to improve the ability to differentiate these evolutionary processes have largely been focused on the WF model and Kingman coalescent, and only recently have efforts been made to extend this focus to the type of non-WF MMC models of relevance to most human pathogens, and *M.TB* in particular. For example, Eldon et al. (2015) demonstrated that population-size change and Ψ may be differentiated based on expectations in the SFS; Matuszewski et al. (2018) derived analytical expectations for such SFS and demonstrated an ability to coestimate size change and Ψ in a maximum-likelihood framework, and Sackman et al. (2019) utilized an approximate Bayesian framework to additionally identify positively selected mutations, though this approach requires the presence of time-sampled patient data.

Non-patient-associated *M.TB* samples—the emerging role of ancient DNA

Importantly, clinical research on *M.TB* remains necessarily focused on resistant strains. Thus, historical and ancient samples are highly valuable for characterizing levels and patterns of variation from the pre-antibiotic era. Owing to the above-described difficulties in untangling the contributions of various evolutionary processes, ancient DNA is beginning to be utilized to provide a deeper temporal perspective as well. The first ancient genome-level data belonging to members of the MTBC were released in 2014, produced from human remains discovered in the Osmore River valley in Peru, dated to ~1000 years before present (Bos et al. 2014). These and the ancient *M.TB* genomes produced since (Kay et al. 2015; Sabin et al. 2020) have offered valuable evolutionary insights, potentially better-informing mutation rates across the MTBC for example, though they also present unique challenges. Specifically, pathogen genomes from archeological remains are typically lower coverage than their modern, clinical counterparts. In addition, ancient samples are vulnerable to contamination by environmental microbes, which include many benign species within the genus *Mycobacterium* that share substantial nucleotide identity with species in the MTBC (Warinner et al. 2017). As such, in addition to the suboptimal nature of working with consensus sequences described above, there is an issue associated with the filtering of multi-allelic sites for which there is no clearly common allele (Bos et al. 2014; Sabin et al. 2020). Furthermore, in many genetic studies of ancient microbes, heterogeneity among variants is considered to be a sign of exogenous contamination (Bos et al. 2019). While the fraction of heterozygous sites and the allele-frequency distribution may act together as a marker of quality under this rationale, it also inherently neglects potentially significant population-specific variation. Kay et al. (2015) implemented an alternative approach, in which variant genotypes were treated as true within-individual variation in the form of mixed-strain infections. To orient the strains within the MTBC, sequencing reads were mapped to phylogenetically relevant nodes; however, any variation unique to the samples (i.e., not represented by the existing phylogeny) was not analyzed.

To overcome these difficulties in order to leverage true within-host variation of *M.TB* from ancient samples, and thus better quantify the above-described variation-determining processes, variant sites thus must have high coverage and be rigorously authenticated. Though many studies present high-coverage pathogen genomes produced through shotgun sequencing alone, such a strategy may be cost-prohibitive, depending on the metagenomic makeup of the sequencing library. Alternatively, investigators may use a targeted enrichment approach on a pathogen-positive library to achieve sufficient coverage for confident variant analysis. Either strategy carries the risk of including

exogenous contamination from closely related taxa, however, which must be detected and excluded if possible. Taxonomic binning tools can be used to assess the overall genetic diversity in the library (Warinner et al. 2017; Bos et al. 2019) and extract taxon-specific reads for downstream analysis, competitive mapping strategies can be used to eliminate reads that better align to known contaminant reference sequences (Andrades Valtueña et al. 2017), and mapping stringency can be relaxed and enhanced to determine the impact of poor-quality alignments on the mean coverage of the target reference sequence (Bos et al. 2019; see Pfeifer 2017 for a general overview of these bioinformatic pipelines). Researchers could also use fragment-misincorporation plots to determine if an alignment to a target reference sequence has the damage pattern expected in ancient DNA (i.e., C>T transitions in the sequencing reads due to deamination of cytosine to uracil, which is read by the Illumina sequencer as thymine). One may also, with sufficient coverage, only utilize reads with damage for downstream analysis, clipping the damaged ends for genotyping, as has been done with some ancient human datasets (Skoglund et al. 2014; Posth et al. 2018).

An additional consideration for future analyses of ancient *M.TB* DNA is within-host population structure, as discussed above. Successful next-generation sequencing (NGS) investigations of ancient *M.TB* DNA have largely been restricted to a few sampling sites in the body. In skeletonized remains, which make up the majority of samples that are screened, vertebral bodies with lesions suggestive of spinal TB have been the only positive source (excluding PCR-based studies) published thus far. In mummified remains, ribs, lung tissue, abdominal tissue, or calcified nodules from the lungs have yielded positive results (Kay et al. 2015; Sabin et al. 2020). These samples represent only one compartment per individual (Fig. 3). This sheds light on another difficulty of ancient DNA studies of TB, which is that the stochasticity of preservation of DNA throughout different elements (as demonstrated thus far with human DNA; Damgaard et al. 2015), or of human remains more generally, limits our ability to gather genetic data from the whole infectious community within an individual. This unpredictability is unavoidable when working with archeological remains. However, incorporating the biological reality of structuring between tuberculous granulomas and other lesions into discussions contextualizing genetic variation in ancient specimens will provide more nuance to our understanding of *M.TB* evolution.

Conclusion

The main challenge of studying drug-resistance evolution in *M.TB*, and other pathogens, is to disentangle the contribution of specific population-level evolutionary processes,

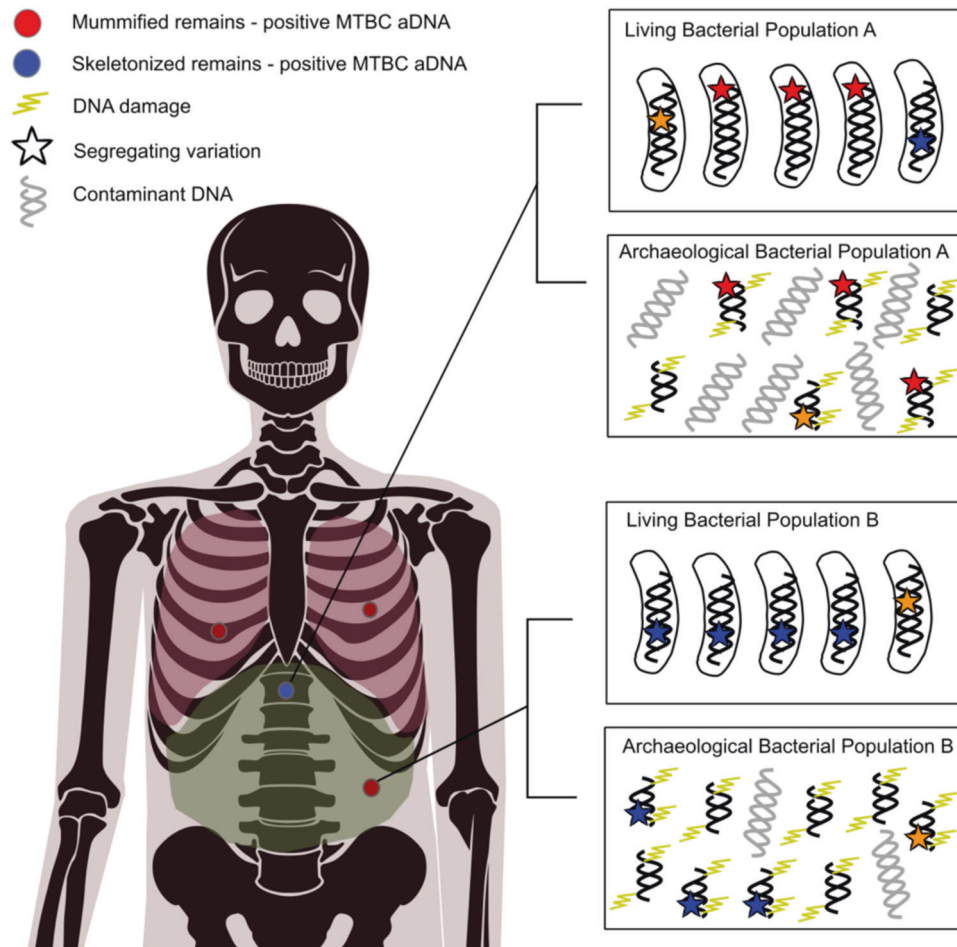


Fig. 3 *M.TB* structuring and preservation in archeological human remains. The red and blue dots indicate body sites from which ancient *M.TB* DNA has been successfully recovered using NGS techniques. Red dots represent successful recoveries from mummified remains where soft tissue had been preserved, with positive samples taken from a rib bone (Kay et al. 2015), a calcified lung nodule (Sabin et al. 2020), soft tissue from the lungs (Kay et al. 2015), and soft tissue taken from the abdominal cavity (Kay et al. 2015). Blue dots represent successful recoveries from skeletonized remains (Bos et al. 2014). Published

including positive selection and selective sweeps, purifying and background selection, population bottlenecks and structuring, and the underlying mutation rates and progeny distributions. Importantly, recent theoretical and statistical results are beginning to explore the abilities and limitations of joint parameter estimation (Eldon et al. 2015; Matuszewski et al. 2018; Sackman et al. 2019), and to construct an appropriate evolutionary null model for *M.TB*. Yet, these efforts in MMC models are in their infancy compared with the decades of important developments achieved under the more common Kingman coalescent, and continued theory development in this area is greatly needed (Wakeley 2013; Irwin et al. 2016). What has become clear thus far however is that background selection and progeny-skew effects are likely much more dominant in shaping the patterns of

positive findings from skeletonized remains have been limited to vertebrae as of the writing of this paper. In the majority of cases, only one sampling site is represented per individual. The diversity discovered in an individual sample may not be representative of the total infection population across different subpopulations within a host. In addition, the stochasticity of DNA preservation in terms of ubiquitous contamination, fragmentation, and cytosine-to-uracil deamination poses barriers to accurate reconstruction of bacterial population diversity during life.

variation in *M.TB* than previously appreciated (Morales-Arce et al. 2020; Menardo et al. 2020), and ought to be included in future null modeling. Fortunately, advanced simulation tools (e.g., SLiM; Haller and Messer 2019) readily allow for the modeling of these processes, as well as for the incorporation of population-size change and positive selection as well. Furthermore, experimental approaches are also beginning to quantify these parameters better—with mutation-accumulation studies better informing the underlying mutation rates (Ford et al. 2011; 2013), and digital barcoding offering insights on progeny distributions (Martin et al. 2017)—again highlighting the value of better integrating experimental and natural population studies (Bank et al. 2014). More mutation-accumulation studies in the genus *Mycobacterium* (e.g., Kucukyildirim et al. 2016) and

in members of MTBC specifically could shed light on the controversies surrounding mutation-rate evolution in *M.TB*. Such experimental evolution approaches could also prove invaluable for better-characterizing mutational interactions, the distribution of fitness effects, and progeny distributions. On the empirical side, new sequencing technologies combined with time-sampled data can detect lower-frequency genetic variation and provide improved statistical power to quantify these different evolutionary processes. Importantly, this quantification of within-patient variation will allow the field to move away from the common practice of only comparing per-individual consensus sequences (see Séraphin et al. 2019; Lee et al. 2020)—an approach that neglects the vast majority of segregating variation, thus hampering scans for resistance mutations, mixed-strain infections, and transmission chains. In addition to this enhanced view of modern variation, recent studies also provided insight into ancient within-host variation. As such, future research on the evolution of *M.TB* will benefit from both shallow-time serial patient sampling and the deep-time serial sampling of ancient genomes.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

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

References

- Alanis AJ (2005) Resistance to antibiotics: are we in the post-antibiotic era? *Arch Med Res* 36:697–705
- Andrades Valtueña A, Mittnik A, Key FM, Haak W, Allmae R et al. (2017) The stone age plague and its persistence in Eurasia. *Curr Biol* 27:3683–3691
- Bank C, Ewing G, Ferrer-Admetlla A, Foll M, Jensen JD (2014) Thinking too positive? Revisiting current methods of population genetic selection inference. *Trends Genet* 30:540–546
- Bos KI, Harkins K, Herbig A, Coscolla M, Weber N, Comas I et al. (2014) Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature* 514:494–497
- Bos KI, Kuhnert D, Herbig A, Esquivel-Gomez L, Andrades Valtueña A, Barquera R et al. (2019) Paleomicrobiology: diagnosis and evolution of ancient pathogens. *Annu Rev Microbiol* 73:639–666
- Cadena AM, Fortune S, Flynn J (2017) Heterogeneity in tuberculosis. *Nat Rev Immunol* 17:691–702
- Centers for Disease Control and Prevention (CDC) (1989) Tuberculosis and human immunodeficiency virus infection: recommendations of the Advisory Committee for the Elimination of Tuberculosis (ACET). *MMWR* 38:236–250
- Charlesworth B, Charlesworth D (2010) Elements of evolutionary genetics. Roberts and Company, USA
- Charlesworth B, Morgan M, Charlesworth D (1993) The effects of deleterious mutations on neutral molecular variation. *Genetics* 134:1289–1303
- Chiner-Oms A, Sanchez-Buso L, Corander J, Gagneux S, Harris S, Young D et al. (2019) Genomic determinants of speciation and spread of the *Mycobacterium tuberculosis* complex. *Sci Adv* 5: eaaw3307
- Cohen KA, Manson A, Desjardins C, Abeel T, Earl A (2019) Deciphering drug resistance in *Mycobacterium tuberculosis* using whole-genome sequencing: progress, promise, and challenges. *Genome Med* 11:45
- Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D et al. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393:537–544
- Crisci JL, Poh YP, Mahajan S, Jensen JD (2013) The impact of equilibrium assumptions on tests of selection. *Front Genet* 4:1–7
- Damgaard PB, Margaryan A, Schroeder H, Orlando L, Willerslev E, Allentoft M (2015) Improving access to endogenous DNA in ancient bones and teeth. *Sci Rep* 5:11184
- Dean GS, Rhodes S, Coad M, Whelan A, Cockle P, Clifford D et al. (2005) Minimum infective dose of *Mycobacterium bovis* in cattle. *Infect Immun* 73:6467–6471
- Dheda K, Gumbo T, Maartens G, Dooley K, McNerney R, Murray M et al. (2017) The epidemiology, pathogenesis, transmission, diagnosis, and management of multidrug-resistant, extensively drug-resistant, and incurable tuberculosis. *Lancet Respir Med* 5:291–360
- Dheda K, Lenders L, Magomedze G, Srivastava S, Raj P, Arning E et al. (2018) Drug-penetration gradients associated with acquired drug resistance in patients with tuberculosis. *Am J Respir Crit Care* 198:1208–1219
- Dicks KV, Stout JE (2019) Molecular diagnostics for *Mycobacterium tuberculosis* infection. *Annu Rev Med* 70:77–90
- Dookie N, Rambaran S, Padayatchi N, Mahomed S, Naidoo K (2018) Evolution of drug resistance in *Mycobacterium tuberculosis*: a review on the molecular determinants of resistance and implications for personalized care. *J Antimicrob Chemother* 73:1138–1151
- Dos Vultos T, Mestre O, Rauzier J, Golec M, Rastogi N, Rasolofo V et al. (2008) Evolution and diversity of clonal bacteria: The paradigm of *Mycobacterium tuberculosis*. *PLoS ONE* 3:e1538
- Eldholm V, Pettersson J, Brynildsrud O, Kitchen A, Rasmussen E, Lillebaek T et al. (2016) Armed conflict and population displacement as drivers of the evolution and dispersal of *Mycobacterium tuberculosis*. *Proc Natl Acad Sci USA* 113: 13881–13886
- Eldon B, Birkner M, Blath J, Freund F (2015) Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? *Genetics* 199:841–856
- Ford CB, Lin P, Chase M, Shah R, Iartchouk O, Galagan J et al. (2011) Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. *Nat Genet* 43:482–486
- Ford C, Yusim K, Ioerger T, Feng S, Chase M, Greene M et al. (2012) *Mycobacterium tuberculosis* – heterogeneity revealed through whole genome sequencing. *Tuberculosis (Edinb)* 92:194–201
- Ford CB, Shah R, Maeda M, Gagneux S, Murray M, Cohen T et al. (2013) *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nat Genet* 45:784–790
- Galagan JE (2014) Genomic insights into tuberculosis. *Nat Rev Genet* 15:307–320
- Ghajavand H, Kamakoli M, Khanipour S, Dizaji S, Masoumi M, Jamnani F et al. (2019) Scrutinizing the drug resistance mechanism of multi- and extensively-drug resistant *Mycobacterium tuberculosis*: mutations versus efflux pumps. *Antimicrob Resist Infect Control* 8:70

- Gygli SM, Borrell S, Trauner A, Gagneux S (2017) Antimicrobial resistance in *Mycobacterium tuberculosis*: mechanistic and evolutionary perspectives. *FEMS Microbiol Rev* 41:354–373
- Haller BC, Messer PW (2019) SLiM 3: forward genetic simulations beyond the Wright–Fisher model. *Mol Biol Evol* 36:632–637
- Harris RB, Sackman A, Jensen JD (2018) On the unfounded enthusiasm for soft selective sweeps II: examining recent evidence from humans, flies, and viruses. *PLoS Genet* 14:e1007859
- Hershberg R, Lipatov M, Small P, Sheffer H, Niemann S, Homolka S et al. (2008) High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. *PLoS Biol* 6:e311
- Irwin KK, Laurent S, Matuszewski S, Vuilleumier S, Ormond L, Shim H et al. (2016) On the importance of skewed offspring distributions and background selection in virus population genetics. *Heredity* 117:393–399
- Jensen JD, Payseur BA, Stephan W, Aquadro CF, Lynch M, Charlesworth D et al. (2019) The importance of the Neutral Theory in 1968 and 50 years on: a response to Kern and Hahn 2018. *Evolution* 73:111–114
- Kay GL, Sergeant M, Zhou Z, Chan J, Millard A, Quick J et al. (2015) Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. *Nat Commun* 6:6717
- Kondrashov AS, Crow JF (1991) Haploidy or diploidy: which is better? *Nature* 351:314–315
- Köser CU, Ellington M, Peacock S (2014) Whole-genome sequencing to control antimicrobial resistance. *Trends Genet* 30:401–407
- Kucukyildirim S, Long H, Sung W, Miller S, Doak T, Lynch M (2016) The rate and spectrum of spontaneous mutations in *Mycobacterium smegmatis*, a bacterium naturally devoid of the post-replicative mismatch repair pathway. *G3* 6:2157–2163
- Kurz SG, Furin J, Bank C (2016) Drug-resistant tuberculosis: challenges and progress. *Infect Dis Clin N Am* 30:509–522
- Leanerts A, Barry III C, Dartois V (2015) Heterogeneity in tuberculosis pathology, microenvironments and therapeutic responses. *Immunol Rev* 264:288–307
- Lee RS, Radomski N, Proulx J-F, Levade I, Shapiro B, McIntosh F et al. (2015) Population genomics of *Mycobacterium tuberculosis* in the Inuit. *Proc Natl Acad Sci USA* 112:13609–13614
- Lee RS, Proulx JF, McIntosh F, Behr M, Hanage W (2020) Previously undetected super-spreading of *Mycobacterium tuberculosis* revealed by deep sequencing. *Elife* 9:e53245
- Ley SD, de Vos M, Van Rie A, Warren R (2019) Deciphering within-host microevolution of *Mycobacterium tuberculosis* through whole-genome sequencing: the phenotypic impact and way forward. *Microbiol Mol Biol Rev* 83:e00062–18
- Lieberman T, Wilson D, Misra R, Xiong L, Moodley P, Cohen T et al. (2016) Genomic diversity in autopsy samples reveals within-host dissemination of HIV-associated *Mycobacterium tuberculosis*. *Nat Med* 22:1470–1474
- Liu F, Hu Y, Wang Q, Min Li H, Gao G, Liu C et al. (2014) Comparative genomic analysis of *Mycobacterium tuberculosis* clinical isolates. *BMC Genom* 15:469
- Liu Q, Ma A, Wie L, Pang Y, Wu B, Luo T et al. (2018) China's tuberculosis epidemic stems from historical expansion of four strains of *Mycobacterium tuberculosis*. *Nat Ecol Evol* 2:1982–1992
- Liu Q, Via L, Luo T, Liang L, Liu X, Wu S et al. (2015) Within patient microevolution of *Mycobacterium tuberculosis* correlates with heterogeneous responses to treatment. *Sci Rep* 5:17507
- Martin CJ, Cadena A, Leung V, Lin P, Maiello P, Hicks N et al. (2017) Digitally barcoding *Mycobacterium tuberculosis* reveals in vivo infection dynamics in the macaque model of tuberculosis. *mBio* 8:e00312–e00317
- Matuszewski S, Hildebrandt M, Achaz G, Jensen JD (2018) Coalescent processes with skewed offspring distributions and nonequilibrium demography. *Genetics* 208:323–338
- Meehan C, Goig G, Kohl T, Verboven L, Dippenaar A, Ezewudo M et al. (2019) Whole genome sequencing of *Mycobacterium tuberculosis*: current standards and open issues. *Nat Rev Genet* 17:533–545
- Menardo F, Gagneux S, Freund F (2020) Multiple merger genealogies in outbreaks of *Mycobacterium tuberculosis*. *Mol. Biol. Evol.*, in press.
- Minias A, Minias P, Czubat B, Dziadek J (2018) Purifying selective pressure suggests the functionality of a vitamin B12 biosynthesis pathway in a global population of *Mycobacterium tuberculosis*. *Genome Biol Evol* 10:2326–2337
- Morales-Arce AY, Harris R, Stone A, Jensen JD (2020) Evaluating the contributions of purifying selection and progeny-skew in dictating within-host *Mycobacterium tuberculosis* evolution. *Evolution* 74:5:992–1001
- Moreno-Gamez S, Hill A, Rosenblum D, Petrov D, Nowak M, Pennings P (2015) Imperfect drug penetration leads to spatial monotherapy and rapid evolution of multidrug resistance. *Proc Natl Acad Sci USA* 112:E2874–E2883
- Müller B, Borrell S, Rose G, Gagneux S (2013) The heterogeneous evolution of multidrug-resistant *Mycobacterium tuberculosis*. *Trends Genet* 29:160–169
- Namouchi A, Didelot X, Schock U, Gicquel B, Rocha E (2012) After the bottleneck: genome-wide diversification of the *Mycobacterium tuberculosis* complex by mutation, recombination, and natural selection. *Genome Res* 22:721–734
- Navarro Y, Perez-Lago L, Herranz M, Sierra O, Comas I, Sicilia J et al. (2017) In-depth characterization and functional analysis of clonal variants in a *Mycobacterium tuberculosis* strain prone to microevolution. *Front Microbiol* 8:694
- Nimmo C, Shaw L, Doyle R, Williams R, Brien K, Burgess C et al. (2019) Whole genome sequencing *Mycobacterium tuberculosis* directly from sputum identifies more genetic diversity than sequencing from culture. *BMC Genom* 20:389
- O'Donnell MR, Larsen M, Brown T, Jain P, Munsamy V, Wolf A et al. (2019) Early detection of emergent extensively drug-resistant tuberculosis by flow cytometry-based phenotyping and whole-genome sequencing. *Antimicrob Agents Chemother* 63:e01834–18
- O'Grady F, Riley RL (1963) Experimental airborne tuberculosis. *Adv Tuberc Res* 12:150–190
- Ohta T (1973) Slightly deleterious mutant substitutions in evolution. *Nature* 246:96–98
- O'Neill MB, Mortimer T, Pepperell C (2015) Diversity of *Mycobacterium tuberculosis* across evolutionary scales. *PLoS Pathog* 11:e1005257
- O'Neill MB, Shockey A, Zarley A, Aylward W, Eldholm V, Kitchen A et al. (2019) Lineage specific histories of *Mycobacterium tuberculosis* dispersal in Africa and Eurasia. *Mol Ecol* 28:3241–3256
- Operario DJ, Koeppel A, Turner S, Bao Y, Pholwat S, Banu S et al. (2017) Prevalence and extent of heteroresistance by next generation sequencing of multidrug-resistant tuberculosis. *PLoS ONE* 12:e0176522
- Otto SP, Gerstein AC (2008) The evolution of haploidy and diploidy. *Curr Biol* 18:R1121–R1124
- Payne JL, Menardo F, Trauner A, Borrell S, Gygli S, Loiseau C et al. (2019) Transition bias influences the evolution of antibiotic resistance in *Mycobacterium tuberculosis*. *PLoS Biol* 17:e3000265
- Pepperell C, Hoepfner V, Lipatov M, Wobeser W, Schoolnik G, Feldman M (2010) Bacterial genetic signatures of human social

- phenomena among *M. tuberculosis* from an Aboriginal Canadian population. *Mol Biol Evol* 27:427–440
- Pepperell CS, Casto A, Kitchen A, Granka J, Comejo O, Holmes E et al. (2013) The role of selection in shaping diversity of natural *M. tuberculosis* populations. *PLoS Pathog* 9:e1003543
- Pfeifer SP (2017) From next-generation resequencing reads to a high-quality variant data set. *Heredity* 118:111–124
- Posth C, Nakatsuka N, Lazaridis I, Skoglund P, Mallick S, Lamnidis T et al. (2018) Reconstructing the deep population history of central and South America. *Cell* 175:1185–1197
- Renzette N, Pfeifer SP, Matuszewski S, Kowalik TF, Jensen JD (2017) On the analysis of intrahost and interhost viral populations: human cytomegalovirus as a case study of pitfalls and expectations. *J Virol* 5:e01976–16
- Rich AR (1946) Pathogenesis of tuberculosis. Thomas, Springfield, IL
- Ruesen C, Chaidir L, van Laarhoven A, Dian S, Rizal Ganiem A, Nebenzahl-Guimaraes H et al. (2018) Large-scale genomic analysis shows association between homoplastic genetic variation in *Mycobacterium tuberculosis* genes and meningeal or pulmonary tuberculosis. *BMC Genom* 19:122
- Ryndak MB, Laal S (2019) *Mycobacterium tuberculosis* primary infection and dissemination: a critical role for alveolar epithelial cells. *Front Cell Infect Microbiol* 9:299
- Sabin S, Herbig A, Vagene A, Ahlstrom T, Bozovic G, Arcini C et al. (2020) A seventeenth-century *Mycobacterium tuberculosis* genome supports a Neolithic emergence of the *Mycobacterium tuberculosis* complex. *Genome Biol.*, in press.
- Sackman AM, Harris RB, Jensen JD (2019) Inferring demography and selection in organisms characterized by skewed offspring distributions. *Genetics* 211:1019–1028
- S raphin MN, Norman A, Rasmussen E, Gerace A, Chiribau C, Rowlinson M-C et al. (2019) Direct transmission of within-host *Mycobacterium tuberculosis* diversity to secondary cases can lead to variable between-host heterogeneity without *de novo* mutation: a genomic investigation. *EBioMedicine* 47:293–300
- Singh R, Dwivedi S, Gaharwar U, Meena R, Rajamani P, Prasad T (2020) Recent updates on drug resistance in *Mycobacterium tuberculosis*. *J Appl Microbiol* 128:1547–1567
- Shockey AC, Dabney J, Pepperell C (2019) Effects of host, sample, and in vitro culture on genomic diversity of pathogenic *Mycobacteria*. *Front Genet* 10:477
- Skoglund P, Northoff B, Shunkov M, Derevianko A, Paabo S, Krause J et al. (2014) Separating ancient DNA from modern contamination in a Siberian Neandertal. *Proc Natl Acad Sci USA* 111:2229–2234
- Sonin LS (1951) The role of particle size in experimental airborne infection. *Am J Hyg* 53:337–354
- Tellier A, Lemaire C (2014) Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. *Mol Ecol* 23(11):2637–2652
- Thornton KR, Jensen JD (2007) Controlling the false-positive rate in multilocus genome scans for selection. *Genetics* 175:735–750
- Toungoussova OS, Bjune G, Caugant D (2006) Epidemic of tuberculosis in the former Soviet Union: Social and biological reasons. *Tuberculosis* 86:1–10
- Trauner A, Liu Q, Via L, Lin X, Ruan X, Liang L et al. (2017) The within-host population dynamics of *Mycobacterium tuberculosis* vary with treatment efficacy. *Genome Biol* 18:71
- Wakeley J (2013) Coalescent theory has many new branches. *Theor Pop Biol* 87:1–4
- Walker TM, Ip C, Harrell R, Evans J, Kapatai G, Dedicoat M et al. (2013) Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect Dis* 13:137–146
- Walsh B, Lynch M (2018) Evolution and selection of quantitative traits. Oxford University Press, UK
- Warinner C, Herbig A, Mann A, Fellows Yates J, Weiss C, Burbano H et al. (2017) A robust framework for microbial archaeology. *Annu Rev Genomics Hum Genet* 18:321–356
- World Health Organization (2019) Global tuberculosis report Geneva. World Health Organization, Geneva, Switzerland. Licence, CCBY-NC-SA3.0IGO
- Zignol M, Cabibbe A, Dean A, Glaziou P, Alikhanova N, Ama C et al. (2018) Genetic sequencing for surveillance of drug resistance in tuberculosis in highly endemic countries: a multi-country population-based surveillance study. *Lancet Infect Dis* 18:675–683