In-host population dynamics of *M. tuberculosis* during treatment failure

Roger Vargas Jr[1,2*], Luca Freschi[2], Maximillian Marin[1,2], L. Elaine Epperson[3], Melissa Smith[4,5], Irina Oussenko[5], David Durbin[6], Michael Strong[3], Max Salfinger[7], and Maha Reda Farhat[2,8*]

[1] Department of Systems Biology, Harvard Medical School

[2] Department of Biomedical Informatics, Harvard Medical School

[3] Center for Genes, Environment, and Health, National Jewish Health

[4] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai

[5] Icahn Institute of Data Sciences and Genomics Technology

[6] Mycobacteriology Reference Laboratory, Advanced Diagnostic Laboratories, National Jewish Health

[7] College of Public Health, University of South Florida

[8] Pulmonary and Critical Care Medicine, Massachusetts General Hospital

[*]Corresponding authors: roger_vargas@g.harvard.edu, Maha_Farhat@hms.harvard.edu

ABSTRACT

Tuberculosis (TB) is the leading cause of death globally from an infectious disease. Understanding the dynamics of TB's causative agent *Mycobacterium tuberculosis* (Mtb) in host is vital for antibiotic treatment and vaccine design. Here we use longitudinally collected clinical Mtb isolates from the sputa of 307 subjects to investigate Mtb diversity during the course of active TB disease. We excluded cases suspected of reinfection or contamination to analyze data from 200 subjects, 167 of which met microbiological criteria for delayed culture conversion, treatment failure or relapse. Using technical and biological replicate samples, we defined an allele frequency threshold attributable to in-host evolution. Of the 167 subjects with unsuccessful treatment outcome, 16% developed resistance amplification between sampling; 74% of amplification occurred among isolates that were genotypically resistant at the outset. Low abundance resistance variants in the first isolate predicts the fixation of these variants in the subsequent sample. We identify in-host variation in resistance and metabolic genes as well as in genes known to modulate host innate immunity by interacting with TLR2. We confirm these genes to be under positive selection by assessing phylogenetic convergence across a genetically diverse independent sample of 10,018 isolates.

INTRODUCTION

Tuberculosis (TB) and its causative pathogen *Mycobacterium tuberculosis* (Mtb) remain a major public health threat[1]. Yet the majority of individuals exposed to Mtb clear or contain the infection, and only 5-10% of those infected develop active TB disease at some point in their lifetime[2]. While basic human immune mechanisms to Mtb have been identified, attempts at effective vaccine development guided by these mechanisms have repeatedly failed[3]. Global efforts that include scale up of directly observed therapy have also been challenged by rising estimates of multidrug resistance. Mtb is an obligate human pathogen that has co-evolved with its human host over millennia[4]. Infection and disease involves a complex human host-pathogen interaction that is both physically and temporally heterogeneous[5]. Consequently all selective forces acting on Mtb will originate within the host, and the study of temporal dynamics of this is likely to inform antibiotic treatment[6] and rational vaccine design[3].

At long timescales, signatures of positive selection associated with antibiotic resistance have been characterized, but epitope regions appear to be under purifying selection[7–10] calling into question how Mtb interacts with host adaptive immunity. Little is known about selection at short timescales, such as within single infections. Drug pressure may select for resistance-conferring mutations, thus an understanding of how the frequency of minor alleles changes longitudinally can inform optimal drug treatment[6,11,12]. A recent study found treatment relapse to be strongly associated with bacterial factors[13]; therefore there is a need to better characterize these as predictors of treatment response. Bacterial factors of interest include not only low frequency resistance variants but also variants that may induce other phenotypes, such as drug tolerance or more effective immune evasion[14]. To elucidate these processes, we aimed to study how genomic diversity arises in-host in Mtb populations, employing a longitudinal sampling scheme from patients with active TB disease.

Allele frequencies within bacterial populations may differ between pooled samples (**Fig. 1a**) because they represent a difference in the genetic composition of the infecting population, commonly referred to as heterogeneity. Mtb population heterogeneity might be present within a host because (1) the host is infected with multiple strains or is re-infected by a new strain (consistent with mixed infection or re-infection) or (2) genetic diversity arises within the Mtb population during infection[15–17]. WGS of pooled sputum samples has been used extensively to investigate the metagenomic diversity of bacterial pathogens in humans[12,18–21]. However, non-uniform sampling[22], genetic drift and selection during *in vitro* expansion[22], laboratory contamination[23,24], sequencing error and mapping error all represent examples of experimental error that give rise to erroneous variant calls. This is especially problematic when calling variants at low[23] and mixed[15] allele frequencies, or sampling repeatedly from the same source[22]. Here, we present a framework to overcome these barriers and demonstrate the use of longitudinally collected isolates to investigate true in-host diversity with implications for Mtb treatment. We analyzed 614 paired longitudinal isolates representing 307 subjects from eight studies[17,22,25–29]. We find a high turnover of low-frequency alleles in loci associated with antibiotic resistance but that mutant alleles in these loci that rise to a frequency of 19% are predicted to fix in-host with a sensitivity of 27.0% and specificity of 95.6%. We show that changes in allele frequency are common among replicate isolates and that changes in frequency of 70% are indicative of in-host evolution using archived MTB isolates. We demonstrate that many loci involved the acquisition of antibiotic resistance and modulation of innate host-immunity appear to be under positive selection.

RESULTS

Identifying clonal Mtb populations in-host

To isolate the *in vivo* clonal dynamics of Mtb during infection among the 307 subjects with longitudinal samples, we excluded 32 subjects with isolate microbiological contamination at any time point[23], and 31 subjects with evidence for mixed infection with two or more Mtb lineages[24] (**Fig. 1b, Supplementary Fig. 2**). We also excluded 44 subjects with evidence for re-infection with a different Mtb strain between the first and second time points, using a pairwise genetic distance >7 fixed SNPs (fSNPs) (**Methods, Fig. 1c, Supplementary Fig. 2**). We implemented WGS SNP calling filters to minimize the likelihood of false positives and estimated the error rate of our analysis pipeline using a control dataset of 82 isolate pairs (162 total) that were *in vitro* technical or biological replicates (**Methods**, **Supplementary Fig. 2-3**). Of the 307 subjects, 200 had isolate pairs that passed all filters, with an estimated false positive SNP rate of 0.0513 or less. The 200 isolates represented the five main Mtb lineages.

In-host pathogen dynamics in antibiotic resistance loci

The presence of minor resistance alleles in-host has implications for the development of resistance amplification and has previously been studied for small sample sizes using WGS[11,22]. To investigate temporal dynamics related to antibiotic pressure[6,11,22], we identified non-synonymous and intergenic SNPs within a set of 36 predetermined resistance loci associated with antibiotic resistance[7,30] (**Supplementary Table 4**) that changed in allele frequency by more than 5%[6] between the first and second sampling time point (**Methods**). We detected 1,964 such SNPs across our sample of 200 subjects, 1,799 were non-synonymous, 91 were intergenic, and 74 occurred within the *rrs* region. (**Supplementary Table 5**).

We searched for evidence for competition between Mtb strains with different drug resistance mutations[6,11,22], or clonal interference, by characterizing longitudinal isolates fulfilling the following three criteria: (i) isolates contain multiple resistance SNPs in the same gene within the same subject, (ii) at alternate allele frequencies that change in opposing directions over time and (iii) the alternate (mutant) allele frequency was intermediate to high at ≥ 40% in at least 1 isolate[30] for at least one of the co-occurring SNPs. This identified 11 cases of clonal interference (**Fig. 2a**, **Supplementary Fig. 4**), demonstrating most often the fixation of a single allele in the second isolate from a mixture of multiple alleles at lower frequencies in the first isolate collected.

Antibiotic Resistance mutations are associated with delayed culture conversion and begets resistance amplification

Although detailed data on treatment regimens for the study subjects was not available to us, the source studies[17,22,25–29] indicated that all subjects had either recently completed treatment or were receiving treatment when samples were collected. Microbiological criteria for treatment failure include persistent positive sputum culture between 2 to 5 months from treatment initiation varying by treatment program[31]. We considered subjects with samples collected ≥60 days apart, by definition culture positive at sample collection time, as *delayed culture conversion, failure* or *relapse cases* (hitherto failure for brevity) (**Supplementary Fig. 1a**). Of the 270 subjects with mixed or clonal infection and reinfection, 5 had incomplete isolate collection dates (**Supplementary Table 2**). Of the remaining 265 subjects, 230 had samples collected ≥60 days apart and consisted of 35 reinfections (13%), 28 mixed infections at one or two time points (11%) and the majority, 167, had clonal infection (**Supplementary Fig. 1a, 2**).

To identify antibiotic resistance (AR) acquisition among subjects with clonal infection, we defined an AR SNP as one of the previously identified 1,964 SNPs with moderate to high ΔAF ≥ 40% based on prior evidence of association between such SNPs and phenotypic resistance[30]. Forty-one AR SNPs were detected across our sample. The acquisition of AR SNPs was significantly associated with failure ($P = 0.017$ Fisher's exact test); 16.2% of failures acquired at least 1 AR SNP while none of the other 28 subjects acquired an AR SNP during treatment (**Supplementary Fig. 1a**). We examined genotypic resistance to any drug, or multidrug resistance (MDR *i.e.* resistance to at least isoniazid and a rifamycin) by interrogating the first isolate collected from each subject for fixed AR SNPs[30] (**Methods**). Using this approach, we identified 230 pre-existing AR SNPs in 39% (65/167) of the failure subjects with 23% (39/167) being MDR (**Supplementary Fig. 1b-c, Supplementary Tables 6 and 7**). The acquisition of additional resistance mutations was significantly associated with pre-existing AR ($OR = 6.03, P = 6.8 \times 10^{-5}$ Fisher's exact test) or pre-existing MDR ($OR = 4.95, P = 3.8 \times 10^{-4}$ Fisher's exact test) with 20/27 (74%) of AR SNP acquisition among failure cases occurring in subjects with pre-existing resistance.

Allele frequency >19% predicts subsequent fixation of resistance variants.

We determined the lowest AR allele frequency that can accurately predict the development of fixed resistance alleles later in time[6,11] (**Fig. 2b**). We studied the AF trajectories of 1,964 AR SNPs detected with an AF_1 >5% at the first time point. We calculated the true positive rate (TPR) and false positive rates (FPR) for varying values of $AF_1 \in \{0,1,2,\cdots,99,100\}\%$ (**Supplementary Fig. 1d**, **Fig. 2b, Methods**). Allowing a maximum FPR of 5%, we found the optimal classification threshold to be $AF_1^* = 19\%$ with an associated sensitivity of 27.0% and a specificity of 95.6%. Ten mutant alleles across 14 isolates from 7 subjects had a frequency between 19% and 75% at the first time point and rose to fixation at the second time point (mean ΔAF 41%).

Genome-wide in-host diversity

Beyond antibiotic pressure, selective forces acting on the infecting Mtb strain in-host are largely unknown. To investigate this reliably across the entire Mtb genome, we first examined the genome-wide allele frequency distribution for both technical replicates (*in vitro* technical or biological replicates, sample size m=62 after exclusions, **Supplementary Figure 2**) and in-host longitudinal pairs (**Supplementary Fig. 2-3**). We detected five SNPs in *glpK* (with ΔAF ≥ 25%) among five replicate pairs (mean ΔAF=45%) consistent with an adaptive role for *glpK* mutations *in vitro*[32] and accordingly excluded this gene from further analysis (**Methods**). The genome-wide AF distribution demonstrated an abundance of SNPs with small changes in AF among both replicate and longitudinal pairs likely resulting from technical factors or noise. To clearly distinguish signal related to in-host factors from noise, we determined the ΔAF threshold above which SNPs/isolate-pair were rare among technical replicates *i.e.* constituted 5% or less of the total SNPs when replicate and longitudinal pairs were pooled (**Supplementary Fig. 3**). We determined this ΔAF threshold to be 70% and selected 178 SNPs that developed in-host among the 200 TB cases (**Supplementary Fig. 3c, Supplementary Table 10**).

Characteristics of mutations in-host

Of the 178 SNPs, 115 were non-synonymous, 42 synonymous, and 21 were intergenic (**Fig. 3c**). The 157/178 coding SNPs were distributed across 129/3,886 genes and were observed in 71/200 subjects (**Fig. 3b,d**). The preponderance of non-synonymous SNPs is as previously observed for Mtb[9,33,34]. We analyzed the spectrum of mutations and found the GC > AT nucleotide transition to be the most common. The GC > AT transition is putatively due to oxidative damage including the deamination of cytosine/5-methyl-cytosine or the formation of 8-oxoguanine[35,36]. The transversion AT > TA was the least common substitution (**Fig. 4a**). We expected the number of SNPs detected between longitudinal isolates to increase with time between isolate collection. Regressing the number of SNPs per subject on the timing between isolate collection (for 195 subjects with isolate collection dates) (**Fig. 4b**), we found SNPs to accumulate at an average rate of 0.57 SNPs per genome per year ($P = 4.8 \times 10^{-11}$) consistent with prior *in vivo* estimates[26,35].

Antibiotic Resistance and PE/PPE genes vary while antigens remain conserved

To understand how different classes of proteins evolve in-host, we separated Mtb genes into five non-redundant categories (**Methods**): *Antibiotic resistance* - genes as defined above[7], *PE/PPE* – gene family unique to pathogenic mycobacteria, thought to influence immunopathogenicity and is characterized by conserved proline-glutamate (PE) and proline-proline-glutamate (PPE) motifs at the N protein termini[10,34,37], *Antigen* - genes encoding a CD4$^+$ or CD8$^+$ T-cell epitope[8,10] (excluding PE/PPE genes), *Essential* - genes required for growth *in vitro* and *in vivo*[10,38,39], and *Non-Essential* - genes not categorized into one of the aforementioned categories. The vast majority of genes in each category did not vary within subject (**Fig. 4c**). Antibiotic resistance genes were on average the most diverse category while Essential genes varied the least (**Fig. 4d**). Antigen genes appeared to be as conserved as both Essential ($P = 0.49$ Mann-Whitney U-test) and Non-Essential genes ($P = 0.45$ Mann-Whitney U-test) while PE/PPE genes showed higher levels of nucleotide diversity than both Essential ($P = 0.0038$ Mann-Whitney U-test) and Non-Essential genes ($P = 0.0012$ Mann-Whitney U-test) (**Fig. 4d**).

PE/PPE variation is independent of T-cell recognition

To test whether variation in Antigen or PE/PPE genes occurred in response to T-cell recognition, we separated each gene in these categories into ($CD4^+$ and $CD8^+$ T-cell) epitope and non-epitope concatenates and recalculated nucleotide diversity for these concatenates (**Fig. 4e-h**). For both Antigen and PE/PPE genes (**Fig. 4f,h**), epitope concatenates were less diverse than non-epitope concatenates ($P = 0.018$ and $P = 0.028$ respectively, Whitney U-test). Only one in-host SNP was detected within an epitope-encoding region in the gene *PPE18* (**Fig. 4g**, **Supplementary Fig. 6**, **Supplementary Table 9**). This suggests that T-cell recognition does not drive diversity in these regions.

The PE/PPE genes consist of 3 sub-families (**Fig. 4i-j**), PE-PGRS genes with PE motifs at the N-terminus along with redundant polymorphic GC-rich repetitive sequence, PE genes with PE motifs but without redundant polymorphic GC-rich repetitive sequence, and PPE genes with proline-proline-glutamate motifs at the N-terminus[40]. On average, PPE and PE-PGRS genes appeared more diverse in-host than PE genes ($P = 0.019$ and $P = 0.068$ respectively, Mann-Whitney U-test).

Identifying candidate pathoadaptive loci from genome-wide variation

To identify genes involved in pathogen adaptation[18,19], we applied a test of mutational density[41] (**Methods**) by pooling variation across all 200 pairs of genomes and identifying those genes with more mutations than expected under a neutral model of evolution where variants are

Poisson distributed across the genome[42] (**Fig. 3b, Supplementary Table 11**). We also searched for evidence of convergent evolution *i.e.* genes or pathways where in-host SNPs developed in ≥ 2 subjects (**Methods**). Seven known antibiotic resistance genes[7,12] had significant mutational density ($\alpha = 0.05$, Bonferroni correction) or were convergent across patients: *rpoB*, *gyrA*, *katG*, *rpoC*, *embB, ethA* and *pncA* (mutated in six, four, four, three, three, two and one subject respectively) (**Fig. 3b,d**). Single in-host SNPs occurred in eight additional known resistance loci including three intergenic regions, and in *prpR,* a gene recently implicated with drug tolerance[43] (**Supplementary Table 10**). Three genes with unknown function: Rv0139, Rv0895, and Rv1543 were convergent in two subjects each, two of which (Rv0139, Rv1543) had significant mutational density ($P<2x10^{-5}$) and; three additional genes including *PPE60* displayed significant mutational density ($P<2x10^{-5}$) (**Fig. 3b**). We found evidence for convergence in six pathways not known to result in antibiotic resistance. These pathways are involved with biotin biosynthesis (*fadD23*, *fadD29*, and *fadD30*), ribosomal large subunit proteins (*rpmB1*, *rplE*, and *rplY*), glycerolipid and glycerophosolipid metabolism (*aldA* and Rv2974c), ESAT-6 protein secretion (Rv3870 and Rv3877), coenzyme B12/cobalamin synthesis (*cobH* and *cobK*) and the uncharacterized pathway CBSS-164757.7.peg.5020 (*fdxB* and *PPE18*) (**Supplementary Table 14**).

In-host mutations display phylogenetic convergence across multiple global lineages

We reasoned that pathoadaptive mutations observed to sweep to fixation in-host and not compromise pathogen transmissibility are likely to arise independently within other subjects and in separate geographic regions in a convergent manner[7]. With the exception of *rpoBC*, we excluded regions known to encode antibiotic resistance and screened a genetically and geographically diverse set of 10,018 sequenced clinical isolates for mutations occurring in the same gene identified in the tests for mutational density or convergence at the gene/pathway level described above (22 genes total, **Methods**, **Fig. 5**, **Supplementary Table 17-18**). A mutation was characterized as phylogenetically convergent if it was present ≥10 isolates within at least two Mtb lineages (Lineages 1-4) (**Methods**, **Fig. 5a**).

We identified 67 sites within five of the 22 genes to be phylogenetically convergent (**Fig. 5b**, **Supplementary Table 19**). These included the conserved protein of unknown function *Rv0095c* (18 sites), the PPE genes *PPE60* (9 sites) and *PPE18* (22 sites). We included two genes associated with antibiotic resistance that are known targets of positive selection for comparison to our other hits, *rpoB,* known to encode resistance to rifampicin[7]*,* (12 sites) and *rpoC*, known to encode compensatory rifampicin mutations[44] (6 sites). We manually inspected the alignments corresponding to the four in-host SNPs in PPE genes *PPE60* and *PPE18* (**Supplementary Fig. 9,11**) and performed *in silico* read simulations to confirm SNP calls including repetitive and PE/PPE regions and finally confirmed calls using PacBio sequence data on a subset of isolates (**Methods**).

DISCUSSION

This is the first study examining in-host longitudinal Mtb diversity at scale. To understand how Mtb populations change over time, we sought to investigate changes in the genetic composition of Mtb populations in-host by searching for changes in allele frequencies in serially collected isolates[11,22,29]. In our 400 Mtb whole genomes sampled from 200 active TB patients heavily enriched for delayed culture conversion, treatment failure and relapse, we find a wealth of dynamics in genetic loci associated with antibiotic resistance, including a high turnover of minor variants[22]. Of patients with delayed culture conversion, treatment failure or relapse, we observe a relatively high percentage, 16%, to develop antibiotic resistance over time. The rate of *in vivo* resistance acquisition is higher for the subset of patients with MDR at the outset and negative outcomes, estimated here at 36%. This not only emphasizes the importance of appropriately tailoring treatment regimens but also the need for close surveillance for resistance acquisition by phenotypic or genotypic means. The observed high rate of resistance acquisition also emphasizes Mtb's biological adaptability to drug pressure *in vivo.* For most other pathogens, resistance acquisition in the course of one infection is very rare[45]. In addition to clonal acquisition of resistance and of clinical relevance, we found 27% of patients with unsuccessful treatment outcomes to have mixed infection or reinfection with different Mtb strains. This high percentage suggests that care of these patients and control of disease transmission can be better guided if pathogen sequencing is routinely performed for cases meeting these microbiological criteria especially in high TB prevalence settings.

Under drug selective pressure, we show that clonal interference purges diversity as only a subset of co-existing minor antibiotic resistance alleles reach fixation in many loci. Detection of several minor alleles within an antibiotic resistance locus may thus hint at eventual fixation of one of the alleles within the Mtb population and resistance amplification. We provide a proof of concept that minor alleles can predict future antibiotic resistance, by demonstrating that canonical antibiotic resistance variants occurring at a frequency as low as 19% accurately predict fixation of the variant in >95% of mutations in-host. Yet we find the sensitivity of this threshold to be low, with 73% of new fixed resistance variants not initially observed at an abundance of ≥19%. This is likely related to our simplistic assumption that selective forces are more or less similar between patients, time intervals, drugs and mutations and hence our threshold was estimated by averaging over these variables. In reality predictive minor allele frequencies will vary by drug, type of mutation, patient and treatment variables and these variables can be investigated further for improved sensitivity as more data on this question becomes available.

While various sources of error prevent making inferences on changing bacterial composition (genome-wide) when allele frequencies between samples change by small magnitudes, we determined an appropriate threshold for identifying mutations in-host using archived or frozen Mtb isolates. This demonstrates the importance of including replicate clinical isolates in WGS studies with longitudinal sampling schemes from the same hosts. While culturing sputa from subjects followed by *in vitro* expansion of bacterial pathogens creates experimental noise, other methods of sample extraction, such as DNA extraction directly from MGIT subject samples[46] and higher sequencing depth, may allow for calling relevant changes in allele frequencies at lower thresholds. This would permit the unbiased study of loci that may be under frequency-dependent selection, where changes in allele frequencies would unlikely change by as much as 70% as we used here.

We detected 178 alleles rising to near fixation in-host across our sample of 200 subjects. The observed distribution of variants including the high rate of non-synonymous substitutions and the predominance of GC > AT variants are consistent with other sequencing studies of in-host or clinical MTB[16,35] and adds validity to our analysis approach. The underlying mechanism explaining these observations in Mtb have included purifying pressure on synonymous variants and oxidative DNA damage respectively[33,35]. Overall the observed diversity spared the $CD4^+$ and $CD8^+$ T cell epitope encoding regions of the genome, consistent with prior studies[8,10,47] and adding to the existing literature describing that host adaptive immunity does not drive directional selection in Mtb genomes. Diversity was concentrated in both antibiotic resistance regions and to an even larger extent in PE/PPE genes. Although previous studies have generally avoided reporting short-read variant calls in PE/PPE regions, we demonstrate using read simulation, visualization of illumina read alignments and comparison with long-read sequencing data that the SNPs captured in our study are highly unlikely to be false positive calls. We found PPE and PE-PGRS genes to be more diverse in-host than PE genes and detected a signal of positive selection acting on two genes belonging to the PPE genes but no genes belonging PE or PE-PGRS sub-families (**Fig. 5**). This indicates that PPE genes may be more functionally relevant in the process of host-adaptation.

Evidence of directional selection in Mtb genomes have thus far been largely restricted to adaptation to antibiotic treatment[9,12,22]. We identified six genes and six pathways displaying diversity in-host and not known to be associated with antibiotic resistance (**Fig. 3d**). For a subset we demonstrate similar diversity has arisen independently in separate hosts and in strains with different genetic backgrounds suggesting positive selection (**Fig. 5**). We also identify in-host variation in 12 loci known to be involved in the acquisition of antibiotic resistance[7,44] (**Fig. 3d**) and this lends further validity in identifying genes under selective pressure *in vivo*. The pathways showing in-host convergence may be important for interactions between host-and pathogen arising from either metabolic or immune pressure. Mtb is one of a few types of bacteria that possess the capacity for *de novo* coenzyme B12/cobalamin synthesis, and this pathway has been implicated in Mtb survival in-host and Mtb growth[48]. We identified four genetic variants that developed in three separate patients and in three consecutive genes from the same locus *cobG,* intergenic *cobG-cobH, cobH* and *cobK* (Rv2064-Rv2067). This observation contributes to mounting evidence on the importance of this pathway for *in vivo* Mtb survival and may have implications for drug development[49,50]. Biotin biosynthesis is also relatively unique to mycobacteria and plays an important role in Mtb growth, infection and host survival during latency[51].The other identified pathways include ESAT-6 protein secretion known to play a role in the modulation of host innate immune response[52].

Three additional loci not known to be associated with antibiotic resistance and found to be phylogenetically convergent, include the genes Rv0095c, *PPE18* and *PPE60*. Although of unknown function Rv0095c (SNP A85V) was recently associated with transmission success of an Mtb cluster in Peru[53]. Both *PPE18* and *PPE60* have been shown to interact with toll-like receptor 2 (TLR2)[54, 55]. Additionally, *PPE18* was the only gene to encode an epitope containing a SNP in-host; mutations in the epitope-encoding regions of this gene have previously been described in a set of geographically separated clinical isolates[56]. We also observed one variant arise in-host in *PPE54,* a gene implicated in Mtb's ability to arrest macrophage phagosomal maturation (phagosome-lysosome fusion) and thought to be vital for intracellular persistence[57]. The mechanism by which *PPE54* accomplishes this is unknown, but Mtb modification of phagosomal function is thought to be TLR2/TLR4-dependent[58].

Mtb is known to disrupt numerous *innate* immune mechanisms including phagosome maturation, apoptosis, autophagy as well as inhibition of MHC II expression through prolonged engagement with innate sensor toll-like receptor 2 (TLR2) among others[14]. SNPs in human genes involved with innate-immune pathways have been implicated in-host susceptibility to TB[59–61]. Specifically, SNPs in TLR2 (thought to be the most important TLR in Mtb recognition)[60] and TLR4 have been associated with susceptibility to TB disease[59,61]. Overall, these observations and our results are consistent with ongoing co-evolution between humans and Mtb. It appears that both human (e.g. immune receptors and cytokines) and Mtb (e.g. surface proteins) genetic loci may interact and respond to reciprocal adaptive changes, leaving a signature of selection in the genetic diversity of both humans and Mtb populations[9]. Most co-evolution between Mtb and humans, the main reciprocal adaptations between host and pathogen are thought to have occurred long ago and as a result of long-term host-pathogen interactions[9,61]. Unexpectedly, we observe these dynamics over the short evolutionary timescale of a single infection which has important implications for vaccine development[40].

METHODS

Sequence Data

Longitudinal Isolate Pairs: This study included data for 614 clinical isolates of *M. tuberculosis* that were sampled from the sputum of 307 subjects resulting in n = 307 longitudinal pairs. The sequencing data for 456 publicly available isolates was downloaded from Genbank[62], sequenced using Illumina chemistry to generate paired-end reads and came from previously published studies (T[22], C[25], W[26], B[27], G[17], X[29], H[28], P[63]) (**Supplementary Fig. 2**).

Replicate Isolate Pairs: This study included three types of replicate isolate pairs. (S2 - Sequenced Twice) DNA pooled from a single Mtb clinical isolate that had undergone *in vitro* expansion was sequenced in separate runs on an Illumina sequencing machine (m = 5). (C2 – Cultured & Sequenced Twice) Mtb was cultured from a single frozen clinical sample at separate time points, then sequenced on an Illumina sequencing machine after DNA extraction from culture (m = 73). (P3) Three sputum samples were obtained from a single subject within a 24 hour period[22], cultured separately, underwent DNA extraction and then sequencing on an Illumina sequencing machine. For the purposes of this study, we compared these three isolates pairwise (m = 3).

Public Sequence Data: We downloaded raw sequence data for 10,018 clinical isolates from the public domain[62]. Isolates had to meet the following quality control measures for inclusion in our study: (i) at least 90% of the reads had to be taxonomically classified as belonging to the *Mycobacterium tuberculosis* complex after running the trimmed FASTQ files through Kraken[64], (ii) at least 95% of bases had to have coverage of at least 10x after mapping the processed reads to the H37Rv Reference Genome, and (iii) the global lineage of the isolate was determined via SNP barcoding[65].

DNA extraction for PacBio Sequencing: MTB cultures were allowed to grow for 4-6 weeks. Pellets were heat-killed at 80°C for 20 minutes[66,67], the supernatants were removed, and the enriched cell pellet was subjected to DNA extraction soon after or stored frozen until extraction. Heat-killed cells pellets were immersed and briefly vortexed in 200ul lysis buffer (15% sucrose, 0.05M Tris-Cl pH 8.0, 0.05M EDTA, pH 8.0[68], 50ul of 100mg/ml lysozyme added, and samples were incubated overnight at 37°C. To each sample was added 50ul of 2.5mg/ml proteinase K, 100ul 20% SDS, and 4ul RNaseA/T1, and samples were incubated for 10 minutes at 65°C. 800ul of ChIP DNA binding buffer from Zymo Genomic DNA Clean and Concentrator-25 was added, and the samples were mixed vigorously by hand for at least 60 seconds. The cell debris was pelleted for 2 min at maximum in a microfuge, supernatants were transferred to the Zymo column, and DNA cleaned according to manufacturer's protocol (Zymo Research, Irvine, CA), except that 10mM Tris-Cl pH 8.0 was used for elution to omit EDTA. Yields were determined using fluorescent quantitation (Qubit, Invitrogen/Thermofisher Scientific) and quality was assessed on a 0.8% GelRed agarose gel with 1XTAE, separated for 90 minutes at 80V.

PacBio Sequencing of Mtb Isolates: Approximately 1 mg of high molecular weight genomic DNA was used as input for SMRTbell preparation, according to the manufacturer's specifications (SMRTbell Template Preparation Kit 1.0, Pacific Biosciences, https://www.pacb.com/wp-content/uploads/2015/09/Procedure-Checklist-20-kb-Template-Preparation-Using-BluePippin-Size-Selection.pdf). Briefly, HMW gDNA was sheared to 20kb using the Covaris g-tube at 4500 rpm. Following shearing, gDNA underwent DNA damage repair, ligation to SMRTbell adaptors and exonuclease treatment to remove any unligated gDNA. At least 500 ng final SMRTbell library per sample was cleaned with AMPure PB beads and 3-50 kb fragments were size selected using the BluePippin system on 0.75%

agarose cassettes and S1 ladder, as specified by the manufacturer (Sage Science). Size selected SMRTbell libraries were annealed to sequencing primer and bound to the P6 polymerase prior to loading on the RSII sequencing system (Pacific Biosciences). Sequencing was performed using C4 chemistry and 240-minute movies. Following data collection, raw data was converted into subreads for subsequent analysis using the RS_Subreads.1 pipeline within SMRTPortal (version 2.3), the web-based bioinformatics suite for analysis of RSII data.

Epitope Collection and Analysis

CD4$^+$ T and CD8$^+$ T cell epitope sequences were downloaded from the Immune Epitope Database[69] on May 23rd, 2018 according to criteria described previously[8] [linear peptides, *M. tuberculosis* complex (ID:77643, Mycobacterium complex), positive assays only, T cell assays, any MHC restriction, host: humans, any diseases, any reference type] yielding a set of 2,031 epitope sequences (**Supplementary Table 8**). We mapped each epitope sequence to the genes encoded by the H37Rv Reference Genome[70] using BlastP with an e-value cutoff of 0.01 (**Supplementary Fig. 5**). We retained only epitope sequences that mapped to at least 1 region in H37Rv (due to sequence homology, some epitopes mapped to multiple regions) and whose BlastP peptide start/end coordinates matched those specified in IEDB (n = 1,949 representing 1,505 separate epitope entries in IEDB). We then filtered out any epitopes occurring in Mobile Genetic Elements which resulted in a final set of 1,875 epitope sequences, representing 348 genes (antigens) used for downstream analysis. The distribution of peptide lengths for this final set of epitopes is given in **Supplementary Fig. 5**. Since many of these epitope sequences overlap, we constructed non-redundant epitope concatenate sequences for each antigen (n = 348) gene[8,10,71]. The regions of each antigen not encoding an epitope were concatenated into a non-epitope sequence for that gene.

Gene Sets

Every gene on H37Rv was classified into one of six non-redundant gene categories according to the following criteria: (i) genes identified as belonging to the PE/PPE family of genes[10,37] were classified as *PE/PPE* (n = 167), (ii) genes flagged as being associated with antibiotic resistance were classified into the *Antibiotic Resistance* category (n = 28), (iii) genes encoding a T cell epitope (but not already classified as a PE/PPE or Antibiotic Resistance gene) were classified as an *Antigen* (n = 257), (iv) genes required for growth *in vitro*[38] and *in vivo*[39] and not already placed into a category above were classified as *Essential* genes (n = 682), (v) genes flagged as transposases, integrases, phages or insertion sequences were classified as *Mobile Genetic Elements*[10] (n = 108), (vi) any remaining genes not already classified above were placed into the *Non-Essential* category (n = 2752) (**Supplementary Table 3**).

Variant Calling

Illumina FastQ Processing and Mapping to H37Rv: The raw sequence reads from all sequenced isolates were trimmed with Prinseq[72] (settings: -min_qual_mean 20) (version 0.20.4) then aligned to the H37Rv Reference Genome (Genbank accession: NC_000962) with the BWA mem[73] algorithm (settings: -M) (version 0.7.15). The resulting SAM files were then sorted (settings: SORT_ORDER = coordinate), converted to BAM format and processed for duplicate removal with Picard (http://broadinstitute.github.io/picard/) (version 2.8.0) (settings: REMOVE_DUPLICATES = true, ASSUME_SORT_ORDER = coordinate). The processed

BAM files were then indexed with Samtools[74]. We used Pilon[75] on the resulting BAM files to call bases for all reference positions corresponding to H37Rv as well as micro-Indels from pileup (settings: --variant).

Single Nucleotide Polymorphism (SNP) Calling: To prune out low-quality base calls that may have arisen due to sequencing or mapping error, we dropped any base calls that did not meet any of the following criteria[21]: (i) the call was flagged as either *Pass* or *Ambiguous* by Pilon, (ii) the reads aligning to that position supported at most 2 alleles (ensuring that 1 allele matched the reference allele if there were 2), (iii) the mean base quality at the locus was > 20, (iv) the mean mapping quality at the locus was > 30, (v) none of the reads aligning to the locus supported an insertion or deletion, (vi) a minimum coverage of 25 reads at the position, and (vii) the position is not located in a mobile genetic element region of the reference genome. We then used the Pilon-generated[75] VCF files to calculate the frequencies for both the reference and alternate alleles, using the *INFO.QP* field (which gives the proportion of reads supporting each base weighted by the base and mapping quality of the reads, *BQ* and *MQ* respectively, at the specific position) to determine the proportion of reads supporting each base for each locus of interest.

Additional SNP Filtering for Isolate Pairs: To call SNPs (and corresponding changes in allele frequencies) between pairs of isolates (Replicate and Longitudinal pairs), we required: (i) *SNP Calling* filters be met, (ii) the number of reads aligning to the position is below the 99th percentile for all of the calls made for that isolate, (iii) the call at that position passes all filters for each isolate in the pair, and (iv) SNPs in *glpK* were dropped as mutants arising in this gene are thought to be an artifact of *in vitro* expansion[32]; we detected four non-synonymous SNPs in *glpK* between three longitudinal pairs (mean ΔAF=64%).

Additional SNP Filtering for Antibiotic Resistance Loci Analysis: To call SNPs (and corresponding minor changes in allele frequencies) between pairs of isolates (Longitudinal Pairs), we required: (i) *SNP Calling* filters be met, (ii) *Additional SNP Filtering for Isolate Pairs* filters be met, (iii) $\left|AF_1^{alt} - AF_2^{alt}\right| = \Delta\text{AF} \geq 5\%$, (iv) if $5\% \leq \Delta\text{AF} < 20\%$, then the SNP was only retained if each allele (across both isolates) with $\text{AF} > 0\%$ was supported by at least 5 reads (ensuring that at least 5 reads supported each minor allele at lower values of ΔAF), (v) the SNP was classified as either intergenic or non-synonymous, (vi) the SNP was located in a gene, intergenic region or rRNA coding region associated with antibiotic resistance (**Supplementary Table 4**).

Additional SNP Filtering for Public Isolates: We screened a set of 10,018 public isolates for the same SNPs detected in our in-host analysis. In these isolates, we evaluated the base calls at the same reference positions for which we detected in-host SNPs and required that the calls be flagged as *Pass* by Pilon in addition to our other filters for SNP calling. This ensured that at least 75% of reads at a given position supported the same alternate allele detected in-host.

PacBio *de novo* Assembly, Genome Polishing, and Variant Calling: PacBio and Illumina sequencing data was available for 19 clinical Mtb isolates. We used Canu[76] to *de novo* assemble the raw PacBio subreads from these 19 isolates (settings: genomeSize=4.4m -pacbio-raw) (version 1.8). We used Circlator[77] to close the resulting assembly using the corrected-trimmed reads provided by Canu. PacBio's bax2bam function (settings: --subread) was used to convert PacBio legacy BAX files to BAM format. We ran PacBio's implementation of Minimap2[78] (pbmm2) to map and sort raw PacBio subreads to the closed genome from Circlator. We iteratively polished the assembly three times by running the Quiver algorithm[79] and used Samtools[74] to index the fasta files from the resulting assemblies. Fifteen of our samples assembled into a single contig, 2 samples assembled into 2 contigs each, 1 assembled into 4 contigs and 1 assembled into 24 contigs (**Supplementary Table 20**). To call SNPs relative to the H37Rv reference, we used Minimap2[80] to align each PacBio assembly to the H37Rv reference sequence. We used the *paftools.js call* utility included with Minimap2 to generate variant calls from each assembly to reference alignment. We excluded samples that assembled into more than a single contig from downstream analysis. Additionally, we excluded samples: M0018577_8a, M0013712_6, and M0002959_6 due to having a pairwise genetic distance > 100 SNVs with their corresponding Illumina sequenced samples. This large number of SNVs between PacBio and Illumina sequences originating from the same Mtb isolate was likely due to contamination or mislabeling of samples.

Mixed Lineage and Contamination Detection for Isolate Pairs

Kraken: To filter out samples that may have been contaminated by foreign DNA during sample preparation, we ran the trimmed reads for each longitudinal and replicate isolate through Kraken2[64] against a database[23] containing all of the sequences of bacteria, archaea, virus, protozoa, plasmids and fungi in RefSeq (release 90) and the human genome (GRCh38). We calculated the proportion reads that were taxonomically classified under the *Mycobacterium tuberculosis* Complex (MTBC) for each isolate and implemented a threshold of 95%. An isolate pair was dropped if either isolate had less than 95% of reads aligning to MTBC.

F2: To further reduce the effects of contamination, we aimed to identify samples that may have been subject to inter-lineage mixture samples resulting from of a co-infection (F2). We computed the F2 lineage-mixture metric for each longitudinal and replicate isolate (**Fig. 1**). We wrote a custom script to carry out the same protocol for computing F2 as previously described[24]. Briefly, the method involves calculating the minor allele frequencies at lineage-defining SNPs[65]. From 64 sets of SNPs that define the deep branches of the MTBC[65], we considered the 57 sets that contain more than 20 SNPs to obtain better estimates of minor variation[24,65]. For each SNP set i, (i) we summed the total depth and (ii) the number of reads supporting the most abundant base (at each position) over all of the reference positions (SNPs) that met our mapping quality, base quality and insertion/deletion filters, which yields d_i and x_i respectively. Subtracting these two quantities yields the minor depth for SNP set i, $m_i = d_i - x_i$. The minor allele frequency estimate for SNP set i is then defined as $p_i = m_i \,/\, d_i$. Doing this for all 57 SNP sets gives $\{p_1, p_2, \cdots, p_{57}\}$. We then sorted $\{p_1, p_2, \cdots, p_{57}\}$ in descending order and estimated the minor variant frequency for all of the reference positions (SNPs) corresponding to the top 2 sets (highest p_i values) which yields the F2 metric. Letting $n2$ be the number of SNPs in the top 2 sets, then $F2 = \sum_{j=1}^{n2} m_j \,/\, \sum_{i=1}^{n2} d_i$. Isolate pairs were dropped if the F2 metric for either isolate passed the F2 threshold set for mixed lineage detection (**Fig. 1, Supplementary Fig. 2**).

Pre-existing Genotypic Resistance

We determined pre-existing resistance for a subject (with a pair of longitudinal isolates) by scanning the first isolate for the detection of at least 1 of 177 SNPs predictive of resistance with AF ≥ 75% (from a minimal set of 238 variants[30]). We excluded predictive indels and the *gid* E92D variant as the latter is likely a lineage marking variant that is not indicative of antibiotic resistance. We defined pre-existing multidrug resistance for a subject by scanning the first isolate collected for detection of at least 1 SNP predictive of Rifampicin resistance (14/178 predictive SNPs) and at least 1 SNP predictive of Isoniazid resistance (18/178 predictive SNPs).

True & False Positive Rate Analysis for Heteroresistant Mutations

To determine the predictive value of low-frequency heteroresistant alleles, we classified SNPs as fixed if the alternate allele frequency in the 2nd isolate collected from the subject was at least 75% (alt $AF_2 \geq 75\%$). We first dropped SNPs for which alt $AF_1 \geq 75\%$ and alt $AF_2 \geq 75\%$ (high frequency mutant alleles in both isolates). We then set a threshold (F_i) for the alternate allele frequency detected in the 1st isolate collected from the subject (alt AF_1) and predicted whether an alternate allele would rise to a substantial proportion of the sample (alt $AF_2 \geq 75\%$) as follows:

$$alt\ AF_1 < F_i \longrightarrow alt\ AF_2 < 75\%$$
$$alt\ AF_1 \geq F_i \longrightarrow alt\ AF_2 \geq 75\%$$

We classified every SNP as True Positive (TP), False Positive (FP), True Negative (TN) or False Negative (FN) according to:

$$TP\text{:}\ \ alt\ AF_1 \geq F_i\ \ \&\ \ alt\ AF_2 \geq 75\%$$
$$FP\text{:}\ \ alt\ AF_1 \geq F_i\ \ \&\ \ alt\ AF_2 < 75\%$$
$$TN\text{:}\ \ alt\ AF_1 < F_i\ \ \&\ \ alt\ AF_2 < 75\%$$
$$FN\text{:}\ \ alt\ AF_1 < F_i\ \ \&\ \ alt\ AF_2 \geq 75\%$$

True Positive Rates (TPR) and False Positive Rates (FPR) were calculated as:

$$TPR = \frac{\#TP}{\#TP + \#FN} \quad FPR = \frac{\#FP}{\#FP + \#TN}$$

Finally, we made predictions for all SNPs and calculated the TPR and FPR for all values of $F_i \in \{0\%, 1\%, 2\%, \cdots, 98\%, 99\%, 100\%\}$.

Mutation Density Test

The method to detect significant variation for a given locus amongst pairs of sequenced isolates has been described previously[41]. Briefly, let $\mathcal{N}_j \sim Pois(\lambda_j)$ be a random variable for the number of SNPs detected across all isolate pairs (for the in-host analysis this is the collection of longitudinal isolate pairs for all subjects) for gene j. Let (i) N_i = number of SNPs across all pairs for gene i, (ii) $|g_i|$ = length of gene i, (iii) P = number of genome pairs and (iv) G = the number of genes across the genome being analyzed (all genes in the essential, non-essential, antigen, antibiotic resistant and family protein categories).

Then the length of the genome (concatenate of all genes being analyzed) is given by $\sum_{i=1}^{G}|g_i|$ and the number of SNPs across all genes and genome pairs is given by $\sum_{i=1}^{G} N_i$. The null rate for $\mathcal{N}_j$ is given by the mean SNP distance between all pairs of isolates, weighted by the length of gene j as a fraction of the genome concatenate and number of isolate pairs:

$$\lambda_j = \left(\frac{\sum_{i=1}^{G} N_i}{P}\right)\left(\frac{|g_i|}{\sum_{i=1}^{G}|g_i|}\right)\left(\frac{1}{P}\right)$$

The p-value for gene j is then calculated as $\Pr\left(N_i > \mathcal{N}_j\right)$. We tested 3,386 genes for mutational density and applied Bonferroni correction to determine a significance threshold. We determine a gene to have a significant amount of variation if the assigned p-value $< \frac{0.05}{3{,}386} \approx 1.477 \times 10^{-5}$.

Nucleotide Diversity

We define the nucleotide diversity $\left(\pi_g\right)$ for a given gene g as follows: (i) let $\left|gene_g\right|$ = base-pair length of the gene, (ii) $N_{i,j}$ = number of in-host SNPs (independent of the change in allele frequency for each SNP) between the longitudinal isolates for subject i occurring on gene j and (iii) P = number of subjects. Then

$$\pi_g = \left(\frac{1}{P}\right)\left(\frac{1}{\left|gene_g\right|}\right)\sum_{i=1}^{P} N_{i,g}$$

Correspondingly, let G be a category consisting of M genes, then the average nucleotide diversity for G is given by:

$$\pi_G = \left(\frac{1}{M}\right)\left(\frac{1}{P}\right)\sum_{j=1}^{M}\left(\frac{1}{|gene_j|}\right)\left(\sum_{i=1}^{P} N_{i,j}\right)$$

SNP confirmation in repetitive genomic regions

Several of the SNPs detected belong to the GC-rich repetitive PE/PPE gene category[37]. Variants called on these genes are commonly excluded from comparative genomic analyses[8,10,16,21,25] due to the limitations of short-read sequencing data and the possibility of making spurious variant calls in these regions of the genome, however the rates at which these false calls occur has not been evaluated. We reasoned that our stringent filtering criteria, quality of sequencing data and depth of coverage allowed us to reliably detect variants in these regions of the genome.

SNP Calling Simulations: Certain repetitive regions of the *Mycobacterium tuberculosis* genome (ESX, PE/PPE loci) may give rise to false positive and false negative variant calls due to the mis-alignment of short-read sequencing data. To test the rate of false negative and false positive SNP calls in loci with in-host SNPs (**Fig. 5**) we collected the set of non-redundant SNPs observed in these loci (**Supplementary Tables 16, 19**). Next, we collected a set of publicly available reference genomes (**Supplementary Table 15**) and introduced these mutations into the respective loci positions in the reference genomes. We then simulated short-read Illumina sequencing data of comparable quality to our sequencing data from these altered reference genomes. Using our variant-calling pipeline to call polymorphisms, we then estimated the number of true and false positive SNP calls for each gene, based off of how many introduced SNPs were called (true positives), how many introduced SNPs were not called (false negatives) and how many spurious SNPs were called (false positives). A schematic of our simulation methodology is given in **Supplementary Fig. 5**, a detailed explanation is given in the **Supplementary Note** and the results of our simulations (given in **Supplementary Fig. 8)** confirm a low false-positive rate.

PacBio Assembly vs. Illumina Mapping SNP Calling: We compared SNP calling for the genes Rv0095c, *PPE18*, *PPE54* and *PPE60* between 12 isolates for which we had a complete PacBio assembly and Illumina sequencing data (**Supplementary Table 20**). Unlike Illumina generated reads, PacBio reads are much longer and have randomly distributed error profiles[81] which makes PacBio sequencing ideal for constructing microbial genomes and identifying variants in repetitive regions given high coverage. We used our variant calling procedures as outlined above to call SNPs from assemblies constructed from *de novo* assembly of PacBio reads (**A**) and from mapping Illumina reads to the H37Rv reference genome (**B**) for the four genes of interest (**Supplementary Table 21**). We then calculated the number of SNPs that were detected by both methods $|\mathbf{A} \cap \mathbf{B}|$, the number of SNPs detected only from mapping Illumina reads $|\mathbf{A} \setminus \mathbf{B}|$ and the number of SNPs detected only in the PacBio assemblies $|\mathbf{B} \setminus \mathbf{A}|$ (**Supplementary Fig. 12**). In these four genes, we observed that a large proportion of SNPs were detected by both sequencing methods ($|\mathbf{A} \cap \mathbf{B}|$), and that the number of SNPs falsely detected by Illumina ($|\mathbf{A} \setminus \mathbf{B}|$) was zero or extremely low across all samples.

We found that 17/178 in-host SNPs and 31/68 phylogenetically convergent SNPs were present in at least 1/12 of our PacBio *de novo* assembled genomes (**Supplementary Table 22**), including SNPs within repetitive genes Rv0095c, *PPE18*, *PPE54* and *PPE60*. We evaluated the capacity of Illumina short-read sequencing technology to detect our in-host SNPs of interest in repetitive genes. For each SNP we measured: (1) the number of times our Illumina SNP calling pipeline correctly identified a SNP when it was present ($|\mathbf{A} \cap \mathbf{B}|$), and (2) the number of times Illumina falsely called a SNP ($|\mathbf{A} \backslash \mathbf{B}|$). All five of our detected in-host SNPs present in *PPE18*, *PPE54* and *PPE60* were always called correctly by Illumina sequencing ($|\mathbf{A} \cap \mathbf{B}|$). Furthermore, no in-host SNPs nor any phylogenetically convergent SNPs were spuriously called via Illumina sequencing and mapping ($|\mathbf{A} \backslash \mathbf{B}|$). The only in-host or phylogenetically convergent SNPs displaying any inconsistent Illumina variant calling were in the Rv0095c gene as some SNPs were called from PacBio sequencing data but not Illumina data. Overall, we detect the presence of many in-host and phylogenetically convergent SNPs in Mtb clinical isolates demonstrating that these SNP calls (from Illumina reads) are unlikely to have resulted from erroneous variant calling.

Global Lineage Typing

We determined the global lineage of each longitudinal ($N = 614$) and public isolate ($N = 10{,}018$) using base calls from Pilon-generated VCF files and a subset of 413 previously established lineage-defining diagnostic SNPs[65].

Phylogenetic Convergence Analysis

We selected a set of genes to test for phylogenetic convergence based on the following criteria: (i) in-host SNPs were detected within the gene across multiple hosts (in-host convergence at the gene level), (ii) the gene was classified as mutationally dense (**Supplementary Table 11**), (iii) the gene belonged to a pathway in which in-host SNPs were detected across multiple hosts (**Supplementary Table 14**) and at least one in-host SNP was detected within the gene (in-host convergence at the pathway level). Twenty-two genes fit at least one of these criteria

(**Supplementary Table 17**). We then scanned 10,018 genetically diverse isolates for SNPs within these genes according to our SNP calling methodology above (**Supplementary Table 18**). To determine phylogenetic convergence for a given SNP site, we required the detection of the alternate allele in at least 10 isolates for at least two global lineages. Sixty-eight SNP sites across six genes were detected as having a signal of phylogenetic convergence (**Supplementary Table 19**). A single SNP site, in which the alternate allele was present in 9,775/10,108 isolates, reflected a rare allele in the reference genome and was dropped from further analysis yielding a set of 67 phylogenetically convergent SNP sites detected across five genes (**Fig. 5**).

Data Analysis and Variant Annotation

Data analysis was performed using custom scripts run in Python and interfaced with iPython[82]. Statistical tests were run with Statsmodels[83] and figures were plotted using Matplotlib[84]. Numpy[85], Biopython[86] and Pandas[87] were all used extensively in data cleaning and manipulation. Functional annotation of SNPs was done in Biopython[86] using the H37Rv reference genome and the corresponding genome annotation. For every SNP called, we used the H37Rv reference position provided by the Pilon[75] generated VCF file to extract any overlapping CDS region and annotated SNPs accordingly. Each overlapping CDS regions was then translated into its corresponding peptide sequence with both the reference and alternate allele. SNPs in which the peptide sequences did not differ between alleles were labeled *synonymous*, SNPs in which the peptide sequences did differ were labeled *non-synonymous* and if there were no overlapping CDS regions for that reference position, then the SNP was labeled *intergenic*.

Pathway Definitions

We used SEED[88] subsystem annotation to conduct pathway analysis and downloaded the subsystem classification for all features of *Mycobacterium tuberculosis* H37Rv (id: 83332.1) (**Supplementary Table 12**). We mapped all of the annotated features from SEED to the annotation for H37Rv. Due to the slight inconsistency between the start and end chromosomal coordinates for features from SEED and our H37Rv annotation, we assigned a locus from H37Rv to a subsystem if both the start and end coordinates for this locus fell within a 20 base-pair window of the start and end coordinates for a feature in the SEED annotation (**Supplementary Table 13**).

ACKNOWLEDGEMENTS

We thank the members of the Farhat lab for helpful discussions and comments on the research project and manuscript. We thank S. Fortune, N. Hicks & D. Warner for helpful suggestions on the manuscript. R.V.J. was supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE1745303. Portions of this research were conducted on the O2 High Performance Compute Cluster, supported by the Research Computing Group, at Harvard Medical School.

AUTHOR CONTRIBUTIONS

R.V.J. and M.R.F. conceived, designed and conducted the study. R.V.J. and M.R.F. drafted the manuscript with input from all authors. L.F. and M.M. provided bioinformatics support. L.E.E., D.D., M. Salfinger and M. Strong cultured Mtb isolates and performed DNA extraction in preparation for PacBio sequencing. M.Smith and I.O. prepared libraries and performed PacBio sequencing runs.

COMPETING INTERESTS

The authors declare no competing interests.

REFERENCES

1. WHO | Global tuberculosis report 2017. Available at: http://www.who.int/tb/publications/global_report/en/. (Accessed: 1st August 2018)
2. Pai, M. *et al.* Tuberculosis. *Nat. Rev. Dis. Primer* **2**, 16076 (2016).
3. Ernst, J. D. Mechanisms of M. tuberculosis Immune Evasion as Challenges to TB Vaccine Design. *Cell Host Microbe* **24**, 34–42 (2018).
4. Gagneux, S. Ecology and evolution of *Mycobacterium tuberculosis*. *Nat. Rev. Microbiol.* **16**, 202–213 (2018).
5. Lin, P. L. *et al.* Sterilization of granulomas is common in active and latent tuberculosis despite within-host variability in bacterial killing. *Nat. Med.* **20**, 75–79 (2014).
6. Sun, G. *et al.* Dynamic population changes in Mycobacterium tuberculosis during acquisition and fixation of drug resistance in patients. *J. Infect. Dis.* **206**, 1724–1733 (2012).
7. Farhat, M. R. *et al.* Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. *Nat. Genet.* **45**, 1183 (2013).
8. Coscolla, M. *et al.* M. tuberculosis T cell epitope analysis reveals paucity of antigenic variation and identifies rare variable TB antigens. *Cell Host Microbe* **18**, 538–548 (2015).
9. Brites, D. & Gagneux, S. Co-evolution of M ycobacterium tuberculosis and H omo sapiens. *Immunol. Rev.* **264**, 6–24 (2015).
10. Comas, I. *et al.* Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. *Nat. Genet.* **42**, 498–498 (2010).

11. Zhang, D. *et al.* Genomic analysis of the evolution of fluoroquinolone resistance in Mycobacterium tuberculosis prior to tuberculosis diagnosis. *Antimicrob. Agents Chemother.* **60**, 6600–6608 (2016).
12. Didelot, X., Walker, A. S., Peto, T. E., Crook, D. W. & Wilson, D. J. Within-host evolution of bacterial pathogens. *Nat. Rev. Microbiol.* **14**, 150 (2016).
13. Colangeli, R. *et al.* Bacterial factors that predict relapse after tuberculosis therapy. *N. Engl. J. Med.* **379**, 823–833 (2018).
14. Ernst, J. D. Mechanisms of M. tuberculosis Immune Evasion as Challenges to TB Vaccine Design. *Cell Host Microbe* **24**, 34–42 (2018).
15. Ford, C. *et al.* Mycobacterium tuberculosis--heterogeneity revealed through whole genome sequencing. *Tuberculosis* **92**, 194–201 (2012).
16. Lieberman, T. D. *et al.* Genomic diversity in autopsy samples reveals within-host dissemination of HIV-associated Mycobacterium tuberculosis. *Nat. Med.* **22**, 1470 (2016).
17. Guerra-Assunção, J. A. *et al.* Recurrence due to relapse or reinfection with Mycobacterium tuberculosis: a whole-genome sequencing approach in a large, population-based cohort with a high HIV infection prevalence and active follow-up. *J. Infect. Dis.* **211**, 1154–1163 (2014).
18. Marvig, R. L., Sommer, L. M., Molin, S. & Johansen, H. K. Convergent evolution and adaptation of Pseudomonas aeruginosa within patients with cystic fibrosis. *Nat. Genet.* **47**, 57 (2015).
19. Lieberman, T. D. *et al.* Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nat. Genet.* **43**, 1275 (2011).

20. Lieberman, T. D. *et al.* Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. *Nat. Genet.* **46**, 82 (2014).
21. Copin, R. *et al.* Within host evolution selects for a dominant genotype of Mycobacterium tuberculosis while T cells increase pathogen genetic diversity. *PLoS Pathog.* **12**, e1006111 (2016).
22. Trauner, A. *et al.* The within-host population dynamics of Mycobacterium tuberculosis vary with treatment efficacy. *Genome Biol.* **18**, 71 (2017).
23. Goig, G. A., Blanco, S., Garcia-Basteiro, A. & Comas, I. Pervasive contaminations in sequencing experiments are a major source of false genetic variability: a Mycobacterium tuberculosis meta-analysis. *bioRxiv* (2018). doi:10.1101/403824
24. Wyllie, D. H. *et al.* Identifying Mixed Mycobacterium tuberculosis Infection and Laboratory Cross-Contamination during Mycobacterial Sequencing Programs. *J. Clin. Microbiol.* **56**, (2018).
25. Casali, N. *et al.* Whole genome sequence analysis of a large isoniazid-resistant tuberculosis outbreak in London: a retrospective observational study. *PLoS Med.* **13**, e1002137 (2016).
26. Walker, T. M. *et al.* Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *Lancet Infect. Dis.* **13**, 137–146 (2013).
27. Bryant, J. M. *et al.* Whole-genome sequencing to establish relapse or re-infection with Mycobacterium tuberculosis: a retrospective observational study. *Lancet Respir. Med.* **1**, 786–792 (2013).
28. Witney, A. A. *et al.* Use of whole-genome sequencing to distinguish relapse from reinfection in a completed tuberculosis clinical trial. *BMC Med.* **15**, 71 (2017).

29. Xu, Y. *et al.* In vivo evolution of drug-resistant Mycobacterium tuberculosis in patients during long-term treatment. *BMC Genomics* **19**, 640 (2018).

30. Farhat, M. R. *et al.* Genetic determinants of drug resistance in Mycobacterium tuberculosis and their diagnostic value. *Am. J. Respir. Crit. Care Med.* **194**, 621–630 (2016).

31. WHO. *Definitions and reporting framework for tuberculosis*.

32. Pethe, K. *et al.* A chemical genetic screen in Mycobacterium tuberculosis identifies carbon-source-dependent growth inhibitors devoid of in vivo efficacy. *Nat. Commun.* **1**, 57 (2010).

33. Namouchi, A., Didelot, X., Schöck, U., Gicquel, B. & Rocha, E. P. C. After the bottleneck: Genome-wide diversification of the Mycobacterium tuberculosis complex by mutation, recombination, and natural selection. *Genome Res.* (2012).

34. Phelan, J. E. *et al.* Recombination in pe/ppe genes contributes to genetic variation in Mycobacterium tuberculosis lineages. *BMC Genomics* **17**, 151 (2016).

35. Ford, C. B. *et al.* Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. *Nat. Genet.* **43**, 482–482 (2011).

36. Dillon, M. M., Sung, W., Lynch, M. & Cooper, V. S. The rate and molecular spectrum of spontaneous mutations in the GC-rich multichromosome genome of Burkholderia cenocepacia. *Genetics* **200**, 935–946 (2015).

37. Brennan, M. J. & Delogu, G. The PE multigene family: a 'molecular mantra'for mycobacteria. *Trends Microbiol.* **10**, 246–249 (2002).

38. Sassetti, C. M., Boyd, D. H. & Rubin, E. J. Genes required for mycobacterial growth defined by high density mutagenesis. *Mol. Microbiol.* **48**, 77–84 (2003).

39. Sassetti, C. M. & Rubin, E. J. Genetic requirements for mycobacterial survival during infection. *Proc. Natl. Acad. Sci.* **100**, 12989–12994 (2003).

40. Brennan, M. J. The enigmatic PE/PPE multigene family of mycobacteria and tuberculosis vaccination. *Infect. Immun.* **85**, e00969-16 (2017).

41. Farhat, M. R., Shapiro, B. J., Sheppard, S. K., Colijn, C. & Murray, M. A phylogeny-based sampling strategy and power calculator informs genome-wide associations study design for microbial pathogens. *Genome Med.* **6**, 101 (2014).

42. Holt, K. E. *et al.* High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi. *Nat. Genet.* **40**, 987 (2008).

43. Hicks, N. D. *et al.* Clinically prevalent mutations in Mycobacterium tuberculosis alter propionate metabolism and mediate multidrug tolerance. *Nat. Microbiol.* **3**, 1032 (2018).

44. Comas, I. *et al.* Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. *Nat. Genet.* **44**, 106 (2012).

45. Llewelyn, M. J. *et al.* The antibiotic course has had its day. *BMJ* j3418 (2017). doi:10.1136/bmj.j3418

46. Votintseva, A. A. *et al.* Same-day diagnostic and surveillance data for tuberculosis via whole-genome sequencing of direct respiratory samples. *J. Clin. Microbiol.* **55**, 1285–1298 (2017).

47. Copin, R. *et al.* Sequence diversity in the pe_pgrs genes of Mycobacterium tuberculosis is independent of human T cell recognition. *MBio* **5**, e00960—-13 (2014).

48. Rowley, C. A. & Kendall, M. M. To B12 or not to B12: Five questions on the role of cobalamin in host-microbial interactions. *PLoS Pathog.* **15**, e1007479 (2019).

49. Minias, A., Minias, P., Czubat, B. & Dziadek, J. Purifying selective pressure suggests the functionality of a vitamin B12 biosynthesis pathway in a global population of Mycobacterium tuberculosis. *Genome Biol. Evol.* **10**, 2326–2337 (2018).

50. Gopinath, K., Moosa, A., Mizrahi, V. & Warner, D. F. Vitamin B12 metabolism in Mycobacterium tuberculosis. *Future Microbiol.* **8**, 1405–1418 (2013).

51. Salaemae, W., Azhar, A., Booker, G. W. & Polyak, S. W. Biotin biosynthesis in Mycobacterium tuberculosis: physiology, biochemistry and molecular intervention. *Protein Cell* **2**, 691–695 (2011).

52. Pathak, S. K. *et al.* Direct extracellular interaction between the early secreted antigen ESAT-6 of Mycobacterium tuberculosis and TLR2 inhibits TLR signaling in macrophages. *Nat. Immunol.* **8**, 610 (2007).

53. Dixit, A. *et al.* Whole genome sequencing identifies bacterial factors affecting transmission of multidrug-resistant tuberculosis in a high-prevalence setting. *Sci. Rep.* **9**, 5602 (2019).

54. Nair, S. *et al.* The PPE18 of Mycobacterium tuberculosis interacts with TLR2 and activates IL-10 induction in macrophage. *J. Immunol.* jimmunol--0901367 (2009).

55. Su, H. *et al.* Mycobacterium tuberculosis PPE60 antigen drives Th1/Th17 responses via Toll-like receptor 2--dependent maturation of dendritic cells. *J. Biol. Chem.* jbc--RA118 (2018).

56. Hebert, A. M. *et al.* DNA polymorphisms in the pepA and PPE18 genes among clinical strains of Mycobacterium tuberculosis: implications for vaccine efficacy. *Infect. Immun.* **75**, 5798–5805 (2007).

57. Brodin, P. *et al.* High content phenotypic cell-based visual screen identifies Mycobacterium tuberculosis acyltrehalose-containing glycolipids involved in phagosome remodeling. *PLoS Pathog.* **6**, e1001100 (2010).
58. Podinovskaia, M., Lee, W., Caldwell, S. & Russell, D. G. Infection of macrophages with M ycobacterium tuberculosis induces global modifications to phagosomal function. *Cell. Microbiol.* **15**, 843–859 (2013).
59. Kleinnijenhuis, J., Oosting, M., Joosten, L. A. B., Netea, M. G. & Van Crevel, R. Innate immune recognition of Mycobacterium tuberculosis. *Clin. Dev. Immunol.* **2011**, (2011).
60. Tientcheu, L. D. *et al.* Immunological consequences of strain variation within the Mycobacterium tuberculosis complex. *Eur. J. Immunol.* **47**, 432–445 (2017).
61. Azad, A. K., Sadee, W. & Schlesinger, L. S. Innate immune gene polymorphisms in tuberculosis. *Infect. Immun.* IAI--00443 (2012).
62. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* **37**, D26–D31 (2008).
63. Farhat, M. R. *et al.* GWAS for quantitative resistance phenotypes in Mycobacterium tuberculosis reveals resistance genes and regulatory regions. *Nat. Commun.* **10**, 2128 (2019).
64. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
65. Coll, F. *et al.* A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nat. Commun.* **5**, 4812 (2014).

66. Van Embden, J. *et al.* Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology. *J. Clin. Microbiol.* **31**, 406–409 (1993).

67. Doig, C., Seagar, A., Watt, B. & Forbes, K. The efficacy of the heat killing of Mycobacterium tuberculosis. *J. Clin. Pathol.* **55**, 778–779 (2002).

68. Käser, M., Ruf, M.-T., Hauser, J. & Pluschke, G. Optimized DNA preparation from mycobacteria. *Cold Spring Harb. Protoc.* **2010**, pdb-prot5408 (2010).

69. Vita, R. *et al.* The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* **43**, D405--D412 (2014).

70. Cole, St. *et al.* Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* **393**, 537 (1998).

71. Stucki, D. *et al.* Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. *Nat. Genet.* **48**, 1535 (2016).

72. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).

73. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows--Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

74. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

75. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One* **9**, e112963 (2014).

76. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).

77. Hunt, M. *et al.* Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol.* **16**, 294 (2015).

78. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

79. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563 (2013).

80. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

81. Rhoads, A. & Au, K. F. PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics* **13**, 278–289 (2015).

82. Pérez, F. & Granger, B. E. IPython: a system for interactive scientific computing. *Comput. Sci. Eng.* **9**, (2007).

83. Seabold, S. & Perktold, J. Statsmodels: Econometric and statistical modeling with python. in *Proceedings of the 9th Python in Science Conference* **57**, 61 (2010).

84. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

85. Van Der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng.* **13**, 22 (2011).

86. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

87. McKinney, W. & others. Data structures for statistical computing in python. in *Proceedings of the 9th Python in Science Conference* **445**, 51–56 (2010).

88. Overbeek, R. *et al.* The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* **42**, D206–D214 (2013).

FIGURE LEGENDS

Figure 1 – **Selection of patients with longitudinal clonal infection** (**a**) Allele frequency change between paired isolates $(\Delta AF) = |AF_1^A - AF_2^A| = |AF_1^B - AF_2^B|$. (**b**) The F2 measure >0.04 (**Methods**) was used to identify and exclude isolate pairs with evidence for mixed strain growth at any time point. (**c**) Replicate and longitudinal pairs with fixed SNP (fSNP) distance of >7 were excluded. For longitudinal isolates fSNP>7 was assessed as consistent with Mtb reinfection with a different strain.

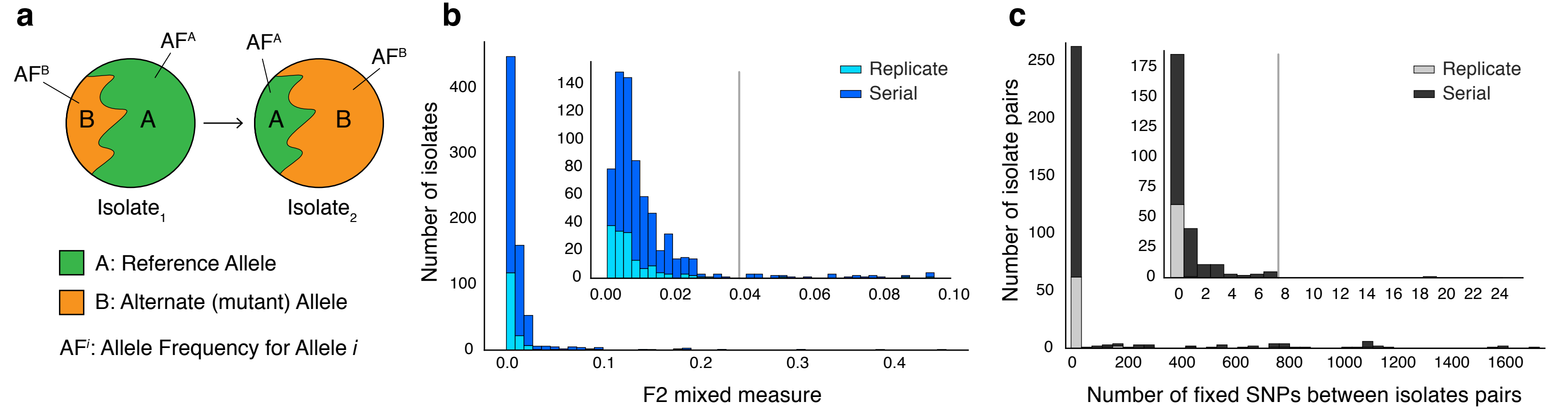
Figure 2 – **Allele frequency dynamics within antibiotic resistance loci.** (**a**) The antibiotic resistance genes *embB*, *katG*, and *gyrA* demonstrate evidence for competing clones during infection (patterns for other genes in **Supplementary Figure 4**). (**b**) Plot of true positive rate (TPR) and false positive rate (FPR) for detecting eventual fixation of a resistance allele as a function of initial allele frequency (AF_1>5%).

Figure 3 – **Genome-wide diversity in 200 clonal Mtb infections.** (**a**) Distribution of five major Mtb lineages among the 200 clonal Mtb infections. (**b**) Distribution of 178 in-host SNPs among the 200 longitudinal isolate pairs across the 4.41 Mbp Mtb genome (blue circles: synonymous, red circles: non-synonymous). Blue and red circles on the innermost black ring indicate the locations of SNPs detected in one patient; circles on the next ring represent SNPs detected in two patients. The $-\log_{10}$(p-value) of the mutational density test (**Methods**) by gene is plotted in the outermost, red and green, regions. Labeled yellow circles represent genes significant at the bonferroni-corrected cutoff ($\alpha = 0.05/3{,}886$). (**c**) Distribution of ΔAF by SNP type: sSNP: synonymous, nSNP: non-synonymous, iSNP: intergenic. (**c**) Heat-map of SNPs per gene (rows)

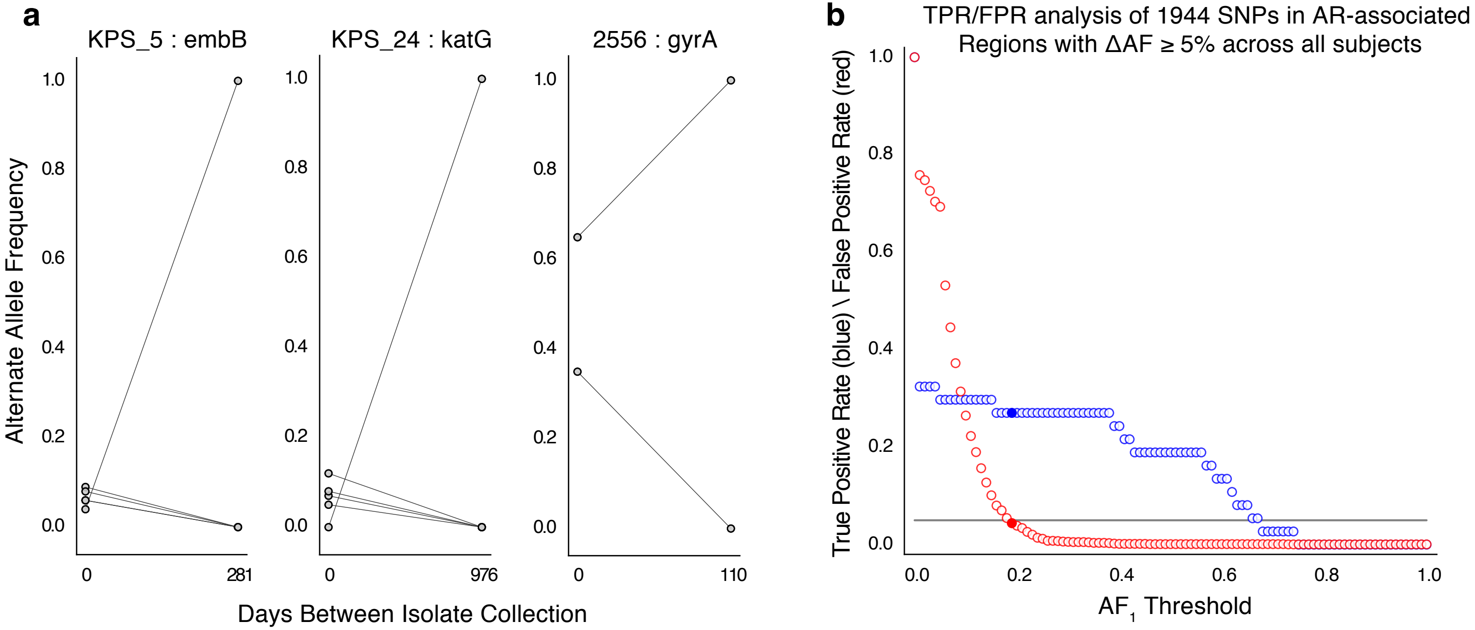
and patient (columns). Colored circles across columns indicate the strain phylogenetic lineage (as represented in (**a**)). Gene names colored according to gene category (**Fig 4d**) with parentheses indicating the number of subjects with a SNP in a given gene. *Indicates genes in which SNPs are detected within multiple hosts.

Figure 4 - **PE/PPE genes vary considerably within host while putative antigens remain conserved.** (**a**) Mutational spectrum of in-host SNPs. (**b**) In-host SNP counts *vs.* time between isolate collection (195/200 subjects with dates shown, *W[26] isolates only had year of collection). (**c**) boxplots of nucleotide diversity by gene within each of 5 non redundant categories (see text). (n = number of genes). (**d**) Average nucleotide diversity across genes by category. Nucleotide diversity in epitope and non-epitope region (**Methods**) of each gene in the Antigen (**e**,**f**) and PE/PPE (**g**,**h**) gene categories. (**i**,**j**) PE/PPE genes separated into three non-redundant categories: PE, PE-PGRS, and PPE. (**i**) The average nucleotide diversity by category. (**j**) box plot of nucleotide diversity by gene.

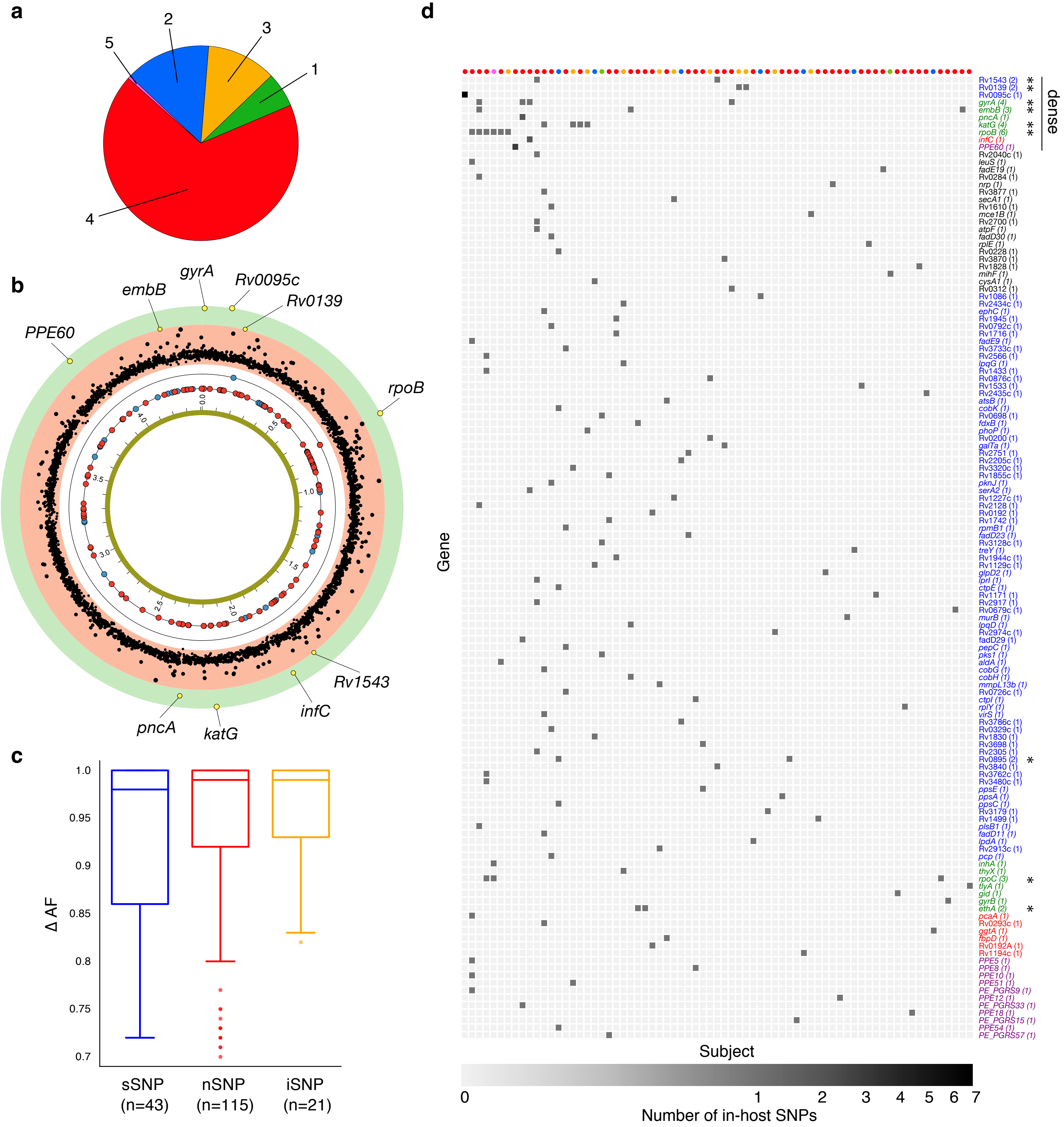
Figure 5 – **Alleles acquired in-host show evidence of phylogenetic convergence across 10,018 clinical Mtb isolates.** (**a**) Global lineage distribution among 10,018 clinical Mtb isolates. (**b**) Sixty-seven SNP sites detected within five genes displayed evidence of phylogenetic convergence (**Methods**). The alternate allele for each SNP was detected in at least 10 isolates within at least two global lineages. The number of isolates with each alternate allele, broken down by global lineage, is displayed. Each mutation is labeled with the reference allele, H37Rv coordinate, and alternate allele (blue:synonymous, red:non-synonymous). The gene name or H37Rv locus tag each mutation occurs within is indicated at the bottom.



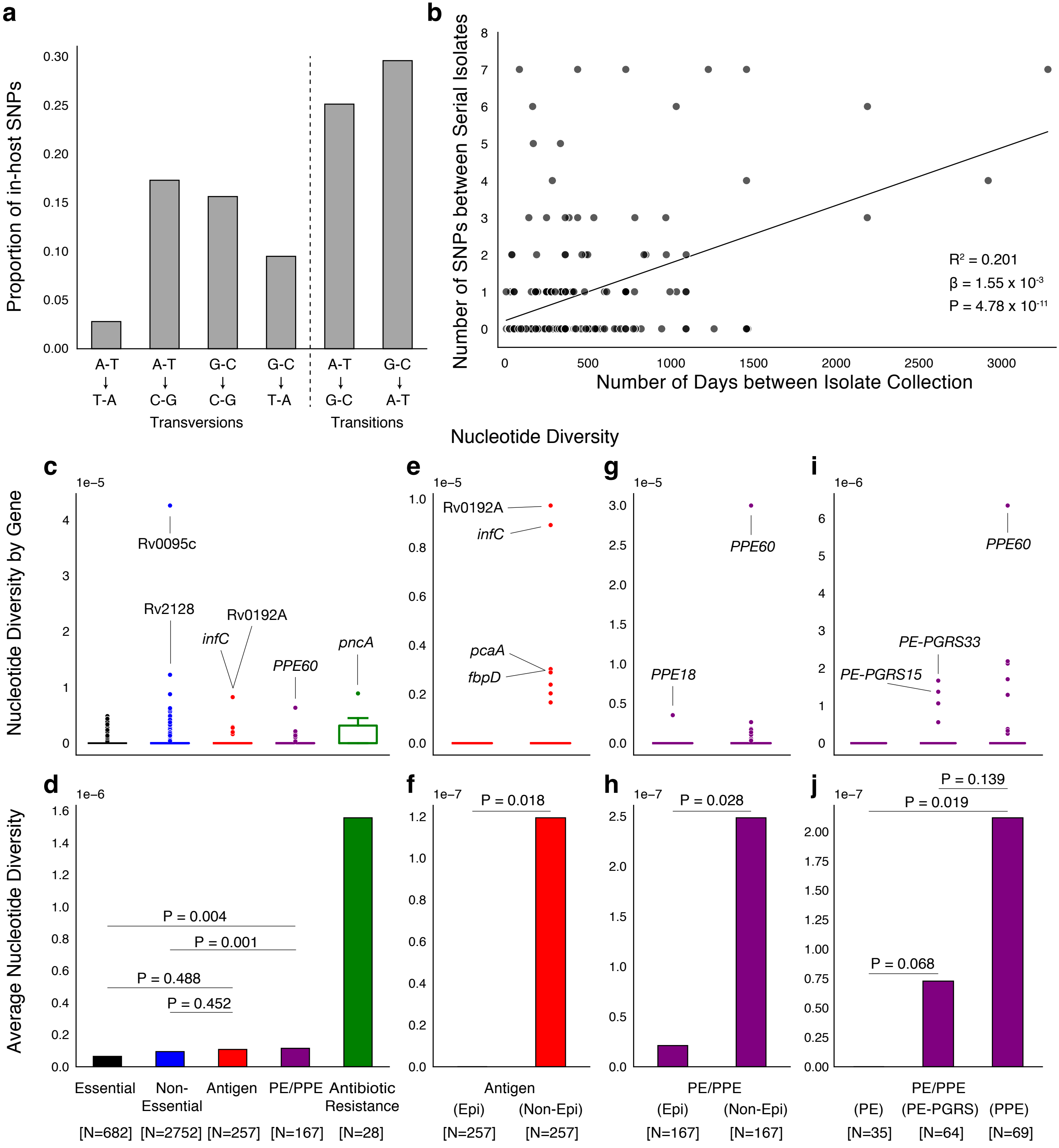
Main Figure 1



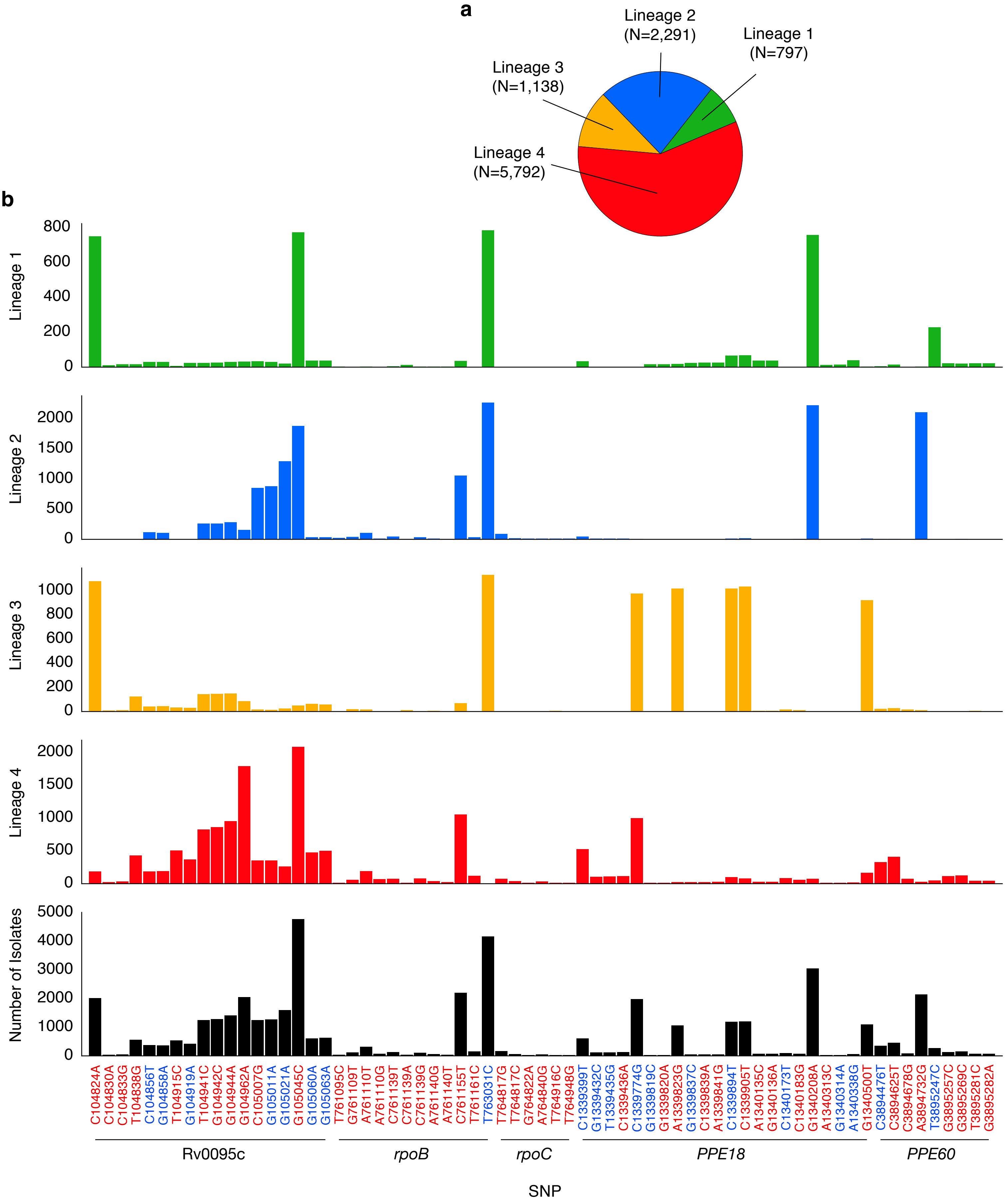
Main Figure 2



Main Figure 3



Main Figure 4



Main Figure 5

SUPPLEMENTARY INFORMATION

In-host population dynamics of M. tuberculosis during treatment failure

Roger Vargas Jr[1,2*], Luca Freschi[2], Maximillian Marin[1,2], L. Elaine Epperson[3], Melissa Smith[4,5], Irina Oussenko[5], David Durbin[6], Michael Strong[3], Max Salfinger[7], and Maha Reda Farhat[2,8*]

[1] Department of Systems Biology, Harvard Medical School

[2] Department of Biomedical Informatics, Harvard Medical School

[3] Center for Genes, Environment, and Health, National Jewish Health

[4] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai

[5] Icahn Institute of Data Sciences and Genomics Technology

[6] Mycobacteriology Reference Laboratory, Advanced Diagnostic Laboratories, National Jewish Health

[7] College of Public Health, University of South Florida

[8] Pulmonary and Critical Care Medicine, Massachusetts General Hospital

[*]Corresponding authors: roger_vargas@g.harvard.edu, Maha_Farhat@hms.harvard.edu

TABLE OF CONTENTS

- Supplementary Note
    - Reference Genome Collection
    - Mapping CDS regions from Reference Genomes to H37Rv
    - Filtering Low-Quality Mapped Reference Genomes
    - Altering RefGenomes at SNP Test Sites
    - Simulating Reads from Complete Genomes
    - Mapping Simulated Reads to H37Rv and Calling SNPs
    - Calling SNPs with MUMmer
    - True & False Positive SNP Call Analysis
- Supplementary Figures 1-12
- Supplementary Table Descriptions 1-22
- References for Supplementary Information

SUPPLEMENTARY NOTE

Reference Genome Collection

We downloaded 60 reference genomes (RefGenome) (i.e. completely assembled *Mycobacterium tuberculosis* genomes) from NCBI (Genbank accession IDs can be found in **Supplementary Table 15**). We limited our collection to genomes for which there were corresponding annotation files.

Mapping CDS regions from Reference Genomes to H37Rv

Since the regions of interest were repetitive loci that have many homologies elsewhere in the genome, we were unable to use traditional alignment methods to map the genes of interest from H37Rv to the other RefGenomes. Instead, we made use of the clonal structure of the Mtb genome to construct gene mappings from H37Rv to the RefGenomes as follows (**Supplementary Figure 7a**):

1. For each gene g annotated in H37Rv, collect the set of gene lengths 5 genes upstream and 5 genes downstream of g from H37Rv. Compare the set of 11 H37Rv gene lengths to every set of 11 consecutive gene neighborhoods on the RefGenome and assign a score based off of the intersection of each pair of sets.
2. Look at the gene neighborhood(s) with the top score after scanning the RefGenome and pairwise globally align[1] g to every gene in the top scoring neighborhood using the following criteria: (i) identical characters are given 2 points, (ii) 1 point is deducted for each non-identical character, (iii) 2 points are deducted for opening a gap, (iv) 2 points are deducted for extending a gap.

3. Take the top scoring alignment r and assign a mapping from H37Rv gene g to RefGenome gene r if (i) the pairwise alignment score is > 0 and (ii) the base pair length of g and r are equivalent (the latter ensures correct placement of mutations in downstream analysis). If either of these criteria is not met, then we do not assign a mapping from g to any CDS region on that RefGenome.

Filtering Low-Quality Mapped Reference Genomes

To assess the quality of the mappings from H37Rv to the set of RefGenomes, we compared the reference position start coordinates of each assigned mapping between each RefGenome and H37Rv. Again making use of Mtb clonality, we reasoned that the genomic structure of each pair of genomes is similar (if each RefGenome is indexed to start at the first gene on H37Rv *Rv0001*, then well mapped RefGenomes will have mapped genes that are located within a neighborhood of the coordinates from H37Rv). To test this (for each RefGenome), we took the absolute difference between the start coordinates for all of the mapped genes between the RefGenome and H37Rv. We then averaged these differences across all gene mappings between both genomes. This measures the conservation (of the ordering) of the mapped genes between each pair of genomes (H37Rv & RefGenome) and gives an indication of how successful the mappings were on a global scale. We downloaded and mapped genes for 60 Genome Assemblies from GenBank[2] and assessed the quality of each set of mappings using the measure described above (**Supplementary Fig. 7b-c**). We excluded 6 RefGenomes on the basis of sporadic gene mappings against H37Rv which was determined by looking at the distribution of the mapping measure for all 60 assemblies. We kept the remaining 54 genomes for use in the simulations.

Altering RefGenomes at SNP Test Sites

We make use of the set of the (non-redundant) observed in-host SNPs across all genes (**Fig. 3d, Supplementary Table 16**) and set of phylogenetically convergent SNPs (**Fig. 5b, Supplementary Table 19**). We alter each RefGenome by introducing mutations (that correspond to the aforementioned SNPs) into the genes successfully mapped to H37Rv, ensuring that the new bases differ from the corresponding base positions on H37Rv. Since successful mappings require that the mapped genes be the same length, the mutations are introduced into the same site on the RefGenome with respect to the gene specific coordinates (i.e. a gene n bp long will have coordinates $\{1, 2, \cdots, n-1, n\}$ from $5' \rightarrow 3'$). We store information pertaining to which bases were altered for each RefGenome $\{SNP\ set\ \boldsymbol{\beta}\}$. No simulations are run for genes on RefGenomes that are not successfully mapped to H37Rv.

Simulating Reads from Complete Genomes

To validate our SNP calling methodology using the set of RefGenomes, we used ART[3] to simulate short-read sequencing data altered versions of the RefGenomes (**Supplementary Fig. 7b**). Since the aim of our simulations was to study the quality of our variant calls on our real data, we simulated data for each (altered) RefGenome that was of comparable quality to our real sequencing data: Illumina HiSeq 1000, read length of 100bp, mean coverage of 80x, paired end reads, 200bp mean size of DNA fragments, 25bp standard deviation of DNA fragment size (settings: -ss HS10 -l 100 -f 80 -p -m 200 -s 25).

Mapping Simulated Reads to H37Rv and Calling SNPs

Next we mapped the pool of simulated reads from the altered RefGenomes against the H37Rv reference genome and called SNPs according to most of the same procedures and WGS filters outlined in **Methods**. However, in this instance we called SNPs at reference positions that supported an alternate allele and required that calls were flagged as *Pass* by Pilon (where the alternate allele frequency was ≥ 75% and no *Ambiguous*, *Low Coverage*, or *Deletion* flags were present at that position). For each RefGenome, this yielded the set of SNPs (between the altered RefGenome and H37Rv) called by our pipeline $\{SNP\ set\ \mathbf{B}\}$ (**Supplementary Fig. 7b**).

Calling SNPs with MUMmer

We used Mummer3[4] to call SNPs between H37Rv and each (unaltered) RefGenome. We aligned each pair of genomes and called SNPs between the alignments using the following commands:

1) nucmer -mum H37Rv.fasta RefGenome.fasta
2) delta-filter -r -q H37Rv_RefGenome.delta > H37Rv_RefGenome.filter
3) show-snps -Clr -T H37Rv_RefGenome.filter > H37Rv_RefGenome.snps

The resulting SNP calls yielded the set of SNPs between each of the unmodified (unaltered) RefGenomes and H37Rv $\{SNP\ set\ \mathbf{A}\}$ (**Supplementary Fig. 7b**).

True & False Positive SNP Call Analysis

To calculate the number of *true positives* and *false positives* with regard to our SNP calling pipeline for each gene g of interest (**Supplementary Fig. 8**), we define the following sets of H37Rv coordinates for each RefGenome:

- $\boldsymbol{\beta}$ - SNPs introduced into (altered) RefGenome
- $\boldsymbol{A}$ - SNPs called between (unaltered) RefGenome & H37Rv
- $\boldsymbol{B}$ - SNPs called between (altered) RefGenome & H37Rv

- $\boldsymbol{C}$ - all reference positions (or coordinates) on H37Rv

The set of coordinates where an alternate allele was introduced into the RefGenome and called by the pipeline (true positive SNPs for gene g) is given by:

$$TP_g = \left(B_g \setminus A_g\right) \cap \left(\beta_g \setminus A_g\right)$$

where we normalize by SNP set A_g to make sure we're only accounting for test SNPs in our computations. The set of coordinates where an alternate allele was note introduced and called by the pipeline (false positive SNPs for gene g) is given by:

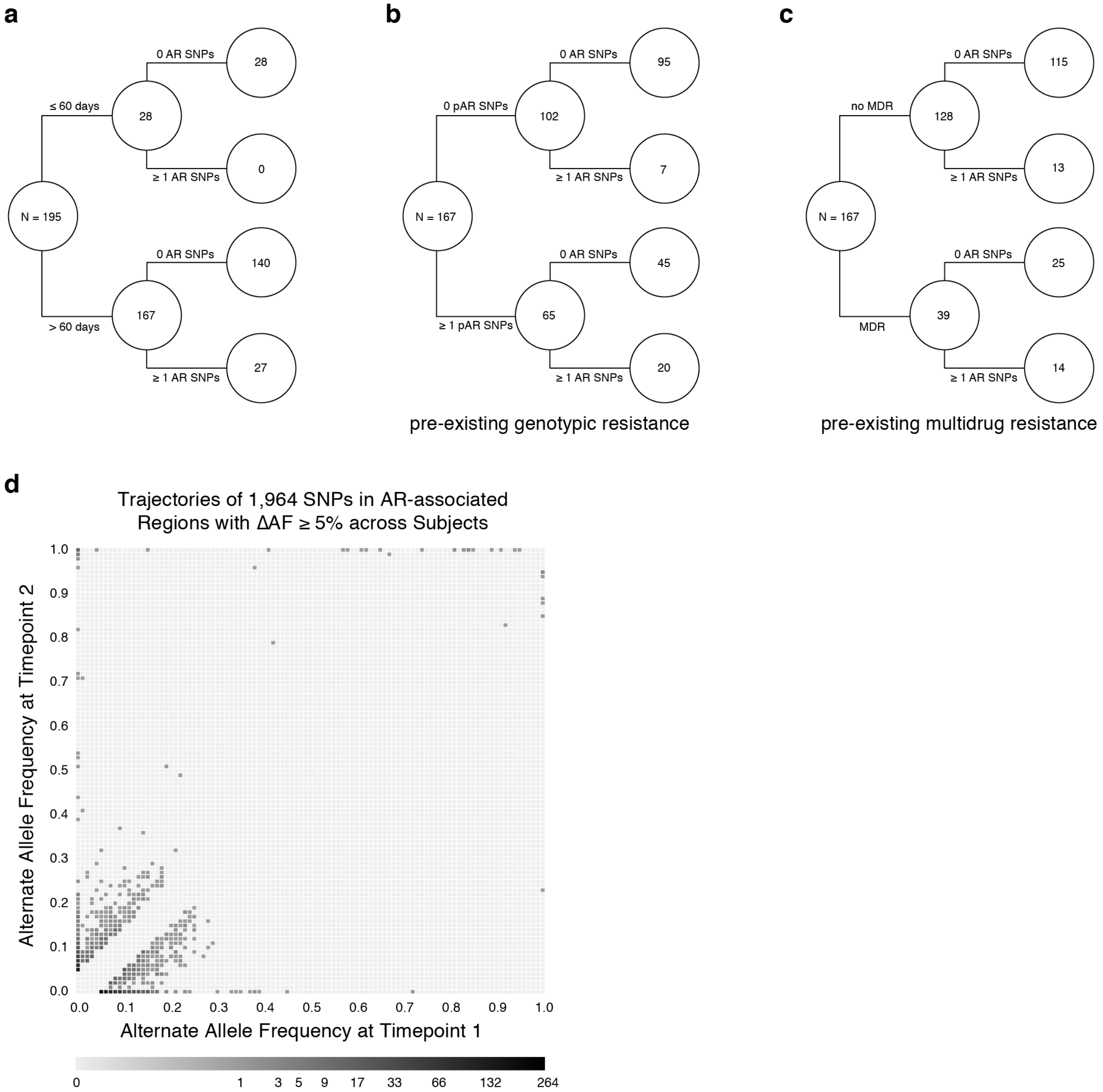
$$FP_g = \left(\left(B_g \setminus A_g\right) \cap C_g\right) \setminus TP_g$$

The set of coordinates where an alternate allele was introduced but was not called by the pipeline (false negative SNPs for gene g) is given by:

$$FN_g = \left(\beta_g \setminus A_g\right) \setminus TP_g$$

The results of our simulations (**Supplementary Fig. 8**) indicate that the number of true positive calls is consistent with the number of known SNPs across all genes and simulations. Perhaps more importantly, our results also suggest that false positive calls are rarely made for any SNP in our sample. Thus, while we may not have called all of the existing variation between paired isolates (false negative calls), it is unlikely that we called non-existing variation between any pair of isolates (false positives). That is, false-positive SNPs are rarely called, even in repetitive loci such as the PE/PPE gene family, supporting our decision to keep all SNP calls for downstream analysis.

139 SUPPLEMENTARY FIGURES



140 **Supplementary Figure 1**

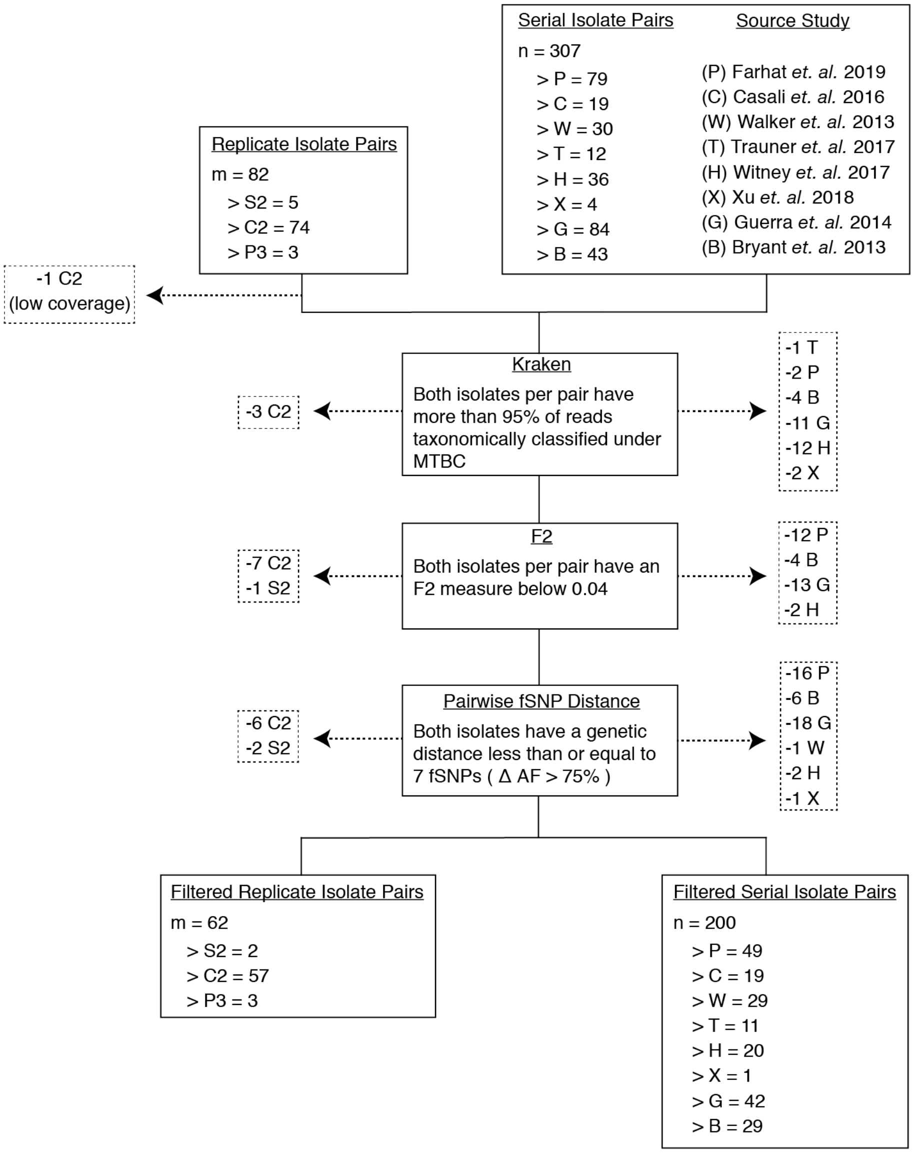
141 Supplementary Figure 1 - **Pre-existing resistance is associated with resistance amplification.**

142 (**a**) The acquisition of AR SNPs is associated with subjects who fail treatment. (**b**-**c**) Among

143 subjects who fail treatment, (**b**) subjects with pre-existing mutations that confer antibiotic

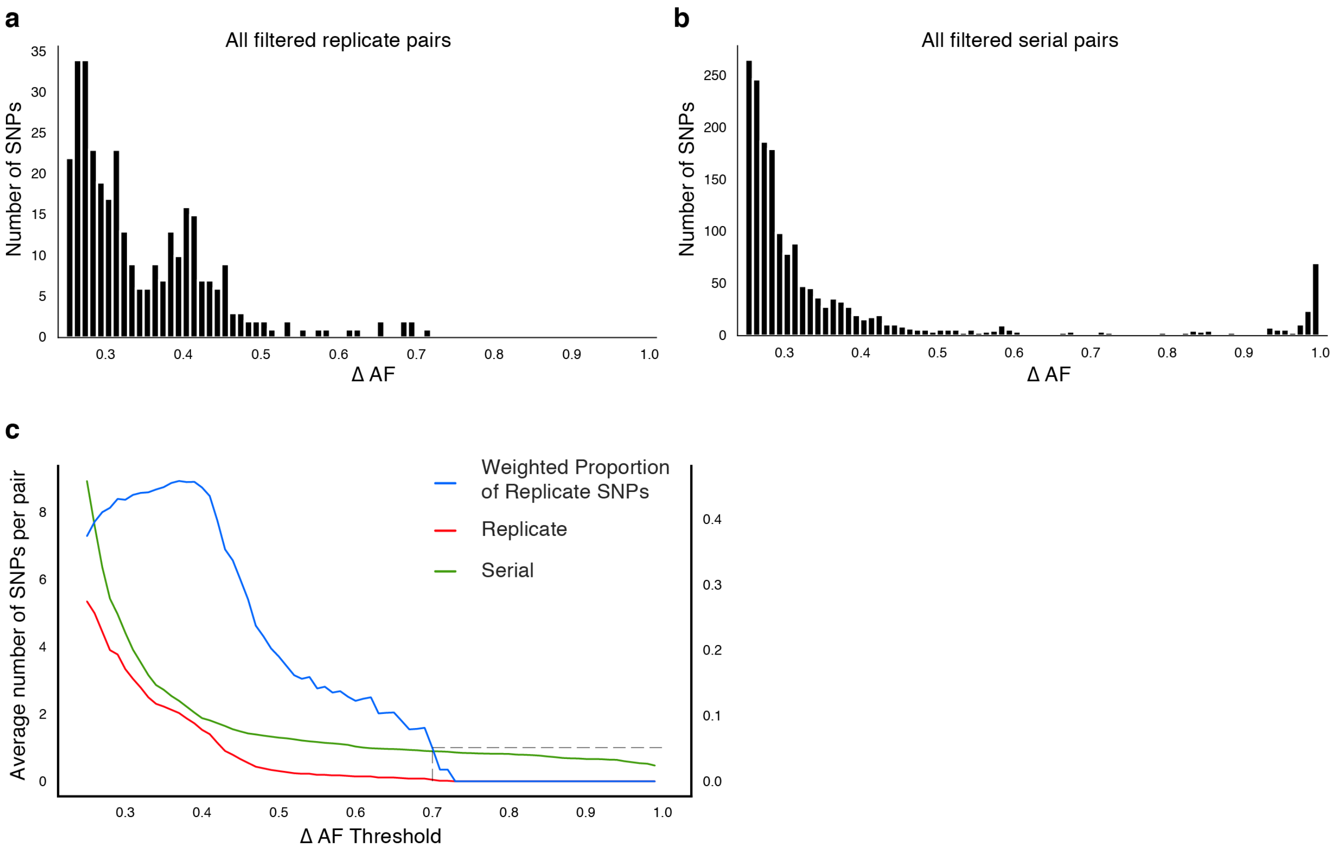
144 resistance and (**c**) those that have pre-existing MDR are more likely to acquire antibiotic

resistance mutations throughout the course of infection. (**d**) The allele frequency trajectories for SNPs that occur in subjects over the course of infection can be used to study the prediction of further antibiotic resistance using the frequency of alternate alleles detected in the longitudinal isolates collected from subjects.



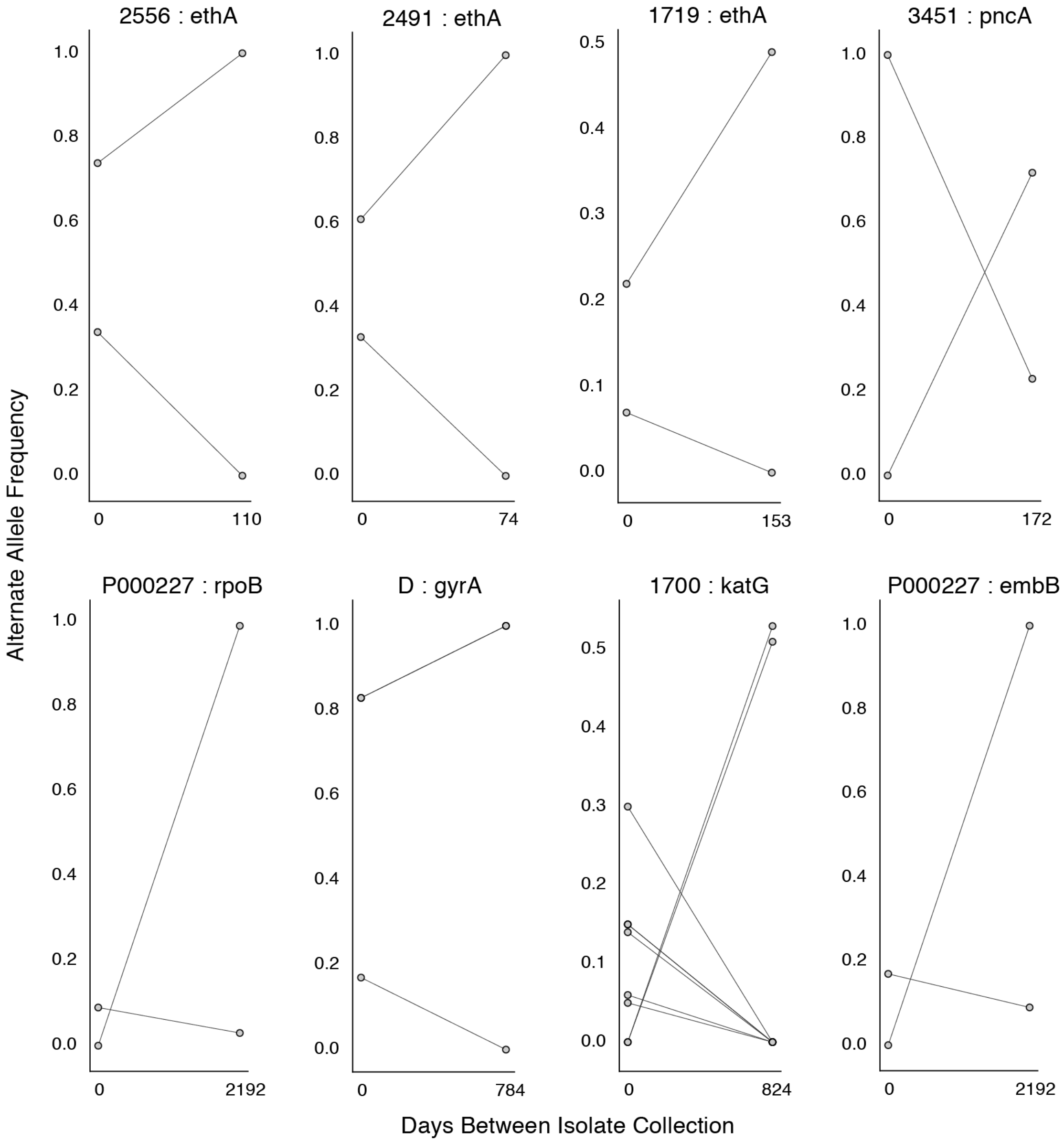
Supplementary Figure 2

Supplementary Figure 2 - **Filtering out laboratory-contaminated samples and subjects with mixed infections.** We implemented several filters to mitigate the effects of contamination from laboratory error or samples from co-infected hosts (**Methods**). Our analysis included three types of replicate pairs (S2, C2, P3) and serial pairs from eight studies (P, C, W, T, B, G, X, H) (**Methods**). At each step, we filtered out any pair of isolates if at least one isolate failed to pass the filter in place (indicated by dashed arrows). First, we used Kraken to filter out isolates that had less than 95% of reads taxonomically classified under MTBC. Second, we filtered out isolates that did not meet the F2 threshold. Third, we filtered out isolate pairs that had a genetic distance greater than 7 fixed SNPs. Our final filtered isolate pair sets included 62 replicate isolate pairs and 200 serial isolate pairs.



Supplementary Figure 3

Supplementary Figure 3 - **Replicate pairs reveal levels of biological noise associated with repeated sampling.** (**a**,**b**) We analyzed the distribution of ΔAF for all SNPs detected across all replicate pairs ($m = 62$) and longitudinal pairs ($n = 200$) for SNPs where ΔAF $\geq$ 25%. (**b**) SNPs were detectable at lower levels of ΔAF for both types of isolate pairs, but SNPs with higher values of ΔAF were only found in longitudinal pairs. (**c**) To determine a ΔAF threshold for calling SNPs representative of changes in bacterial population composition in-host, we calculated the average number of SNPs per pair of isolates at different ΔAF thresholds for both replicate and longitudinal pairs. At a ΔAF threshold of 70% the number of SNPs between replicate pairs represents $\approx$ 5% of the SNPs detected amongst all replicate and longitudinal pairs, weighted by the number of pairs in each group.



173 **Supplementary Figure 4**

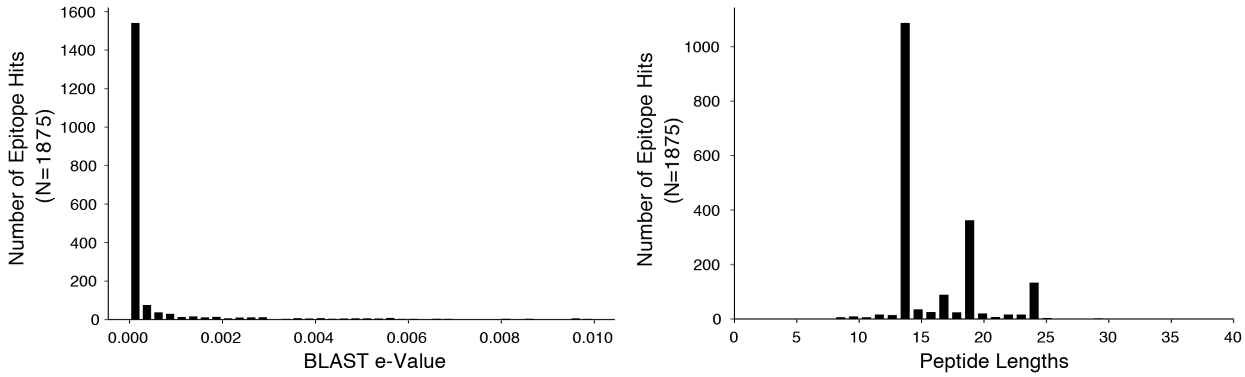
174 Supplementary Figure 4 – **Mutant allele trajectories consistent with clonal interference.**

175 Several examples of co-occurring mutant alleles and their allele frequency trajectories between

176 serial isolate collection demonstrate genetic diversity patterns consistent with competing clones

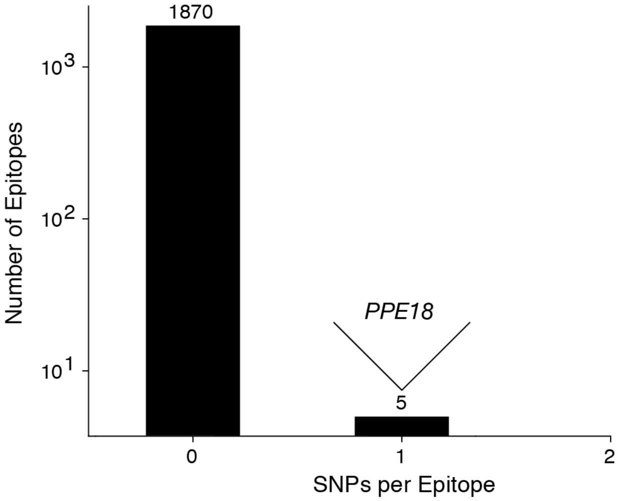
177 in-host.

178



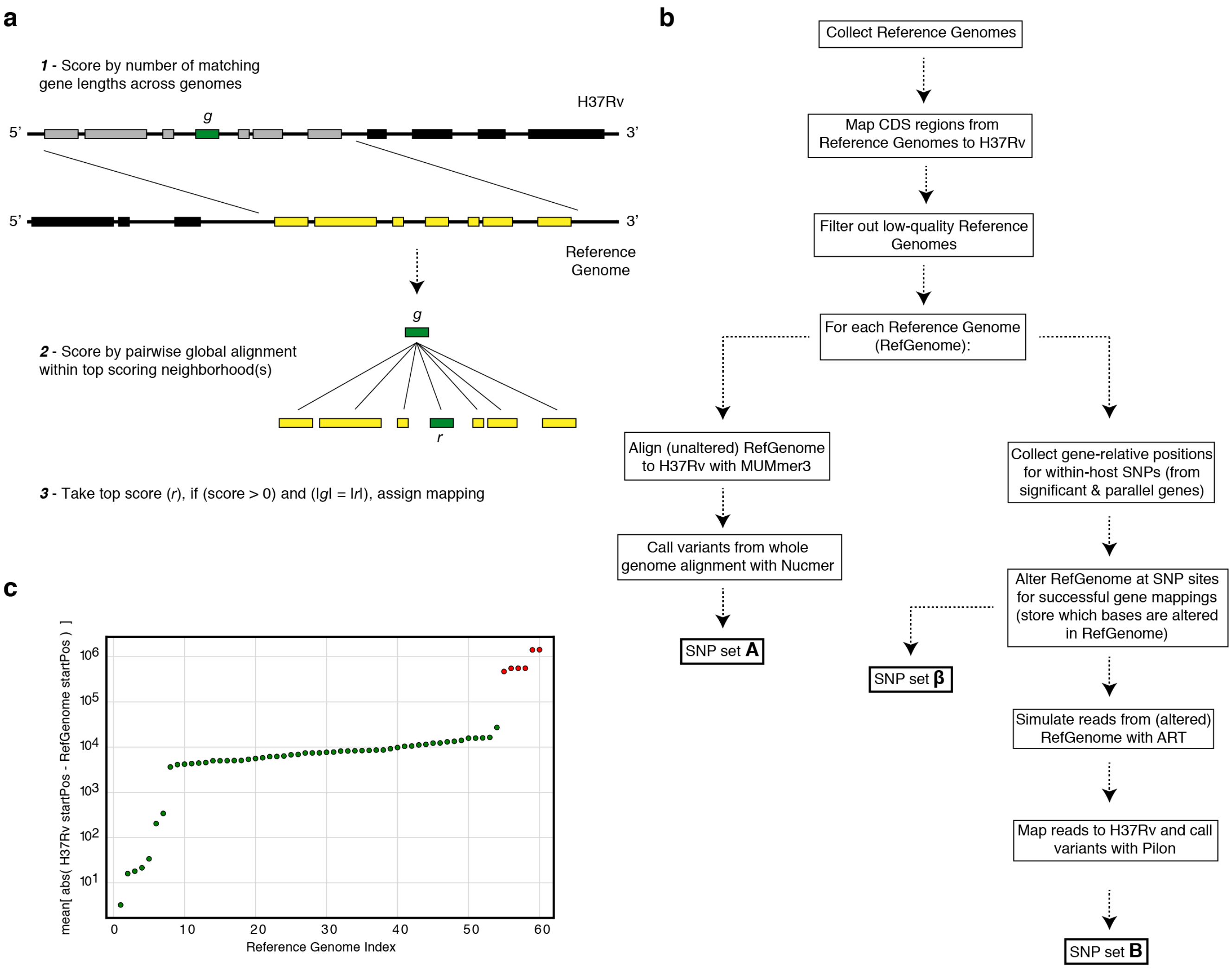
Supplementary Figure 5

Supplementary Figure 5 - **Basic characteristics of epitopes used in analysis.** We downloaded a set of 2,031 epitope peptide sequences from IEDB[7] and used BLASTP to map these peptide sequences to H37Rv imposing an e-value cut-off of 0.01 (**Methods**). (**a**) The distribution of e-values and (**b**) distribution of peptide lengths for the retained epitope mappings.



Supplementary Figure 6

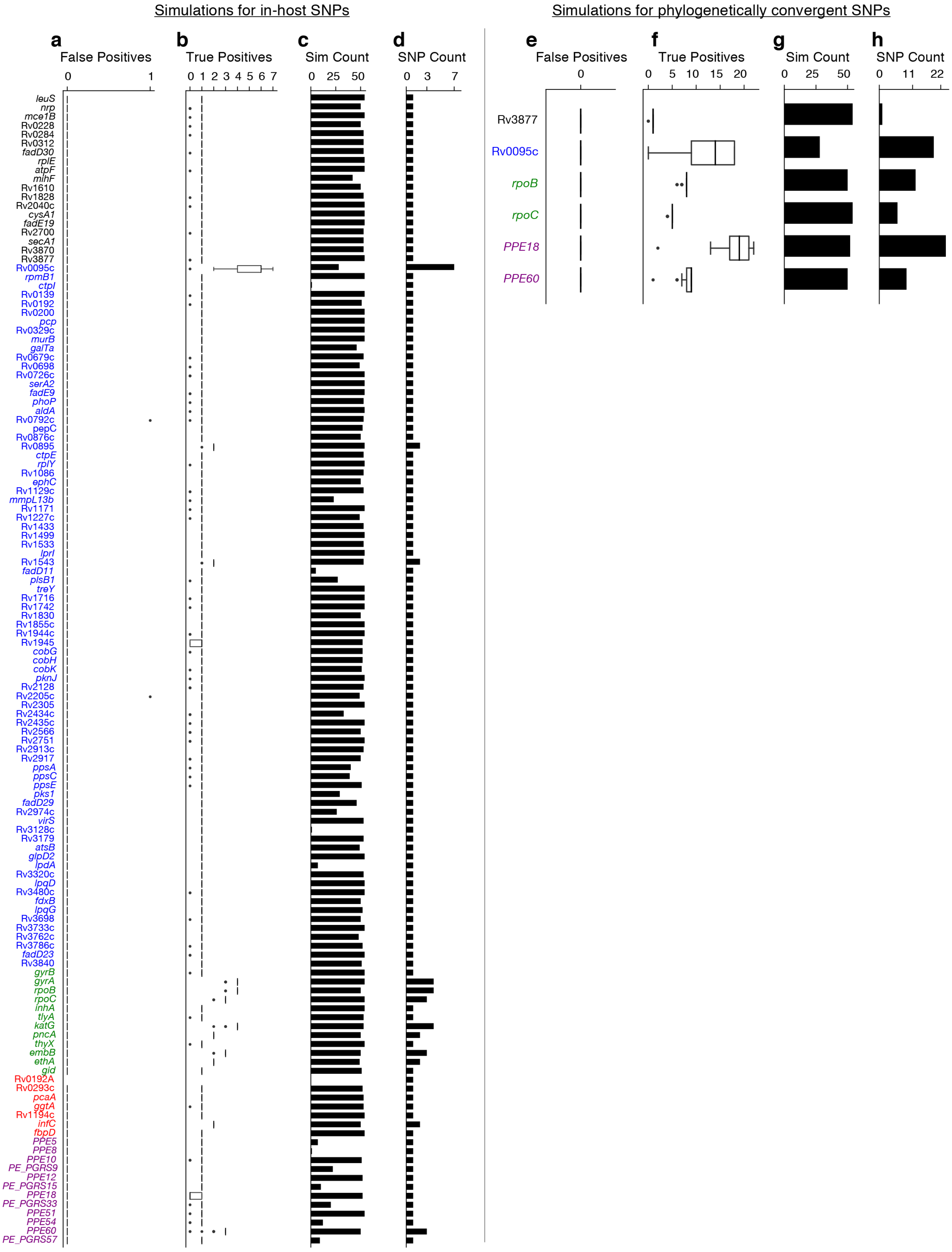
Supplementary Figure 6 - **Most T cell epitopes remain conserved in-host during active TB disease.** No SNPs were detected in-host for a vast majority of CD4$^+$ and CD8$^+$ T cell epitopes, however 1 SNP was detected in a small number $(n = 5)$ of overlapping epitopes in PPE18. A list of these epitopes is given in **Supplementary Table 9**.


a
1 - Score by number of matching gene lengths across genomes
H37Rv
5'
g
3'
5'
3'
Reference Genome
g
2 - Score by pairwise global alignment within top scoring neighborhood(s)
r
3 - Take top score (r), if (score > 0) and (|g| = |r|), assign mapping
b
Collect Reference Genomes
Map CDS regions from Reference Genomes to H37Rv
Filter out low-quality Reference Genomes
For each Reference Genome (RefGenome):
Align (unaltered) RefGenome to H37Rv with MUMmer3
Call variants from whole genome alignment with Nucmer
SNP set A
Collect gene-relative positions for within-host SNPs (from significant & parallel genes)
Alter RefGenome at SNP sites for successful gene mappings (store which bases are altered in RefGenome)
SNP set β
Simulate reads from (altered) RefGenome with ART
Map reads to H37Rv and call variants with Pilon
SNP set B
c
mean[abs(H37Rv startPos - RefGenome startPos)]
10^6
10^5
10^4
10^3
10^2
10^1
0
10
20
30
40
50
60
Reference Genome Index

Supplementary Figure 7

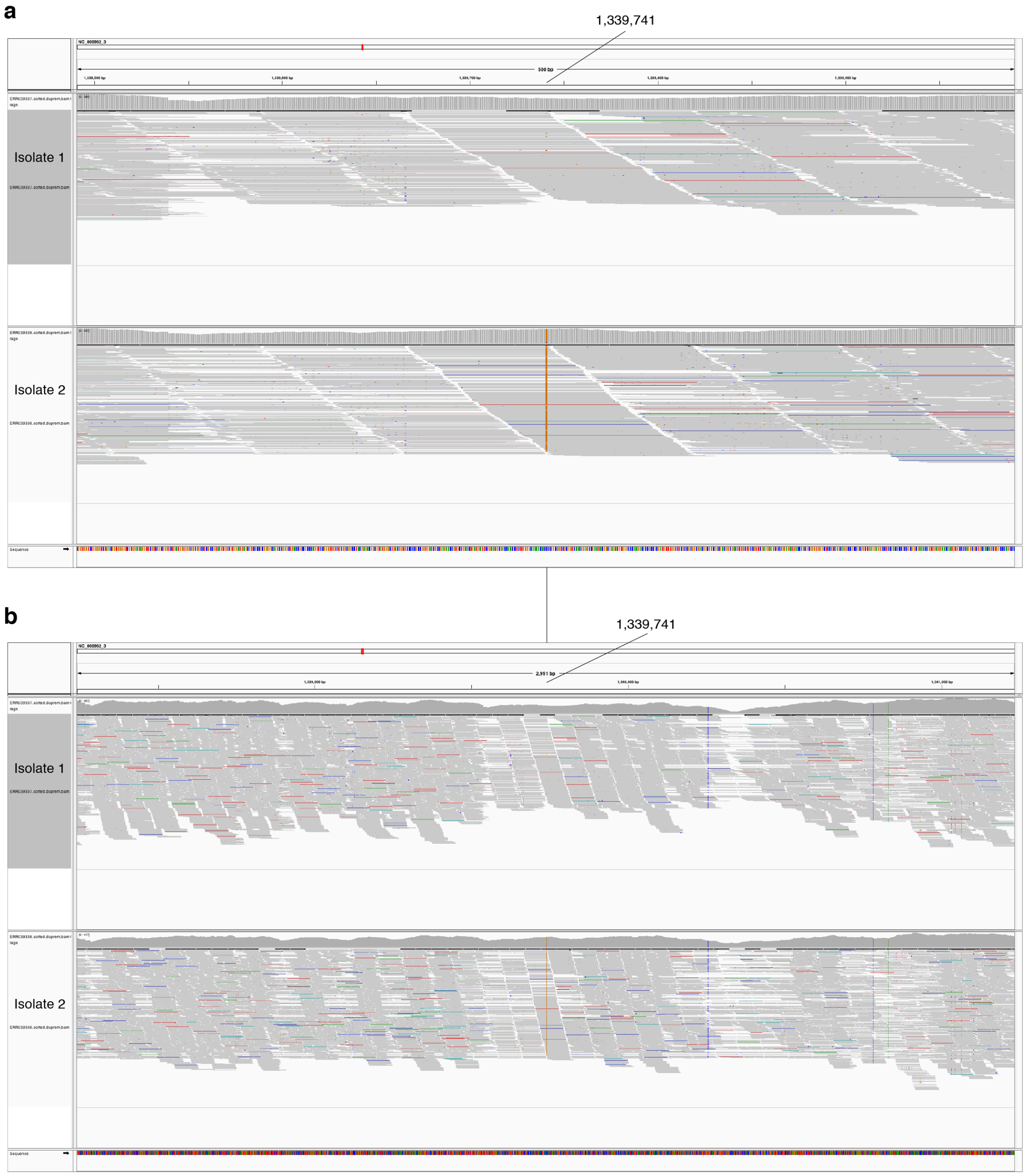
Supplementary Figure 7 - **Overview of simulation methodology.** To test the accuracy of calling SNPs in repetitive regions with our workflow, we introduced mutations into complete *Mycobacterium tuberculosis* genomes (Reference Genomes), simulated reads from those genomes and assessed the accuracy recalling the mutations from the simulated reads while not introducing spurious mutations (**Supplementary Note**). (**a**) We used a sliding window of gene lengths along with a local alignment algorithm to map genes from the H37Rv reference genome to the set Reference Genomes. (**c**) We discarded Reference Genomes that mapped poorly (gene-to-gene) to the H37Rv reference genome (green-RefGenomes kept for simulations, red-discarded

RefGenomes). (**b**) A schematic of our simulation methodology from Reference Genome collection to obtaining SNP sets $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{\beta}$ which are used in our calculations of true positive and false positive calls for each gene (**Supplementary Note**).



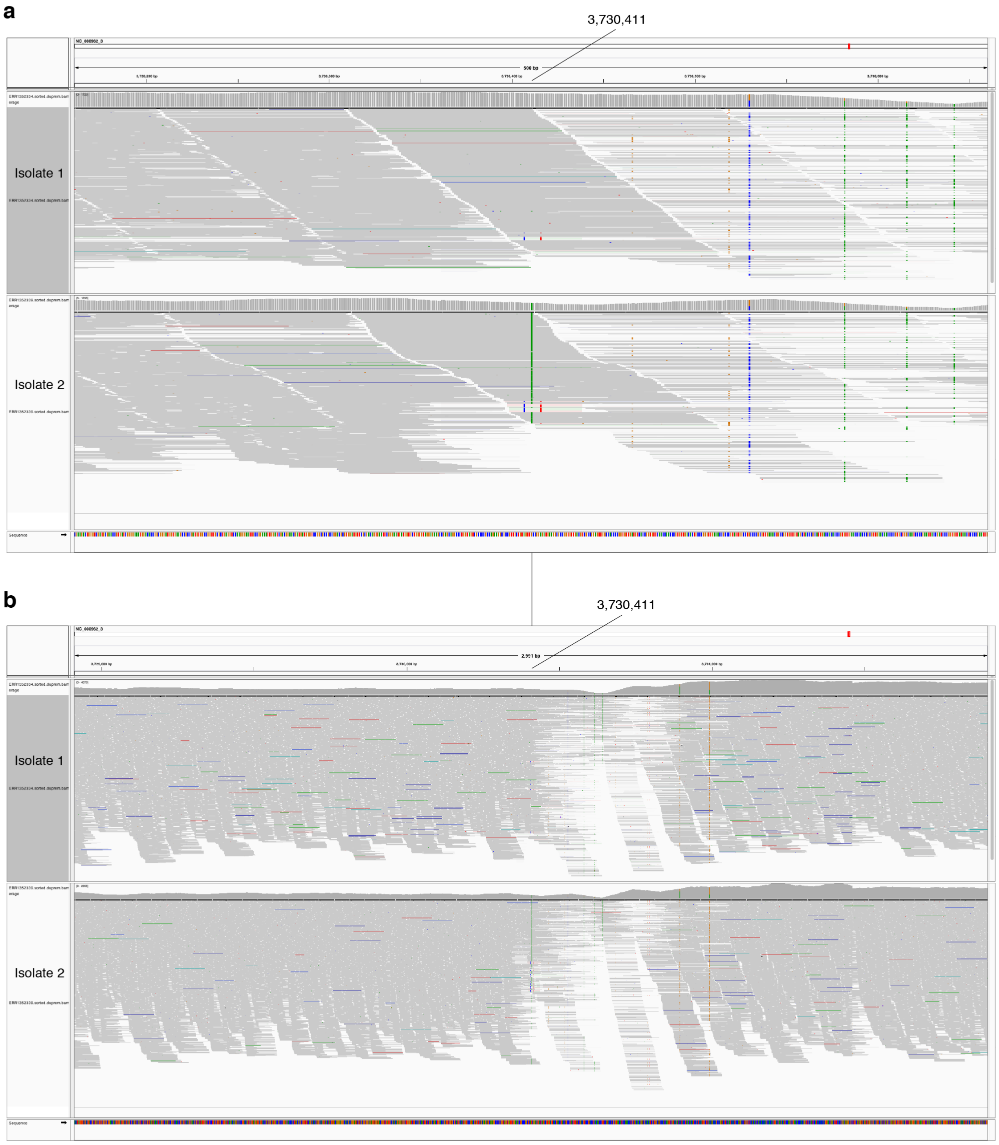
Supplementary Figure 8

Supplementary Figure 8 - **Simulations indicate that we can accurately recall most introduced SNPs while rarely making spurious SNP calls.** We tested the number of true and false positives for each gene with detectable in-host SNPs (**Fig. 3d**) and for each gene with phylogenetically convergent SNPs (**Fig. 5b**). For each gene we collected a set of non-redundant SNPs (genomic positions at which these SNPs were called) observed across all subjects (**Supplementary Table 16**) and SNPs observed to have a signal of phylogenetic convergence (**Supplementary Table 19**), the number of SNPs collected for each gene is given in (**d, h**). We then introduced these mutations into 54 complete genomes (RefGenomes) and simulated reads after introducing the respective mutations (**Supplementary Note**). Only genes that were mapped from H37Rv to a given RefGenome were part of the simulation for that RefGenome. (**c, g**) The number of successful mappings for each gene (i.e. the number of times each gene was part of a simulation). This is also the number of times true and false positive estimates were calculated for each gene (1 estimate / simulation). (**a, e**) False positive calls were rarely made across all genes and simulation runs indicating the rarity of false positive SNP calls (calling a mutation that wasn't introduced) made by our pipeline for observed in-host SNPs and SNPs displaying a signal of phylogenetic convergence, even in repetitive regions. (**d**) The number of true positive calls across all genes (across most simulation runs) closely matched the number of introduced SNPs for each gene indicating the rarity of False Negative SNP calls (not calling a mutation that was introduced). We note that no true or false positive estimates for *Rv0192A* were computed since this gene did not map to H37Rv for any of the 54 Reference Genomes used for the simulations.



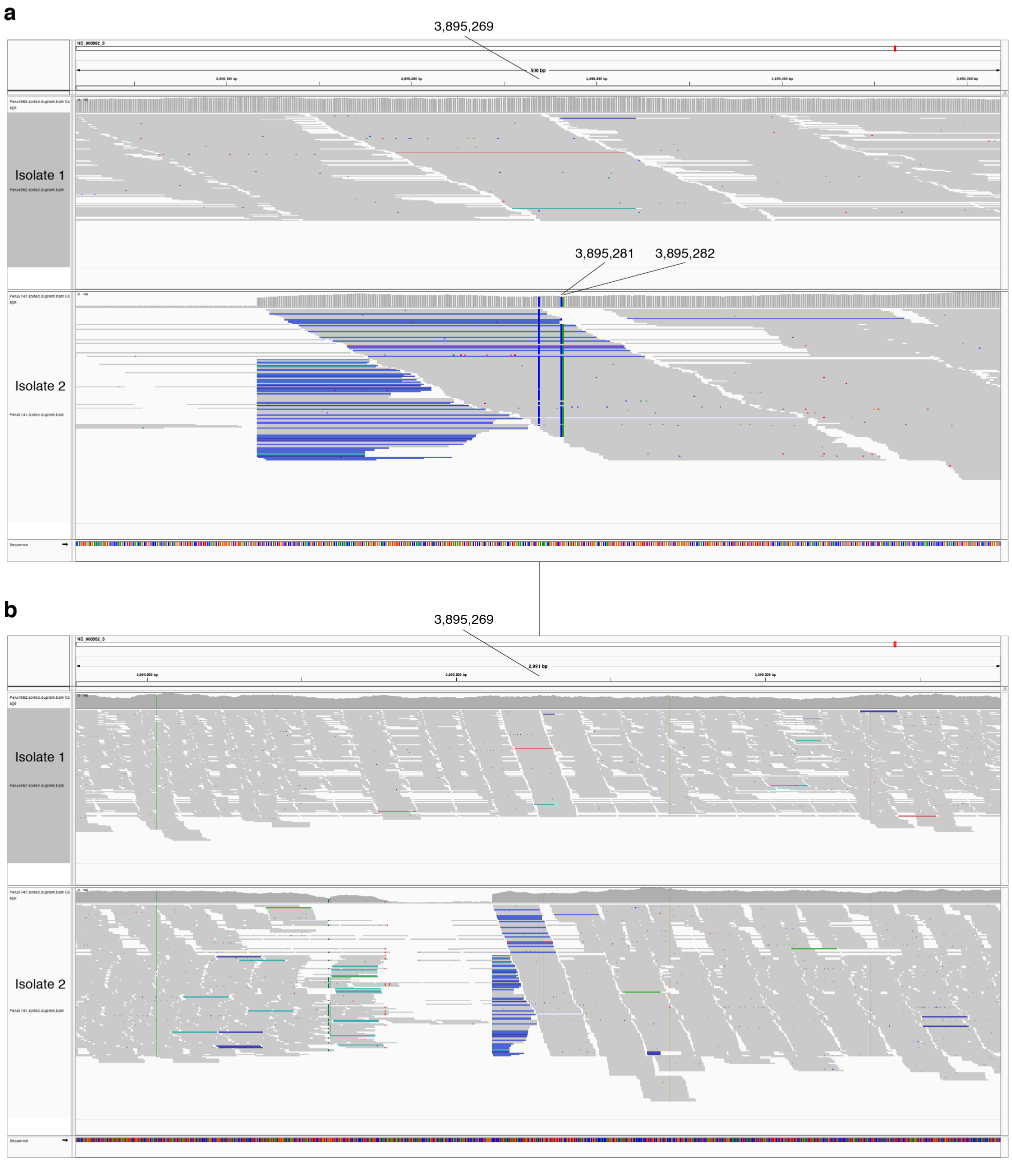
Supplementary Figure 9

Supplementary Figure 9 - **In-host SNP detected in *PPE18*.** IGV[8] image of BAM alignment (reads sorted by start location) for serial clinical isolates that were cultured from sputum collected from patient P000183[9]. **(a)** 500 and **(b)** 3000 basepair windows centered at reference position 1339741. Isolate 1 is the BAM alignment for the isolate collected in 2003 and the reference position 1339741 matches the reference allele (C). Isolate 2 is the BAM alignment for isolate collected in 2008 and reference position 1339741 supports an alternate allele (G).



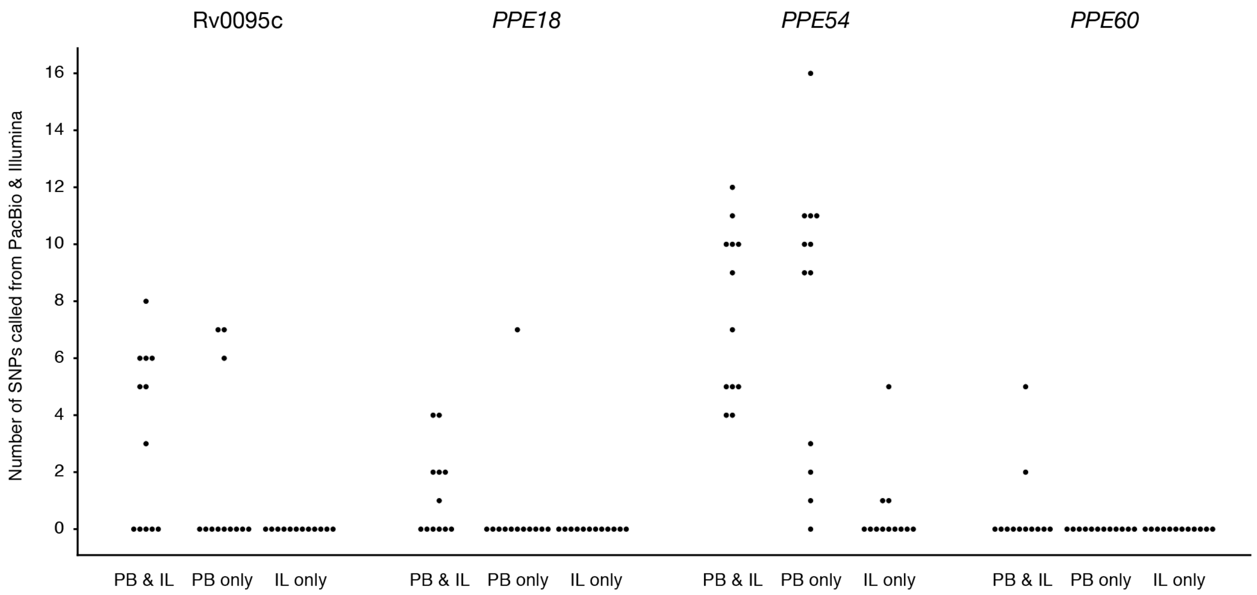
Supplementary Figure 10

Supplementary Figure 10 - **In-host SNP detected in *PPE54*.** IGV[8] image of BAM alignment (reads sorted by start location) for serial clinical isolates that were cultured from sputum collected from patient P09[10]. **(a)** 500 and **(b)** 3000 basepair windows centered at reference position 3730411. Isolate 1 is the BAM alignment for the isolate collected first and the reference position 3730411 matches the reference allele (G). Isolate 2 is the BAM alignment for isolate collected 24 weeks after isolate 1 and reference position 3730411 supports an alternate allele (A).



Supplementary Figure 11

Supplementary Figure 11 - **In-host SNPs detected in *PPE60*.** IGV[8] image of BAM alignment (reads sorted by start location) for serial clinical isolates that were cultured from sputum collected from patient 3096[11]. **(a)** 500 and **(b)** 3000 basepair windows centered at reference position 3895269. Isolate 1 is the BAM alignment for the isolate collected on September 25, 2001 and the reference positions 3895269, 3895281, 3895282 match the reference alleles (G,T,G respectively). Isolate 2 is the BAM alignment for isolate collected on October 18, 2002 and reference positions 3895269, 3895281, 3895282 support alternate alleles (C,C,A respectively).



Supplementary Figure 12

Supplementary Figure 12 – **Most SNP calls are congruent between Illumina and PacBio in Rv0095c, *PPE18*, *PPE54* and *PPE60*.** For each gene, the number of SNPs classified as **|A ∩ B|** (PB & IL), **|B\A|** (PB only), and **|A\B|** (IL only) was plotted for each of our 12 isolates that underwent PacBio and Illumina Sequencing. No calls were made solely from Illumina read mapping (**|A\B|**) in **(a)** Rv0095c, (**b**) *PPE18* or (**d**) *PPE60* demonstrating conservative SNP calling from Illumina reads in these genes. **(c)** In *PPE54*, 3/12 isolates had a small number of SNPs that were called only by Illumina (**|A\B|**). It is important to note that these false positive SNPs (**|A\B|**) did not include any of the 178 in-host SNPs (**Supplementary Table 10**). For the other 9/12 isolates, SNP calls in *PPE54* were either congruent between Illumina and PacBio (**|A ∩ B|**) or only called by PacBio (**|B\A|**) demonstrating a low number of false positives across our 12 Illumina – PacBio sample pairs.

SUPPLEMENTARY TABLE DESCRIPTIONS

Supplementary Table 1: A separate XLSX file containing details for all replicate and serial isolates before Kraken, F2, or pairwise SNP filtering.

Supplementary Table 2: A separate XLSX file containing details for all $(n = 400)$ serial isolates used for in-host analysis after filtering for contaminated & mixed isolate pairs.

Supplementary Table 3: A separate XLSX file with the gene categories assigned to each H37Rv locus tag.

Supplementary Table 4: A separate XLSX file containing a list of genomic regions (with H37Rv coordinates) associated with antibiotic resistance.

Supplementary Table 5: A separate XLSX file containing all SNPs (with $\Delta AF \geq 5\%$) in loci associated with antibiotic resistance (**Supplementary Table 4**) across our sample of 200 serial isolate pairs.

Supplementary Table 6: A separate XLSX file containing all pre-existing antibiotic resistant SNPs detected in the 1st isolate collected from each subject with collection dates $>$ 60 days apart.

Supplementary Table 7: A separate XLSX file containing all pre-existing antibiotic resistant SNPs detected in the 1st isolate collected from each subject with collection dates $\leq$ 60 days apart.

Supplementary Table 8: A separate CSV file containing all of the epitopes downloaded from IEDB on May 23, 2018.

Supplementary Table 9: A separate XLSX file containing the epitopes belonging to *PPE18* where an in-host SNP was detected.

Supplementary Table 10: A separate XLSX file containing information for all 179 in-host SNPs detected across all serial isolate pairs.

Supplementary Table 11: A separate XLSX of all genes identified as *dense*, along with assigned gene category and p-value from mutation density test.

Supplementary Table 12: A separate TSV file containing the downloaded SEED annotation for H37Rv.

Supplementary Table 13: A separate CSV file containing the list of H37Rv locus tags corresponding to each subsystem classified by SEED.

Supplementary Table 14: A separate XLSX file containing the pathways and (corresponding in-host SNPs) displaying evidence of parallel evolution.

Supplementary Table 15: A separate XLSX file with details for the publicly available completed genomes used in our simulations.

Supplementary Table 16: A separate XLSX file with the non-redundant in-host SNPs identified in genes and used for SNP calling simulations.

Supplementary Table 17: A separate XLSX file with details for all genes that were evaluated for a signal of phylogenetic convergence in 10,018 publicly available isolates.

Supplementary Table 18: A separate XLSX file with details for all SNPs that were found in 10,018 publicly available isolates after screening for SNPs occurring within (a) mutationally dense genes, (b) genes convergent in-host & (c) genes belonging to pathways that were convergent in-host.

Supplementary Table 19: A separate XLSX file with details for SNP sites occurring within the genes in **Supplementary Table 17** displayed a signature of phylogenetic convergence after screening 10,018 publicly available isolates. The number of isolates with each unique mutation (broken down by global lineage) is given.

Supplementary Table 20: A separate XLSX file containing details for isolates that underwent Illumina and PacBio sequencing.

Supplementary Table 21: A separate XLSX file containing all 80 SNPs called from the PacBio assemblies and from mapping Illumina reads for Rv0095c, *PPE18*, *PPE54* and *PPE60* across the 12 isolates with both PacBio and Illumina Sequencing data. Each SNP is annotated with the: (1) number of samples where Illumina SNP calling correctly identified the SNP when the SNP was also present in the paired PacBio assembly, (2) number of samples where Illumina SNP calling falsely identified the SNP when the SNP was *not* present in the paired PacBio assembly.

Supplementary Table 22: A separate XLSX file containing a list of the 17/178 in-host SNPs and 31/68 phylogenetically convergent SNPs present in at least 1/12 isolates with both PacBio and Illumina sequencing data. Each SNP is annotated with the: (1) presence of this SNP within our 12 complete PacBio assemblies, (2) number of samples where Illumina SNP calling correctly identified when the SNP also present in the paired PacBio assembly, (3) number of samples where Illumina SNP calling falsely identified the SNP when the SNP was *not* present in the paired PacBio assembly.

REFERENCES

1. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
2. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic acids research* **37**, D26–D31 (2008).
3. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2011).
4. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome biology* **5**, R12–R12 (2004).
5. Wyllie, D. H. *et al.* Identifying Mixed Mycobacterium tuberculosis Infection and Laboratory Cross-Contamination during Mycobacterial Sequencing Programs. *Journal of Clinical Microbiology* **56**, (2018).
6. Goig, G. A., Blanco, S., Garcia-Basteiro, A. & Comas, I. Pervasive contaminations in sequencing experiments are a major source of false genetic variability: a Mycobacterium tuberculosis meta-analysis. *bioRxiv* (2018). doi:10.1101/403824
7. Vita, R. *et al.* The immune epitope database (IEDB) 3.0. *Nucleic acids research* **43**, D405--D412 (2014).
8. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics* **14**, 178–192 (2013).
9. Walker, T. M. *et al.* Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *The Lancet infectious diseases* **13**, 137–146 (2013).

10. Trauner, A. *et al.* The within-host population dynamics of Mycobacterium tuberculosis vary with treatment efficacy. *Genome biology* **18**, 71–71 (2017).

11. Farhat, M. R. *et al.* GWAS for quantitative resistance phenotypes in Mycobacterium tuberculosis reveals resistance genes and regulatory regions. *Nature communications* **10**, 2128 (2019).