How do monomorphic bacteria evolve? The *Mycobacterium tuberculosis* complex and the awkward population genetics of extreme clonality

Christoph Stritt, Sebastien Gagneux

Swiss Tropical and Public Health Institute, Allschwil, Switzerland; University of Basel, Basel, Switzerland

December 14, 2022

Abstract

Exchange of genetic material through sexual reproduction or horizontal gene transfer is ubiquitous in nature. Among the few outliers that rarely recombine and mainly evolve by de novo mutation are a group of deadly bacterial pathogens, including the causative agents of leprosy, plague, typhoid, and tuberculosis. The interplay of evolutionary processes is poorly understood in these organisms. Population genetic methods allowing to infer mutation, recombination, genetic drift, and natural selection make strong assumptions that are difficult to reconcile with clonal reproduction and fully linked genomes consisting mainly of coding regions. In this review, we highlight the challenges of extreme clonality by discussing population genetic inference with the Mycobacterium tuberculosis *complex, a group of closely related obligate bacterial pathogens of mammals. We show how uncertainties underlying quantitative models and verbal arguments affect previous conclusions about the way these organisms evolve. A question mark remains behind various quantities of applied and theoretical interest, including mutation rates, the interpretation of nonsynonymous polymorphisms, or the role of genetic bottlenecks. Looking ahead, we discuss how new tools for evolutionary simulations, going beyond the traditional Wright-Fisher framework, promise a more rigorous treatment of basic evolutionary processes in clonal bacteria.*

Introduction

Mutation, recombination, genetic drift, and natural selection are the basic evolutionary processes that drive the evolution of life. It is the aim and "great obsession" of population genetics to infer these processes from patterns of genetic variation observed in nature (Gillespie, 2004). Since the Modern Synthesis of evolutionary biology in the 1930s, a variety of mathematical models have been developed for this purpose, which today are in wide use in the analysis of genome sequencing data (Templeton, 2021).

A problem in the application of population genetic models to empirical data is that modeling assumptions can be a far cry from the biology and life history of real organisms. Archea and bacteria reproduce clonally through binary fission, frequently undergo horizontal gene transfer (HGT), and have genomes consisting mainly of coding regions. These characteristics are difficult to reconcile with models that are tailored to animals and plants (Woese and Goldenfeld, 2009) and commonly assume random mating, linkage equilibrium, and neutrality (Maynard Smith, 1995; Rocha, 2018). As a consequence, outside the laboratory, studies of bacterial population genetics have either remained descriptive, with much effort going into understanding the extent and effects of HGT (e.g. Denamur et al., 2021); or have resorted to models whose applicability remains an open question (discussed by Johri et al., 2022).

While the opportunistic, hardly predictable process of HGT has been highlighted as the most problematic breach of assumptions (Maynard Smith, 1995), a different, less frequently discussed challenge arises from the opposite extreme of the recombination spectrum: strictly clonal evolution, or the absence of any gene flow. HGT is not a general characteristic of bacteria (Hanage, 2016). Some bacteria are "monomorphic", that is, characterized by low levels of sequence diversity and an apparent absence of genetic exchange (Achtman, 2008). The causative agents of several devastating bacterial diseases of humans and animals belong to this group, including *Bacillus anthracis* (anthrax), *Salmonella enterica* serotype typhi (typhoid), *Yersinia pestis* (plague), *Mycobacterium leprae* (leprosy), and the members of the *Mycobacterium tuberculosis* complex (tuberculosis). Our understanding of the evolution of these bacteria is hampered not only by the low information content in their genomes, but also because there is little theoretical and conceptual work on population genetic inference under extreme clonality.

It has been suggested that phylogenies are all that is needed to study non-recombining bacteria, bacterial population genetics thus becoming "a branch of cladistics" (Maynard Smith, 1995). In the absence of recombination, genetic linkage is complete and a genome is behaving as a single non-recombining locus. This makes for neat phylogenies, since every part of the genome has the same genealogical history. But it also complicates the inference of the processes underlying the observed tree. Under strict clonality, the fate of a mutation arising in any of the few thousand genes present in a typical bacterial genome is tied to all other sites in the genome. Selection acting on this mutation affects the fixation probability of linked variation and interferes with selection at other sites (Charlesworth, 2012; Neher, 2013). The dynamics and outcome of such linked selection depend on a parameter that is usually unknown: the distribution of fitness effects of new mutations (Eyre-Walker and Keightley, 2007). Periodic selection, for instance, results when beneficial mutations are rare enough such that selective sweeps are well separated in time. More complex dynamics emerge when beneficial mutations are frequent and co-occur in the same population or on the same chromosome (Sniegowski and Gerrish, 2010).

Linked selection is rarely mentioned or investigated in the context of extremely clonal bacteria, as already observed ten years ago (Charlesworth, 2012). This is an important omission, as it is not all that clear how one would go about inferring evolutionary processes from fully linked genomes. What biases are introduced when linkage equilibrium and neutrality are assumed when analyzing clonal genomes? Can we meaningfully talk of "populations" when each bacterial cell is a genetically isolated island?

Mycobacterium tuberculosis as a model for clonal evolution

In this review, we highlight the obligate pathogens of the *Mycobacterium tuberculosis* complex (MTBC) as a model to study evolution under extreme clonality. The MTBC comprises a group of closely related obligate pathogens that cause tuberculosis (TB) in humans and a range of wild and domestic animals (Figure 1). Human TB mainly affects the global poor and has killed more than 1.6 million people in 2021 (World Health Organization, 2022), while it also affected large parts of society in Europe and Northern America up to the early 20th century (Dubos and Dubos, 1952). Today, the evolution of antibiotic resistance is a main challenge and focus of research in TB. The genomes of thousands of MTBC strains from around the world have been sequenced, mainly to study epidemiological dynamics and drug resistance evolution, but also to infer the origin and biogeographic history of the species (Gagneux, 2018).

Members of the MTBC are among the more diverse of the predominantly clonal bacteria

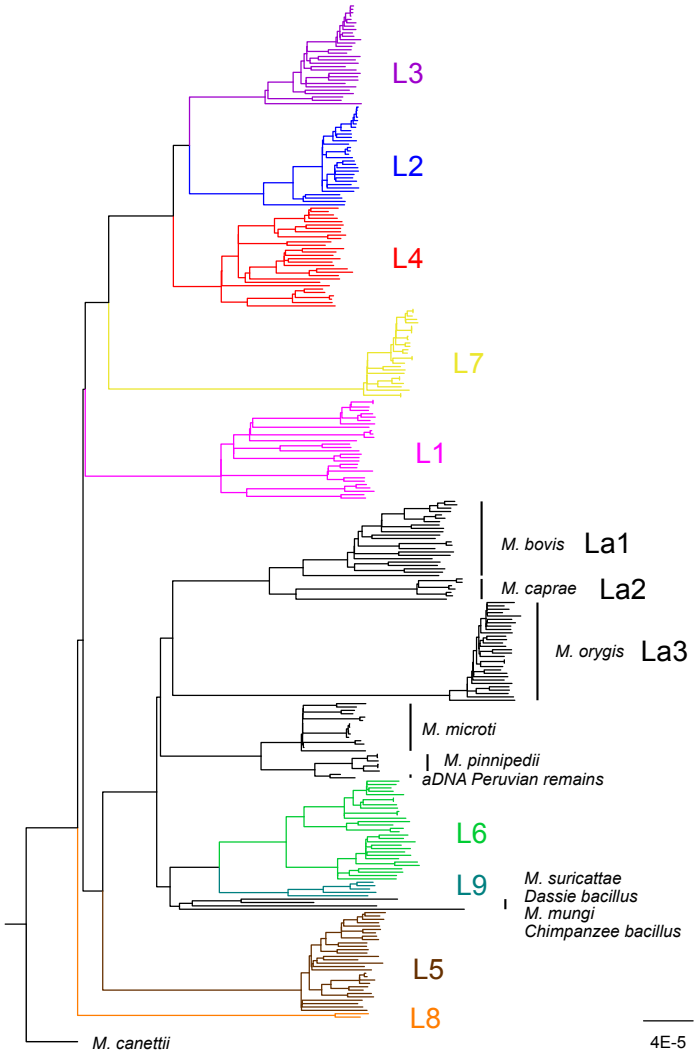

Figure 1: *Rooted maximum likelihood phylogeny of the MTBC estimated from genome-wide SNPs (tree adapted from Zwyer et al. 2021).* M. canettii *is the outgroup, human-adapted lineages (L1 to L9) are shown in colors, animal-adapted lineages in black. Species names represent the historically grown nomenclature, lineage names are a more recent classification based on genomic data. Lineages 1 to 4 and 7 are also referred to as* M. tuberculosis sensu stricto, *lineages 5 and 6 as* M. africanum. *Bootstrap supports for the lineages are above 0.95 and are not displayed in the figure.*

(Achtman, 2012), even though individual strains differ only by a maximum of ca. 2,400 SNPs across the 4.4 Mb genome (Figure 2a). At the molecular level, the MTBC is further characterized by a high GC content, a high proportion of nonsynomyous polymorphisms, and a low proportion of homoplastic mutations (Figure 2b-d). It seems that the low diversity of the MTBC has deterred evolutionary biologists from engaging with this bacterium. Many studies content with speculative invocations of genetic drift and natural selection, typically referring to the triad Hershberg et al. (2008), Namouchi et al. (2012) and Pepperell et al. (2013), who are among the few studies in the

large MTBC literature that have put basic evolutionary processes into focus.

In this review, we present the main hypotheses about what drives the evolution of the MTBC, and how they have been arrived at. Particular attention is paid to models, their assumptions, and the traits of the MTBC that might conflict with the latter. Evolutionary simulations are discussed as a way to achieve a more quantitative treatment of frequently invoked processes such as purifying selection or periodic bottlenecks.

Recombination

How "strict" is clonality in the MTBC? In the past, bacteria were classified as "clonal" or "monomorphic" based on a handful of housekeeping genes (Maynard Smith et al., 1993; Selander et al., 1987). With the full resolution of whole genome sequences, this classification needs to be reassessed. As discussed in the following, experimental and observational evidence agree that the MTBC is predominantly clonal, and that few to no new genes have found their way into the MTBC since the most recent common ancestor of the currently existing lineages. In contrast to interstrain recombination, intrachromosomal recombination is common and increasingly recognized as an important source of genetic variation.

Experimental evidence: genetic factors versus lack of opportunity

Most of the knowledge about the molecular mechanisms of HGT in mycobacteria stems from research with *Mycobacterium smegmatis*, a fast-growing, non-pathogenic mycobacterium more easily amenable to cultivation and genetic engineering than the bacteria of the MTBC. Mycobacteria lack the traditional components of HGT, possibly because transfer through the complex cell envelopes of these diderm bacteria requires other mechanisms (Madacki et al., 2021). Investigations of gene transfer in *M. smegmatis* have led to the description of a previously unknown form of bacterial conjugation: distributive conjugal transfer (DCT, reviewed by Gray and Derbyshire, 2018). DCT involves the transfer of chromosomal DNA and gives rise to mosaic genomes, with hundreds of pieces of DNA of variable sizes dispersed in the receiver genome. DCT thus challenges the paradigm of "localized sex" in bacteria (Smith et al., 1991) and might explain the recombinogenic population structure of many mycobacteria (Panda et al., 2018).

Of particular interest regarding the evolution of the MTBC is the observation of DCT in the closely related *Mycobacterium canettii*. *M. canettii* shares an average nucleotide identity of 97.5% with the MTBC yet is strikingly more diverse: a handful of *M. canettii* strains from eastern Africa harbor more genetic diversity than the whole MTBC (Supply et al., 2013). Mating assays have shown that DCT occurs in *M. canettii*, while no DCT was observed between three MTBC strains (Boritsch et al., 2016). The same assays combining *M. canettii* and MTBC strains revealed that the latter can act as donors but not as receivers of DNA during DCT, as pieces of MTBC DNA were integrated into *M. canettii* genomes but not vice versa (Madacki et al., 2021). In *M. smegmatis*, polymorphisms in the *esxI* secretion locus underlay self identity and conjugal compatibility (Clark et al., 2022). In *M. canettii* and the MTBC, the molecular mechanisms underlying conjugal compatibility do not depend on *esxI* and remain to be elucidated (Madacki et al., 2021).

Lack of opportunity has been proposed to explain why intracellular pathogens such as the MTBC do not seem to recombine (Casadevall, 2008; Chiner-Oms et al., 2019b). Alternatively, avoidance of HGT could be an evolutionary strategy with a genetic basis, an adaptation to parasitism (Tibayrenc and Ayala, 2017). Against the first scenario, it can be argued that there

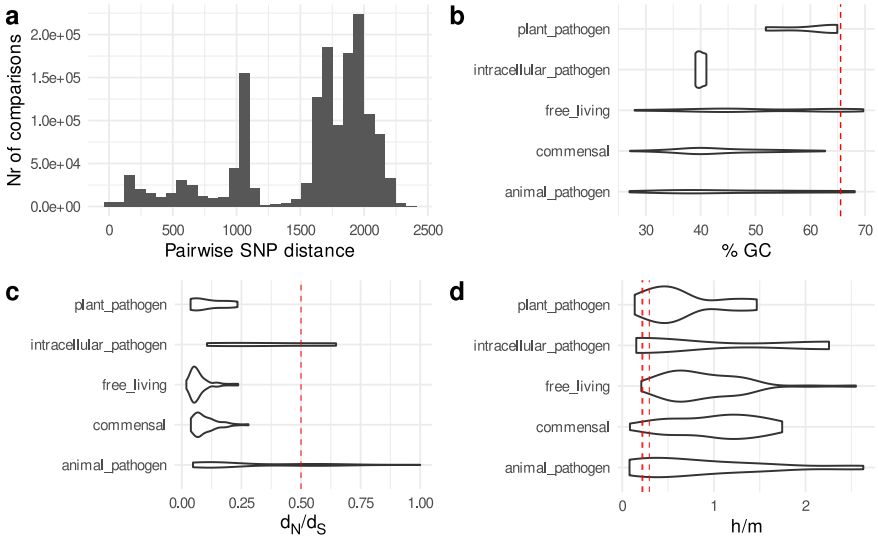

Figure 2: *Genetic diversity and molecular characteristics of the MTBC. a) Pairwise genetic differences between the strains shown in Figure 1, based on single nucleotide polymorphisms. b) to d) show molecular characteristics of the MTBC compared to 150 other bacterial species with diverse lifestyles (data from Bobay and Ochman, 2018). Red lines show the values for the bacteria of the MTBC (*M. tuberculosis sensu stricto, M. bovis, *and* M. africanum*) along the distributions. b) GC content, c)* d_N/d_S*, the genome-wide ratio of nonsynonymous to synonymous polymorphisms, d) the ratio of homoplastic to non-homoplastic mutations, a proxy for recombination.*

is indeed more opportunity to recombine than the label "intracellular pathogen" might suggest. The bacteria of the MTBC are not confined to intracellular environments, but are also present in large extracellular populations after the induction of necrosis (Orme, 2014). Furthermore, mixed infections do occur (Moreno-Molina et al., 2021; Tarashi et al., 2017), such that diverged strains might have the opportunity to recombine. Further investigation into the genetic and environmental determinants of extreme clonality would be worthwhile, and the *M. canettii*-MTBC system provides a great opportunity to elucidate the poorly understood evolutionary transition to extreme clonality characteristic of many obligate pathogens.

Recombination between closely related strains: how strict is clonality?

Genome sequences from diverse MTBC strains are an important complement to experimental data, which leave open the question how far the observed outcome depends on the specific conditions and strains used in the laboratory. Various studies have investigated the extent of HGT in natural strains of the MTBC, motivated by the observation how HGT accelerates resistance evolution in other bacterial pathogens (Davies and Davis, 2010). Some have suggested that interstrain recombination does occur. Liu et al. (2006), using datasets of 36 synonymous SNPs in 3,320 strains and 407 SNPs in 37 strains, found that mutation alone cannot explain the observed haplotype diversity, and identified a mosaic region in front of a *PPE* gene suggesting a recombination hotspot. They also point out the possibility that the pattern may have arisen through recombination between homologous sequences in the same genome. Namouchi et al. (2012) investigated 24

sequenced MTBC genomes and reported that "four different approaches showed evident signs of recombination in *M. tuberculosis*", with recombination typically involving small tracts of around 50 bp. On the other hand, the most extensive investigation to date, using different methods on genome-wide SNPs in 1,591 diverse strains, found "no measurable ongoing recombination among the MTBC strains" (Chiner-Oms et al., 2019b).

Generalizing from these studies is difficult due to the diversity of datasets and methods used. It has been suggested that the signs of recombination described by Namouchi et al. are mainly artefacts as they are overrepresented in regions difficult to align or assemble, in particular repetitive and low-complexity regions in insertion sequences and the expanded *PE/PPE* gene families (Godfroid et al., 2018). Alternatively, signs of recombination can arise from gene conversion during intrachromosomal recombination, to which these repetitive sequences are prone (Liu et al., 2006). Gene conversion is the non-reciprocal transfer of DNA from one homologous sequence to another, which in the MTBC might account for recombination signatures in *ESX*, *PE*, *PPE*, *PE/PGRS* gene families (Karboul et al., 2008; Phelan et al., 2016; Uplekar et al., 2011). Intrachromosomal recombination can also have more dramatic outcomes. More and more structural variants are described in MTBC genomes, ranging from insertion sequence (McEvoy et al., 2007) and gene copy number polymorphisms (Fishbein et al., 2015) to massive inversions (Merrikh and Merrikh, 2018) and tandem duplications (Wang et al., 2022). This is a vast topic deserving a dedicated review. It is brought up here to emphasize that recombination is an umbrella term for diverse processes of inter- and intrachromosomal exchange; and that clonality does therefore not imply absence of recombination, strictly speaking, but only of HGT. In the near future, long-read sequencing should allow more extensive studies of the repetitive "dark matter" in the MTBC genome and how it generates genetic variation intrachromosomally.

A basic limitation of methods to infer recombination is that they cannot distinguish de novo mutations from allelic recombination between closely related individuals, which might involve the exchange of a single nucleotide (Martin et al., 2011). Allelic recombination does not introduce new genes, but it can affect the nucleotide landscape through recombination-associated processes like biased gene conversion (Duret and Galtier, 2009) or increased mutation rates around strand breaks (Fitzgerald and Rosenberg, 2019). While HGT between close relatives would be less restricted by opportunity, genetic incompatibilies might prevent gene transfer between close relatives, as in *M. smegmatis* (Clark et al., 2022).

Mutation

While in some bacteria new variants are more likely to be generated by HGT than by mutation (Vos and Didelot, 2009), under extreme clonality *de novo* mutations are the main source of genetic diversity and adaptation. The speed and direction in which a clonal prokaryote evolves is thus determined by the rate and spectrum of new mutations and by their effect on fitness. Numerous studies have investigated mutagenesis in the MTBC (reviewed by Mcgrath et al., 2014). As discussed below, in addition to methodological issues in estimating mutation rates, the life history of the bacteria, which can include extended periods of dormancy, poses a main challenge in understanding the rate at which variation originates *in vivo*.

In the MTBC literature, as elsewhere, the mutation rate is sometimes confounded with the molecular clock rate. While the former refers to the rate at which mutations appear in the genome, the latter stands for an allegedly constant rate at which mutations accumulate through time (Ho et al., 2011). Both rates are subsumed in the more general concept of evolutionary rates. As discussed

below, the power law that describes the slowing of evolutionary rates as one considers longer timescales is not as clear in the MTBC as in other bacteria: *in vitro* mutation rate estimates can be similar to clock rate estimates from datasets including ancient DNA. How far methodological biases or evolutionary processes underly this surprising finding remains to be understood.

Plasticity of mutation rates and generation times

In the model mycobacterium *M. smegmatis*, a mutation rate of 5.27×10^{-10} mutations per site per generation was inferred in a mutation accumulation experiment (Kucukyildirim et al., 2016). For the MTBC itself, no such experiment has been conducted yet. Fluctuation assays suggest that point mutations in the MTBC appear at a rate of about 2.1×10^{-10} mutations per site per generation and at a similar rate during active disease in macaques if a generation time of 20 h is assumed (Ford et al., 2011, Figure 3). A later study, using the same fluctuation assay, found *in vitro* rates of 6.01×10^{-10} in a lineage 4 and 2.16×10^{-9} in a lineage 2 strain, suggesting somewhat faster and variable mutation rates within the MTBC (Ford et al., 2013). Comparatively fast rates were also proposed in two additional experimental evolution studies. After serial passaging of a MTBC strain through macrophage-like THP1 cells for 80 generations, Guerrini et al. (2016) inferred a rate of 5.7×10^{-9} per bp per generation. Copin et al. (2016), passaging bacteria in mice and assuming a generation time of 20 h, estimated a mutation rate of 3.8×10^{-9} in wild type mice and of 7.7×10^{-10} in T cell-deficient mice, suggesting that the presence of T cells leads to elevated mutation rates.

Overall, per-generation mutation rates estimated for the MTBC are well within the range of those in other bacteria, which typically are in the order 10^{-10} (reviewed by Katju and Bergthorsson, 2019). When trying to scale mutation rates to calendar time, however, complications due to the complex life history of these bacteria become apparent. The bacteria of the MTBC have long generation times ranging from 18 h in nutrient rich medium to potentially much longer time-spans *in vivo* (Colangeli et al., 2020). Assuming a generation time of 24 hours, a mutation rate of 2.1×10^{-10} translates to 7.7×10^{-8} mutations per site per year, or about 0.34 per genome per year, which is indeed low compared to other bacteria (Duchêne et al., 2016; Lynch, 2010).

In contrast to pathogens employing a "hit and run" strategy, bacteria of the MTBC can enter a state of reduced activity and persist for years in latent infections (Dutta and Karakousis, 2014). It is unclear whether latency and longer generation times imply a reduced mutation rate, as expected if mutation is driven by replication, or not, as expected if environmental stress drives mutation (Weller and Wu, 2015). Ford et al. (2011), in their experimental infection of macaques, found similar rates in latent and active disease (Figure 3), supporting stress-induced mutagenesis. A more complex, two-phased scenario was suggested by Colangeli et al. (2020), who investigated 24 paired TB cases with latently infected household contacts: mutation rates remained high up to two years, but then decrease with longer latency as the bacteria enter a quiescent state with longer generation times (Figure 3).

In summary, mutation rates estimated for the MTBC should be interpreted with some caution. Generation times are only known with confidence *in vitro*. At the same time, fluctuation assays reflect the mutation rate of a single gene (*rpoB*, the main drug resistance target of rifampicin) that might not be representative for the whole genome (Katju and Bergthorsson, 2019); and in the absence of stress, which *in vivo* might alter both the rate and the spectrum of new mutations (Fitzgerald and Rosenberg, 2019).

Why are MTBC genomes so GC-rich?

In bacteria, newly arising mutations are biased towards adenines and thymines (Hershberg and Petrov, 2010; Hildebrand et al., 2010), which in the MTBC might reflect stress-induced mutagenesis in an intracellular environment rich in reactive oxgen and nitrogen species (Chiner-Oms et al., 2019a; Liu et al., 2020). If mutation bias and genetic drift alone would determine the nucleotide landscape (mutation-drift equilibrium), the expected GC content in the MTBC would be 41.5% (Hershberg and Petrov, 2010). MTBC genomes, however, consist to 65.6% of guanines and cytosines (Figure 2b; Cole et al., 1998), with values of 80% at synonymous and 60% at nonsynonymous sites. Such a discrepancy between observed and expected GC content is observed in many prokaryotes, whose genomes vary hugely in GC content (Figure 2b). It implies that an unknown process, unaccounted for in standard models of molecular evolution, affects the segregation of polymorphisms through time (Rocha and Feil, 2010).

Several large-scale comparative studies have attempted to find a general explanation for the discordance between expected and observed GC content in prokaryotes. One prominent hypothesis is that nucleotide composition reflects adaptation to environmental conditions, for example through selection for thermal stability of DNA (e.g. Reichenberger et al., 2015). An intriguing twist to this idea was recently added by Weissman et al. (2019), who described a correlation between GC content, environmental variables, and the presence of *Ku*, the key gene in the non-homologous end-joining (NHEJ) pathway for DNA break repair. The authors propose that high GC content could be beneficial in bacteria suffering stress-induced double strand breaks in periods of slow or no growth, when NHEJ is required for repair because only a single copy of the genome is present. This is an interesting scenario for the MTBC, where long periods of latency can occur (see above) and the *Ku* gene is highly conserved.

An alternative explanation for GC bias that does not imply a selective advantage is GC-biased gene conversion (gBGC). This process occurs during homologous recombination when mismatches in heteroduplex DNA are preferentially resolved into guanines and cytosines (reviewed by Duret and Galtier, 2009). The gBGC hypothesis predicts that GC content is higher in regions with high recombination rates, which is observed in mammalian genomes. In bacteria, the role of gBGC is contested. Whether comparative studies find associations between GC content and recombination depends on the method used to infer recombination, and exceptions to general trends are common (Bobay and Ochman, 2017; Lassalle et al., 2015).

With its numerous genome sequences that can be placed in a robust phylogenetic framework, the MTBC provides an opportunity to study the evolution of base composition in detail and thus to complement comparative studies. A hypothesis to test is that the MTBC is evolving from the generally GC-rich state of mycobacteria (58 to 70%, *Mycobacterium* sp. genomes on NCBI) to a more AT-rich state characteristic of obligate pathogens (Rocha and Danchin, 2002, Figure 2b), including *Mycobacterium leprae* (58%).

The time (in)dependence of evolutionary rates in the MTBC

Molecular dating has led to a re-evaluation of the origin and history of the MTBC, as for many other organisms. Earlier studies, assuming a synonymous mutation rate or a co-diversification of humans and the MTBC, located the most recent common ancestor of the existing lineages in Africa and suggested a scenario according to which humans and the MTBC have co-diversified across the globe (Comas et al., 2013; Hughes et al., 2002; Kapur et al., 1994). Recent estimates,

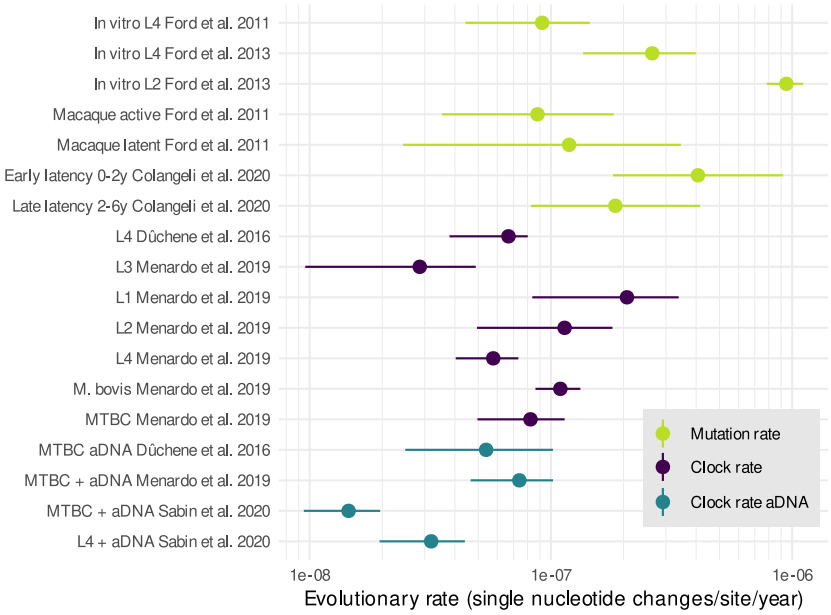

Figure 3: *Evolutionary rates in the MTBC. Only studies that report confidence intervals were considered. For the fluctuation assay estimates in Ford et al. (2011, 2013), a generation time $g = 20$ was assumed to translate rates to calendar time. The rates of Colangeli et al. (2020) were translated back to calendar time by assuming $g = 18h$, as reported by the authors. From the molecular clock study of Menardo et al. (2019), BEAST estimates are reported for a $1/x$ clock rate prior and constant population size. For the BEAST analysis of Sabin et al. (2020), results for the birth-death skyline model with an uncorrelated lognormal clock are reported.*

making use of tip dating, ancient DNA (aDNA) samples, and Bayesian phylogenetics, propose a more recent common ancestor in the Neolithic ca. 6,000 years ago (Bos et al., 2014; Kay et al., 2015; Sabin et al., 2020).

One caveat regarding these estimates is the poorly understood variability of evolutionary rates in the MTBC through time. For mitochondrial DNA, viruses, and bacteria, evolutionary rates usually appear faster when estimated from recent polymorphisms (Ho et al., 2011). For bacteria, Duchêne et al. (2016) found a clear negative association, described by an exponential decay curve, between clock rates and sampling time spans in 16 bacterial species, with an order of magnitude difference between a 10 year and a 100 year sampling period. The delayed effect of purifying selection is the most prominent explanation for this time dependence of evolutionary rates, although methodological biases might also contribute (Emerson and Hickerson, 2015; Ho et al., 2015). Time dependence can have a large effect on molecular dating: Membrebe et al. (2019) showed that accounting for purifying selection by using relaxed clock or epoch models can shift divergence times one order of magnitude back in time. Could this explain the surprisingly recent time to the most recent common ancestor estimated by the aDNA studies?

In the study of Duchêne et al. (2016), the MTBC does not follow the general pattern of time depence: almost identical rates were obtained from samples spanning 15 and 895 years. Similarly, Menardo et al. (2019) found only marginally lower rates when calibrating the clock with the same three samples of ancient DNA from Precolumbian human remains and an extensive MTBC dataset

covering a sampling period of 30 years. An overview of evolutionary rates estimated for the MTBC illustrates the large variability and uncertainty of rate estimates, but also suggest an overall trend of time dependence (Figure 3). As Menardo et al. (2019) showed in their extensive study of the molecular clock in the MTBC, clock rates vary substantially among lineages and clades of the MTBC and have large confidence intervals. Lineage 1, for instance, seems to have evolved faster than other lineages, and indeed faster than the L4 strain accumulated mutations in the fluctuation assay of Ford et al. (2011). On the slow end of the spectrum is the long-term clock rate estimated by Sabin et al. (2020), for which all six aDNA samples available so far were included (1.4×10^{-8}, 95% HPD 9.46×10^{-9}, 1.96×10^{-8}).

A possible methodological bias underlying *in vivo* mutation rate estimates was recently suggested in a simulation study of within-host evolution. Morales-Arce et al., 2020 suggested that the genome-wide mutation rate of the MTBC might be two orders of magnitude faster, in the order 10^{-8}/bp/generation, if one accounts for progeny skew (Box 2) and the removal of mutations through purifying selection during within-host evolution. The authors simulated a population undergoing a transmission bottleneck, followed by a recovery to a large population size and within-host evolution under purifying selection and with per-generation progeny skew. Comparing the resulting patterns of diversity with the empirical within-host data of Trauner et al., 2017, they found that mutation rates in the order of 1×10^{-8} to 9×10^{-8} result in similar levels of variation as described by Trauner et al.

Box 1: Simulating bacterial populations

Simulations are an invaluable tool in evolutionary genetics: they allow to test intuitions and methods, to compare alternative scenarios, and to fit models to data (Hoban et al., 2012; Johri et al., 2022). For bacterial population genetics, the use of simulations was so far rather limited. On the one hand, most simulators are based on the coalescent – the backwards-in-time variant of the Wright-Fisher model. These are fast, but usually limited to neutral scenarios of population size changes and migration. More flexible forward simulators, on the other hand, are much slower because they track the fate of all individuals of the simulated population rather than just of a sample, as in the coalescent (Hoban et al., 2012).

Recent advances in forward simulation, however, make it possible to simulate ever more realistic scenarios through improved computational efficiency (Haller et al., 2019) and more flexible non-Wright-Fisher models (Haller and Messer, 2019). The simulation framework of SLiM was recently used to simulate bacteria evolving in a Petri dish in the presence of an antibiotic (Cury et al., 2021). This individual-based forward simulation was spatially explicit and modelled clonal reproduction through binary fission, gene conversion, density-dependent selection, and positive selection for antibiotic resistance. The scriptability of SLiM allows to incorporate more or less aribtrarily complex genetic architectures and life histories, although computational time still sets boundaries.

Genetic drift and purifying selection

Once a mutation appears in a genome, its fate depends on the selective advantage or disadvantage it confers – and on chance. Genetic drift is the "chance factor" in evolution: it describes the undirected, stochastic change of allele frequencies due to sampling effects (Plutynski, 2007). The

biological relevance of genetic drift is that it sets limits to natural selection (Kimura, 1983; Lynch, 2007). The efficacy of natural selection is inversely related to the strength of drift: when genetic drift is strong, changes in the frequencies of alleles depend less on their effect on fitness, such that, by chance, deleterious alleles can increase and beneficial ones decrease in frequency (Kimura, 1983; Ohta, 1992).

Genetic drift is frequently invoked as an ad hoc explanation, but actually inferring and quantifying it is difficult. In the standard Wright-Fisher (WF) model with panmixia, discrete generations, and no selection, drift occurs when the alleles to form the next generation are randomly sampled from the parental population (Fisher, 1930; Wright, 1931). In this idealized lottery-like scenario, the strength of drift simply depends on the size of the sample, with less drift in larger samples according to the law of large numbers. Natural populations deviate from the WF model in numerous ways, yet population size remains a useful measure for drift when it is rescaled to account for these deviations (Charlesworth, 2009). The resulting effective population size N_e can be interpreted as the size of an idealized WF population that experiences the same amount of drift as the real population in question (e.g. Gillespie, 2004). In bacteria, population subdivision, linked selection, and demographic changes all imply that sampling effects are stronger than under panmixia (Price and Arkin, 2015), and that effective population sizes are orders of magnitude smaller than census sizes (Bobay and Ochman, 2018).

As discussed in this section, arguments about the strength of drift in the MTBC are largely based on indirect evidence in the form of low diversity and overabundant nonsynonymous polymorphisms. Estimates of N_e are sometimes obtained in Bayesian skyline analyses, but their underlying assumptions are problematic. Finally, we discuss transmission bottlenecks in the MTBC, a main mechanism of stochastic sampling whose mid- and long-term consequences go beyond simple reductions in genetic diversity and remain to be understood.

Do overabundant nonsynonymous polymorphisms indicate strong genetic drift?

In the MTBC, the drift-versus-selection discussion has mainly revolved around the large proportion of nonsynonymous polymorphisms observed in the species. The MTBC has a genome-wide ratio of nonsynonymous to synonymous polymorphisms (d_N/d_S) of around 0.5 when diverse strains from across the phylogeny are considered (Figure 2c). This is one third higher than in the closely related *M. canettii* (Supply et al., 2013) and more than six times higher than the median (0.076) of the 153 diverse species studied by (Bobay and Ochman, 2018).

Hershberg et al. (2008) have interpreted the high d_N/d_S in the MTBC as evidence for "extremely reduced purifying selection" – in other words strong genetic drift – which would allow the accumulation of deleterious nonsynonymous mutations. The authors refute the alternative explanation that nonsynonymous changes are due to positive selection by pointing out that d_N/d_S does not differ between housekeeping, surface-exposed, and virulence genes, as might be expected if host immunity would drive adaptive diversification. This interpretation of d_N/d_S fits well with the generalization that the intracellular niche of pathogens and symbionts implies smaller population sizes and stronger drift. Kuo et al. (2009) inferred strong drift in human pathogens including the MTBC and reported a strong inverse relationship between drift and genome size. A similar conclusion is reached by Balbi et al. (2009), who compared *E. coli* with the closely related pathogenic *Shigella* and found signs of increased drift in the latter, including an excess of nonsynonymous mutations and of transversions, which are proportionally more nonsynonymous and thus deleterious than transitions.

Different studies have challenged the view that purifying selection is "extremely reduced" in the MTBC. Bringing in a temporal perspective on d_N/d_S, Namouchi et al. (2012) found 25% more nonsynonymous SNPs on terminal branches in their tree of 22 globally diverse strains. This suggests that deleterious nonsynonymous mutations are purged through selection over time, such that they become scarce in deeper parts of the phylogeny (Rocha et al., 2006). In general, SNPs are strongly skewed towards rare alleles in the MTBC, be it at the global or the within-host level (O'Neill et al., 2015; Trauner et al., 2017). SNPs are thus not only few in the MTBC, but also to a large proportion singletons (Chiner-Oms et al., 2019b), that is, present in one single strain. While this is consistent with purifying selection preventing variants to rise in frequency, other processes can cause the same pattern, in particular the dynamics of clonal growth. Furthermore, it remains to be understood what biases are introduced by the punctual sampling of highly structured and dynamic within-host populations (Morales-Arce et al., 2021).

In the so far only attempt to quantify the strength of purifying selection across the genome, Pepperell et al. (2013) fitted a model including demographic expansion and a fraction of sites under selection to the site frequency spectrum obtained from a global sample of the MTBC. They infer purifying selection at nonsynonymous sites across 95% of the genome, with a selection coefficient s of -9.5×10^{-4}. This value is interpreted as "strong" compared to values in humans and *Drosophila*. The authors used simulations of completely linked genomes to evaluate their models, which assume linkage equilibrium between sites. They find that their best model performs poorly in some scenarios; specifically, strong selection can be misinferred when complete linkage is combined with weak purifying selection, which might thus confound their estimate of s. Other model assumptions were not tested, for example the absence of population subdivision or that the population follows a simple demographic model of exponential growth.

Strong genetic drift leaves other signs than an excess of nonsynonymous mutations, including pseudogenization, proliferation of selfish genetic elements, or an increased proportion of transversions. With strong drift and asexual reproduction, such signatures can accumulate through Muller's ratchet, where lack of recombination and reduced efficacy of purifying selection lead to a build-up of deleterious mutations (Muller, 1964). As pointed out by Namouchi et al. (2012), these signatures are hardly evident in the MTBC. There are 30 pseudogenes in the H37Rv reference genome (Cole et al., 1998), in line with the generally low number of pseudogenes in bacterial genomes (Lawrence et al., 2001) and contrasting with the more than 1,000 pseudogenes described in the genome of *M. leprae* (Gómez-Valero et al., 2007). Also insertion sequences do not thrive in the MTBC: almost all IS activity is due to a single active element, IS6110, which is over-represented in intergenic regions, occurs at low frequencies, and thus probably evolves under strong purifying selection (McEvoy et al., 2007). Finally, transitions occur well in excess of transversions (Payne et al., 2019). Taken together, there is scant evidence for genome erosion driven by Muller's ratchet in the MTBC.

Drift is expected to dominate allele frequency changes when $|N_e \times s| << 1$ (Kimura, 1983; Ohta, 1992). Thus, rather than small population sizes (N_e), reduced selection coefficients (s), as they might arise when many genes are not required anymore after the transition to an intracellular niche, could explain genome erosion in obligate pathogens. Applied to the MTBC, the absence of genome erosion could indicate that these bacteria still require a large complement of genes, which thus remain under strong purifying selection. Alternatively, the MTBC might be a young pathogen in an early phase of genome degradation, where nonsynonymous mutations are only starting to accumulate (Kuo et al., 2009).

Are synonymous sites under selection?

How could the high genome-wide d_N/d_S in the MTBC be explained if not by strong drift? An intriguing alternative scenario is purifying selection at synonymous sites (Namouchi et al., 2012). High d_N/d_S can reflect an overabundance of nonsynonymous mutations (numerator), but also a lower number of synonymous mutations (denominator) than in other species. Fitness effects of synonymous mutations can arise when different codons result in variation in RNA stability, protein folding, and translation efficiency and accuracy (reviewed by Hershberg and Petrov, 2008). Already weak selection on synonymous sites can inflate d_N/d_S, as shown in a recent study of codon usage in 13 bacterial genomes (Rahman et al., 2021).

In the MTBC, codon frequencies are associated with gene expression (Andersson and Sharp, 1996; Pan et al., 1998), but also with the hydrophobicity of proteins and sequence conservation (De Miranda et al., 2000). As suggested in the latter study, a combination of selective pressures may thus act on synonymous sites in the MTBC, including the more efficient and accurate translation of certain codons and constraints on protein folding. Wang and Chen (2013) assessed possible selection on synonymous sites by comparing synonymous (d_s) to intergenic (d_I) diversity across 13 diverse MTBC genomes. Diversity varies strongly depending on the genomic position, suggesting variation in mutation rates or selective pressures across the genome. In the majority of windows, however, d_S is higher than d_I. Under the assumption that intergenic regions are free from selection pressures, these results are interpreted as evidence for positive selection on synonymous sites, specifically for increased translational efficiency.

The alternative explanation, mentioned but not favored by Wang & Chen, is that purifying selection is stronger in intergenic regions than at synonymous sites. Intergenic regions in bacteria are packed with regulatory motives and can hardly be assumed to evolve neutrally (Molina and Van Nimwegen, 2008; Rocha, 2018). Rather than comparing synonymous against assumed neutral sites, Thorpe et al. (2017) assessed the relative strength of purifying selection by comparing the proportion of singleton mutations among different site categories, reflecting that a higher proportion of singletons indicates stronger purifying selection. In five out of six species, site categories show a clear ranking, with the proportion of singletons increasing from synonymous, intergenic, non-synonymous, to non-sense mutations. In the MTBC, however, no differences between categories are apparent: there are similar proportions of singletons in all four categories. This surprising observation can at least partly be explained by the dataset used by the authors, which includes many near-identical MTBC strains sampled in a single country. Still, that even at short timescales non-sense mutations in the MTBC do not appear to be under stronger selection than synonymous mutations asks for clarification in future studies.

In summary, synonymous sites are frequently assumed to be neutral, but studies on codon frequencies and comparisons of synonymous with other sites in the genome suggest a more complex picture. This is a topic deserving a focused study, applying the measures developed in previous work to a large dataset covering different timescales.

Bayesian skyline plots and the issue of storytelling

Neutral sites are in short supply in prokaryotes (Rocha, 2018). In contrast to eukaryotes, the streamlined genomes of archea and bacteria do not contain large swaths of decaying repeats and other DNA debris which can be assumed to be non-functional. This poses a particular challenge for the estimation effective population sizes and the quantification of genetic drift, which traditionally

relies on the availability of sites not affected by natural selection (Charlesworth, 2009).

A popular approach to estimate effective population sizes and their change through time are Bayesian skylines (Ho and Shapiro, 2011). These models are frequently used in Bayesian phylogenetics, where N_e is treated as a nuisance parameter. Many studies, however, interpret N_e literally as historical change in population size and thereby provide instructive examples of how strong assumptions are ignored for the sake of storytelling.

Bayesian skyline models assume neutrality in order to translate coalescence times into population sizes. Several studies have shown that non-neutral processes confound demographic inference and should not simply be assumed away. Recombination (Hedge and Wilson, 2014), population structure (Heller et al., 2013), sampling design, gene conversion, and selection (Lapierre et al., 2016), as well as the skewness of reproductive success (Menardo et al., 2021a) all create spurious signs of population size changes. As observed by Lapierre et al., 2016, such methodological biases might explain why population size trajectories look suspiciously similar for a wide range of species.

Despite these caveats, Bayesian skyline plots continue to be used and interpreted liberally in the MTBC literature. Skyline plots were presented as evidence for a Neolithic expansion (Comas et al., 2013), expansions of specific lineages (Merker et al., 2022; Mulholland et al., 2019; O'Neill et al., 2019), or a recent co-expansion with humans in Tibet (Liu et al., 2021). That population size trajectories "make sense" in the historical narratives of these articles does not add to their credibility, but rather puts into question the way results are made sense of (Katz, 2013). Instead of literal interpretations of Bayesian skylines, an improved understanding is required of how far the demographic past can be reconstructed from the genomes of extremely clonal bacteria without taking into account different confounding factors.

Box 2: Progeny skew in prokaryotes?

Recently, progeny skew was brought up as a neglected aspect of MTBC evolution with potentially significant effects on genetic diversity (Morales-Arce et al., 2020) and population genetic inference (Menardo et al., 2021a). Progeny skew refers to the unequal distribution of offspring among parental individuals in a population. Frequently mentioned examples are viruses, where a single parental sequence can give rise to numerous copies, or marine organisms reproducing through broadcast spawning. Wright-Fisher and coalescence models assume the variation in offspring number is small (Tellier and Lemaire, 2014), which leads to mis-inference of population genetic statistics when applied to such organisms (Sackman et al., 2019).

While progeny skew in viruses or has a direct interpretation in the way these organisms reproduce, it is less straightforward to apply to prokaryotes. Archea and bacteria reproduce through binary fission, which can be thought of as each parent having two offspring and dying after division (Cury et al., 2021); or, in an age-structured population, as each parent having one offspring and surviving. Progeny skew can arise over multiple generations through rapid adaptation, superspreading events, or repeated bottlenecks, and it is thus a meaningful parameter in population-based models with a continuous timescale (Menardo et al., 2021a). In individual-based, discrete-generation models, it is preferable to simulate the processes giving rise to progeny skew explicitly.

How do bottlenecks affect genetic diversity?

In the MTBC, genetic drift is often associated with transmission bottlenecks or founder events, when few or even single strains initiate an infection or an outbreak (Pepperell et al., 2010; Smith et al., 2006). TB infections can be initiated by single to few cells (Ryndak and Laal, 2019); each transmission might thus be a massive founder event where, from the millions of cells forming a within-host population, only a few cells are sampled to start a new population. Similar, small-scale colonization dynamics occur during within-host dissemination, as single to few cells "found" new granulomas in the highly structured habitat of the lung (Martin et al., 2017).

While genetic bottlenecks entail an immediate loss of genetic diversity, the mid- and long-term effects of periodic bottlenecks on genetic diversity and differentiation in clonal pathogens, where extreme bottlenecks alternate with clonal expansions, are less clear. Periodic bottlenecks have been investigated in the context of experimental evolution, where studies mainly focused on the effects of bottlenecks on the rate of adaptation (e.g. Windels et al., 2021). More general considerations can be found in the population genetics literature. One insight of potential relevance for the evolutionary dynamics of the MTBC is that, under predominant purifying selection, rates of evolution are accelerated when N_e is small because more deleterious mutations fix due to genetic drift (Lanfear et al., 2014). A classic example of this is the increased rate of sequence evolution in aphid endosymbionts versus free-living bacteria of the genus *Buchnera* (Moran, 1996). In the absence of homogenizing gene flow, founder events might thus be expected to increase genetic differentiation and overall diversity among lineages of the MTBC. Following this logic, the low global diversity of the MTBC (Figure 2a) is not evidence for strong bottlenecks. The puzzling observation rather is that there is not more diversity given the repeated bottlenecks during within- and between-host evolution and the absence of gene flow. Low diversity despite frequent bottlenecking could thus indicate purifying selection.

The purpose of these considerations is to show that genetic bottlenecks are more complex and interesting than they appear in the literature, where they often serve as *ad hoc* explanation for low diversity. More work on periodic bottlenecks in bacterial pathogens is needed. This work could take into account some real-world complications such as the unclear number of cells actually transmitted, which is most likely larger than the minimum number required to start an infection (Namouchi et al., 2012). Furthermore, infection might not occur at a single time point, but extend through time as hosts are repeatedly exposed to bacteria-laden aerosol droplets (Ryndak and Laal, 2019). This situation resembles the source-sink dynamics of metapopulation models with repeated colonization events rather than a single bottleneck.

Positive selection

As unclear as the role of genetic drift and purifying selection in the evolution of the MTBC is the role of positive selection. Most insights about how the MTBC has adapted to environmental challenges either regard pathoadaptation in the distant past before the MRCA, as revealed through comparative genomics (reviewed by Pepperell, 2022), or the recent evolution of antibiotic resistance (reviewed by Gygli et al., 2017). Much less is known about the genetics underlying adaptation to different mammalian host species, evident in host tropism (Brites et al., 2018; Zwyer et al., 2021), or about local adaptation to different human populations, as suggested by sympatric patient-pathogen associations observed in cosmopolitan settings (Gagneux et al., 2006).

Identifying signatures of positive selection in linked genomes is challenging since most tests

rely on the comparison of haplotypes within genomes (Shapiro et al., 2009). Two diversity-based signatures that are not haplotype-based have been used extensively to identify positive selection in MTBC genomes: homoplasy and excess of nonsynonymous polymorphisms. In the following, we discuss the properties of these measures and whether they can be used to elucidate the role of positive selection beyond the case of antibiotic resistance, which so far provides the confirmed cases of adaptive evolution in the MTBC.

Homoplasies: how common is convergent adaptation?

Molecular homoplasy designates the independent appearance of identical mutations in different parts of a phylogeny through chance, recombination, or convergent selection (Stern, 2013). Chance homoplasy between genomes showing so little overall diversity is rare (Comas et al., 2009, Figure 2d), and its probability can be assessed through permutation tests (Farhat et al., 2013). Mutation hotspots can facilitate chance homoplasy (Galtier et al., 2006): in the MTBC, highly mutable tandem repeats frequently cause homoplasy (Outhred et al., 2020), while it is not known how rates of point mutations vary along the genome. Recombination has been argued against as a cause of homoplasies because homoplasies in the MTBC do not occur in clusters, as would be expected when recombination involves diverged DNA (Chiner-Oms et al., 2019b). Non-clustering homoplasies, however, are also expected when recombinant genomes are similar (Bobay et al., 2015). Furthermore, intrachromosomal recombination can generate homoplasies, as suggested by their increased occurrence in homologous *PE/PPE* genes (Tantivitayakul et al., 2020).

Clear examples of convergent selection as a cause of homoplasy have been presented for genes involved in antimicrobial resistance (Comas et al., 2012; Farhat et al., 2013). Against a background of low diversity and rare homoplasy, some of these genes show exceptional patterns. In 1,161 strains sampled in Russia and South Africa, one specific mutation in the *katG* gene, which confers isoniazid resistance, has originated more than 70 times independently (Mortimer et al., 2017). This is an extreme pattern that arises because *katG* is a "tight target" of selection, that is, only single to few mutations can cause resistance without incurring high fitness costs. In other genes ("sloppy targets"), fewer homoplasies are observed but in more positions. The high incidence of parallelism in resistance evolution, in combination with large datasets, allows the use of genome wide association approaches to identify new drug resistance loci and to elucidate the genetic architecture of resistance phenotypes (e.g. Crook et al., 2022).

The basic limitation of homoplasies as a signature of selection is that they only reveal cases of convergent evolution. In the case of antibiotic resistance, convergence is ubiquitous. Thousands of parallel evolutionary experiments are conducted when people around the world are treated with the same antibiotics proposed by the WHO, imposing strong selective pressures with high rewards for resistance mutations in target genes (Walker et al., 2022). For other selective pressures, things are less clear. Recently, two cases of convergent selection were shown in studies of experimental evolution with *M. canettii* and the MTBC. Selecting *M. canettii* strains for *in vivo* persistence in mice, Allen et al. (2021) identified two parallel mutations and demonstrated their effect on persistence through gene knock-out and complementation. Smith et al. (2022) selected for biofilm formation in experimentally evolved MTBC strains and identified two loci that mutated independently and are associated to biofilm-associated traits and fitness proxies. Both studies found that parallel mutations emerged in similar strains, suggesting that the genetic background constrains evolutionary trajectories. These studies also illustrate the rapidity with which mutations otherwise rare or absent can prevail in the presence of new selective pressures; and the significance

of structural variation, as convergent evolution involved a large duplication (Smith et al., 2022) and a deletion of two genes (Allen et al., 2021).

Convergence might not only be favored by strong selective pressures, but also through human demography and migration. Repeated introductions of sublineages into a region, as described for Tibet (Liu et al., 2021), are natural experiments where genetically highly similar strains are confronted with a new environment. Liu et al. identified several genes that accumulate mutations independently after repeated introductions to the Tibetan Plateau, including *sseA*, a gene involved in the detoxification of reactive oxygen species, and three genes involved in DNA repair (*dnaE2*, *recB*, *mfd*). With the already large and still growing amount of data on MTBC outbreaks, such natural experiments of parallel evolution can provide valuable insights into the dynamics and genes involved in local adaptation.

Nonsynonymous polymorphisms: how frequent is positive selection?

The second widely used statistic to infer selection and its direction is the ratio of non-synonymous to synonymous polymorphisms d_N/d_S. The intuition behind this measure is that an increased rate of nonsynonymous compared to synonymous changes indicates positive selection. As for homoplasies, genes involved in antibiotic resistance provide the clearest examples (Osório et al., 2013; Wilson et al., 2020), and indeed the two signatures often co-occur.

Compared to homoplasy, which is a fairly intuitive heuristic for convergent selection, d_N/d_S is a more complicated statistic that can be estimated in different ways and whose properties and limitations have been explored in numerous studies (overview in Yang, Ziheng, 2014). Frequently, d_N/d_S is estimated by comparing pairs of sequences (e.g. with the method of Nei and Gojobori, 1986). This is e.g. the case for d_N/d_S in Figure 2c or in the study of Hershberg et al. (2008) discussed above, who presented genome-wide average pairwise d_N/d_S as evidence for reduced purifying selection. Although average pairwise d_N/d_S is sometimes used gene-wise in selection sans, it is a coarse measure. The ratio averages over the sites of a locus and the branches in a phylogeny. It thus has low sensitivity, as only in loci with strong signals and multiple sites under selection will the signal not be canceled by sites under purifying selection (Yang and Bielawski, 2000). A signal for positive selection may also be canceled if it is only present on a specific branch (Yang and Nielsent, 2002).

A family of more versatile maximum likelihood models have been developed that incorporate explicit models of codon evolution and allow to test for increased rates of nonsynonymous changes on particular branches or in particular codons of a gene (implemented in PAML; Yang, 2007). These methods are computationally intensive and not suitable for exploratory analyses on large phylogenies, while small MTBC datasets might not contain enough diversity to estimate parameters. They can be used, however, to obtain a more detailed picture of selective pressures in genes of interest and to formally test for selection using model comparisons (Yang, 1998). A recent example of an exploratory selection scan followed by more rigorous statistical testing is the study of Menardo et al. (2021b). In a first step, they identified a hypervariable epitope at the *esxH* locus, which codes for a secreted effector interacting with the human immune system. Codon models were then used to test for site- and branch-specific selection. Significant signatures were found in MTBC lineage 1 but not in other lineages and located to the N-terminal epitope of the gene. Further dissection of these signatures showed that they occur in strains collected in South and Southeast Asia, suggesting that this locus might be involved in adaptation to regional human host populations.

Two recent studies have proposed methods to estimate d_N/d_S for large datasets while avoiding site and branch averaging, respectively. Wilson et al. (2020) present a phylogeny-free (and thus fast) method to infer selection at the codon level. Applying their method to more than 10,000 MTBC genomes, they found a d_N/d_S significantly larger than 1 in 2,729 out of 3,979 genes. Chiner-Oms et al. (2022) investigated the temporal trajectories of p_N/p_S in a large phylogeny of 5,000 strains (p_N/p_S is based on simple counts while d_N/d_S includes correction through a substitution model, Yang, Ziheng, 2014, p. 47ff). Focusing on shifts in p_N/p_S along the tree, they found evidence for elevated nonsynonymous changes at some point in time in almost half the genes of the MTBC. While both studies generate long lists of candidate genes, they also lead to the inevitable follow-up question of exploratory selection scans: what to do with these candidates. Considering the difficulty of experimental validation in a human pathogen, further characterization of the candidates with the phylogenetically explicit methods of PAML could be useful.

Overall, homoplasies and d_N/d_S tell us little about the big unknown of clonal evolution: the distribution of fitness effects (see introduction). Methods exist to infer the distribution of fitness effects from d_N/d_S using population genetic (reviewed by Eyre-Walker and Keightley, 2007) or phylogenetic (e.g. Tamuri et al., 2012) models. Recently, the relationship between selection coefficients and d_N/d_S under clonal reproduction were explored in the context of somatic evolution (Williams et al., 2020). The model developed in the study relaxes some of the strong assumptions of previous approaches, in particular constant population sizes and evolution over long timescales, by integrating d_N/d_S and the clone size distribution. It would be worthwile to explore whether this approach can be applied to bacterial within-host populations in order to learn more about the distribution of fitness effects *in vivo*.

Discussion

What evolutionary processes drive and have driven the evolution of the MTBC? The most robust insight, forming the premise of this review, is that horizontal gene transfer is negligible, although recombination more generally is not, considering the mutagenic role of intrachromosomal recombination. Regarding mutation, genetic drift, and natural selection, much remains unclear.

The current understanding of evolutionary processes in the MTBC is based on a complex mesh of indirect evidence, intuitions, deductions from general principles, and assumption-rich models. Studies focusing on population genetic processes are few and far apart, their methods and datasets heterogeneous. Some of the hypotheses developed in these articles, for example that the evolution of the MTBC is driven by genetic drift (Hershberg et al., 2008) or purifying selection (Pepperell et al., 2013), have solidified into strong beliefs through repetition, even though the original studies have pointed out caveats and the subtler meanings of "is driven by" remain unexplored. With the large amount of sequencing data now available, covering evolutionary timescales from within-host evolution to global patterns of diversity, it would be a good moment to revisit some past hypotheses. We envisage focused studies that – in contrast to the typically broad scope of studies of the "early" genomics era – address specific hypotheses and pay more attention to methodological limitations.

While studying methods is less interesting that studying organisms, the bottleneck in data analysis increasingly lies in the comprehension of complex methods rather than the availability of data (Johri et al., 2022). New tools for evolutionary simulations, such as the versatile forward simulation tool SLiM (Box 1), could provide a long-needed crutch to move forward by allowing to simulate ever more realistic biologies and life histories.

To illustrate the utility of simulations, we used SLiM to simulate the within-host dissemination dynamics of a clonal pathogen (script available on https://git.scicore.unibas.ch/TBRU/slim_simulations). The model is inspired by the study of Martin et al., 2017, who used DNA barcoding and infection mapping to infer the temporal and spatial dynamics of an MTBC infection in macaques. Populations in the simulation might be thought of as granulomas that grow and give rise to new granulomas – a metapopulation model with unidirectional migration from "full" to "emtpy" populations. Infection begins with a single bacterium giving rise to an exponentially growing population through clonal reproduction and 19 "empty" populations. Once this population reaches carrying capacity $K = 20,000$, it can seed new populations (Figure B1a), which again grow and can seed new populations when K is reached. More specifically, each generation a number of n migrants is drawn from a Poisson distribution with mean 1; n individuals are then drawn from a random population that has reached K and migrated to a random empty population until all populations are occupied. Exemplary growth dynamics of the simulation are depicted in Figure B1b. Mutations are simulated at a rate $\mu = 5 \times 10^{-10}$/bp/gen in a genome of 4 Mb. Selection is either assumed to be absent ($s = 0$) or purifying ($s = -9.5e-4$), as proposed by Pepperell et al., 2013. The simulation ends after 70 generations, which with a generation time of 24 h corresponds to a 10 week infection. For both selection coefficients, the simulation was replicated 100 times.

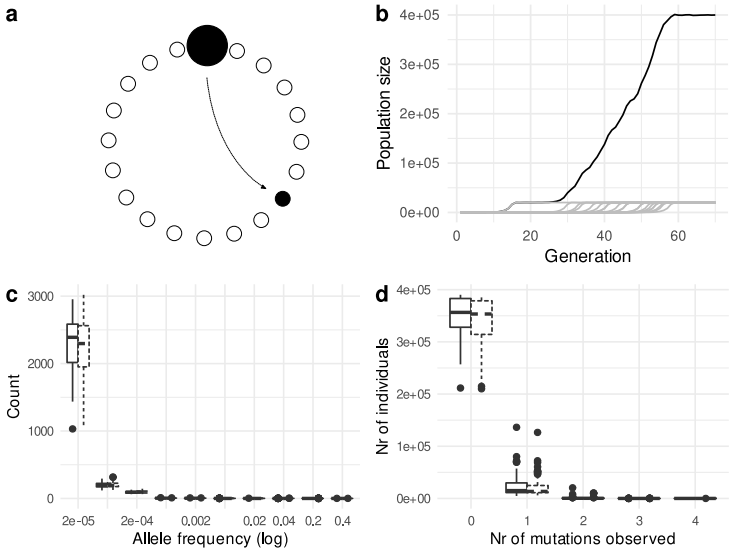

Figure 4: *A metapopulation model for within-host evolution. a) Infection begins with a single bacterium giving rise to an exponentially growing population through clonal reproduction. Once this population reaches carrying capacity $K = 20,000$, it can seed new populations which again grow exponentially. b) Exemplary growth dynamics of the model, the solid line showing total population size, dashed lines showing subpopulation sizes. c) Site frequency spectrum at generation 70. d) Number of individuals with 0 to 4 SNPs at generation 70.*

Independently of purifying selection, the dynamics of clonal growth and dissemination over 70 bacterial generations give rise to an extreme skew towards rare alleles (Figure B1c). A large proportion of the mutations are in fact singletons, that is, only present in a single individual. At generation 70, the vast majority of individuals have no mutation, except in few instances where

a mutation arose early (Figure B1d). Some simulations produced outlier values because not all populations were "filled" after 70 generations.

The purpose of this simulation is to illustrate the simulation approach. Some potential applications of evolutionary simulations for the MTBC are listed in the following.

Simulations of structured within-host populations could be used to investigate the implications of sampling and culturing for our understanding of within-host diversity, and to develop new experimental designs and sampling strategies. How representative, for example, can sputum samples possibly be of within-host diversity? Coupling within- and between-host evolution, periodic bottlenecking could be simulated to study how diversity accumulates through time as a function of bottleneck size, purifying selection, or mutation rates. This would lead to a more nuanced understanding of transmission bottlenecks, which have more complex consequences than simple reduction of diversity.

Gene conversion between closely related strains could be simulated to test different methods to infer recombination. More generally, methods should be tested on simulated data to understand their behavior and make an informed choice, instead of resorting to the typical bioinformatics approach of using multiple methods and reporting intersecting results, which leaves the door open to confirmation bias. Finally, the ultimate challenge would be to try to simultaneously infer demography and selection using approximate Bayesian computation (Johri et al., 2022). It is difficult, however, to conceive what kind of data would be suitable for this. At the microevolutionary scale that is most straightforward to simulate, there is just so little diversity that it is dubious that parameter-rich models could be fitted with any confidence.

Simulations are not a panacea, but they allow to raise the debate to a more transparent, quantitative level than achieved by the so far largely verbal arguments. If nothing else, they could allow to better understand what kind of inference is at all possible, given the lack of HGT and the low levels of genetic diversity in monomorphic bacteria.

Acknowledgments

Our best thanks for their comments on earlier versions of the manuscripts to Daniela Brites, Etthel Windels, Michaela Zwyer, Selim Bouaouina, Fabrizio Menardo, Ana Morales-Arce, and the members of the Gagneux group. This work was funded through grants from the European Research Council, grant number 883582, and the Swiss National Science Foundation, grant numbers 310030_188888 and CRSII5_177163.

References

Achtman M (2008). Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annual Review of Microbiology* 62, 53–70. DOI: 10.1146/annurev.micro.62.081307.162832.

— (2012). Insights from genomic comparisons of genetically monomorphic bacterial pathogens. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367, 860–867. DOI: 10.1098/rstb.2011.0303.

Allen AC et al. (2021). Parallel *in vivo* experimental evolution reveals that increased stress resistance was key for the emergence of persistent tuberculosis bacilli. *Nature Microbiology* 6, 1082–1093. DOI: 10.1038/s41564-021-00938-4.

Andersson SG and PM Sharp (1996). Codon usage in the *Mycobacterium tuberculosis* complex. *Microbiology* 142, 915–925. DOI: 10.1099/00221287-142-4-915.

Balbi KJ, EP Rocha, and EJ Feil (2009). The temporal dynamics of slightly deleterious mutations in *Escherichia coli* and *Shigella* spp. *Molecular Biology and Evolution* 26, 345–355. DOI: 10.1093/molbev/msn252.

Bobay LM and H Ochman (2017). Impact of recombination on the base composition of bacteria and archaea. *Molecular Biology and Evolution* 34, 2627–2636. DOI: 10.1093/molbev/msx189.

Bobay LM, CC Traverse, and H Ochman (2015). Impermanence of bacterial clones. *PNAS* 112, 8893–8900. DOI: 10.1073/pnas.1501724112.

Bobay LM and H Ochman (2018). Factors driving effective population size and pan-genome evolution in bacteria. *BMC Evolutionary Biology* 18, 1–12. DOI: 10.1186/s12862-018-1272-4.

Boritsch EC et al. (2016). Key experimental evidence of chromosomal DNA transfer among selected tuberculosis-causing mycobacteria. *PNAS* 113, 9876–9881. DOI: 10.1073/pnas.1604921113.

Bos KI et al. (2014). Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature* 514, 494–497. DOI: 10.1038/nature13591.

Brites D et al. (2018). A new phylogenetic framework for the animal-adapted *Mycobacterium tuberculosis* complex. *Frontiers in Microbiology* 9 (NOV), 1–14. DOI: 10.3389/fmicb.2018.02820.

Casadevall A (2008). Evolution of intracellular pathogens. *Annual Review of Microbiology* 62, 19–33. DOI: 10.1146/annurev.micro.61.080706.093305.

Charlesworth B (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics* 10, 195–205. DOI: 10.1038/nrg2526.

— (2012). The effects of deleterious mutations on evolution at linked sites. *Genetics* 190, 5–22. DOI: 10.1534/genetics.111.134288.

Chiner-Oms Á, MG López, M Moreno-Molina, V Furió, and I Comas (2022). Gene evolutionary trajectories in *Mycobacterium tuberculosis* reveal temporal signs of selection. *PNAS* 119, e2113600119–e2113600119. DOI: 10.1073/pnas.2113600119.

Chiner-Oms Á et al. (2019a). Genome-wide mutational biases fuel transcriptional diversity in the *Mycobacterium tuberculosis* complex. *Nature Communications* 10, 1–11. DOI: 10.1038/s41467-019-11948-6.

Chiner-Oms Á et al. (2019b). Genomic determinants of speciation and spread of the *Mycobacterium tuberculosis* complex. *Science Advances* (June). DOI: 10.1101/314559.

Clark RR, P Lapierre, E Lasek-Nesselquist, TA Gray, and KM Derbyshire (2022). A polymorphic gene within the *Mycobacterium smegmatis* esx1 locus determines mycobacterial self-identity and conjugal compatibility. *mBio* 13. DOI: 10.1128/mbio.00213-22.

Colangeli R et al. (2020). *Mycobacterium tuberculosis* progresses through two phases of latent infection in humans. *Nature Communications* 11, 1–10. DOI: 10.1038/s41467-020-18699-9.

Cole S et al. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393 (NOVEMBER), 537–544. DOI: 10.1038/29241.

Comas I, S Homolka, S Niemann, and S Gagneux (2009). Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. *PLoS ONE* 4. DOI: 10.1371/journal.pone.0007815.

Comas I et al. (2012). Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. *Nature Genetics* 44, 106–110. DOI: 10.1038/ng.1038.

Comas I et al. (2013). Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nature Genetics* 45, 1176–1182. DOI: 10.1038/ng.2744.

Copin R et al. (2016). Within-host evolution selects for a dominant genotype of *Mycobacterium tuberculosis* while T cells increase pathogen genetic diversity. *PLoS Pathogens* 12, 1–16. DOI: 10.1371/journal.ppat.1006111.

Crook DW et al. (2022). Genome-wide association studies of global *Mycobacterium tuberculosis* resistance to 13 antimicrobials in 10,228 genomes identify new resistance mechanisms, 1–27. DOI: 10.1371/journal.pbio.3001755.

Cury J, BC Haller, G Achaz, and F Jay (2021). Simulation of bacterial populations with SLiM. *Peer Community in Evolutionary Biology*, 1–36. DOI: 10.1101/2020.09.28.316869.

Davies J and D Davis (2010). Origins and evolution of antibiotic resistance. *Microbiología (Madrid, Spain)* 12, 9–16. DOI: 10.1128/mmbr.00016-10.

De Miranda AB, F Alvarez-Valin, K Jabbari, WM Degrave, and G Bernardi (2000). Gene expression, amino acid conservation, and hydrophobicity are the main factors shaping codon preferences in *Mycobacterium tuberculosis* and *Mycobacterium leprae*. *Journal of Molecular Evolution* 50, 45–55. DOI: 10.1007/s002399910006.

Denamur E, O Clermont, S Bonacorsi, and D Gordon (2021). The population genetics of pathogenic *Escherichia coli*. *Nature Reviews Microbiology* 19, 37–54. DOI: 10.1038/s41579-020-0416-x.

Dubos R and J Dubos (1952). *The White Plague*. Little, Brown and Company.

Duchêne S et al. (2016). Genome-scale rates of evolutionary change in bacteria. *Microbial genomics* 2, e000094–e000094. DOI: 10.1099/mgen.0.000094.

Duret L and N Galtier (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual Review of Genomics and Human Genetics* 10, 285–311. DOI: 10.1146/annurev-genom-082908-150001.

Dutta NK and PC Karakousis (Sept. 2014). Latent tuberculosis infection: myths, models, and molecular mechanisms. *Microbiology and Molecular Biology Reviews* 78, 343–371. DOI: 10.1128/MMBR.00010-14.

Emerson BC and MJ Hickerson (2015). Lack of support for the time-dependent molecular evolution hypothesis. *Molecular Ecology* 24, 702–709. DOI: 10.1111/mec.13070.

Eyre-Walker A and PD Keightley (2007). The distribution of fitness effects of new mutations. *Nature Reviews Genetics* 8. DOI: 10.1038/nrg2146.

Farhat MR et al. (2013). Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nature Genetics* 45, 1183–1189. DOI: 10.1038/ng.2747.

Fishbein S, N van Wyk, RM Warren, and SL Sampson (2015). Phylogeny to function: PE/PPE protein evolution and impact on *Mycobacterium tuberculosis* pathogenicity. *Molecular Microbiology* 96, 901–916. DOI: 10.1111/mmi.12981.

Fisher RA (1930). *The Genetical Theory of Natural Selection*. Pages: 308. 308 pp.

Fitzgerald DM and SM Rosenberg (2019). What is mutation? A chapter in the series: how microbes "jeopardize" the modern synthesis. *PLoS Genetics* 15, 1–14. DOI: 10.1371/journal.pgen.1007995.

Ford CB et al. (2011). Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. *Nature Genetics* 43, 482–488. DOI: 10.1038/ng.811.

Ford CB et al. (2013). *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nature Genetics* 45, 784–790. DOI: 10.1038/ng.2656.

Gagneux S (2018). Ecology and evolution of *Mycobacterium tuberculosis*. *Nature Reviews Microbiology* 16, 202–213. DOI: 10.1038/nrmicro.2018.8.

Gagneux S et al. (2006). Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *PNAS* 103, 2869–2873. DOI: 10.1073/pnas.0511240103.

Galtier N, D Enard, Y Radondy, E Bazin, and K Belkhir (2006). Mutation hot spots in mammalian mitochondrial DNA. *Genome Research* 16, 215–222. DOI: 10.1101/gr.4305906.

Gillespie JH (2004). *Population Genetics – A Concise Guide*. The Johns Hopkins University Press.

Godfroid M, T Dagan, and A Kupczok (2018). Recombination signal in *Mycobacterium tuberculosis* stems from reference-guided assemblies and alignment artefacts. *Genome Biology and Evolution* 10, 1920–1926. DOI: 10.1093/gbe/evy143.

Gómez-Valero L, EP Rocha, A Latorre, and FJ Silva (2007). Reconstructing the ancestor of *Mycobacterium leprae*: the dynamics of gene loss and genome reduction. *Genome Research* 17, 1178–1185. DOI: 10.1101/gr.6360207.

Gray TA and KM Derbyshire (2018). Blending genomes: distributive conjugal transfer in mycobacteria, a sexier form of HGT. *Molecular Microbiology* 108, 601–613. DOI: 10.1111/mmi.13971.

Guerrini V et al. (2016). Experimental evolution of *Mycobacterium tuberculosis* in human macrophages results in low-frequency mutations not associated with selective advantage. *PLoS ONE* 11, 1–15. DOI: 10.1371/journal.pone.0167989.

Gygli SM, S Borrell, A Trauner, and S Gagneux (2017). Antimicrobial resistance in *Mycobacterium tuberculosis*: mechanistic and evolutionary perspectives. *FEMS Microbiology Reviews* 41, 354–373. DOI: 10.1093/femsre/fux011.

Haller BC, J Galloway, J Kelleher, PW Messer, and PL Ralph (2019). Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Molecular Ecology Resources* 19, 552–566. DOI: 10.1111/1755-0998.12968.

Haller BC and PW Messer (2019). SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model. *Molecular Biology and Evolution* 36, 632–637. DOI: 10.1093/molbev/msy228.

Hanage WP (2016). Not so simple after all: bacteria, their population genetics, and recombination. *Cold Spring Harbor Perspectives in Biology* 8. DOI: 10.1101/cshperspect.a018069.

Hedge J and DJ Wilson (2014). Bacterial phylogenetic reconstruction from whole genomes is robust to recombination but demographic inference is not. *mBio* 5, 5–8. DOI: 10.1128/mBio.02158-14.

Heller R, L Chikhi, and HR Siegismund (2013). The confounding effect of population structure on Bayesian skyline plot inferences of demographic history. *PLoS ONE* 8. DOI: 10.1371/journal.pone.0062992.

Hershberg R and DA Petrov (2008). Selection on codon bias. *Annual Review of Genetics* 42, 287–299. DOI: 10.1146/annurev.genet.42.110807.091442.

— (2010). Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genetics* 6. DOI: 10.1371/journal.pgen.1001115.

Hershberg R et al. (2008). High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. *PLoS Biology* 6, 2658–2671. DOI: 10.1371/journal.pbio.0060311.

Hildebrand F, A Meyer, and A Eyre-Walker (2010). Evidence of selection upon genomic GC-content in bacteria. *PLoS Genetics* 6. DOI: 10.1371/journal.pgen.1001107.

Ho SY, S Duchêne, M Molak, and B Shapiro (2015). Time-dependent estimates of molecular evolutionary rates: evidence and causes. *Molecular Ecology* 24, 6007–6012. DOI: 10.1111/mec.13450.

Ho SYW and B Shapiro (2011). Skyline-plot methods for estimating demographic history from nucleotide sequences. *Molecular Ecology Resources* 11, 423–434. DOI: 10.1111/j.1755-0998.2011.02988.x.

Ho SYW et al. (2011). Time-dependent rates of molecular evolution. *Molecular Ecology* 20, 3087–3101. DOI: 10.1111/j.1365-294X.2011.05178.x.

Hoban S, G Bertorelle, and OE Gaggiotti (2012). Computer simulations: tools for population and evolutionary genetics. *Nature Reviews Genetics* 13. DOI: 10.1038/nrg3130.

Hughes AL, R Friedman, and M Murray (2002). Genomewide pattern of synonymous nucleotide substitution in two complete genomes of *Mycobacterium tuberculosis*. *Emerging Infectious Diseases* 8, 1342–1346. DOI: 10.3201/eid0811.020064.

Johri P et al. (2022). Recommendations for improving statistical inference in population genomics. *PLoS Biology* 20, 1–23. DOI: 10.1371/journal.pbio.3001669.

Kapur V, TS Whittam, and JM Musser (1994). Is *Mycobacterium tuberculosis* 15,000 years old? *Journal of Infectious Diseases* 170, 1348–1349. DOI: 10.1093/infdis/170.5.1348.

Karboul A et al. (2008). Frequent homologous recombination events in *Mycobacterium tuberculosis* PE/PPE multigene families: potential role in antigenic variability. *Journal of Bacteriology* 190, 7838–7846. DOI: 10.1128/JB.00827-08.

Katju V and U Bergthorsson (2019). Old trade, new tricks: Insights into the spontaneous mutation process from the partnering of classical mutation accumulation experiments with high-throughput genomic approaches. *Genome Biology and Evolution* 11, 136–165. DOI: 10.1093/gbe/evy252.

Katz Y (2013). Against storytelling of scientific results. *Nature Methods* 10, 1045–1045. DOI: 10.1038/nmeth.2699.

Kay GL et al. (2015). Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. *Nature Communications* 6. DOI: 10.1038/ncomms7717.

Kimura M (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press. DOI: 10.1017/CBO9781107415324.004.

Kucukyildirim S et al. (2016). The rate and spectrum of spontaneous mutations in *Mycobacterium smegmatis*, a bacterium naturally devoid of the postreplicative mismatch repair pathway. *G3: Genes, Genomes, Genetics* 6, 2157–2163. DOI: 10.1534/g3.116.030130.

Kuo CH, NA Moran, and H Ochman (2009). The consequences of genetic drift for bacterial genome complexity. *Genome Research* 19, 1450–1454. DOI: 10.1101/gr.091785.109.

Lanfear R, H Kokko, and A Eyre-Walker (Jan. 2014). Population size and the rate of evolution. *Trends in Ecology & Evolution* 29, 33–41. DOI: 10.1016/j.tree.2013.09.009.

Lapierre M, C Blin, A Lambert, G Achaz, and EP Rocha (2016). The impact of selection, gene conversion, and biased sampling on the assessment of microbial demography. *Molecular biology and evolution* 33, 1711–1725. DOI: 10.1093/molbev/msw048.

Lassalle F et al. (2015). GC-content evolution in bacterial genomes: the biased gene conversion hypothesis expands. *PLoS Genetics* 11, 1–20. DOI: 10.1371/journal.pgen.1004941.

Lawrence JG, RW Hendrix, and S Casjens (2001). Where are the pseudogenes in bacterial genomes? *Trends in Microbiology* 9, 535–540. DOI: 10.1016/S0966-842X(01)02198-9.

Liu Q et al. (May 29, 2020). *Mycobacterium tuberculosis* clinical isolates carry mutational signatures of host immune environments. *Science Advances* 6, eaba4901. DOI: 10.1126/sciadv.aba4901.

Liu Q et al. (2021). Local adaptation of *Mycobacterium tuberculosis* on the Tibetan Plateau. *PNAS* 118, 1–10. DOI: 10.1073/pnas.2017831118.

Liu X, MM Gutacker, JM Musser, and YX Fu (2006). Evidence for recombination in *Mycobacterium tuberculosis*. *Journal of Bacteriology* 188, 8169–8177. DOI: 10.1128/JB.01062-06.

Lynch M (2007). *The Origins of Genome Architecture*. Sinauer Associates Sunderland, MA.

— (2010). Evolution of the mutation rate. *Trends in Genetics* 26, 345–352. DOI: 10.1016/j.tig.2010.05.003.

Madacki J et al. (2021). ESX-1-Independent horizontal gene transfer by *Mycobacterium tuberculosis* complex strains. *mBio* 12, 1–19. DOI: 10.1128/mbio.00965-21.

Martin CJ et al. (2017). Digitally barcoding *Mycobacterium tuberculosis* reveals *in vivo* infection dynamics in the macaque model of tuberculosis. *mBio* 8, 1–12. DOI: 10.1128/mBio.00312-17.

Martin DP, P Lemey, and D Posada (2011). Analysing recombination in nucleotide sequences. *Molecular Ecology Resources* 11, 943–955. DOI: 10.1111/j.1755-0998.2011.03026.x.

Maynard Smith J (1995). Do bacteria have population genetics? In: *Population genetics of bacteria: Symposium 52*. Cambridge University Press, pp. 1–12.

Maynard Smith J, NH Smith, M O'Rourke, and BG Spratt (1993). How clonal are bacteria? *PNAS* 90, 4384–4388. DOI: 10.1073/pnas.90.10.4384.

McEvoy CR et al. (2007). The role of IS6110 in the evolution of *Mycobacterium tuberculosis*. *Tuberculosis* 87, 393–404. DOI: 10.1016/j.tube.2007.05.010.

Mcgrath M, NC Gey van pittius, PD Van helden, RM Warren, and DF Warner (2014). Mutation rate and the emergence of drug resistance in *Mycobacterium tuberculosis*. *Journal of Antimicrobial Chemotherapy* 69, 292–302. DOI: 10.1093/jac/dkt364.

Membrebe JV et al. (2019). Bayesian inference of evolutionary histories under time-dependent substitution rates. *Molecular Biology and Evolution* 36, 1793–1803. DOI: 10.1093/molbev/msz094.

Menardo F, S Gagneux, and F Freund (2021a). Multiple merger genealogies in outbreaks of *Mycobacterium tuberculosis*. *Molecular Biology and Evolution* 38, 290–306. DOI: 10.1101/2019.12.21.885723.

Menardo F, S Duchêne, D Brites, and S Gagneux (2019). The molecular clock of *Mycobacterium tuberculosis*. *PLoS Pathogens* 15, 1–24. DOI: 10.1371/journal.ppat.1008067.

Menardo F et al. (2021b). Local adaptation in populations of *Mycobacterium tuberculosis* endemic to the Indian Ocean Rim. *F1000Research*. DOI: 10.1101/2020.10.20.346866.

Merker M et al. (Aug. 30, 2022). Transcontinental spread and evolution of *Mycobacterium tuberculosis* W148 European/Russian clade toward extensively drug resistant tuberculosis. *Nature Communications* 13, 5105. DOI: 10.1038/s41467-022-32455-1.

Merrikh CN and H Merrikh (2018). Gene inversion potentiates bacterial evolvability and virulence. *Nature Communications* 9. DOI: 10.1038/s41467-018-07110-3.

Molina N and E Van Nimwegen (2008). Universal patterns of purifying selection at noncoding positions in bacteria. *Genome Research* 18, 148–160. DOI: 10.1101/gr.6759507.

Morales-Arce AY, RB Harris, AC Stone, and JD Jensen (2020). Evaluating the contributions of purifying selection and progeny-skew in dictating within-host *Mycobacterium tuberculosis* evolution. *Evolution* 74, 992–1001. DOI: 10.1111/evo.13954.

Morales-Arce AY, SJ Sabin, AC Stone, and JD Jensen (2021). The population genomics of within-host *Mycobacterium tuberculosis*. *Heredity* 126, 1–9. DOI: 10.1038/s41437-020-00377-7.

Moran NA (1996). Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *PNAS* 93, 2873–2878. DOI: 10.1073/pnas.93.7.2873.

Moreno-Molina M et al. (2021). Genomic analyses of *Mycobacterium tuberculosis* from human lung resections reveal a high frequency of polyclonal infections. *Nature Communications* 12, 1–11. DOI: 10.1038/s41467-021-22705-z.

Mortimer TD, AM Weber, and CS Pepperell (2017). Signatures of selection at drug resistance loci in *Mycobacterium tuberculosis*. *mSystems* 8, 1–11. DOI: 10.1101/173229.

Mulholland CV et al. (2019). Dispersal of *Mycobacterium tuberculosis* driven by historical European trade in the South Pacific. *Frontiers in Microbiology* 10 (December), 1–13. DOI: 10.3389/fmicb.2019.02778.

Muller HJ (1964). The relation of recombination to mutational advance. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis* 1, 2–9. DOI: 10.1016/0027-5107(64)90047-8.

Namouchi A, X Didelot, U Schöck, B Gicquel, and EP Rocha (2012). After the bottleneck: genome-wide diversification of the *Mycobacterium tuberculosis* complex by mutation, recombination, and natural selection. *Genome Research* 22, 721–734. DOI: 10.1101/gr.129544.111.

Neher RA (2013). Genetic draft, selective interference, and population genetics of rapid adaptation. *Annual Review of Ecology, Evolution, and Systematics* 44, 195–215. DOI: 10.1146/annurev-ecolsys-110512-135920.

Nei M and T Gojobori (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* 3, 418–426. DOI: 10.1093/oxfordjournals.molbev.a040410.

O'Neill MB et al. (2019). Lineage specific histories of *Mycobacterium tuberculosis* dispersal in Africa and Eurasia. *Molecular Ecology* 28, 3241–3256. DOI: 10.1111/mec.15120.

O'Neill MB, TD Mortimer, and CS Pepperell (2015). Diversity of *Mycobacterium tuberculosis* across evolutionary scales. *PLoS Pathogens* 11, 1–29. DOI: 10.1371/journal.ppat.1005257.

Ohta T (1992). The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics* 23, 263–286. DOI: 10.1146/annurev.es.23.110192.001403.

Orme IM (2014). A new unifying theory of the pathogenesis of tuberculosis. *Tuberculosis* 94, 8–14. DOI: 10.1016/j.tube.2013.07.004.

Osório NS et al. (2013). Evidence for diversifying selection in a set of *Mycobacterium tuberculosis* genes in response to antibiotic- and nonantibiotic-related pressure. *Molecular Biology and Evolution* 30, 1326–1336. DOI: 10.1093/molbev/mst038.

Outhred AC et al. (2020). Extensive homoplasy but no evidence of convergent evolution of repeat numbers at MIRU loci in modern *Mycobacterium tuberculosis* lineages. *Frontiers in Public Health* 8 (August), 1–12. DOI: 10.3389/fpubh.2020.00455.

Pan A, C Dutta, and J Das (1998). Codon usage in highly expressed genes of *Haemophillus influenzae* and *Mycobacterium tuberculosis*: translational selection versus mutational bias. *Gene* 215, 405–413. DOI: 10.1016/S0378-1119(98)00257-1.

Panda A, M Drancourt, T Tuller, and P Pontarotti (2018). Genome-wide analysis of horizontally acquired genes in the genus *Mycobacterium*. *Scientific Reports* 8, 1–13. DOI: 10.1038/s41598-018-33261-w.

Payne JL et al. (2019). Transition bias influences the evolution of antibiotic resistance in *Mycobacterium tuberculosis*. *PLoS Biology* 17, 1–23. DOI: 10.1371/journal.pbio.3000265.

Pepperell C et al. (2010). Bacterial genetic signatures of human social phenomena among *M. tuberculosis* from an aboriginal canadian population. *Molecular Biology and Evolution* 27, 427–440. DOI: 10.1093/molbev/msp261.

Pepperell CS (2022). Evolution of tuberculosis pathogenesis. *Annual Review of Microbiology* 76, 661–680.

Pepperell CS et al. (2013). The role of selection in shaping diversity of natural *M. tuberculosis* populations. *PLoS Pathogens* 9. DOI: 10.1371/journal.ppat.1003543.

Phelan JE et al. (2016). Recombination in pe/ppe genes contributes to genetic variation in *Mycobacterium tuberculosis* lineages. *BMC Genomics* 17, 1–12. DOI: 10.1186/s12864-016-2467-y.

Plutynski A (2007). Drift: a historical and conceptual overview. *Biological Theory* 2, 156–167. DOI: 10.1162/biot.2007.2.2.156.

Price MN and AP Arkin (2015). Weakly deleterious mutations and low rates of recombination limit the impact of natural selection on bacterial genomes. *mBio* 6. DOI: 10.1128/mBio.01302-15.

Rahman S, SL Kosakovsky Pond, A Webb, and J Hey (May 18, 2021). Weak selection on synonymous codons substantially inflates *dN/dS* estimates in bacteria. *PNAS* 118, e2023575118. DOI: 10.1073/pnas.2023575118.

Reichenberger ER, G Rosen, U Hershberg, and R Hershberg (2015). Prokaryotic nucleotide composition is shaped by both phylogeny and the environment. *Genome Biology and Evolution* 7, 1380–1389. DOI: 10.1093/gbe/evv063.

Rocha C and A Danchin (2002). Base composition bias might result from competition for metabolic resources. *TRENDS in genetics* 18, 291–294.

Rocha EP (2018). Neutral theory, microbial practice: challenges in bacterial population genetics. *Molecular Biology and Evolution* 35, 1338–1347. DOI: 10.1093/molbev/msy078.

Rocha EP and EJ Feil (2010). Mutational patterns cannot explain genome composition: are there any neutral sites in the genomes of bacteria? *PLoS Genetics* 6, 1–4. DOI: 10.1371/journal.pgen.1001104.

Rocha EP et al. (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. *Journal of Theoretical Biology* 239, 226–235. DOI: 10.1016/j.jtbi.2005.08.037.

Ryndak MB and S Laal (2019). *Mycobacterium tuberculosis* primary infection and dissemination: a critical role for alveolar epithelial cells. *Frontiers in Cellular and Infection Microbiology* 9. DOI: 10.3389/fcimb.2019.00299.

Sabin S et al. (2020). A seventeenth-century *Mycobacterium tuberculosis* genome supports a Neolithic emergence of the *Mycobacterium tuberculosis* complex. *Genome Biology* 21, 1–24. DOI: 10.1186/s13059-020-02112-1.

Sackman AM, RB Harris, and JD Jensen (2019). Inferring demography and selection in organisms characterized by skewed offspring distributions. *Genetics* 211, 1019–1028. DOI: 10.1534/genetics.118.301684.

Selander RK, JM Musser, DA Caugant, MN Gilmour, and TS Whittam (1987). Population genetics of pathogenic bacteria. *Microbial Pathogenesis* 3, 1–7. DOI: 10.1016/0882-4010(87)90032-5.

Shapiro BJ, LA David, J Friedman, and EJ Alm (2009). Looking for Darwin's footprints in the microbial world. *Trends in Microbiology* 17, 196–204. DOI: 10.1016/j.tim.2009.02.002.

Smith JM, CG Dowson, and BG Spratt (1991). Localized sex in bacteria. *Nature* 349, 29–31. DOI: 10.1038/349029a0.

Smith NH, SV Gordon, R de la Rua-Domenech, RS Clifton-Hadley, and RG Hewinson (2006). Bottlenecks and broomsticks: the molecular evolution of *Mycobacterium bovis*. *Nature Reviews Microbiology* 4, 670–681. DOI: 10.1038/nrmicro1472.

Smith TM et al. (2022). Rapid adaptation of a complex trait during experimental evolution of *Mycobacterium tuberculosis*. *Elife* 11, e78454.

Sniegowski PD and PJ Gerrish (2010). Beneficial mutations and the dynamics of adaptation in asexual populations. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365, 1255–1263. DOI: 10.1098/rstb.2009.0290.

Stern DL (2013). The genetic causes of convergent evolution. *Nature Reviews Genetics* 14, 751–764. DOI: 10.1038/nrg3483.

Supply P et al. (2013). Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of Mycobacterium tuberculosis. *Nature Genetics* 45, 172–179. DOI: 10.1038/ng.2517.

Tamuri AU, M dos Reis, and RA Goldstein (Mar. 1, 2012). Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics* 190, 1101–1115. DOI: 10.1534/genetics.111.136432.

Tantivitayakul P et al. (2020). Homoplastic single nucleotide polymorphisms contributed to phenotypic diversity in *Mycobacterium tuberculosis*. *Scientific Reports* 10, 1–10. DOI: 10.1038/s41598-020-64895-4.

Tarashi S, A Fateh, M Mirsaeidi, SD Siadat, and F Vaziri (2017). Mixed infections in tuberculosis: the missing part in a puzzle. *Tuberculosis* 107, 168–174. DOI: 10.1016/j.tube.2017.09.004.

Tellier A and C Lemaire (2014). Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. *Molecular Ecology* 23, 2637–2652. DOI: 10.1111/mec.12755.

Templeton AR (2021). *Population genetics and microevolutionary theory*. Second edition. Hoboken, NJ: John Wiley & Sons.

Thorpe HA, SC Bayliss, LD Hurst, and EJ Feil (2017). Comparative analyses of selection operating on nontranslated intergenic regions of diverse bacterial species. *Genetics* 206, 363–376. DOI: 10.1534/genetics.116.195784.

Tibayrenc M and FJ Ayala (2017). Is predominant clonal evolution a common evolutionary adaptation to parasitism in pathogenic parasitic protozoa, fungi, bacteria, and viruses? *Advances in Parasitology* 97, 243–325. DOI: 10.1016/bs.apar.2016.08.007.

Trauner A et al. (2017). The within-host population dynamics of *Mycobacterium tuberculosis* vary with treatment efficacy. *Genome Biology* 18, 1–17. DOI: 10.1186/s13059-017-1196-0.

Uplekar S, B Heym, V Friocourt, J Rougemont, and ST Cole (2011). Comparative genomics of ESX genes from clinical isolates of *Mycobacterium tuberculosis* provides evidence for gene conversion and epitope variation. *Infection and Immunity* 79, 4042–4049. DOI: 10.1128/IAI.05344-11.

Vos M and X Didelot (2009). A comparison of homologous recombination rates in bacteria and archaea. *ISME Journal* 3, 199–208. DOI: 10.1038/ismej.2008.93.

Walker TM et al. (2022). The 2021 WHO catalogue of *Mycobacterium tuberculosis* complex mutations associated with drug resistance: a genotypic analysis. *The Lancet Microbe* 3, e265–e273. DOI: 10.1016/s2666-5247(21)00301-3.

Wang L et al. (2022). Multiple genetic paths including massive gene amplification allow *Mycobacterium tuberculosis* to overcome loss of ESX-3 secretion system substrates. *PNAS* 119. DOI: 10.1073/pnas.2112608119.

Wang TC and FC Chen (2013). The evolutionary landscape of the *Mycobacterium tuberculosis* genome. *Gene* 518, 187–193. DOI: 10.1016/j.gene.2012.11.033.

Weissman JL, WF Fagan, and PL Johnson (2019). Linking high GC content to the repair of double strand breaks in prokaryotic genomes. *PLoS Genetics* 15, 1–19. DOI: 10.1371/journal.pgen.1008493.

Weller C and M Wu (2015). A generation-time effect on the rate of molecular evolution in bacteria. *Evolution* 69, 643–652. DOI: 10.1111/evo.12597.

Williams MJ et al. (2020). Measuring the distribution of fitness effects in somatic evolution by combining clonal dynamics with dN/dS ratios. *eLife* 9, 1–19. DOI: 10.7554/eLife.48714.

Wilson DJ et al. (2020). GenomegaMap: within-species genome-wide dN/dS estimation from over 10,000 genomes. *Molecular Biology and Evolution*, 1–11. DOI: 10.1093/molbev/msaa069.

Windels EM et al. (2021). Population bottlenecks strongly affect the evolutionary dynamics of antibiotic persistence. *Molecular Biology and Evolution* 38, 3345–3357. DOI: 10.1093/molbev/msab107.

Woese CR and N Goldenfeld (2009). How the microbial world saved evolution from the Scylla of molecular biology and the charybdis of the Modern Synthesis. *Microbiology and Molecular Biology Reviews* 73, 14–21. DOI: 10.1128/mmbr.00002-09.

World Health Organization (2022). *Global tuberculosis report 2022*. Section: xiii, 51 p. Geneva: World Health Organization.
Wright S (1931). Evolution in mendelian populations. *Genetics* 16. DOI: 10.4161/hv.21408.
Yang Z (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution* 15, 568–573. DOI: 10.1093/oxfordjournals.molbev.a025957.
— (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24, 1586–1591. DOI: 10.1093/molbev/msm088.
Yang Z and JR Bielawski (2000). Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution* 15, 496–503. DOI: 10.1016/S0169-5347(00)01994-7.
Yang Z and R Nielsent (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution* 19, 908–917. DOI: 10.1093/oxfordjournals.molbev.a004148.
Yang, Ziheng (2014). *Molecular Evolution – A Statistical Approach*. Oxford University Press.
Zwyer M et al. (2021). A new nomenclature for the livestock-associated *Mycobacterium tuberculosis* complex based on phylogenomics. *Open Research Europe* 1, 100–100. DOI: 10.12688/openreseurope.14029.1.