Genomic analysis of *Mycobacterium tuberculosis* reveals complex etiology of tuberculosis in Vietnam including frequent transmission of Beijing lineage and positive selection for EsxW Beijing variant

Authors

Kathryn E Holt[1,*], Paul McAdam[1,^], Phan Vuong Khac Thai[2,^], Nguyen Thuy Thuong Thuong[3,4], Dang Thi Minh Ha[2], Nguyen Ngoc Lan[2], Nguyen Huu Lan[2], Nguyen Thi Quynh Nhu[3], Hoang Thanh Hai[3], Vu Thi Ngoc Ha[3], Guy Thwaites[3,4], David J Edwards[1], Artika P Nath[5,9], Kym Pham[5], David B Ascher[1], Jeremy Farrar[3,4], Chiea Chuen Khor[6], Yik Ying Teo[7,8], Michael Inouye[5,9,10], Maxine Caws[11,^], Sarah J Dunstan[12,*,^]

[1]Department of Biochemistry and Molecular Biology, Bio 21 Molecular Science and Biotechnology Institute, University of Melbourne, Parkville, Victoria 3010, Australia

[2]Pham Ngoc Thach Hospital for Tuberculosis and Lung Disease, Ho Chi Minh City, District 5, Viet Nam

[3]Oxford University Clinical Research Unit, Ho Chi Minh City, District 5, Viet Nam

[4]Centre for Tropical Medicine, Nuffield Department of Clinical Medicine, Oxford University, Oxford, UK.

[5]Department of Pathology, University of Melbourne, Parkville, Victoria 3010, Australia

[6]Genome Institute of Singapore, Singapore

[7]Department of Statistics and Applied Probability, National University of Singapore, Singapore.

[8]Saw Swee Hock School of Public Health, National University of Singapore, Singapore

[9]Systems Genomics Lab, Baker Heart and Diabetes Institute, Melbourne 3004, Victoria, Australia

[10]Department of Public Health and Primary Care, University of Cambridge, Strangeways Research Laboratories, Cambridge CB1 8RN, United Kingdom

[11]Liverpool School of Tropical Medicine, Liverpool, UK

[12]Peter Doherty Institute for Infection and Immunity, University of Melbourne, Parkville, Victoria 3010, Australia

^ These authors contributed equally to this work

* Corresponding authors

Correspondence should be addressed to KEH (kholt@unimelb.edu.au) and SJD (sarah.dunstan@unimelb.edu.au)

Introduction

Tuberculosis (TB) is a leading cause of death from infectious disease and the global burden is now higher than at any point in history [1,2]. Despite coordinated efforts to control TB transmission, the factors contributing to its successful spread remain poorly understood. Vietnam is identified as one of 30 high burden countries for TB and MDR-TB with an incidence of 137 TB cases per 100,000 individuals in 2015 [2]. Recent phylogenomic analyses of the causative agent *Mycobacterium tuberculosis* (*Mtb*) in other high-prevalence regions have provided insights into the complex processes underlying TB transmission [3-5]. Here we examine the transmission dynamics of *Mtb* isolated from TB patients in Ho Chi Minh City (HCMC), Vietnam via whole genome analysis of 1,635 isolates and comparison with 3,144 isolates from other locations. The genomic data reveal an underlying burden of disease caused by endemic *Mtb* Lineage 1 associated with activation of long-term latent infection, on top of which is overlaid a three-fold higher burden associated with more recently introduced Lineage 2 and 4 *Mtb* strains. We find that Beijing lineage *Mtb* is frequently transferred between Vietnam and other countries, and detect higher levels of transmission of Beijing lineage strains within this host population than endemic Lineage 1 *Mtb*. Screening for parallel evolution of Beijing lineage-associated SNPs in other *Mtb* lineages as a signal of positive selection, we identify a mutation in the ESX-5 type VII secreted protein EsxW, which could potentially contribute to the enhanced transmission of the Beijing lineage in Vietnamese and other host populations.

To characterize the diversity of *Mtb* circulating in HCMC, we sequenced the genomes of 1,635 isolates (**Supplementary Table 1**) obtained from 2,091 HIV uninfected, smear positive adults (≥18 years) commencing anti-TB therapy at district TB units (DTUs) in eight districts of HCMC between December 2008 and July 2011 (see **Methods**). A total of 73,718 high quality SNPs were identified and used to reconstruct a maximum likelihood (ML) phylogeny (**Fig. 1a**) and to assign lineages [6]. Four major lineages (Lineages 1-4) were present within the study population. The majority of isolates (n=957, 59%) belonged to lineage 2.2.1, a subgroup of the Beijing lineage (2.2). Lineage 1 (Indo-Oceanic lineage; n=388, 23.7%) and Lineage 4 (Euro-American lineage; n=192, 11.7%) were also common. A single isolate belonged to Lineage 3 (East African-Indian lineage) and was excluded from further analysis. The distribution of lineages did not change during the 2.5 year period of the study (**Fig. 1b**), and was in agreement with previous genotyping (MIRU-VNTR, spoligotyping) studies in urban areas of Vietnam (≥50% Beijing lineage (2.2) and ~20% Lineage 1.1/EIA in Hanoi and HCMC, 1998-2009) [7-11]. Known antimicrobial resistance mutations were detected in all lineages but were more frequent in Beijing sublineage 2.2.1 (**Table 1**), consistent with earlier reports from Vietnam [7-9,11]. In particular, the Beijing sublineage 2.2.1 was enriched for mutations associated with resistance to streptomycin (OR 4.6 [95% CI 3.6-6.0], $p=10^{-15}$), isoniazid (OR 1.7 [1.3-2.1], $p=3x10^{-5}$), rifampicin (OR 5.4 [2.5-13.2], $p=1x10^{-7}$) and ethambutol (OR 5.6 [2.6-13.7], $p=1x10^{-7}$), using Fisher's exact test to compare to all non-2.2.1 *Mtb* isolates.

Whilst the majority of TB patients were male (74%, typical for TB studies in Vietnam and elsewhere [8-10,12]), the Beijing sublineage was significantly associated with TB in females (OR 1.28 [95% CI 1.01-1.62], p=0.043 using Fisher's exact test; see **Table**

1), consistent with prior observations in Vietnam[8] and Nepal[13]. Beijing sublineage 2.2.1 was also significantly associated with younger people: its frequency declined with age, from 74% of cases in <20 year olds vs 50% in ≥60 year olds (p=0.0023 Fisher's exact test, p=0.0024 linear trend test; **Fig. 1c**). In contrast, Lineage 1 was significantly associated with males (25% of male cases vs 19% of females, p=0.017) and increased with age regardless of gender, from 12% in <20 year olds vs 35% in ≥60 year olds (p=0.0007 Fisher's exact test, p=0.0014 linear trend test; **Fig. 1c**). These data confirm that Beijing sublineage 2.2.1 is capable of causing active disease in a wider demographic range of the Vietnamese host population, particularly among females and younger people, than is the endemic Lineage 1 [7-10], which is associated with the more typical profile of TB susceptibility that is skewed towards males [M:F prevalence ratio for smear-positive TB was recently estimated at 2.5 (95% CI 2.07–3.04)] based on 40 surveys in 22 countries[12]; and older people [2,14].

Consequently, we hypothesised that Beijing lineage or sublineage 2.2.1 was also more transmissible than Lineage 1, and/or more capable of causing active disease in infected hosts, in the Vietnamese Kinh host population. We used the whole genome phylogeny to investigate this possibility in more detail, comparing several diversity metrics for each lineage (**Fig. 2**). Terminal branch lengths, which represent the upper bound of evolutionary time since transmission for each *Mtb* case, were significantly shorter for Beijing sublineage 2.2.1 *Mtb* isolates (median 8 SNPs) than for non-Beijing lineage isolates (Lineage 1: median 53 SNPs, $p < 1 \times 10^{-15}$ using Kolmogorov-Smirnov test; Lineage 2.1: 30 SNPs, $p < 1 \times 10^{-6}$; Lineage 4: 17 SNPs, $p < 1 \times 10^{-9}$), and slightly shorter than Beijing sublineage 2.2.2 isolates (9 SNPs, p=0.02) (**Fig. 2a**). Indeed the distribution of mean node-to-tip distances for all internal

nodes was skewed significantly lower within the Beijing sublineage 2.2.1 compared to the rest of the tree (median 16 SNPs compared to 62, 57, 39 and 60 SNPs for Lineages 1, 2.1, 2.2.2 and 4, respectively; $p<0.0015$ in all cases). These observations could be explained by (i) a slower mutation rate in the Beijing lineage, resulting in slower accumulation of SNPs over time; (ii) sampling bias, whereby new cases of active TB arising in the study population were more likely to be detected and included in the study if they were caused by Beijing lineage strains; or (iii) strain-specific transmission dynamics in the study population, whereby the average time taken to progress to active disease is shorter for Beijing lineage than for other *Mtb* strains. The mutation rate for *Mtb* has been estimated at ~0.5 SNPs per year for Lineage 4, and ~2x faster for Beijing strains [15] (Lineage 1 has not been specifically measured but is assumed to be similar to Lineage 4, which is considered typical, rather than Lineage 2, which is considered an atypically fast growing strain). Hence if all *Mtb* lineages were subject to the same transmission dynamics within the study area, we would expect to see longer terminal branch lengths for Beijing lineage isolates, whereas here we observe the opposite (significantly shorter branch lengths). Patients were recruited into the study following positive diagnosis at eight central district TB units in HCMC (see map in **Supplementary Fig. 1**). The identity of the infecting *Mtb* lineages was not known at the time of recruitment, and we are unaware of any factors that would bias the inclusion of Beijing lineage cases over others presenting to the clinics for treatment; hence we conclude the frequency of infections with Beijing strains, and their shorter branch lengths, reflect differences in the transmission dynamics of these strains within the study population.

To better understand the differences in transmission dynamics, we explored the

distribution of potential transmission clusters using a range of maximum pairwise SNP distance thresholds to define a cluster [16] (**Fig. 2b-c**). Using the smallest cut-off of five SNPs (transmission age of <5 years) there were n=109 clusters, of which 76 (70%) belonged to Beijing sublineage 2.2.1; these had a mean size of 2.4 strains per cluster, compared to mean 2.1 for other clusters (**Fig. 2b-c**). Using cut-offs of 10 and 20 SNPs there were n=164 and n=220 clusters, respectively; of these, 118 (72%) and 156 (71%), respectively, were Beijing 2.2.1 and these showed significantly larger cluster sizes (means, 2.5 and 3.2 strains per cluster) than those of other lineages (means, 2.2 and 2.4). Notably, the proportion of cases that belonged to transmission clusters was significantly lower amongst lineage 1 cases (7.7% at the 10 SNP threshold) compared to lineage 4 cases (20.3%, $p=2x10^{-5}$), which in turn was significantly lower than among Beijing lineage cases (31.5%, $p=3x10^{-3}$; see **Fig. 2b**). Therefore, not only was sublineage 2.2.1 the most common cause of TB in the HCMC study population, but these infections were more likely to result from recent detectable transmission within the local population, despite our study having low power to detect recent transmission due to sampling only a small fraction of all incident cases in HCMC (~30% of those in the study districts, and <10% across HCMC, and restricted to new cases of smear-positive, culture-positive HIV-negative TB only; see **Methods**). As we do not have household data for the study cohort, we could not examine spatial distributions of the transmission clusters in detail; however pairs of TB cases whose infecting *Mtb* strains were separated by ≤10 or ≤20 SNPs were significantly more likely to be diagnosed in the same HCMC district tuberculosis unit than more distantly related pairs of the same lineage (27% or 21%, vs 15% amongst strains separated by >20 SNPs; $p<1x10^{-12}$, see **Supplementary Fig. 2**). This phenomenon was observed for both Beijing 2.2.1 and non-Beijing clusters,

but the effect was significantly less for the Beijing lineage (of strain pairs separated by <20 SNPs, 21% of Beijing pairs vs 29% of non-Beijing pairs were diagnosed in the same district tuberculosis units; p=0.036, see **Supplementary Fig. 2**; note there were too few transmission clusters from other lineages to test for differences between the various non-Beijing lineages). This suggests that Beijing sublineage 2.2.1 strains may be more readily transmitted across greater geographical distances within HCMC than other *Mtb* strains. This could potentially be associated with the higher frequency of Beijing lineage TB amongst younger adults (**Fig. 1c**), who may be more likely to travel regularly between districts, promoting onward transmission; however the age distribution amongst transmission clusters did not differ by lineage. Taken together, the phylogenomic data reveal significant differences in transmission dynamics between the various *Mtb* lineages circulating in HCMC, and suggest that newly diagnosed cases of Lineage 1 *Mtb* in this population often result from activation of longer-term latent infections with many private SNPs and no evidence of recent transmission, whilst new cases of Lineage 2.2.1 *Mtb* often result from more recent transmission and shorter time to develop active disease.

It has been suggested previously that the Beijing lineage is slowly displacing the resident Lineage 1 strains in Vietnam, following the introduction of the Beijing strain into urban areas and subsequent spread to rural areas where Lineage 1 still dominates [8,10]. Our data are consistent with this, showing a higher frequency of Beijing lineage (65%) amongst *Mtb* infections in HCMC in 2009-2010 compared to the frequencies reported in the city a decade earlier [7] (53% in 1998), or in rural areas outside the city [8] (32-37% in 2003-2005). We therefore hypothesized that Beijing sublineage 2.2.1 isolates from HCMC may represent a locally established epidemic subclade of the

Beijing lineage, similar to that previously described in Russia [3]. To investigate this we combined our HCMC *Mtb* genome data with 3,146 publicly available *Mtb* whole genome sequences from Russia [3], Malawi [4,5], Argentina [17], and China [18], as well as globally dispersed Lineage 1 and 2 genomes [19] [20,21] (**Supplementary Table 2**), then inferred phylogenies for Lineages 1, 2 and 4 (**Fig. 3**). HCMC Lineage 1 strains were quite distinct from those found in other locations (**Fig. 3a**), with little evidence of transfer between Vietnam and other regions. The vast majority (n=319, 82%) belonged to a localised subclade (designated 1.1.1.1 [6]) that included only seven previously sequenced strains, all of which were from Vietnam. A further 46 (12%) of the HCMC strains belonged to a related sister clade (1.1.1) which also included strains from neighbouring Cambodia (n=1) and Thailand (n=2); and the remainder (n=23, 6%) belonged to a distinct subclade 1.2.1, in which the HCMC strains were intermingled with others from the Philippines (n=9) and China (n=1). These data suggest that Lineage 1 associated TB in HCMC results mainly from the local endemic *Mtb* population. In contrast, Beijing sublineage 2.2.1 isolates from HCMC formed several distinct clusters that each shared a recent common ancestor with Lineage 2.2.1 isolates from outside Vietnam (**Fig. 3b**). Notably, isolates from Russia, Malawi, China and numerous other countries were interspersed throughout the HCMC sublineage 2.2.1 population (**Fig 3b**), suggesting multiple, frequent transfers of this lineage between host populations in HCMC and other geographic regions. HCMC Lineage 4 isolates were drawn from eight of the ten recognised sublineages [6] – including those identified as specialist, generalist and intermediate in their geographic range [22] – and were interspersed with isolates from other geographical locations, consistent with multiple imports into HCMC from external populations (**Fig. 3c**). In further support of these observations, stochastic mapping of locations onto the

phylogenies predicted dozens of strain transfer events between Vietnam and other locations for Lineages 2 and 4, but not for Lineage 1 (**Fig. 3d**) (see **Methods**). These results strongly support the notion that *Mtb* sublineage 1.1.1.1 is endemic in Vietnam. The lack of transfer of this clade with other geographical regions may be associated with adaptation to the local host population with which it has co-evolved for centuries, similar to the 'host-specialist' clades recently identified within Lineage 4 [22]. In contrast, whilst the direction of transfer of Lineage 2.2.1 strains between HCMC and other geographical regions cannot be determined from our data, the frequency of transfer events and the scale of diversity amongst the HCMC strains makes it unlikely that the rise of Beijing sublineage 2.2.1 in HCMC represents clonal spread of a locally established subclade. Regardless of direction, the frequency of transfer that we observed between Vietnam and diverse geographically dispersed populations supports previous contentions that the Beijing lineage is a host-generalist, capable of moving between ethnically diverse host populations [23].

Taken together, our data show that the burden of TB in HCMC comprises (i) an underlying burden of disease caused by endemic Lineage 1 *Mtb* strains (24% of all TB cases), which disproportionately affects men and older people and is associated more with activation of long-term latent infection than short-term transmission clusters; and (ii) an additional disease burden caused by more recently introduced Lineage 2 and 4 *Mtb* strains (76% of all TB cases). In particular, the Beijing sublineage 2.2.1 was associated with a wider demographic host range, infecting women and young people significantly more frequently than other lineages,,and was associated with shorter time to active disease and frequent onward local transmission. One third of all Beijing strains were involved in transmission clusters defined using a

maximum pairwise patristic distance threshold of ≤10 SNPs, and these were associated with wider geographic dispersal within HCMC. Notably, 75% of TB cases associated with transmission clusters involved the Beijing lineage, accounting for 20% of all cases included in the genomic study. These findings could be used to inform local TB control programs. For example, TB contact tracing is commonly used in low-incidence high-income countries, but results vary in high-incidence low- and middle-income countries and there is a need to prioritize resources towards cases where contact tracing is most likely to yield results, which includes consideration of the likelihood of transmission and progression to active disease within the time frame of a contact tracing program (1-2 years)[24,25]. In HCMC it may thus be advantageous to direct resources towards contact tracing of Beijing lineage cases, as they pose the greatest risk of onward transmission resulting in new active TB cases that are likely to be detected by contact tracing efforts.

The population structure (**Figs 1-2)** provides evidence that Beijing lineage strains are more transmissible within this HIV-negative HCMC population than are other *Mtb* lineages. Genomic evidence for enhanced transmission of the Beijing lineage has been documented in Russia (associated with antimicrobial resistance) [3] and Malawi (independent of antimicrobial resistance) [4]. While antimicrobial resistance was common amongst HCMC Beijing lineage isolates, the majority of transmission clusters (defined by ≤10 SNP distance) comprised groups of isolates that did not share any known resistance mutations that could account for their transmission success (**Supplementary Fig. 3**). This is consistent with previous reports that the Beijing lineage is highly transmissible and more likely to progress to active disease in various host populations and is also more virulent and less pro-inflammatory in various

cellular assays, independent of antimicrobial resistance [23,26-28]. We therefore aimed to interrogate the *Mtb* genome data to identify mutations that may contribute to the success of the Beijing lineage (2.2). Evolutionary convergence has previously been used as a signal of positive selection to identify mutations associated with antimicrobial resistance in *Mtb*[29,30]. We reasoned that advantageous polymorphisms contributing to the enhanced transmissibility of Lineage 2.2 should be fixed in this lineage, and should also be under positive selection that is detectable as convergent or parallel evolution at the same variant sites in other lineages. We identified a total of 424 homoplasic nonsynonymous SNPs (nsSNPs) across the HCMC phylogeny. The most frequent of these occurred in genes in which convergent evolution has previously been associated with antimicrobial resistance including *gidB, embB, gyrA, rpoB,* and *inhA* [29], which together accounted for 12.4% of all homoplasic nsSNPs. The distribution of common homoplasic nsSNPs in these genes is shown in **Supplementary Figure 4**; in particular, rifampicin resistance-associated mutations in *rpoB* and ethambutol-associated mutations in *embB* arose independently many times in Beijing sublineage 2.2.1 (n=33/35 *rpoB* -450 mutations; n=16/17 *embB*-306 mutations; n=8/10 *embB*-406 mutations). The homoplasic nsSNPs included three that arose on the branch defining Lineage 2.2 and also elsewhere in the HCMC tree (**Table 2**, **Supplementary Fig. 5**). One was a mutation in *esxW* (Rv3620c) codon 2 (EsxW-Thr2Ala), which arose on nine other branches (six times in Lineage 4, three times in Lineage 1; see **Supplementary Fig. 5**) and showed evidence of onward transmission on four occasions. Comparison to the global tree detected the same *EsxW* mutation on a further ten Lineage 4 branches in Malawi and Russia, with onward transmission detected on six occasions. The other two mutations were in *Rv3081* (conserved hypothetical protein) and *GidB* (mutations in which are often

associated with streptomycin resistance) and arose less frequently (**Table 2**). In contrast, homoplasic nsSNPs on the branches defining Lineages 1 or 4 were each detected on only one or two other branches in the HCMC tree and no additional branches of the global tree (**Supplementary Table 3**). No homoplasic SNPs were associated with sublineage 2.2.1, and although synonymous or intergenic SNPs can have functional consequences, we found no such homoplasies associated with Beijing or other lineages.

EsxW is included in multiple *Mtb* vaccines currently under development (including H65, ID83, ID93) [31-34] due to its demonstrated immunogenicity in mice, safety and immunogenicity in non-human primates [34], demonstrated T-cell targeting in humans,[35] and epitopes predicted to bind a wide range of human HLA-DRB1 alleles [31,36]. Hence we considered whether the EsxW-Thr2Ala mutation could affect epitope binding. However, residue two lies in the N-terminal loop of EsxW (see **Fig. 4e**), outside the experimentally demonstrated epitope region (residues 24-34) [37]. We performed *in silico* epitope binding analysis on the Beijing and non-Beijing EsxW protein sequences using the Immune Epitope Database Analysis Resource [38] this predicted HLA binding to the first 9-10 residues of EsxW, but no differences in binding affinities for the wildtype and Thr2Ala mutant alleles. This is consistent with experimental data showing that immunization with ID93 (a recombinant fusion protein containing H37Rv (Lineage 4) wildtype alleles of EsxW/V) is protective against Beijing (2Ala) as well as non-Beijing (2Thr) strains [33,34].

Next we considered whether the EsxW-Thr2Ala mutation could impact on gene expression or protein structure and function. *EsxW* is one of 23 *esx* genes in the *Mtb*

316 genome, including 11 clustered pairs of *esx* genes whose products form heterodimers
317 that are each secreted by one of five type VII secretion systems (T7SS; ESX-1 to
318 ESX-5). The most-studied of these pairs is EsxB (CFP10)/EsxA (ESAT-6), secreted
319 by ESX-1 and encoded in the RD1 locus which also encodes the ESX-1 system.
320 EsxW and its heterodimerization partner EsxV are encoded by adjacent genes in the
321 RD8 locus and secreted by ESX-5. The ESX-5 system is the most recently evolved
322 T7SS in *Mtb* and is present only in the slow-growing Mycobacteria [39,40] (including the
323 *Mtb* complex, *M. leprae*, *M. ulcerans* and *M marinum*). ESX-5 is unique amongst the
324 *Mtb* T7SS in that (a) it secretes most of the PE/PPE proteins [41], which are also unique
325 to slow-growing *Mycobacteria*, comprise a substantial amount of protein coding
326 capacity (~10%) in the *Mtb* genome and play various roles in virulence [42,43]; and (b) it
327 is associated with five pairs of Esx proteins, resulting from duplication and expansion
328 of *esxM* and *esxN* (which are encoded within the locus encoding the ESX-5 T7SS
329 machinery) to create four paralogous copies elsewhere in the *Mtb* genome [44] (**Fig. 4**).
330 Each pair includes a member of the QILSS family (EsxM paralog, including EsxW)
331 and a member of the Mtb9.9 family (EsxN paralog, including EsxV), which partner to
332 form heterodimers. The paralogous proteins differ from one another by just a few
333 amino acids (see tree in **Fig. 4b**), and the reason for this apparent redundancy is not
334 clear. One pair, EsxJ/EsxI, has been shown to play a role in substrate selection for
335 ESX-5 secretion [45], and it is hypothesized that the other paralogs including EsxW/V
336 play a similar role [44]. Notably, two of the loci encoding QILSS paralogs (EsxP/O,
337 RD5; and EsxW/V, RD8) are missing from the *M. bovis* BCG vaccine strain.
338
339 The upstream sequences of *esxW* and its homologs differ substantially (**Fig. 4a**). This
340 suggests their expression is subject to different regulatory controls, which could

provide a reason for their expansion in the *Mtb* genome despite the lack of differentiation at the protein level [46] (tree in **Fig. 4b**); indeed, it has been shown that the different QILSS family members are all expressed at different levels by *Mtb* strain H37Rv during growth in broth, sputum and macrophages [47]. This uniqueness also allows unambiguous read mapping and confident SNP calling at the N-terminal region of EsxW; however we also used PCR and capillary sequencing to confirm the EsxW-2Ala allele in all non-Lineage 2 strains in which the SNP was identified from Illumina reads (see **Methods**). We sought to investigate the expression of QILSS family proteins, and the potential impact of the EsxW-Thr2Ala mutation on gene expression, during growth in macrophages. For these experiments we selected four non-Beijing *Mtb* isolates harbouring the EsxW-2Ala allele, and the closest genetic relative of each with the wildtype allele EsxW-2Thr, and used RNAseq to measure genome-wide *Mtb* expression levels (see **Methods**). The results showed that the ESX-5 system was highly expressed in all isolates following 24h growth in macrophages (mean 0.25% of total *Mtb* RNAseq reads); however the QILSS paralogs were expressed at different levels relative to ESX-5 (**Fig. 4c**). Compared to *esxM*, which is encoded within the ESX-5 locus itself and was expressed at the highest levels in all isolates, *esxK* and *esxP* were expressed at intermediate levels (mean 42% of *esxM* level) and *esxJ* and *esxW* were expressed at low levels (mean 2.2% of *esxM* level). There were no significant differences between the EsxW-2Ala mutant vs wildtype in terms of *Mtb* growth in macrophages (measured at 4, 7 or 11 days post-infection), or in *esxW* (**Fig. 4d**), ESX-5 or global gene expression measured at 24h post-infection. It has been reported that *esxW* was significantly expressed in the lungs of *Mtb* aerosol-infected mice [48]; hence future experiments with wildtype and mutant EsxW in this animal system could potentially help to unravel its functional effects.

Given the proposed role for the EsxW/V heterodimer in substrate selection for ESX-5 secretion [44], we investigated whether the EsxW-Thr2Ala mutation could affect protein structure and function. In *M. canetti*, the reference *Mtb* genome H37Rv, and the majority of non-Beijing *Mtb* isolates, EsxW carries the polar threonine (codon ACC) at residue 2, while the other QILSS proteins in the *Mtb* complex and other slow growing Mycobacteria carry the hydrophobic alanine (GCC) at this position (**Fig. 4b**). In the *Mtb* Beijing lineage, EsxW residue 2 is converted to the more typical residue alanine (GCC), making it identical at the protein level to EsxJ, which has been shown to be involved in substrate selection for ESX-5 secretion [45]. We analysed the quantitative effects of the Thr2Ala mutation on the stability and affinity of the EsxW protomer and the EsxW/V heterodimer, using computational modeling and direct biophysical experiments (see **Methods** and **Fig. 4e**). This indicated that the 2Ala mutation was likely to lead to a mild increase in the affinity and stability of the heterodimer complex (mean $\Delta\Delta G$ 0.28 ± 0.06 Kcal/mol), and this was supported by biophysical measurements that showed the mutant bound slightly stronger to EsxV than the wildtype (K_D^{wt} = 0.6 μM; K_D^{T2A} = 0.4 μM; p=0.05 using two-tailed t-test; **Fig. 4f**). We hypothesise that, if EsxW/V does indeed play a role in substrate selection for ESX-5 secretion, then increased stability of the heterodimer and/or the hydrophobicity of EsxW-2Ala could potentially affect the efficiency of secretion of certain PE/PPE proteins. This could thereby have downstream impacts on one or more of the known functions of ESX-5 including inflammasome activation, IL-1β secretion or escape from macrophages [49], any of which could potentially work to promote transmission between hosts.

While the mechanism remains to be elucidated, our results provide evidence that EsxW is under selection in natural *Mtb* populations. This is consistent with the idea that the protein is important for host interactions, potentially through substrate selection for ESX-5 secretion under certain conditions. Immunizing against EsxW has already been shown to be protective against infection [32-35], and positive selection for this protein suggests vaccines including it are likely to remain effective in the long term. Importantly, our data show that not all *Mtb* are equal: genetically diverse strains display distinct transmission dynamics even within a single localized host population, suggesting that more detailed understanding of lineage-specific variation in *Mtb* could be highly informative for TB control.

Accession codes

Mtb genome data was deposited in NCBI BioProject [ID: PRJNA355614; http://www.ncbi.nlm.nih.gov/bioproject/355614]; individual accession numbers for *Mtb* genomes analysed in this study are given in **Supplementary Tables 1-2** (data from previous studies).

ACKNOWLEDGMENTS

We would like to thank the clinical staff who recruited patients into our study from the following District TB Units (DTU) in HCMC, Viet Nam; District 1, 4, 5, 6, 8, Tan Binh, Binh Thanh and Phu Nhuan DTUs; and also our colleagues from Pham Ngoc Thach Hospital for Tuberculosis and Lung Disease, HCMC Viet Nam. This work was supported by the National Health and Medical Research Council, Australia (Project grant #1056689, Fellowship #1061409 to KEH, Fellowship #1061435 to MI,

Fellowship #1072476 to DBA), A*STAR Biomedical Research Council, Singapore (12/1/21/24/6689) and the Wellcome Trust UK (research training fellowship #081814/Z/06/Z to MC) and as part of their Major Overseas Program in Viet Nam (089276/Z/09/Z).

AUTHOR CONTRIBUTIONS

SJD, KEH, MC, MI, YYT, CCK, are the study principal investigators who conceived and obtained funding for the project. SJD provided overall project co-ordination; MI organized and supervised the DNA sequencing and KEH devised the overall analysis plan and wrote the first draft of the manuscript along with PM. MC and SJD established the TB cohort for this genetics study by working with PVKT, DTMH, NNL, NHL, NTQN, NTTT, GT and JJF to coordinate the collection of clinical samples and phenotypes. KP performed DNA quality checks and genome sequencing on all Vietnamese samples, while VTNH performed Sanger sequencing on selected samples. DBA performed protein structure analyses, and HTH and NTTT performed the macrophage growth and infection experiments of EsxW variants. KEH, PM, MI, DJE, AN analyzed the data. All authors critically reviewed manuscript revisions and contributed intellectual input to the final submission.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Online Methods

Bacterial isolates used in this study. Between December 2008 and July 2011, 2,091 individuals of the Vietnamese Kinh ethnic group attending the outpatient department of Pham Ngoc Thach Hospital or from 8 District Tuberculosis Units (District 1, 4, 5, 6, 8, Tan Binh, Binh Thanh and Phu Nhuan) in HCMC were recruited into a clinical study investigating predictors of failure and relapse in isoniazid resistant TB [50]. The 8 TB units were chosen for inclusion from amongst the 24 servicing HCMC as they are centrally located and close to the TB reference laboratory, which was essential for the logistics of sample collection and processing (see map in **Supplementary Figure 1**). Inclusion criteria were: (1) 18 years or older, (2) negative HIV test, (3) provision of written informed consent, (4) smear positive pulmonary TB. Exclusion criteria were: (1) under 18 years of age, (2) HIV infected, (3) unable or unwilling to provide consent, (5) pregnancy, (6) prior history of TB antibiotic therapy, (7) will receive TB-DOTS (directly observed treatment, short-course) outside the study centres. Over the 2.5 year study period there were N=5036 new smear positive cases of TB (HIV positive and negative) at the district tuberculosis units, of which N=2091 were eligible for recruitment. Of these, N=1822 *Mtb* strains were isolated from the study participants. The annual incidence of pulmonary TB in HCMC is ~82,000, of which ~11,000 are in HCMC; therefore our sample represents ~6.6% of all cases in HCMC during the 2.5-year study period, and >30% of cases in the study districts.

Ethics. The study protocol was approved by the Institutional Research Board of Pham Ngoc Thach Hospital (the supervisory institution of the District TB Units in southern Vietnam), Ho Chi Minh City Health Services and the Oxford University Tropical

461 Research Ethics Committee, UK (Oxtrec 030-07). Written informed consent was
462 obtained from all patients.
463
464 **DNA extraction and sequencing.** *Mtb* isolates were subcultured on Lowenstein
465 Jensen media and DNA extracted at the Oxford University Clinical Research Unit in
466 HCMC using the cetyl trimethylammonium bromide (CTAB) extraction protocol as
467 described previously [51]. DNA was successfully obtained from N=1,728 isolates and
468 shipped to the University of Melbourne for whole genome sequencing. DNA extracts
469 were purified using AxyPrep™ Mag PCR Normalizer Protocol prior to library
470 preparation. A total of N=1,655 DNA samples passed QC, were included for
471 sequencing and subjected to library preparation using the Nextera XT protocol.
472 Libraries were quantified using Quant-iT PicoGreen (dsDNA kit, Invitrogen), then
473 normalised and pooled to 4 nM concentration. DNA underwent 150 bp paired end
474 sequencing (Rapid mode v2) on the Illumina HiSeq 2500 platform (Illumina, San
475 Diego). Sequence data was excluded for N=19 *Mtb* isolates that yielded less than the
476 pre-established criteria of 10x mean read depth across the *Mtb* genome, as SNPs can
477 not be reliably called below this depth. Sequence data was successfully generated for
478 N=1,635 *Mtb* isolates from HCMC (representing 90% of those isolated from eligible
479 patients in the cohort) with median three million reads per sample, providing median
480 99.2% coverage and 86x depth for each *Mtb* genome (**Supplementary Table 1**). To
481 confirm the EsxW codon 2 SNP, we performed PCR and capillary sequencing of the
482 region flanking the SNP in all non-Beijing lineage strains in which the SNP was
483 detected from Illumina data. The primers used were: EsxW F1 (5' TCG CGG TCA
484 ACA CAT CAG AA 3'), EsxW R1 (5' CTG CCC CTC GAT ACG GAT TC 3').
485

Data Availability. *Mtb* genome data generated from 1,635 isolates (Figure 1) has been deposited in NCBI BioProject [accession ID: PRJNA355614; http://www.ncbi.nlm.nih.gov/bioproject/355614]. A total of 3,144 *Mtb* genomes were included in the analysis in Figure 3, comprising data from localized studies: 1,032 from Russia [3], 1,621 from Malawi [4,5], 248 Argentina [17], and 78 from China [18]; as well as 106 globally dispersed Lineage 2 genomes [19] and 59 globally dispersed Lineage 1 genomes [20,21]. Illumina *Mtb* genome sequences from various previously published studies were downloaded from the European Nucleotide Archive (http://www.ebi.ac.uk/ena; individual accession numbers are given in **Supplementary Table 2**). The H37Rv reference genome sequence (https://www.ncbi.nlm.nih.gov/genbank/; accession NC_000962.3) was used for all reference-driven analyses.

SNP analysis. Sequence reads were mapped to the H37Rv reference genome using the RedDog pipeline v0.5 (https://github.com/katholt/RedDog). Briefly, Bowtie2 v2.2.3 was used for read alignment with the sensitive-local algorithm and the maximum insert length set to 2000 (via the -x parameter) [52] and variant sites (i.e. SNPs) were called using SAMTools v0.1.19 [53]. SNPs located in previously reported repetitive regions of the genome were excluded prior to phylogenetic analysis [54,55] (**Supplementary Table 4**); sites for which a definitive allele call could not be made in at least 99.5% of all isolate sequences were also excluded from the set of SNPs used for phylogenetic analysis. Two SNP alignments were compiled for analysis: one comprising the 1,635 HCMC isolates (total 73,718 SNPs), and one comprising all 4,779 isolates (including the HCMC isolates and the global collections downloaded from public data; total 133,492 SNPs).

***In silico* lineage and antimicrobial resistance typing.** Mykrobe Predictor v0.3.6 was used to analyse raw Illumina reads generated from HCMC *Mtb* isolates and (a) assign each isolate to one of the seven *Mtb* lineages, and (b) detect known resistance associated polymorphisms [56] (summarized in **Table 1,** individual mutation calls are provided in **Supplementary Table 1**). All *Mtb* isolates were further assigned to sublineages by comparing SNPs identified using RedDog with those used in the haplotyping scheme defined by Coll *et al.* [6] (lineage assignments are in **Supplementary Tables 1-2**).

Phylogenomic analyses. ML phylogenetic trees were inferred using RAxML v7.7.2 [57] for (a) all HCMC isolates (presented in **Fig. 1**); and (b) each of lineages 1, 2 and 4 using combined data from the HCMC isolates and available public data **(**presented in **Fig. 3;** see isolates list in **Supplementary Tables 1-2**). The trees presented are those with the highest likelihood from 5 replicate runs, constructed using the GTR model of nucleotide substitution and a Gamma model of rate heterogeneity to analyse a concatenated alignment of SNP alleles. An approximate ML tree containing all data (HCMC isolates and available public data) was inferred using FastTree v2.1.8 [58]. Ancestral sequence reconstruction was performed for the HCMC tree and combined tree using FastML v3.1 to infer the sequence alignment at each internal node of the ML phylogeny [59]. Substitution events occurring on each branch of the tree were extracted by comparing the joint reconstruction sequences for the parent and child nodes; these data were used to identify homoplasic SNPs, and to identify lineage-specific polymorphisms as well as independent occurrences of those polymorphisms outside of the lineage of interest (data in **Table 2**). Terminal branch lengths reported

are the number of substitutions (SNPs) mapped to each terminal branch (data in **Fig. 2a**) and were compared to Beijing 2.2.1 sublineage using 2-sample Kolmogorov-Smirnov tests. Metrics for genetic diversity and tree topology were calculated from the phylogenies using R. Node-to-tip distances showed similar variances within groups (standard deviations of 27-75 SNPs) and all groups were compared to Beijing 2.2.1 sublineage using 2-sample Kolmogorov-Smirnov tests. Clusters were defined as subtrees for whom the maximum patristic distance between descendant tips fell below a specified threshold (data in **Fig. 2b-c**). Each cluster was checked to determine whether all members of the cluster shared any of the antimicrobial resistance mutations identified by Mykrobe Predictor; clusters in which no known antimicrobial resistance mutation was conserved in all members of the cluster are reported as not explained by antimicrobial resistance (data in **Supplementary Fig. 3**).

Phylogeography analysis. Transmission between geographical regions was assessed separately for Lineage 1, 2 and 4 trees using an implementation of stochastic mapping on phylogenies (SIMMAP) implemented in the phytools v0.5 package for R [60,61]. Region of origin was treated as a discrete trait and mapped to each tree using the ARD model (which allows each region-to-region transfer rate to vary independently) with 100 replicates. The results reported (**Fig. 3d**) are the median values for the number of transitions to Vietnam from any other region, summarized from 100 replicate mappings for each tree.

Esx sequence analysis. Esx protein sequences were extracted from the H37Rv reference genome using Artemis, aligned using Muscle, and subjected to phylogenetic inference using PhyML v3.0 (tree in **Fig. 4**). DNA sequences flanking the start codon

of each *esx* gene were extracted from the H37Rv reference genome using Artemis and aligned and visualised using JalView v2.6.1 (**Fig. 4**).

Macrophage infections.

Mtb infection of macrophages (THP1 human cell line, 88081201, Sigma-Aldrich) has been previously described [62]. To study *Mtb* growth in macrophages, 2.5 x 10^5 THP1 cells were seeded in 24-well cell culture plates and infected with *Mtb* isolates in triplicate at a multiplicity of infection (MOI) of 1. At 4h post infection extracellular bacteria were removed by washing. At 0, 4, 7 and 11 days post-infection, intracellular bacteria were harvested and plated on Middlebrook 7H10 agar plates. Colony forming unit counts resulted after 3 to 4 weeks incubation at 37^0C. To measure genome-wide *Mtb* expression levels, confluent monolayers of 1.5 x 10^7 THP1 in 75cm^2 cell culture flasks were infected with *Mtb* isolates at MOI 4. At 4h post infection extracellular bacteria were removed by washing. At 24h post-infection, intracellular bacteria were stabilized and collected in guanidine thiocyanate-based lysis buffer (4 M guanidine thiocyanate, 0.5% Na N-lauryl sarcosine, 25 mM sodium citrate, and 0.1 M β-mercaptoethanol). Total RNA from was then extracted using a previously published method [63].

RNAseq analysis. RNA samples were subjected to reverse transcription and the resulting cDNA was sequenced via Illumina HiSeq (100 bp paired end) at Macrogen. The resulting reads were first mapped to the H37Rv *Mtb* reference genome to retrieve bacterial sequences, resulting in 6-19 million bacterial sequence reads per sample. A strain-specific reference genome was prepared for each of the eight *Mtb* isolates, by polishing the H37Rv genome sequence with the Illumina reads previously obtained by

sequencing whole genomic DNA extracted from that isolate (i.e. the data used for phylogenomic analyses) using Pilon (v1.22) with default settings. For each isolate, the RNAseq reads were mapped to the isolate's own reference genome sequence using Bowtie2 (v2.2.3) [52]. Read alignments were visualised using the BamView function in v14.0.0 [64], which was also used to filter reads to include only those with mapping score ≥30 for both forward and reverse reads, and to extract read counts for regions of interest. Expression values for the ESX-5 locus were calculated as the total reads mapping to the region spanning from *pe19* to *eccA5* (coordinates 2029880 to 2038532 in H37Rv). Due to the very close sequence homology in the protein-coding regions of the QILSS genes, it was not possible to assess their expression by counting total reads mapping to the length of the gene. Instead, we assessed relative expression of these genes by extracting the number of reads mapping to a unique 30 bp marker region spanning the start codon of each paralog (coordinates -21 to +9 relative to the first base pair in each protein-coding sequence; see **Fig. 4** for uniqueness of these regions). To facilitate comparison of QILSS gene expression across strains with different RNAseq library sizes, we normalised the read counts for each isolate by the total number of reads mapping to the ESX-5 locus from *pe19* to *eccA5* (coordinates 2029880 to 2038532 in H37Rv). The ESX-5 locus was chosen for this purpose as all QILSS proteins are secreted by ESX-5 and thus depend on its expression (note the locus contains the QILSS gene *esxM*, which was the highest expressed of all QILSS genes in all isolates). The ratios of reads mapping to each QILSS gene vs ESX-5 locus in each strain *i* (x_i/ESX-5_i) is shown in **Figure 4**. To investigate differences in QILSS gene expression between wildtype (*i*) and mutant (*j*) isolate pairs (**Fig. 4**), we calculated the difference in these ratios relative to the wildtype level, as follows: (x_i/ESX-5_i – x_j/ESX-5_j)/(x_i/ESX-5_i).

Computational protein structure analysis. Structure guided approaches can provide valuable insight into the molecular mechanism of mutations and their role in diseases [21,65-74] [75]. To evaluate the structural effects of the Thr2Ala mutation in EsxW, models of EsxW and EsxV were generated using Modeller v9.19[76] and MacroModel (vSchrodinger 2017, New York, NY), based on the experimental structures of homologous ESAT-6-like complexes (PDB IDs: 1WA8, 2KG7, 2VS0, 3OGI, 3ZBH, 4GZR, 4IOG and 4LWS; sequence identities ranging from 20-91%). The models were then minimized using the MMF94s forcefield in Sybyl-X 2.1.1 (Certara L.P., St Louis, MO), with the final structure having more than 95% of residues in the allowed region of a Ramachandran plot. The quality of the models was confirmed with Verify3D v1 (data not shown). Model structures were examined using Pymol v1.9. The model of the EsxW-EsxV complex was built using the homologous complexes to guide protein docking. The EsxW protomer and EsxW-EsxV heterodimer model were subjected to molecular dynamics to generate a series of representative conformations using Desmond (vSchrodinger 2017, New York, NY). Sodium and chloride ions were added to reach a final concentration of 150 mM Na^+ and the system was solvated. The Amber and Charmm36 force fields were applied to the system and the default Desmond minimization and equilibration procedure was followed. Simulations were kept at constant pressure (1 atom) and temperature (335 K). The structural consequences of the EsxW Thr2Ala variant were analysed to account for all the potential effects of the mutations [77]. The effects of the variant upon the stability of EsxW was predicted using SDM v2 [78], mCSM-Stability v1 [79] and DUET v1 [80]. The effect of the mutations upon the affinity of EsxW to bind to EsxV, and the stability of the complex, were predicted using mCSM-PPI v1 [79]. The predictions were analysed

and averaged across the conformational landscape. These computational approaches represent the wild-type structural and chemical environment of a residue as a graph based signature in order to determine the change upon mutation in Gibb's Free Energy of stability or binding. Reported results are based on analysis of EsxW (wildtype and Thr2Ala mutant) and EsxV amino acid sequences encoded in the H37Rv (lineage 4) reference genome; the only difference between these dimers and those encoded in lineage 1 or 2 is at EsxV residues 20 and 23. Structural modelling with the lineage 1/2 EsxV sequence yielded the same results as the analyses conducted with the lineage 4 background (i.e. stronger binding affinity observed with the EsxW Thr2Ala mutant).

Biophysical Measurements. EsxV and wildtype (2Thr) and mutant (2Ala) EsxW were cloned into a pET18 vector with a C-terminal hexahistidine tag (again using the sequences encoded in H37Rv (lineage 4) reference genome). A di-lysine motif was added prior to the hexahistidine tag of EsxW to facilitate labelling on the extended C-terminus. Protein was expressed, purified and refolded following the method of Mahmood and collegues [81]. Solution microscale thermophoresis (MST) binding studies were performed to measure the binding affinity of EsxV for both the wildtype and mutant EsxW, using standard protocols on a Monolith NT.115 (Nanotemper Technologies) [82-84]. The EsxW was chosen to be labelled due to the presence of a lysine in EsxV at the heterodimer binding interface. As there were no lysines present in EsxW, the dilysine motif was added to the C-terminal projection to facilitate labelling using a RED-NHS (Amine Reactive) Protein Labeling Kit (Nanotemper Technologies), which contains an NT-647 dye, as per the manufacturer's instructions. Labelled either wildtype or 2Ala mutant EsxW was mixed with EsxV in PBS with

0.05% Tween-20. Each replicate was undertaken using a 16-step 2-fold serial dilution series. The EsxW protein concentration was chosen such that the observed fluorescence was approximately 400 units at 70% LED power. The samples were loaded into standard capillaries and heated at 40% laser power (48 mW) for 30 s, followed by 5 s cooling. The data were normalized against the baseline obtained in the absence of any EsxV, and the maximal response obtained at the highest concentration of inhibitor. The dissociation constant K_D was obtained by plotting the normalized fluorescence (F_{norm}) against the logarithm of the concentrations of the dilution series and resulted in a sigmoidal binding curve that could be directly fitted with a nonlinear solution of the law of mass action (**Figure 4**). All experiments were performed with four replicates, and the normalized fluorescence thermophoresis curves were analyzed using GraphPad v6 (GraphPad, San Diego, CA, USA).

675 **References**

676

677 1. Zumla, A. *et al.* Eliminating tuberculosis and tuberculosis-HIV co-disease
678 in the 21st century: key perspectives, controversies, unresolved issues,
679 and needs. *J Infect Dis* **205 Suppl 2**, S141-6 (2012).
680 2. World Health Organisation. WHO | Global tuberculosis report 2017.
681 (World Health Organization, 2017).
682 3. Casali, N. *et al.* Evolution and transmission of drug-resistant tuberculosis
683 in a Russian population. *Nat Genet* **46**, 279-86 (2014).
684 4. Guerra-Assuncao, J.A. *et al.* Large-scale whole genome sequencing of M.
685 tuberculosis provides insights into transmission in a high prevalence
686 area. *Elife* **4**(2015).
687 5. Guerra-Assuncao, J.A. *et al.* Recurrence due to Relapse or Reinfection With
688 Mycobacterium tuberculosis: A Whole-Genome Sequencing Approach in a
689 Large, Population-Based Cohort With a High HIV Infection Prevalence and
690 Active Follow-up. *J Infect Dis* (2014).
691 6. Coll, F. *et al.* A robust SNP barcode for typing Mycobacterium tuberculosis
692 complex strains. *Nat Commun* **5**, 4812 (2014).
693 7. Anh, D.D. *et al.* Mycobacterium tuberculosis Beijing genotype emerging in
694 Vietnam. *Emerg Infect Dis* **6**, 302-5 (2000).
695 8. Buu, T.N. *et al.* The Beijing genotype is associated with young age and
696 multidrug-resistant tuberculosis in rural Vietnam. *Int J Tuberc Lung Dis*
697 **13**, 900-6 (2009).
698 9. Maeda, S. *et al.* Mycobacterium tuberculosis strains spreading in Hanoi,
699 Vietnam: Beijing sublineages, genotypes, drug susceptibility patterns, and
700 host factors. *Tuberculosis (Edinb)* **94**, 649-56 (2014).
701 10. Nguyen, V.A. *et al.* High prevalence of Beijing and EAI4-VNM genotypes
702 among M. tuberculosis isolates in northern Vietnam: sampling effect, rural
703 and urban disparities. *PLoS One* **7**, e45553 (2012).
704 11. Nguyen, V.A. *et al.* Mycobacterium tuberculosis lineages and anti-
705 tuberculosis drug resistance in reference hospitals across Viet Nam. *BMC*
706 *Microbiol* **16**, 167 (2016).
707 12. Horton, K.C., MacPherson, P., Houben, R.M., White, R.G. & Corbett, E.L. Sex
708 Differences in Tuberculosis Burden and Notifications in Low- and Middle-
709 Income Countries: A Systematic Review and Meta-analysis. *PLoS Med* **13**,
710 e1002119 (2016).
711 13. Malla, B. *et al.* First insights into the phylogenetic diversity of
712 Mycobacterium tuberculosis in Nepal. *PLoS One* **7**, e52297 (2012).
713 14. Lee, C.H. *et al.* Treatment delay and fatal outcomes of pulmonary
714 tuberculosis in advanced age: a retrospective nationwide cohort study.
715 *BMC Infect Dis* **17**, 449 (2017).
716 15. Ford, C.B. *et al.* Mycobacterium tuberculosis mutation rate estimates from
717 different lineages predict substantial differences in the emergence of
718 drug-resistant tuberculosis. *Nat Genet* **45**, 784-90 (2013).
719 16. Wlodarska, M., Johnston, J.C., Gardy, J.L. & Tang, P. A microbiological
720 revolution meets an ancient disease: improving the management of
721 tuberculosis with genomics. *Clin Microbiol Rev* **28**, 523-39 (2015).

17. Eldholm, V. *et al.* Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. *Nat Commun* **6**, 7119 (2015).
18. Zhang, H. *et al.* Genome sequencing of 161 Mycobacterium tuberculosis isolates from China identifies genes and intergenic regions associated with drug resistance. *Nat Genet* **45**, 1255-60 (2013).
19. Merker, M. *et al.* Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage. *Nat Genet* **47**, 242-9 (2015).
20. Comas, I. *et al.* Population Genomics of Mycobacterium tuberculosis in Ethiopia Contradicts the Virgin Soil Hypothesis for Human Tuberculosis in Sub-Saharan Africa. *Curr Biol* **25**, 3260-6 (2015).
21. Phelan, J. *et al.* Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. *BMC Med* **14**, 31 (2016).
22. Stucki, D. *et al.* Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. *Nat Genet* (2016).
23. Coscolla, M. & Gagneux, S. Consequences of genomic diversity in Mycobacterium tuberculosis. *Semin Immunol* **26**, 431-44 (2014).
24. Fox, G.J., Barry, S.E., Britton, W.J. & Marks, G.B. Contact investigation for tuberculosis: a systematic review and meta-analysis. *Eur Respir J* **41**, 140-56 (2013).
25. Lonnroth, K. *et al.* Systematic screening for active tuberculosis: rationale, definitions and key considerations. *Int J Tuberc Lung Dis* **17**, 289-98 (2013).
26. Hanekom, M. *et al.* Mycobacterium tuberculosis Beijing genotype: a template for success. *Tuberculosis (Edinb)* **91**, 510-23 (2011).
27. Parwati, I., van Crevel, R. & van Soolingen, D. Possible underlying mechanisms for successful emergence of the Mycobacterium tuberculosis Beijing genotype strains. *Lancet Infect Dis* **10**, 103-11 (2010).
28. van Laarhoven, A. *et al.* Low induction of proinflammatory cytokines parallels evolutionary success of modern strains within the Mycobacterium tuberculosis Beijing genotype. *Infect Immun* **81**, 3750-6 (2013).
29. Farhat, M.R. *et al.* Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. *Nat Genet* **45**, 1183-9 (2013).
30. Hazbon, M.H. *et al.* Convergent evolutionary analysis identifies significant mutations in drug resistance targets of Mycobacterium tuberculosis. *Antimicrob Agents Chemother* **52**, 3369-76 (2008).
31. Knudsen, N.P. *et al.* Tuberculosis vaccine with high predicted population coverage and compatibility with modern diagnostics. *Proc Natl Acad Sci U S A* **111**, 1096-101 (2014).
32. Baldwin, S.L. *et al.* Intradermal immunization improves protective efficacy of a novel TB vaccine candidate. *Vaccine* **27**, 3063-71 (2009).
33. Baldwin, S.L. *et al.* Protection and Long-Lived Immunity Induced by the ID93/GLA-SE Vaccine Candidate against a Clinical Mycobacterium tuberculosis Isolate. *Clin Vaccine Immunol* **23**, 137-47 (2015).

34. Bertholet, S. *et al.* A defined tuberculosis vaccine candidate boosts BCG and protects against multidrug-resistant Mycobacterium tuberculosis. *Sci Transl Med* **2**, 53ra74 (2010).
35. Bertholet, S. *et al.* Identification of human T cell antigens for the development of vaccines against Mycobacterium tuberculosis. *J Immunol* **181**, 7948-57 (2008).
36. Uplekar, S., Heym, B., Friocourt, V., Rougemont, J. & Cole, S.T. Comparative genomics of Esx genes from clinical isolates of Mycobacterium tuberculosis provides evidence for gene conversion and epitope variation. *Infect Immun* **79**, 4042-9 (2011).
37. Lewinsohn, D.M. *et al.* Human Mycobacterium tuberculosis CD8 T Cell Antigens/Epitopes Identified by a Proteomic Peptide Library. *PLoS One* **8**, e67016 (2013).
38. Kim, Y. *et al.* Immune epitope database analysis resource. *Nucleic Acids Res* **40**, W525-30 (2012).
39. Mortimer, T.D., Weber, A.M. & Pepperell, C.S. Evolutionary Thrift: Mycobacteria Repurpose Plasmid Diversity during Adaptation of Type VII Secretion Systems. *Genome Biol Evol* **9**, 398-413 (2017).
40. Gey Van Pittius, N.C. *et al.* The ESAT-6 gene cluster of Mycobacterium tuberculosis and other high G+C Gram-positive bacteria. *Genome Biol* **2**, RESEARCH0044 (2001).
41. Abdallah, A.M. *et al.* PPE and PE_PGRS proteins of Mycobacterium marinum are transported via the type VII secretion system ESX-5. *Mol Microbiol* **73**, 329-40 (2009).
42. Fishbein, S., van Wyk, N., Warren, R.M. & Sampson, S.L. Phylogeny to function: PE/PPE protein evolution and impact on Mycobacterium tuberculosis pathogenicity. *Mol Microbiol* **96**, 901-16 (2015).
43. Brennan, M.J. The Enigmatic PE/PPE Multigene Family of Mycobacteria and Tuberculosis Vaccination. *Infect Immun* **85**(2017).
44. Groschel, M.I., Sayes, F., Simeone, R., Majlessi, L. & Brosch, R. ESX secretion systems: mycobacterial evolution to counter host immunity. *Nat Rev Microbiol* **14**, 677-691 (2016).
45. Shah, S., Cannon, J.R., Fenselau, C. & Briken, V. A Duplicated ESAT-6 Region of ESX-5 Is Involved in Protein Export and Virulence of Mycobacteria. *Infect Immun* **83**, 4349-61 (2015).
46. Kumar, A., Chandolia, A., Chaudhry, U., Brahmachari, V. & Bose, M. Comparison of mammalian cell entry operons of mycobacteria: in silico analysis and expression profiling. *FEMS Immunol Med Microbiol* **43**, 185-95 (2005).
47. Bukka, A., Price, C.T., Kernodle, D.S. & Graham, J.E. Mycobacterium tuberculosis RNA Expression Patterns in Sputum Bacteria Indicate Secreted Esx Factors Contributing to Growth are Highly Expressed in Active Disease. *Front Microbiol* **2**, 266 (2012).
48. Coppola, M. *et al.* New Genome-Wide Algorithm Identifies Novel In-Vivo Expressed Mycobacterium Tuberculosis Antigens Inducing Human T-Cell Responses with Classical and Unconventional Cytokine Profiles. *Sci Rep* **6**, 37793 (2016).

49. Abdallah, A.M. *et al.* Mycobacterial secretion systems ESX-1 and ESX-5 play distinct roles in host cell death and inflammasome activation. *J Immunol* **187**, 4744-53 (2011).
50. Thai, P.V.K. *et al.* Bacterial risk factors for treatment failure and relapse among patients with Isoniazid resistant tuberculosis. *BMC Infectious Diseases* **in press**(2018).
51. Caws, M. *et al.* The influence of host and bacterial genotype on the development of disseminated disease with Mycobacterium tuberculosis. *PLoS Pathog* **4**, e1000034 (2008).
52. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-9 (2012).
53. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
54. Comas, I. *et al.* Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nat Genet* **45**, 1176-82 (2013).
55. Pepperell, C.S. *et al.* The role of selection in shaping diversity of natural M. tuberculosis populations. *PLoS Pathog* **9**, e1003543 (2013).
56. Bradley, P. *et al.* Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nat Commun* **6**, 10063 (2015).
57. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-3 (2014).
58. Price, M.N., Dehal, P.S. & Arkin, A.P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
59. Ashkenazy, H. *et al.* FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Res* **40**, W580-4 (2012).
60. Bollback, J.P. SIMMAP: stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics* **7**, 88 (2006).
61. Revell, L. phytools: an R package for phylogenetic compaitive biology (and other things). *Methods in Ecology and Evolution* **3**, 217-223 (2012).
62. Vijay, S.V., D.N; Hai, H.T.; Ha, V.T.H.; Dung, V.T.M.; Dinh, T.D.; Nhung, H.N.; Tram, T.T.B.; Aldridge, B.B.; Hanh, N.T.; Thu, D.D.A.; Phu, N.H.; Thwaites, G.E.; Thuong, N.T.T. Influence of Stress and Antibiotic Resistance on Cell-Length Distribution in Mycobacterium tuberculosis Clinical Isolates. *Frontiers in Microbiology* **8**, 1-12 (2017).
63. Rohde, K.H., Abramovitch, R.B. & Russell, D.G. Mycobacterium tuberculosis invasion of macrophages: linking bacterial gene expression to environmental cues. *Cell Host Microbe* **2**, 352-64 (2007).
64. Carver, T. *et al.* BamView: visualizing and interpretation of next-generation sequencing read alignments. *Brief Bioinform* **14**, 203-12 (2013).
65. Jafri, M. *et al.* Germline Mutations in the CDKN2B Tumor Suppressor Gene Predispose to Renal Cell Carcinoma. *Cancer Discov* **5**, 723-9 (2015).
66. Usher, J.L. *et al.* Analysis of HGD Gene Mutations in Patients with Alkaptonuria from the United Kingdom: Identification of Novel Mutations. *JIMD Rep* **24**, 3-11 (2015).

67. Jubb, H.C. *et al.* Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health. *Prog Biophys Mol Biol* **128**, 3-13 (2017).
68. Kano, F.S. *et al.* The Presence, Persistence and Functional Properties of Plasmodium vivax Duffy Binding Protein II Antibodies Are Influenced by HLA Class II Allelic Variants. *PLoS Negl Trop Dis* **10**, e0005177 (2016).
69. Nemethova, M. *et al.* Twelve novel HGD gene variants identified in 99 alkaptonuria patients: focus on 'black bone disease' in Italy. *Eur J Hum Genet* **24**, 66-72 (2016).
70. Silvino, A.C. *et al.* Variation in Human Cytochrome P-450 Drug-Metabolism Genes: A Gateway to the Understanding of Plasmodium vivax Relapses. *PLoS One* **11**, e0160172 (2016).
71. White, R.R. *et al.* Ubiquitin-Dependent Modification of Skeletal Muscle by the Parasitic Nematode, Trichinella spiralis. *PLoS Pathog* **12**, e1005977 (2016).
72. Albanaz, A.T.S., Rodrigues, C.H.M., Pires, D.E.V. & Ascher, D.B. Combating mutations in genetic disease and drug resistance: understanding molecular mechanisms to guide drug design. *Expert Opin Drug Discov* **12**, 553-563 (2017).
73. Casey, R.T. *et al.* SDHA related tumorigenesis: a new case series and literature review for variant interpretation and pathogenicity. *Mol Genet Genomic Med* **5**, 237-250 (2017).
74. Pandurangan, A.P., Ascher, D.B., Thomas, S.E. & Blundell, T.L. Genomes, structural biology and drug discovery: combating the impacts of mutations in genetic disease and antibiotic resistance. *Biochem Soc Trans* **45**, 303-311 (2017).
75. Soardi, F.C. *et al.* Familial STAG2 germline mutation defines a new human cohesinopathy. *Npj Genomic Medicine* **2**(2017).
76. Sali, A. & Blundell, T.L. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779-815 (1993).
77. Pires, D.E., Chen, J., Blundell, T.L. & Ascher, D.B. In silico functional dissection of saturation mutagenesis: Interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. *Sci Rep* **6**, 19848 (2016).
78. Pandurangan, A.P., Ochoa-Montano, B., Ascher, D.B. & Blundell, T.L. SDM: a server for predicting effects of mutations on protein stability. *Nucleic Acids Res* (2017).
79. Pires, D.E., Ascher, D.B. & Blundell, T.L. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* **30**, 335-42 (2014).
80. Pires, D.E., Ascher, D.B. & Blundell, T.L. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res* **42**, W314-9 (2014).
81. Mahmood, A. *et al.* Molecular characterization of secretory proteins Rv3619c and Rv3620c from Mycobacterium tuberculosis H37Rv. *FEBS J* **278**, 341-53 (2011).
82. Ascher, D.B. *et al.* Potent hepatitis C inhibitors bind directly to NS5A and reduce its affinity for RNA. *Sci Rep* **4**, 4765 (2014).

83. Chan, L.J. *et al.* Conjugation of 10 kDa Linear PEG onto Trastuzumab Fab' Is Sufficient to Significantly Enhance Lymphatic Exposure while Preserving in Vitro Biological Activity. *Mol Pharm* **13**, 1229-41 (2016).
84. Chan, L.J. *et al.* PEGylation does not significantly change the initial intravenous or subcutaneous pharmacokinetics or lymphatic exposure of trastuzumab in rats but increases plasma clearance after subcutaneous administration. *Mol Pharm* **12**, 794-809 (2015).

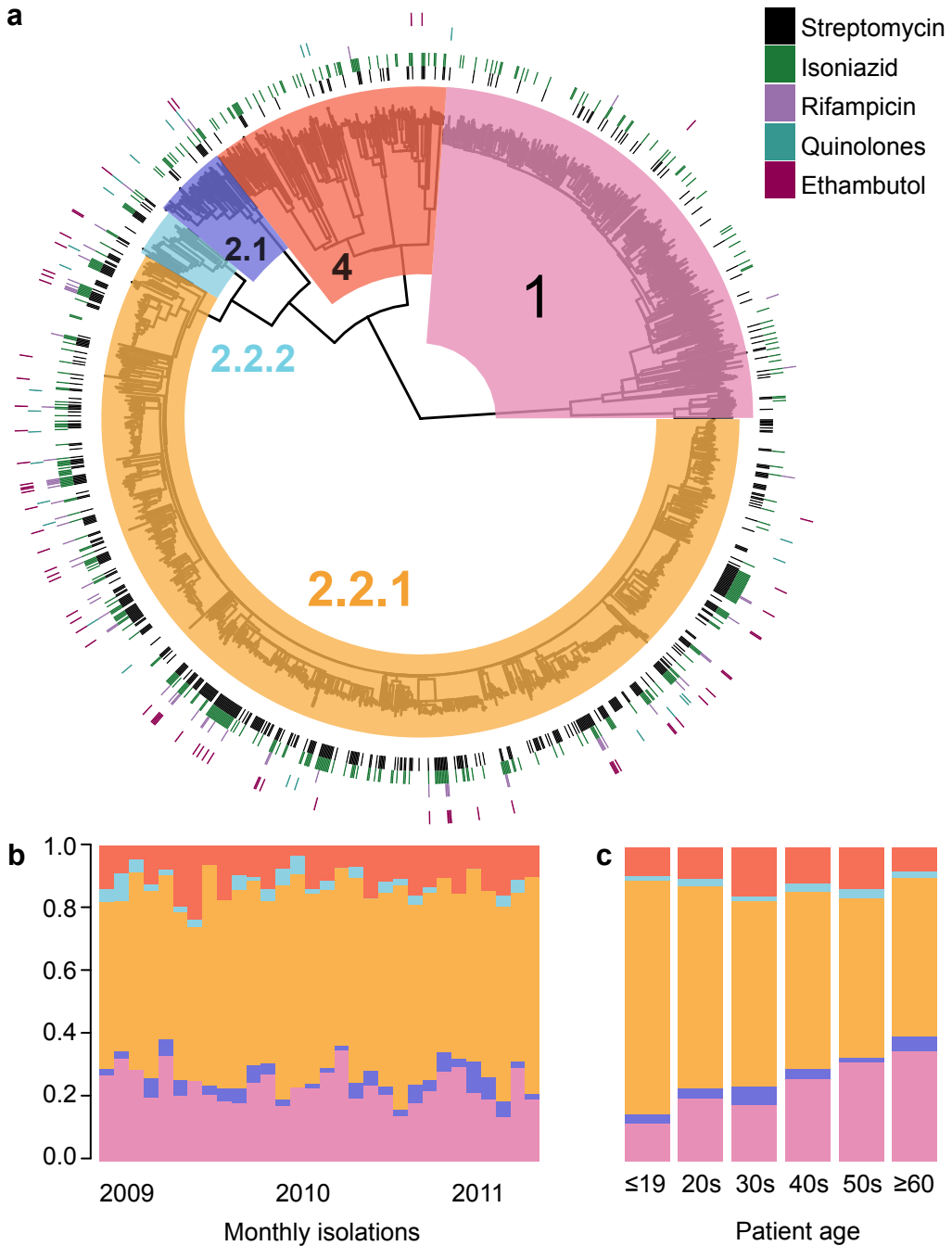
Figure Legends

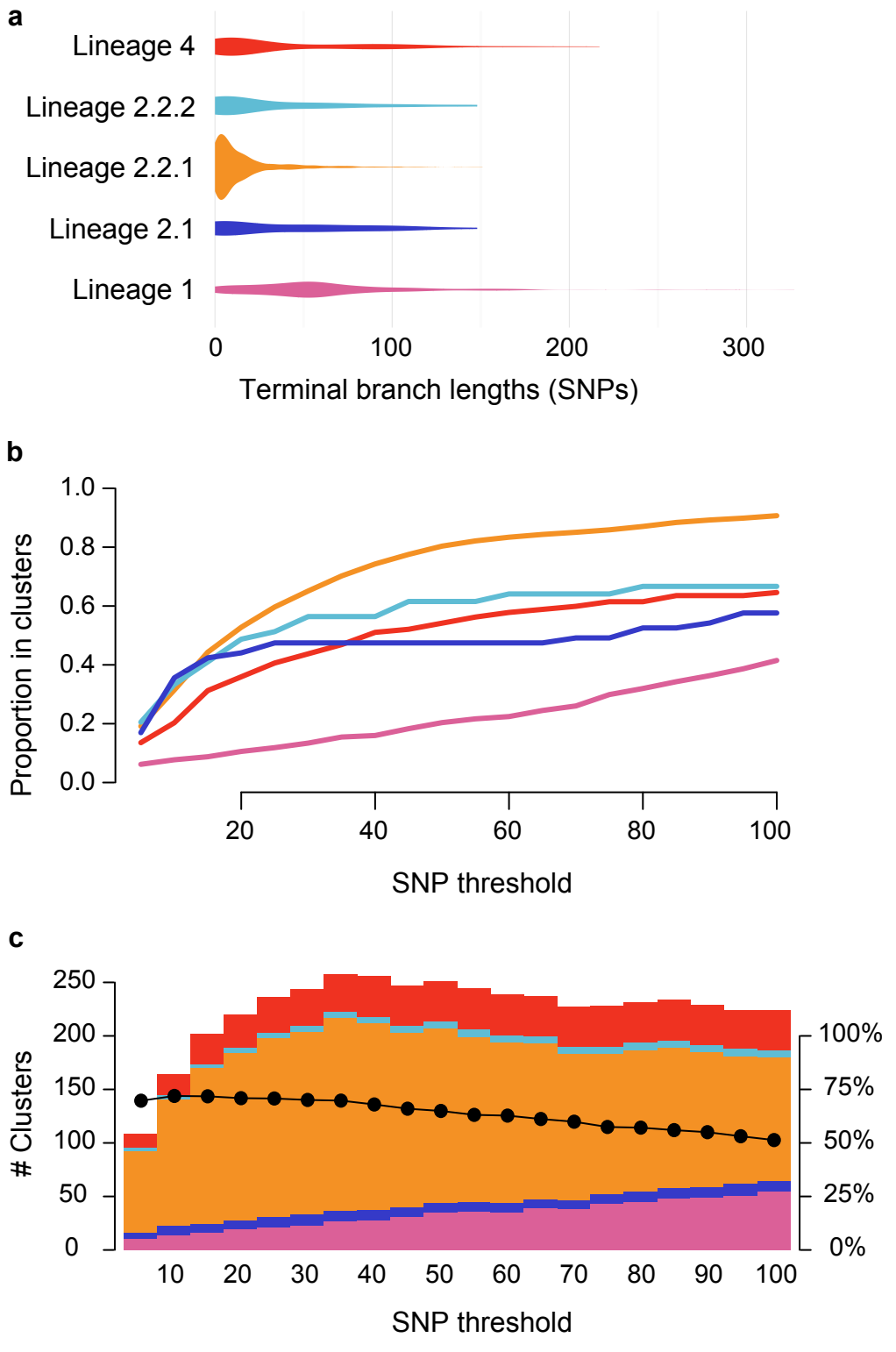
Figure 1. Circulating *M. tuberculosis* strains in HCMC are divided into multiple distinct lineages. **(a)** Maximum-likelihood phylogeny of 1635 *Mtb* isolates collected from TB patients in HCMC, with backgrounds shaded by lineage. Exterior rings indicate presence of known antimicrobial resistance-associated mutations (coloured by drug, according to legend in top right). **(b)** Frequency distribution of lineages by month. **(c)** Frequency distribution of lineages by patient age group.

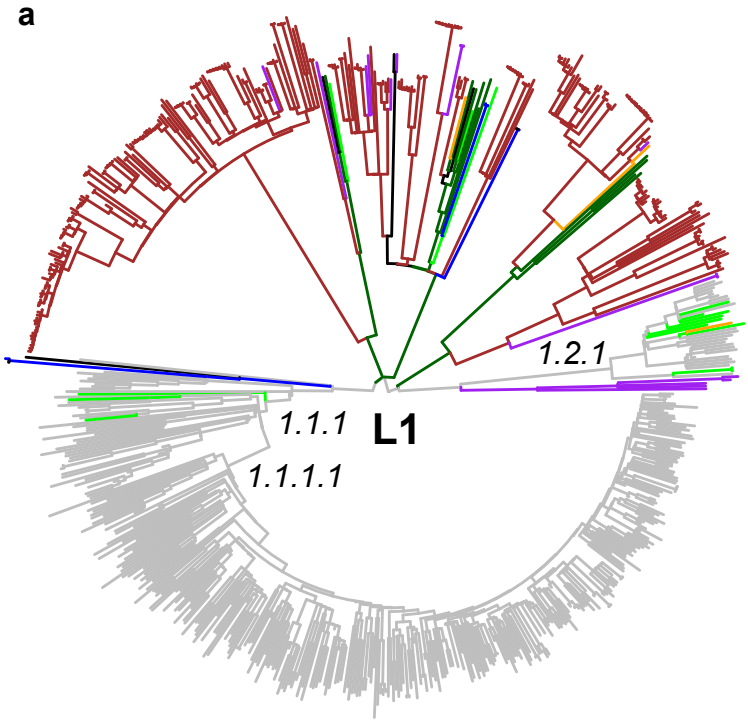
Figure 2. Properties of lineage subtrees for HCMC *M. tuberculosis* genomes. **(a)** Distributions of terminal branch lengths for the 1635-strain phylogeny. **(b)** Mean subtree heights (y-axis; measured as mean node-to-tip distances for each subtree) vs subtree size (x-axis; number of descendant tips). Shaded region indicates standard error of the mean across subtrees of a given size; labels indicate lineage. **(c)** Stacked area plot showing number of clusters (y-axis) within each lineage (coloured as in panel a) identified using different maximum patristic distance thresholds to define clusters (x-axis).

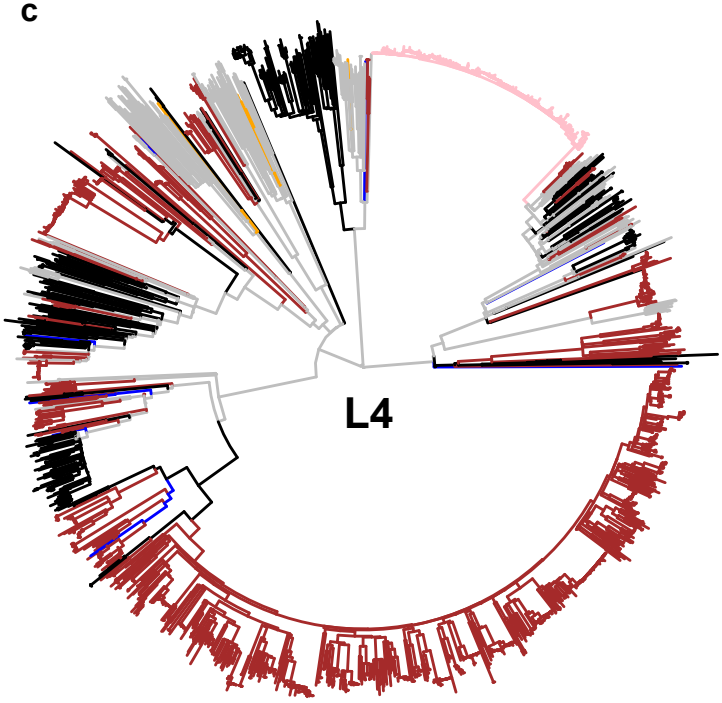
Figure 3. Phylogenies of *M. tuberculosis* showing relationships between isolates from HCMC and other locations. HCMC isolates are coloured grey, isolates from four other localised studies are coloured as in panel (d), other locations are shown in black. **(a)** Lineage 1 (n=675 genomes). **(b)** Lineage 2 (n=1871 genomes). **(c)** Lineage 4 (n=2066 genomes). **(d)** Number of transfers between Vietnam and other locations predicted by stochastic mapping of locations onto the Lineage 2 and 4 trees.

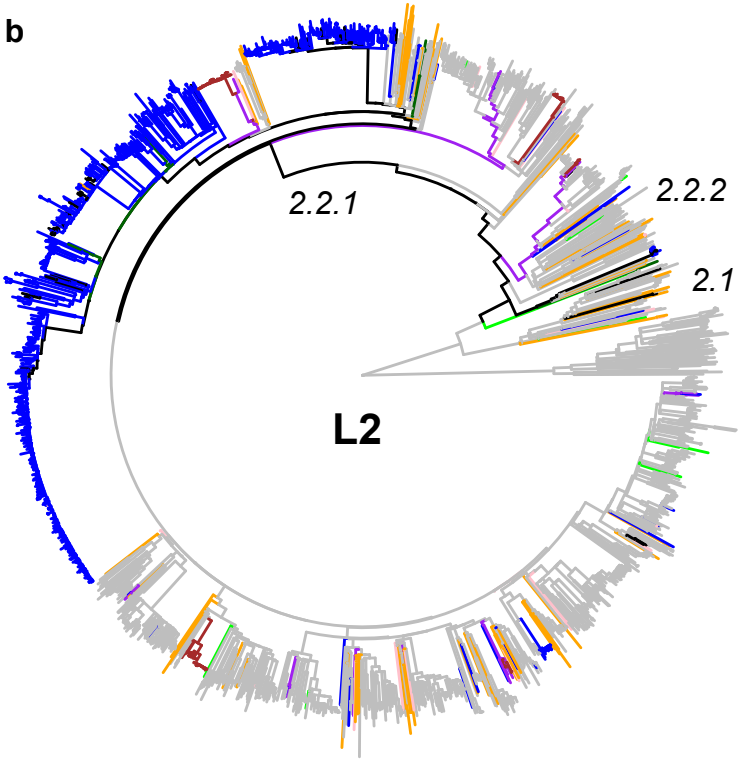
Figure 4. Details of the EsxW mutation at the gene, mRNA, protein and heterodimer level. (a) Variation in the promoter region of *esxW* and the other CFP10 paralogs extracted from the *Mtb* H37Rv (lineage 4) reference genome. Sites are coloured by conservation, coordinates are relative to the start codon. **(b)** Variation in the protein coding region. Tree shows maximum likelihood phylogeny inferred from amino acid sequences; box indicates QILSS proteins, which are secreted via ESX-5. Alignment of the start of the protein-coding DNA sequence is shown, coloured by conservation; arrow indicates site of homoplasic SNP in *esxW*, which in the Beijing lineage and in some lineage 1 or 4 mutants has undergone a substitution of A to G (which appears to be a reversion to the ancestral allele for CFP10 paralogs), resulting in an amino acid substitution of Thr to Ala at residue 2 of the encoded EsxW protein. **(c-d)** RNAseq results for *esxW* and QILSS/ESX-5 paralogs measured in 4 pairs of *Mtb* isolates, each including a lineage 1 or 4 EsxW-2Ala mutant and its genetically closest EsxW-2Thr relative, following 24 h macrophage infection. mRNA expression levels were estimated from read counts uniquely mapping to the region -21 to +9 for each gene; normalized to total reads uniquely mapping to the locus encoding the ESX-5 machinery in each isolate (note *esxM* is encoded within this locus). **(d)** Difference between 2Ala mutant and 2Thr wildtype for each pair, relative to wildtype expression level; and locations of the genes in the reference genome. **(e)** Structural model of the EsxW-EsxV heterodimer. EsxV is shown as a surface (grey) and EsxW as a ribbon (red); key residues in EsxW are shown as sticks and labelled. **(f)** Comparison of biophysical measurements of heterodimer binding affinity between wildtype and mutant EsxW. All binding curves were determined across four replicates by microscale thermophoresis (MST), and are represented as the mean ± standard deviation of 4 replicates.


a
Streptomycin
Isoniazid
Rifampicin
Quinolones
Ethambutol
1
2.1
2.2.2
4
2.2.1
b
1.0
0.8
0.6
0.4
0.2
0.0
2009
2010
2011
Monthly isolations
c
≤19
20s
30s
40s
50s
≥60
Patient age


a
Lineage 4
Lineage 2.2.2
Lineage 2.2.1
Lineage 2.1
Lineage 1
0
100
200
300
Terminal branch lengths (SNPs)
b
Proportion in clusters
1.0
0.8
0.6
0.4
0.2
0.0
20
40
60
80
100
SNP threshold
c
Clusters
250
200
150
100
50
0
100%
75%
50%
25%
0%
10
20
30
40
50
60
70
80
90
100
SNP threshold



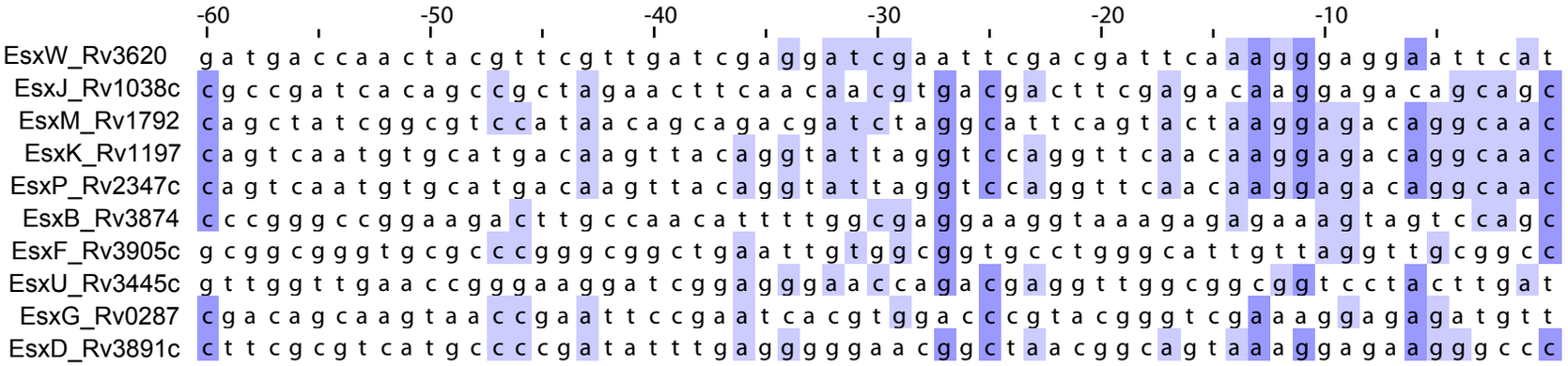




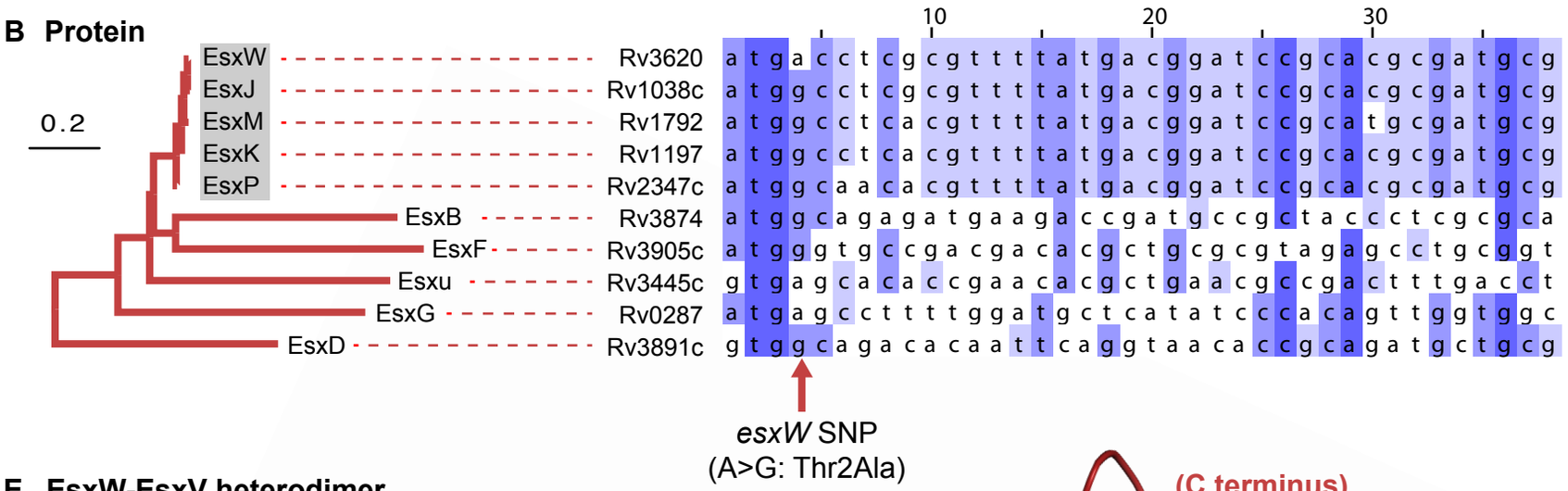
d

Origin	Lineage 1	2	4
China	1	57	4
Europe	0	15	3
Russia	0	10	33
South-East Asia	6	9	0
Malawi	2	3	15
Other Africa	0	7	0
Micronesia	0	2	0
Southern Asia	0	2	0
South America	0	1	1

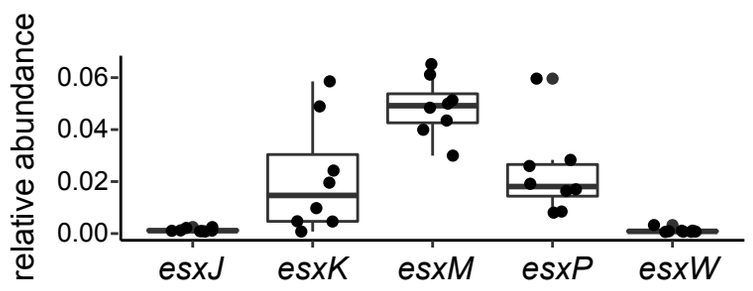
A Promoter region



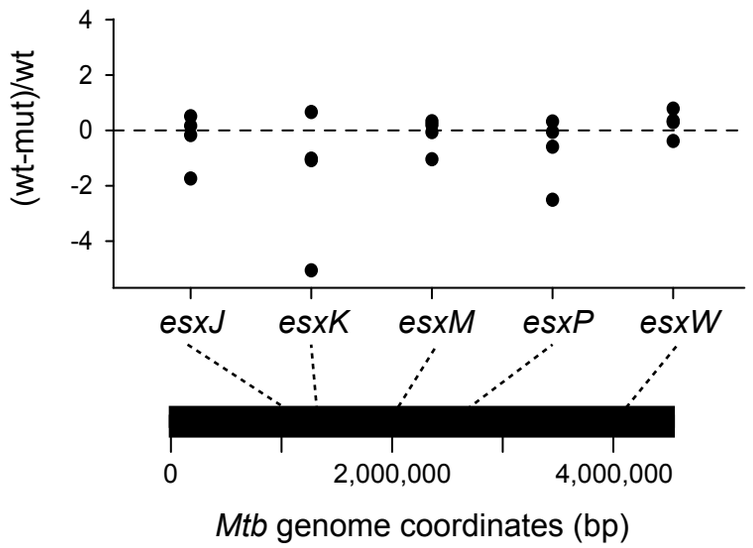
B Protein



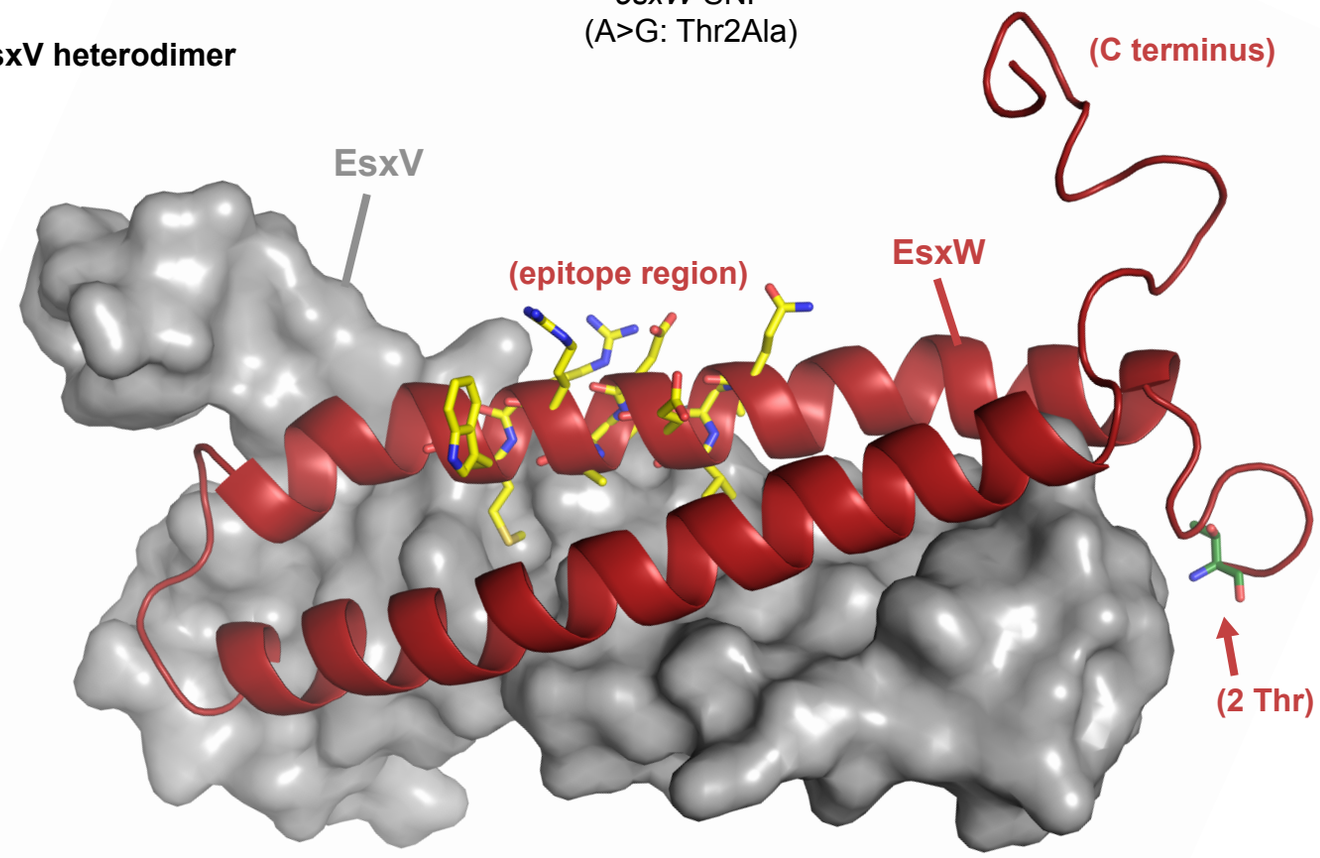
C Expression relative to ESX-5 locus



D Expression (*esxW* wt vs mut)



E EsxW-EsxV heterodimer



F EsxW-EsxV binding affinity

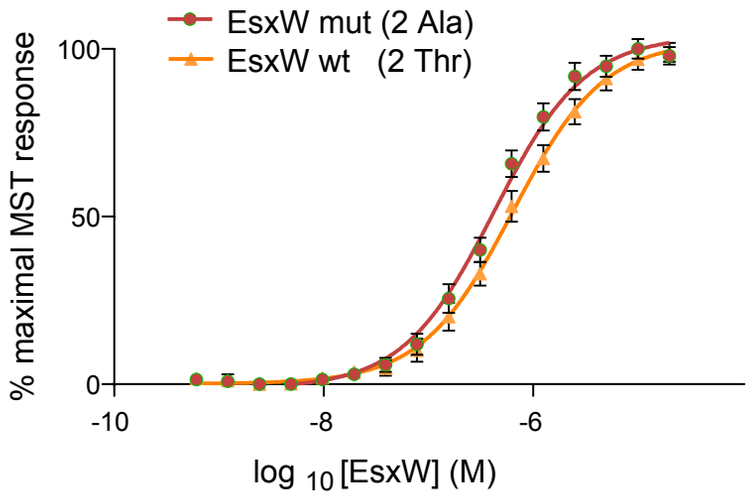


Table 1. Lineage characteristics for HCMC *M. tuberculosis* isolates, including known antimicrobial resistance mutations identified using Mykrobe Predictor.

	Lineage				
	1	**2.1**	**2.2.1**	**2.2.2**	**4**
Gender					
Female	82 (21.1%)	9 (15.3%)	265 (27.8%)	10 (25.6%)	56 (29.2%)
Male	306 (78.9%)	50 (84.7%)	692 (72.3%)	29 (74.4%)	136 (70.8%)
Antimicrobial					
Streptomycin	48 (12.4%)	10 (17.0%)	426 (44.5%)	12 (30.8%)	30 (15.6%)
Isoniazid	57 (14.7%)	12 (20.3%)	269 (28.1%)	9 (23.1%)	52 (27.1%)
Rifampicin	3 (0.8%)	2 (3.4%)	58 (6.1%)	2 (5.1%)	1 (0.5%)
Quinolones	1 (0.3%	3 (5.1%)	18 (1.9%)	2 (5.1%)	2 (1.0%)
Ethambutol	1 (0.3%	2 (3.4%)	60 (6.3%)	3 (7.7%)	2 (1.0%)

Table 2. Homoplasic non-synonymous SNPs identified as occurring on the Beijing lineage-defining branch and also arising independently within other lineages.
The number of branches on which each SNP was identified outside the Beijing lineage-defining branch is shown, and the number of such branches that have multiple descendant tips (indicating onward transmission of the SNP) is shown in brackets. HCMC refers to the 1,635 isolates from HCMC, Vietnam; Elsewhere refers to the 3,146 additional isolates from published studies[3-5, 17-19]; trees are shown in Figure 3.

Mutation	No. branches outside Beijing lineage (no. transmitted)		Function
	HCMC	Elsewhere	
EsxW-T2A	9 (4)	10 (6)	ESX-5 secreted protein (CFP10 homolog)
Rv3081-F220L	2 (1)	7 (3)	hypothetical protein
GidB-E92D	1 (1)	0 (0)	streptomycin resistance