Universidade do Minho
Escola de Ciências da Saúde

Carlos André Rodrigues Magalhães

Evidence for T cell driven diversifying selection in *Mycobacterium tuberculosis in vivo-expressed* genes

Evidências de seleção diversificante mediada por células T em genes de *Mycobacterium tuberculosis* expressos *in vivo*

setembro de 2015

Carlos André Rodrigues Magalhães **Evidence for T cell driven diversifying selection in *Mycobacterium tuberculosis in vivo-expressed* genes**

UMinho | 2015

Universidade do Minho
Escola de Ciências da Saúde

Carlos André Rodrigues Magalhães

Evidence for T cell driven diversifying selection in *Mycobacterium tuberculosis in vivo-expressed* genes

Evidências de seleção diversificante mediada por células T em genes de *Mycobacterium tuberculosis* expressos *in vivo*

Dissertação de Mestrado
Mestrado em Ciências da Saúde

Trabalho efetuado sob a orientação do
Doutor Nuno Miguel Sampaio Osório

setembro de 2015

The work presented in this thesis was done in the Microbiology and Infection Research Domain of the Life and Health Science Research Institute (ICVS), School of Health Science, University of Minho, Braga, Portugal (ICVS/3B's - PT Government Associate Laboratory, Braga/Guimarães, Portugal).

ACKNOWLEDGEMENTS

Várias pessoas contribuíram para a realização desta tese ao longo destes últimos 2 anos, às quais gostaria de expressar o meu agradecimento profundo.

Ao Nuno Osório, um enorme obrigado por tudo! Pela integração, conselhos e vitais ensinamentos que me fizeram evoluir em muitos aspetos. Por acreditares em mim e no meu trabalho. Pela incansável ajuda e força sempre que precisei. Sem o teu contributo, nada disto seria possível. Da minha parte, grande admiração e respeito, não só como cientista mas também como pessoa. Não poderia ter encontrado melhor orientador.

Ao professor Gil Castro e à Margarida Saraiva, os primeiros responsáveis pelo meu ingresso no ICVS. Obrigado por me fazerem crescer com todas as recomendações e palavras de incentivo ao longo destes anos. É muito bom saber que acreditam em mim.

Ao Jérémy, membro emprestado da equipa de bioinformática. Pela amizade e companhia nos bons e maus momentos. E pelas bolachas e comida caseira também, muito importante ☺ Admiro muito o teu carácter e espírito de entrega para com a vida e tenho muita sorte em ser teu amigo.

À Estela, a minha habitante da Catalunha preferida. Pelos conselhos para todas as questões e pelo ânimo que me transmitiste quando dele precisava. Apesar dos poucos meses cá passados, tornaste-te uma amiga para a vida.

À Adriana, amiga crónica de sempre. Que os momentos divertidos e as longas conversas não tenham fim. Obrigado por me distraíres quando precisava de ser distraído e por me acordares quando precisava de ser acordado.

Ao Cristiano, amigo de anos. As tuas palavras de incentivo, força e exemplo de vida foram também contagiantes para mim.

Por último mas não menos importante, aos meus pais e à minha irmã. São os meus alicerces e estão sempre no meu coração.

RESUMO

A tuberculose, uma doença devastadora causada por bactérias do complexo *Mycobacterium tuberculosis* (MTBC), mantém-se vastamente fora de controlo. Embora a imunidade mediada por células T positivas para *cluster* de diferenciação 4 (T $CD4^+$) seja crítica para o controlo da infeção, não existe um conhecimento profundo sobre as características das respostas protetoras mediadas por células T bem como dos fatores que têm impacto no seu estabelecimento. Neste contexto, a diversidade genética do MTBC tem sido subestimada, visto que a maioria dos antigénios conhecidos de *M. tuberculosis* são hiperconservados. A existência de epítopos de células T variáveis foi investigada através da análise da diversidade, evolução e imunogenicidade de uma ilha genómica expressa *in vivo* (iVEGI) em 270 genomas do MTBC. A iVEGI apresentou um nível de diversidade nucleotídica acima da média do genoma, devido à presença de 21 genes altamente diversos. A análise computacional de evolução molecular previu 11 codões de nove genes como estando sob seleção diversificante. De particular interesse foram três substituições nos genes *accD2*, *Rv0987* e *Rv0988*, que alteraram significativamente o número de péptidos codificados com elevada afinidade de ligação prevista para diferentes moléculas de antigénio leucocitário humano (HLA) de classe II. Ensaios *in vitro* realizados com o HLA DRB1*01:01 corroboraram a existência de péptidos com elevada afinidade de ligação e revelaram diferenças de mais de 58% na afinidade de ligação entre péptidos *wild-type* e mutante que sobrepõem AccD2 R233L e Rv0987 V169L. Notavelmente, o péptido $AccD2_{228-241}$ demonstrou níveis de afinidade e estabilidade tão elevados como o controlo positivo ESAT-6_{3-17}, sendo assim um forte candidato para um epítopo de célula T $CD4^+$ restrito a uma linhagem de *M. tuberculosis*. Durante a longa coevolução entre *M. tuberculosis* e populações humanas, a diversidade genética em epítopos específicos de células T pode ter sido selecionada de modo a induzir respostas imunes vantajosas para o agente patogénico. Revelar a trajetória evolutiva de *M. tuberculosis* em resposta à pressão imunológica pode oferecer ferramentas sem precedentes para explorar os seus alvos moleculares em favor do hospedeiro, nomeadamente para o desenvolvimento de vacinas mais eficientes.

Palavras-Chave: Tuberculose; *Mycobacterium tuberculosis*; seleção diversificante; epítopo de células T, antigénio leucocitário humano

ABSTRACT

Tuberculosis, a devastating disease caused by bacteria from the *Mycobacterium tuberculosis* complex (MTBC), remains vastly uncontrolled. Although cluster of differentiation 4 positive ($CD4^+$) T cell-mediated immunity is critical for infection control, a deep knowledge on the characteristics of the protective T cell responses and on the factors impacting their establishment is lacking. In this context, the MTBC genetic diversity has been underappreciated, as most of the known *M. tuberculosis* antigens are hyperconserved. The existence of varying T cell epitopes was investigated by analysing the diversity, evolution and immunogenicity of an *in vivo*-expressed genomic island across 270 MTBC genomes. The *in vivo*-expressed genomic island displayed a level of nucleotide diversity above the whole-genome average due to the presence of 21 highly diverse genes. Computational molecular evolution analysis predicted 11 codons from nine genes to be under diversifying selection. Of particular interest were three substitutions in the genes *accD2*, *Rv0987* and *Rv0988*, which significantly altered the number of encoded peptides with predicted high binding affinity to class II human leukocyte antigen (HLA) molecules. *In vitro* assays performed with HLA DRB1*01:01 corroborated the existence of high binding affinity peptides and revealed differences of more than 58% in the binding affinity between wild-type and variant peptides overlapping AccD2 R233L and Rv0987 V169L. Notably, the $AccD2_{228-241}$ peptide showed affinity and stability levels as high as the positive control ESAT-6_{3-17}, thus being a strong candidate for an *M. tuberculosis* lineage-restricted $CD4^+$ T cell epitope. During the long co-evolution of *M. tuberculosis* and human populations, genetic diversity in specific T cell epitopes might have been selected to induce immune responses advantageous for the pathogen. Unveiling the evolutionary path of *M. tuberculosis* in response to immune pressure might offer unprecedented tools to exploit its molecular targets in favour of the host, namely for the development of more efficient vaccines.

KEYWORDS: Tuberculosis; *Mycobacterium tuberculosis*; diversifying selection; T cell epitope; human leukocyte antigen

TABLE OF CONTENTS

FIGURES INDEX

TABLES INDEX

ABBREVIATIONS

AIDS	Acquired immune deficiency syndrome
ALOX5	*Arachidonate 5-Lipoxygenase*
AM	Alveolar macrophage
APC	Antigen-presenting cells
BCG	Bacille-Calmette-Guérin
BEB	Bayes Empirical Bayes
CD4$^+$	Cluster of differentiation 4 positive
DNA	Deoxyribonucleic acid
FUBAR	Fast Unbiased Bayesian AppRoximation
GWAS	Genome-wide association study
HIV	Human immunodeficiency virus
HLA	Human leukocyte antigen
IFN-γ	Interferon-gamma
IL-10	Interleukin-10
IRGM	*Immunity-related GTPase M*
iVEGI	*in vivo*-expressed genomic island
LRT	Likelihood ratio test
LSP	Large sequence polymorphism
MBL2	*Mannose-binding lectin 2*
MHC	Major histocompatibility complex
MIRU-VNTR	Mycobacterial interspersed repetitive units-variable number of tandem repeats
MRCA	Most recent common ancestral
MSMD	Mendelian susceptibility to mycobacterial diseases
MTBC	*Mycobacterium tuberculosis* complex
PAML	Phylogenetic analysis by maximum likelihood
PCR	Polymerase chain reaction
PRR	Pattern recognition receptor
SNP	Single nucleotide polymorphism
TB	Tuberculosis
TLR	Toll-like receptor

WHO World Health Organization

INTRODUCTION

Tuberculosis: a major human health problem

Tuberculosis (TB) is an infectious disease caused by the bacillus *Mycobacterium tuberculosis*. Human TB affects primarily the lungs, where it leads to progressive tissue necrosis and cavity formation (Figure 1A). Without treatment, patients get increasingly debilitated with fever, weight loss and ultimately respiratory failure [1,2]. In less common cases, the pathogen can disseminate and induce pathology in other parts of the body, such as the central nervous system, bones and the gastrointestinal tract [3]. Only pulmonary TB can lead to aerosol transmission of the tubercle bacilli to other individuals [4]. The single preventive strategy available is the Bacille-Calmette-Guérin vaccine (BCG), first administered in 1921. Unfortunately, this vaccine is only efficient in protecting children against disseminated TB and its efficacy against pulmonary TB in adults is highly variable [5,6]. It is estimated that at least one third of the world's population had already inhaled airborne droplets containing *M. tuberculosis* [7]. Most people are able to control bacterial growth and remain asymptomatic (latent TB). Nevertheless, 5-10% will eventually manifest the active form of the disease [1,2]. This percentage is even higher in immunocompromised patients, such as the ones infected with the human immunodeficiency virus (HIV). According to the World Health Organization (WHO), nine million people developed TB in 2013, resulting in a tragic amount of 1.5 million deaths (0.4 million due to HIV-associated TB) (Figure 1B) [7]. In fact, TB is only seconded by the acquired immune deficiency syndrome (AIDS) in terms of human casualties due to an infectious disease worldwide. Furthermore, TB also poses a high global economic burden, with eight billion dollars spent each year in diagnostic and treatment [7]. Once a person is diagnosed with TB, typical treatment relies on the administration of four antibiotics (isoniazid, rifampicin, ethambutol and pyrazinamide) over six months. Although this regime is highly effective among new cases (85% success), current global death rate decline is a modest 1-2% (Figure 1B) [7]. In addition to this, the emergence of drug-resistant strains threatens the use of antibiotics in the future [7]. Therefore, new therapeutic strategies are urgently needed to tackle this disease.

(A)

(B)

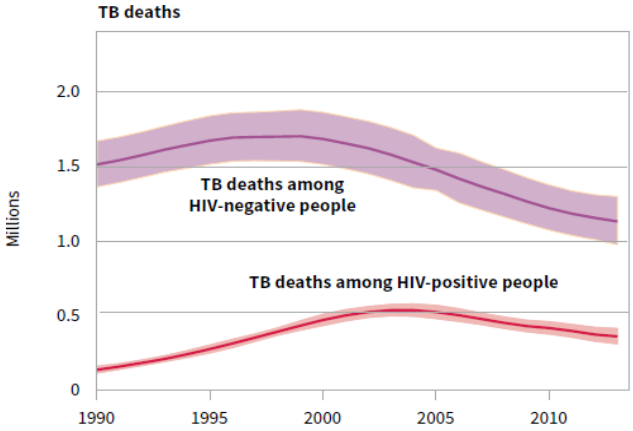


Figure 1. Pathology and death rate of human TB. (A) Gross appearance of TB dissemination in the lungs of a 35-year old male individual infected with HIV [8]. (B) Estimated number of TB deaths in millions per year, 1990–2013 [7]. Reproduced with permission from WHO, ID 182946.

The origin of human TB

TB has been a cause of human death since the antiquity. The earliest written evidence of the disease was found in 4700-year old Chinese medical manuscripts [9]. Paleopathological lesions characteristic of TB were also detected in human remains dated from various historical periods of the human civilization and found in several locations around the globe. These include ancient Syria (8200-7600 BC) [10], Israel (7250-6160 BC) [11], Germany (5400-4800 BC) [12], Egypt (2050-1650 BC) [13], United Kingdom (400–230 BC) [14] and more recently Hungary (1731–1838 AD) [15]. It is estimated that 25% of the Egyptians mummified after 3400 BC suffered from TB [16]. A similar scenario was also true for all adults that died in Europe during the Industrial Revolution [17].

The etiologic agent of human TB remained unknown until the isolation of *M. tuberculosis* in 1882, by Robert Koch [18]. In 1896, taxonomists Lehmann and Neumann included *M. tuberculosis* in the *Mycobacterium* genus, alongside with another important human pathogen, *Mycobacterium leprae* [19]. Two years later, Theobald Smith isolated a similar organism (posteriorly named *Mycobacterium bovis*) from cattle with TB-like clinical presentation [20]. Since then, other TB-causing bacteria have been recovered from wild and domestic mammals, such as goats and sheep (*Mycobacterium caprae*) [21], seals (*Mycobacterium pinnipedii*) [22] and rodents (*Mycobacterium microti*) [23]. Nowadays, it is known that these bacteria share a high level of similarity at the nucleotide level and together constitute the *Mycobacterium tuberculosis* complex (MTBC) [24,25]. Contrarily to the majority of the mycobacteria, members of the MTBC are obligate pathogens with no known environmental reservoir [26,27]. In addition, MTBC bacteria are found in a limited number of different host species [26,27]. Since mycobacteria were

already widely dispersed before the existence of animal kingdom, some of them may have evolved to colonize specific hosts during the early expansion of animal life on Earth [26]. Owing to its broader host spectrum [28], *M. bovis* was believed to be the most recent common ancestral (MRCA) of the MTBC. In that sense, the bacterium was thought to have spread to humans due to close contact with infected cattle [26]. However, analysis of its genome revealed no new genes comparatively to *M. tuberculosis* [29]. Instead, the genomes of *M. bovis* and other animal-adapted MTBC species were found to be smaller due to the accumulation of large deletions [24]. Since horizontal gene transfer is thought to be scarce across MTBC [30,31], these losses of genetic material must have occurred once and could not have been reverted over time. Thus, it was proposed that human and animal-adapted MTBC members were two distinct clades and that the ancestral of the complex would be more similar to *M. tuberculosis* [24]. In support of this notion, skeletal evidence indicated that human TB was present before the early stages of animal domestication [10]. Moreover, although animal-adapted MTBC bacteria can occasionally infect humans [32,33], the ability to transmit TB seems to be restricted to *M. tuberculosis* [26,34]. Significant advances in finding the MRCA of the MTBC were made in 2013, with the inclusion of *Mycobacterium canettii* (an occasional human TB-causing bacterium [35]), as a member of the complex. The genome of *M. canettii* was found to be larger than any other MTBC bacterium, while sharing more than 95% nucleotide sequence identity [36]. It was also reported that *M. canettii* displays ongoing horizontal gene transfer and recombination between strains [36,37], similarly to other environmental mycobacteria [38,39]. Together, all these characteristics placed *M. canettii* in the early branches of the MTBC evolution [36]. The mechanisms underlying the transition of an *M. canettii*-like ancestral to *M. tuberculosis* are not fully clarified [36,40] and will not be discussed in this thesis. In the next section, the focus will be on the population structure of MTBC and its impact on the human population.

MTBC population structure

In clinical settings, detection and classification of *M. tuberculosis* isolates is routinely performed through polymerase chain reaction (PCR) amplification of certain repetitive and polymorphic genomic regions of the bacterium [41]. The occurrence pattern of these deoxyribonucleic acid (DNA) stretches is characteristic of each strain and can be interrogated by two methods, spoligotyping [42] and mycobacterial interspersed repetitive units-variable number of tandem repeats (MIRU-VNTR) [43]. Over the years, the amount of data generated by these techniques lead to the creation of online databases, such as SITVITWEB [44]. Further analysis of the genotyping patterns available at SITVITWEB [44] and

others [45] revealed the existence of distinct groups of strains with variable predominance throughout the world. In other words, each considered geographical region seemed to have a prevailing genotype [44,45]. Other studies harnessing single and large sequence polymorphism (SNP and LSP) data also reached the same conclusions [46–49]. Although PCR-based typing is useful for epidemiological studies [41], it only assess the diversity within small portions of the genome. Alternatively, whole-genome sequencing has a much higher discriminatory power and can be used to measure the evolutionary relationships and distances between different strains [50]. A robust phylogenetic tree of the *M. tuberculosis* population (Figure 2) was constructed in 2013 by Iñaki Comas and colleagues [51]. To achieve that, the authors took advantage of 216 whole-genome sequences of clinical isolates from several regions of the globe. The study highlighted that the global *M. tuberculosis* diversity was distributed by seven main phylogenetic lineages, each one with a strong association for certain world regions. Lineage 1 (also designated as Indo-Oceanic lineage) predominates in East Africa and South-East Asian countries (such as India, Philippines and Vietnam); lineage 2 is highly dispersed through Eastern Asia (e.g. China and Mongolia); lineage 3 is frequently found in Central Asia (e.g. India, Pakistan and Bangladesh) and to a lesser extent in East Africa; lineage 4 (Euro-American lineage) encompasses Europe, America, Africa and Middle-East; lineages 5 and 6 (also known as *Mycobacterium africanum*) are narrowly restricted to West Africa (e.g. Guinea-Bissau) and finally, lineage 7 was only found till date in Ethiopia [52]. It was also verified that the *M. tuberculosis* phylogeny was similar to a human phylogeny based on representative mitochondrial groups [51]. This suggested that the *M. tuberculosis* lineages could have followed the major human migrations out of the African continent, the cradle of mankind [49,51,53–55]. Lineages 1, 5 and 6 were the first to diverge from a 70 000 year-old *M. canetti*-like progenitor and were designated as ‘ancient’ MTBC lineages. In contrast, lineages 2, 3 and 4 (which shared a characteristic large genomic deletion [24]) were classified as ‘modern’ MTBC strains [51]. Additionally, animal-adapted MTBC strains were confirmed to share a common ancestral with *M. tuberculosis*, as aforementioned [24].

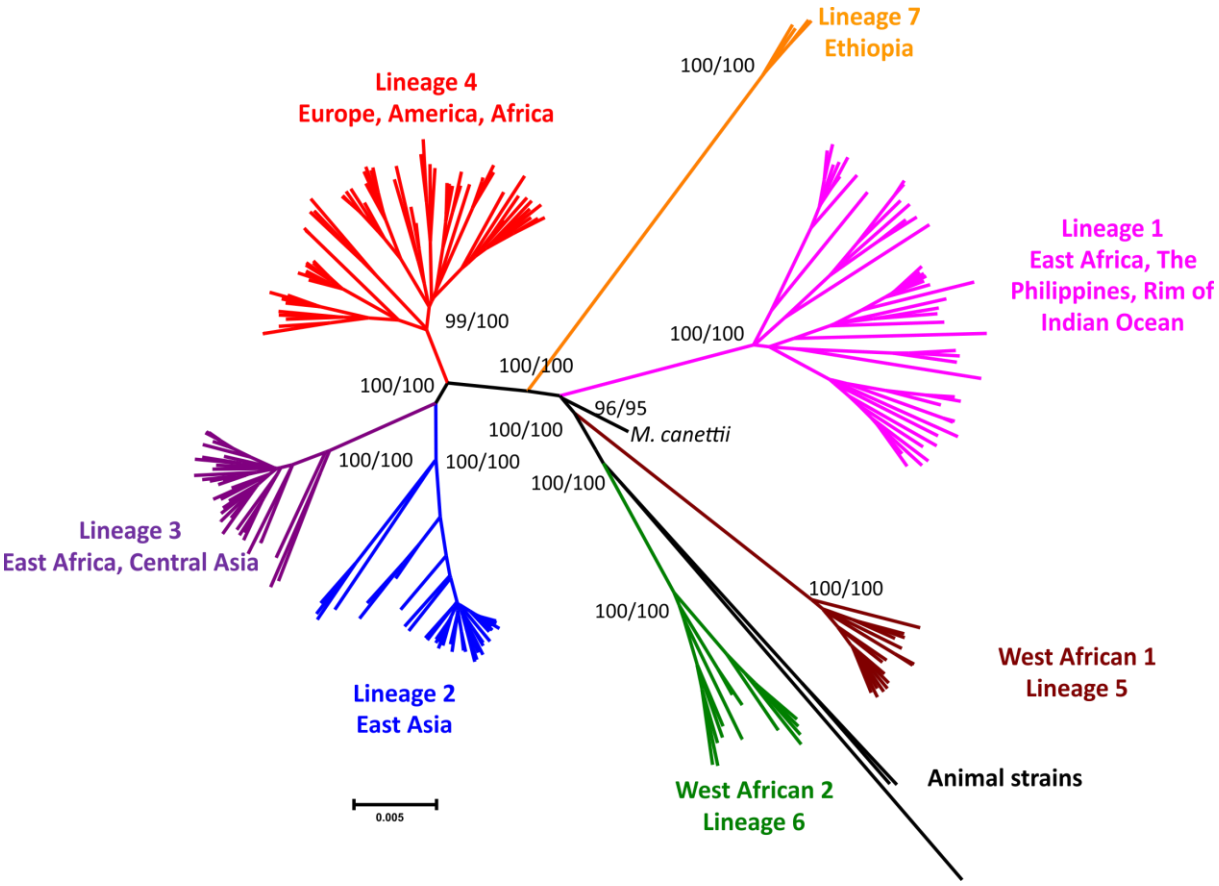


Figure 2. Phylogeny of the MTBC based on whole-genomes of 216 *M. tuberculosis* and 3 animal-adapted strains. Major MTBC lineages are depicted by different colors. Scale bar indicates substitutions per site. Extracted from [51]. Reproduced with permission from Nature Publishing Group, license number 3717280653767.

Genetic diversity among host and pathogen populations in TB

TB is a heterogeneous disease with several outcomes and clinical manifestations upon exposure to its causative agent [1,2]. Much of this heterogeneity was linked with social and environmental factors in the past [56]. However, accumulating evidence has been showing that there is an intrinsic role of biological factors related with the bacterium, the host and the host-pathogen interaction.

On the bacterium's side, it was reported that the diversity among MTBC lineages and sub-lineages has clinical relevance, namely on disease onset, severity and transmission. For instance, a delay in seeking medical care was detected by Yimer et al. (2015) in Ethiopians infected with lineage 7 strains [57]. Parwati et al. (2010) associated lineage 2 with treatment failure [58] and Stavrum et al. (2014) related lineage 4 with more debilitating symptoms (such as weight loss) in Tanzanian individuals [59]. Albanna et al. (2011) found lineage 3 to be less transmissible than lineages 1, 2 and 4 across Montreal, Canada [60], whereas an increasing frequency of lineage 2 was described in South Africa by Cowley et al. (2008) [61]. At a sub-lineage level, Kato-Maeda et al. (2010) identified a highly transmissible sub-lineage of lineage 2 in San Francisco [62] and Gehre et al. (2013) registered striking differences in transmissibility

among sub-lineages of lineage 5 in Benin and Nigeria [63]. Distinct lineage 2 strains were also found to impact the interaction with host innate immune system, either by preferentially activating the pattern recognition receptor (PRR) Toll-like receptor (TLR) 2 or TLR4 [64]. MTBC lineages are also not similar in epidemiological terms, with the 'modern' MTBC lineages 2 and 4 being the most widespread [44]. These have been correlated with a low induction of pro-inflammatory molecules *in vitro* [65,66] and a higher capacity to replicate *in vivo* when compared to 'ancient' MTBC strains [67]. Although the aforementioned studies emphasize the influence of pathogen diversity in TB, some controversial data can be found in the literature. For example, Marais et al. (2009) found the transmissibility not to be significantly associated with the *M. tuberculosis* genotype in South Africa [68] and Krishnan et al. (2011) also reported no significant alterations between lineage 1 and lineage 2 strains regarding the *in vitro* inflammatory profile [69]. These apparent contradictions in studies performed in different geographic areas might be related with the heterogeneity in human populations and its impact in TB. Clinical surveys revealed that a substantial part (30-50%) of household contacts with the disease did not result in measurable infection [70,71], thus suggesting that some people might be intrinsically less susceptible to TB. Indeed, the SNP rs2057178 on chromosome 11p13 was related with protection against TB in Gambia, Indonesia and Russia by a genome-wide association study (GWAS) [72]. Additionally, genetic mutations were found to cause Mendelian susceptibility to mycobacterial diseases (MSMD) [73,74], which is a primary immunodeficiency that leads to complications even with typically non-pathogenic mycobacteria [75]. In addition, susceptibility and immune response to TB were also described to vary according to ethnicity [76,77]. Many other studies identified human genetic variants thought to influence the risk of developing TB (reviewed in [78] and [79]). Nevertheless, the results have been considered inconsistent and difficult to reproduce. Regarding these studies, the lack of incorporation of the nucleotide diversity and population structure of both the host and *M. tuberculosis* was appointed as one of the most prominent weaknesses [80].

Influence of the host-pathogen genetic diversity on TB

Accumulating evidence has been showing that epidemiological and clinical aspects of TB are significantly shaped by the interaction between human and MTBC populations.

At the epidemiological level, Gagneux et al. (2006) and Reed et al. (2009) demonstrated that the transmission dynamics of TB in highly urbanized settings (i.e. San Francisco [48] and Montreal [81]) were not random. Instead, MTBC lineages frequently found in established ethnic communities (namely

Chinese, Filipino and Vietnamese) were the ones that predominated in the corresponding countries of origin (Figure 3) [48,81]. Interestingly, Hirsh et al. (2004) had previously referred that the host's ethnic origin could predict the MTBC lineage of the infecting strain [46].

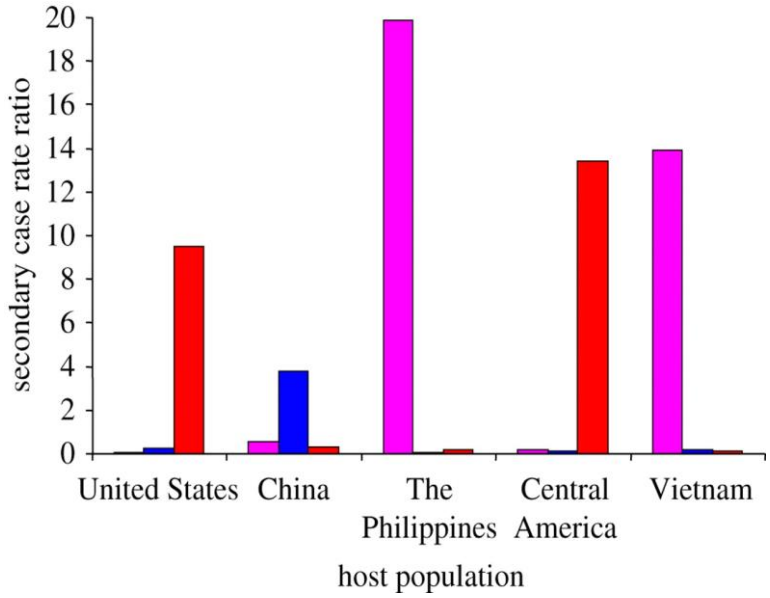


Figure 3. Transmission dynamics of TB in San Francisco, United States of America (2001-2009). The number of secondary TB cases is plotted as a function of the number of index cases (secondary case rate ratio) Pink, lineage 1 or Indo-Oceanic lineage; blue, lineage 2 or East Asian lineage and red, lineage 4 or Euro-American lineage. Adapted from [82].

At the clinical level, de Jong et al. (2008) observed that lineages 2 and 4 had a higher rate of progression to active TB comparatively to lineage 6 in Gambia [83] and a multivariate analysis performed by Caws et al. (2008) showed that lineage 2 strains were more commonly found in Vietnamese patients with a particular polymorphism on the TLR2-encoding gene [84]. In Ghana (West Africa), Herb et al. (2008) reported a statistically significant association between a polymorphism (G254K) on the *Arachidonate 5-Lipoxygenase (ALOX5)* gene and increased predisposition to develop TB due to lineage 6 [85]; Intemann et al. (2009) linked a nucleotide variant upstream of the *Immunity-related GTPase M* (*IRGM*) coding region with protection from TB caused by lineage 4 (and not by other MTBC lineages) [86]; and Thye et al. (2011) found that the LYQC haplotype of the *Mannose-binding lectin 2 (MBL2)* gene conferred an exclusive protective effect against lineages 5 and 6 [86]. Interestingly, the *MBL2* LYQC haplotype is rarely found outside of sub-Saharan Africa (which encompasses West Africa region), suggesting that it could have had increased its frequency over time owing to its protection to endemic MTBC lineages. [87]. On the other hand, the exonic variant G254K of ALOX5 is highly frequent in Africa relative to other world regions [85], indicating that lineage 6 strains (abundant in West Africa) might have a selective advantage in hosts harbouring the referred polymorphism.

Overall, described findings indicated that MTBC lineages tend to be more successful (i.e. more adapted) upon infection of humans that share the same ancestral geographic region (sympatric) than with hosts of foreign ascendency (allopatric) [48,81,82,88].. Furthermore, genetic variants associated with host-pathogen adaptation in TB might have arisen due to natural selection, rather than by chance (i.e. genetic drift) [89]. Consequently, the detection of molecular signatures of adaptation could reveal the basis underlying the success of *M. tuberculosis* and reveal potential targets for therapy design.

Detection of diversifying selection

Diversifying selection (also referred as positive selection [90]), is the directional action of natural selection to increase the prevalence of beneficial genetic variants over time (Figure 4) [91]. More importantly, diversifying selection is thought to be the main mechanism of adaptation in disease contexts [92]. Classical examples are the high frequency of sickle-cell anaemia (caused by a mutation in *Hemoglobin-B* gene) in malaria endemic regions [93] and the development of drug-resistance during treatment in HIV [94].

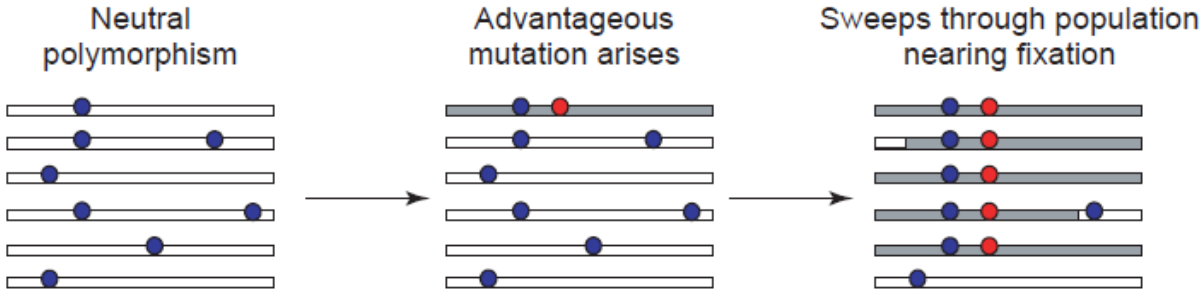


Figure 4. Schematic representation of the action of diversifying selection. Adapted from [95]. Reproduced with permission from Elsevier Ltd, license number 3717271494244.

The detection of diversifying selection is not easy if the selective pressure and its targets are not known. To try to mitigate these issues, numerous computational methods were developed to detect selection in genomic sequences by taking advantage of its increasing availability. Briefly, the methods can be divided in two types: the ones based on population genetics data and comparative genomics [91]. The population genetics-based methods apply statistical models in order to distinguish if variation among a representative sample of the population was randomly generated or fixed by selection. Typical evaluated datasets include linkage disequilibrium [96] or allelic frequency [97], for instance. Despite its popularity in human genetics, these methodologies are highly dependent on assumptions about the population demography, such as population size, that are not easily estimated in pathogen populations [91,98]. Furthermore, the resolution provided is rarely to the codon level [98,99]. In contrast, comparative genomics methods rely

on measuring information within sequence alignments of protein-coding DNA [91]. The main parameter assessed is the ratio (ω) of the number of nonsynonymous nucleotide substitutions (i.e. that change the encoded aminoacid) per site (dN) over the number of synonymous nucleotide substitutions (i.e. with no aminoacid changing) per site (dS) (ω=dN/dS) [90]. The interpretation of the ω ratio is based on the principle that most mutations are synonymous and evolutionary neutral, i.e. without functional consequences. In that sense, a $\omega = 1$ means that an aminoacid change is neutral and a $\omega < 1$ will reflect negative selection (meaning that aminoacid change was probably deleterious and its fixation rate was reduced). Evidence for diversifying selection is inferred when ω is significantly higher than 1 [90]. Initial applications of the method averaged the ω ratio over all codons in a gene and thus only provided information for the whole-protein [100]. Nevertheless, the action of positive selection is thought to be aimed at only a few number of sites [101–103]. To address this matter, new statistical models were developed to allow estimation of different ω ratios in distinct codons [101,104]. In each model, the variation of ω over all sites can be described by a specific statistical distribution. A likelihood ratio test (LRT) is then used to evaluate what model fits better to the data. To check for positive selection, LRT can compare a model that does not admit $\omega \leq 1$ (null hypothesis) with one that admits $\omega > 1$ (alternative hypothesis) [103]. Two pairs of models incorporated in Phylogenetic analysis by maximum likelihood (PAML) package were shown to be the most reliable for this purpose (null versus alternative hypotheses, respectively): M1a versus M2a and M7 versus M8 [105]. The LRT is statistically significant (null hypothesis rejected) when twice the log likelihood difference between the two compared models is above the previously described critical values of a chi-squared distribution with two degrees of freedom [105]. The parameters in the statistical distributions of M2a and M8 models are grouped into classes, with one of them allowing $\omega > 1$ [105]. When this condition is met, a post-inference method designed Bayes Empirical Bayes (BEB) is used to calculate the probability of a codon to come from the class with $\omega > 1$ [106]. A more recent method named Fast Unbiased Bayesian AppRoximation (FUBAR) [107] also applies a codon model to infer the synonymous and the nonsynonymous substitution rates (designated as α and β by the authors, respectively). FUBAR differs from PAML due to the adoption of some computational shortcuts and the use of a distinct method to assess site-specific posterior probabilities for being under diversifying selection, indicated as the posterior probability of $\beta > \alpha$ [107]. Codon models were previously shown to be powerful and robust in the absence of recombination [108]. Indeed, PAML and FUBAR have unveiled positively selected sites in human pathogens, such as in genes encoding for surface glycoproteins of the influenza H5N1 virus [109], as well as a specific siderophore in *Pseudomonas aeruginosa* [110]. PAML has also revealed sites under diversifying selection in the genome of *M.*

tuberculosis related to drug-resistance [111,112] and to other evolutionary pressures [111]. Particularly, evidences of positive selection were identified in a putative antigenic region, suggesting that human immune system can be responsible for the selection of some variants in the genome of the pathogen [111].

The role of T cells in the immune response against TB

The immunity against TB is tightly dependent on T lymphocytes, namely $CD4^+$ and $CD8^+$ T cells [1,113]. Indeed, the low numbers of $CD4^+$ T cells characteristic of HIV-patients are associated with high susceptibility to TB [114] and mice that are deficient in $CD4^+$ T cells die rapidly from the disease [115]. $CD8^+$ T cells also contribute to disease resistance in humans [116], rhesus macaques [117] and mice [115]. Owing to the importance of these two types of T lymphocytes in the control of TB, it is very relevant to detail the characteristics of the T cell-mediated TB protective immune responses. After aerosol inhalation, tubercle bacilli are phagocytized by antigen-presenting cells (APC) such as alveolar macrophages (AM) (Figure 5A) and interstitial dendritic cells [1]. On the surface of APC, peptides generated by proteolysis of *M. tuberculosis* proteins (i.e. the antigens) are bounded to major histocompatibility complex (MHC) proteins (also known as human leukocyte antigen [HLA] in humans). HLA class I molecules present antigenic peptides to $CD8^+$ T cells and HLA class II molecules to $CD4^+$ T cells (Figure 5B) [118]. T cell receptor then recognize a specific complex formed by the HLA molecule and the presented peptide (i.e. the T cell epitope) [119,120]. After this step, T cells can stimulate bacterial killing in macrophages by the secretion of inflammatory mediators such as interferon-gamma (IFN-γ) [1] (Figure 5A). Some T cells also differentiate into a memory phenotype that provide more rapid protection

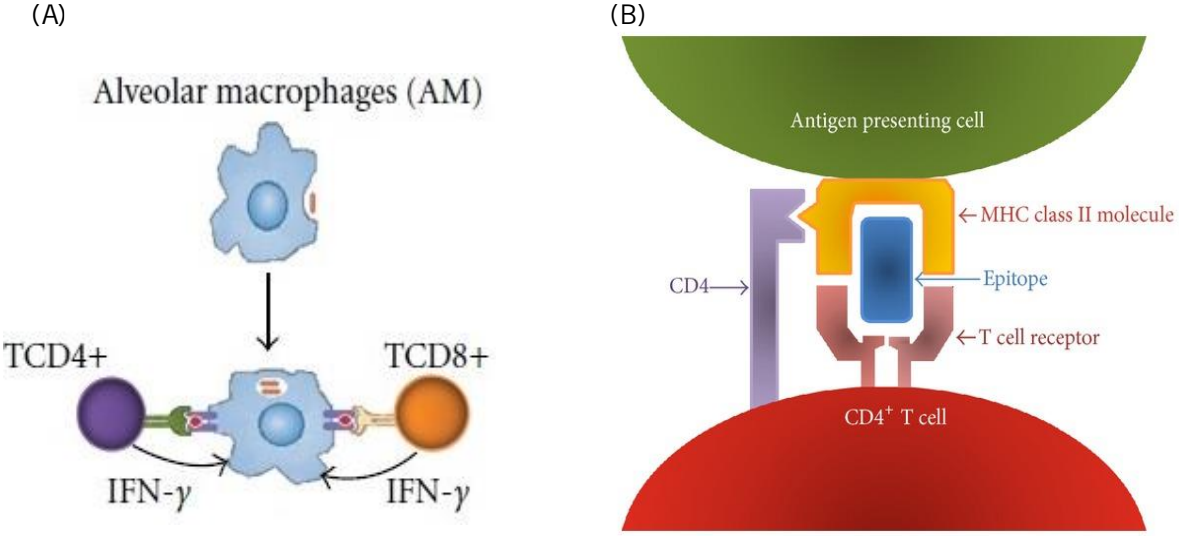


Figure 5. Key players in the immunity against TB. (A) Phagocytosis of *M. tuberculosis* and secretion of IFN-γ by T cells. Adapted from [122]. (B) Recognition of epitope-MHC II complex by a $CD4^+$ T cell. Adapted from [121].

upon new contacts with the same pathogen peptide. This property turns T cells attractive to be explored in the context of vaccine design [113].

Principles and techniques for T cell epitope discovery

The formation of a epitope-HLA complex with the right affinity and stability is a strong requirement to elicit a T cell response [123]. This makes the step of formation of the epitope-HLA complex an important focus of epitope discovery. Nevertheless, identifying the epitope with proper characteristics is a difficult task. First, T cell epitopes can be derived from virtually all proteins expressed in an organism [124]. Second, HLA molecules can accommodate peptides of several lengths in their binding groove, i.e. 8-15 amino acids in the case of HLA class I and 8-20 in HLA class II [125]. Laboratorial approaches to identify T cell epitopes rely on the synthesis of overlapping peptides encompassing antigens of interest. Synthesized peptides are then used to stimulate sensitized T cells *ex vivo* and the best epitope candidates are inferred through the quantification of correlates of T cell immune response, such as cytokine production [126] or T cell proliferation [127]. The complexity, time and cost of these experimental steps are huge constraints to a more widespread application [125].

In the last two decades, several computational methods were developed to analyse and model the properties of the immune system, giving rise to the field of immunoinformatics [124,128]. T cell epitope prediction, also known as reverse vaccinology [129], is a specific branch of immunoinformatics that offers a large potential to accelerate the process of epitope discovery [123]. The main principle of T cell epitope prediction is to find peptides with high binding affinity to HLA molecules, since these have more probability to be presented to T cells [130,131]. This pretension is complicated by the existence of thousands of HLA class I (HLA-DR, -DQ and -DP) and class II (HLA-DR, -DQ and -DP) alleles in the human population, each one with a distinct binding specificity [132]. Furthermore, HLA alleles have different frequencies across the worldwide population [133]. This means that a certain T cell epitope can elicit a response in some human populations but not in others [134]. Taking this into account, Bui et al. (2006) developed an algorithm that calculates the population coverage of an epitope, i.e. the fraction of individuals in a population in which it will be putatively effective [135].

In addition to the HLA molecule, binding affinity also depends on the amino acid sequence of the antigen epitopes [136,137]. In the context of TB, it was previously shown that a large set of experimentally validated MTBC T cell epitopes are evolutionarily conserved [138]. However, the known repertoire of MTBC epitopes is likely to be incomplete, thus impairing a full view of the evolution across the MTBC

immunoproteome [139,140]. In fact, only a small percentage of *M. tuberculosis* proteins has been shown to include epitopes [140,141].

Among the variety of *in silico* algorithms that predict T cell epitopes (reviewed in [123] and [142]), the most accurate ones are trained with large sets of epitopes with experimentally determined binding affinity [143]. Particularly, NetMHCpan [144] and NetMHCIIpan [145] were shown to be the best overall predictors in independent benchmarks [146–148]. Both methods only need the amino acid sequences of interest as input to generate predictions for all known HLA molecules. Quantitative prediction of the binding affinity is extrapolated from a training data set through advanced computational algorithms [144,145]. It is also possible to predict the binding affinity to HLA proteins with no known experimental binding data on the basis of other HLA molecules with similar binding pocket residues [143].

The bulk of the immune response is aimed at a small fraction of the possible epitopes [119,120], a feature called immunodominance [149]. In this regard, it was described that a higher kinetic stability of the HLA-peptide complex is a good predictor for immunodominance in both HLA class I [150] and HLA class II [151–153]. In *M. tuberculosis*, besides antigen 6-kDa early secretory protein antigenic target (ESAT-6) [154], only a few other immunodominant epitopes were validated in TB patients [155]. Another factor that affects immunodominance is the relative abundance of peptides available for binding. Genes with high *in vivo* expression levels in the context of infection constitute good candidates for the identification of relevant T cell epitopes since their encoded peptides are more likely to be presented than other low abundance peptides with similar HLA-binding affinity and stability [156]. In the *M. tuberculosis* genome, Talaat and colleagues found a genomic region of 44,849 nucleotides that harbour several genes highly expressed *in vivo* [157,158]. The region is constituted by 43 contiguous genes and was designated as *in vivo*-expressed genomic island (iVEGI) [157]. Interestingly, iVEGI is upregulated upon infection in wild-type mice models but not in immune compromised animals or during *in vitro* growth of *M. tuberculosis* [157]. This suggests a role for iVEGI in the interaction with host immune cells, highlighting T cell responses as a possible driving force of natural selection in this region.

AIMS

Co-evolution between humans and *M. tuberculosis* strains have produced numerous examples of clinical and epidemiological heterogeneity (reviewed in [159] and [88]). However, the molecular basis underlying host-pathogen interactions in TB remain poorly understood. Indeed, despite the critical importance of T cells to the control of TB, the impact of the genetic diversity selected during the MTBC evolution on the full repertoire of *M. tuberculosis* T cell epitopes is poorly study. Still, the existing studies show evidence for genetic variation in known T cell epitopes [138] and for diversifying selection in a predicted antigenic region [111]. The identification and functional characterization of the immune responses elicited by variable epitopes are of critical importance from a vaccine-development point of view since these might correlate with protective responses. Highly *in vivo-expressed* genes mediate most of the host-pathogen interactions and are thus thought to be major targets of immune-related selective pressure. Therefore, *M. tuberculosis* iVEGI, a genomic region found to be upregulated *in vivo* in immunocompetent mice models of infection [157], is a good candidate region to identify these epitopes.

Taking this into account, the aim of this thesis is to analyse the iVEGI for evidence supporting the presence of *M. tuberculosis* epitopes under T cell driven positive selection. In order to address this goal, a research workflow composed of four main tasks was designed:

(i) Comparative genomics analysis, to investigate the nucleotide diversity of 43 genes highly expressed in *in vivo* models of infection;

(ii) Detection of diversifying selection, to search for signatures of molecular adaptation among the genetic diversity unveiled at (i);

(iii) Prediction of $CD4^+$ and $CD8^+$ T cell epitopes in protein sequences encoded by genes with sites under diversifying selection;

(iv) *In vitro* HLA-peptide binding assays, to provide the experimental validation of the best candidates from preceding tasks.

RESULTS

High nucleotide diversity and evidence for diversifying selection in the iVEGI

The *in vivo* expression pattern of iVEGI suggests a role in host-pathogen interactions during infection and the consequent possibility of evolution in response to host pressures. To investigate this, the sequence diversity on the iVEGI was characterized. This region was extracted from the genomes of 270 isolates belonging to the seven MTBC lineages and comparative genomics analysis was performed. The iVEGI was significantly more diverse ($\pi = 5.04 \times 10^{-4}$, $p < 0.001$) than the rest of the genome ($\pi = 2.65 \times 10^{-4}$), with the majority of the genes analysed (21 out of 40) displaying levels of nucleotide diversity (π) above the genome-wide mean (Figure 6A). The level of genetic diversity in the iVEGI was very similar ($p > 0.05$) between 'ancient' (lineages 1, 5, 6 and 7; $\pi = 3.42 \times 10^{-4}$) and 'modern' (lineages 2, 3 and 4; $\pi = 3.62 \times 10^{-4}$) MTBC strains. This is in contrast with the whole-genome data, which showed that the 'ancient' lineages ($\pi = 3.10 \times 10^{-4}$) had a significantly higher nucleotide diversity level ($p < 0.001$) when compared to the 'modern' lineages ($\pi = 1.82 \times 10^{-4}$) (Figure 6B). A more detailed analysis revealed a total of 320 SNPs, of which 291 were located in coding regions and 29 in intergenic regions (Figure 6C and Supplementary Table 1). SNPs were detected in all the seven MTBC lineages and with highly variable frequencies. Homoplastic SNPs were also found, 13 located in genes and two in intergenic regions (Supplementary Table 1).

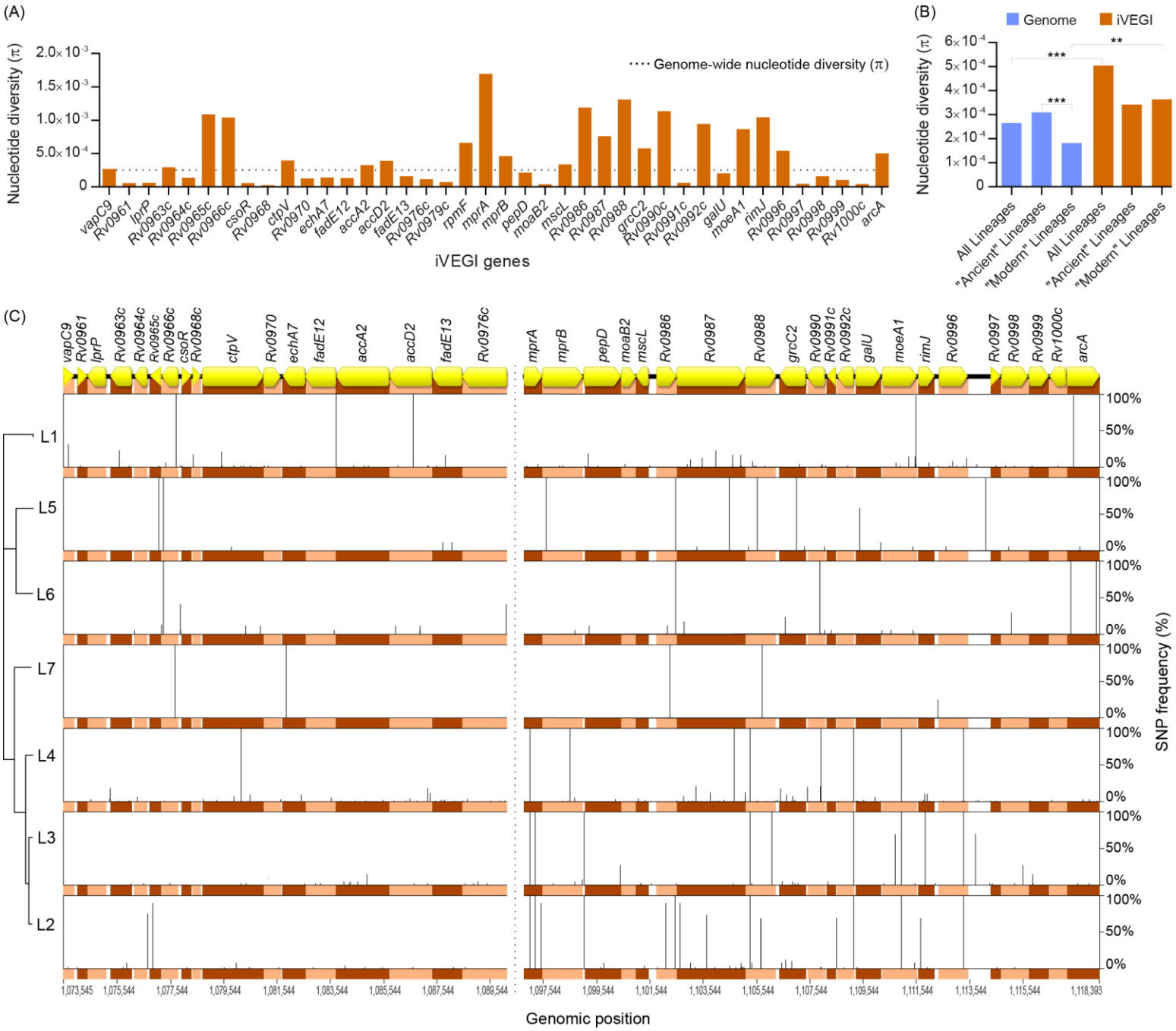


Figure 6. Comparative genomics analysis of the iVEGI. (A) Nucleotide diversity of individual iVEGI genes. (B) Nucleotide diversity (π) of the genome (in blue) versus iVEGI (in orange) in strains from the 'ancient', 'modern' and all the seven MTBC lineages. (C) Schematic representation of the sequence variation found in iVEGI. The relative frequency of the nucleotide substitutions in each of the seven MTBC lineages (right *yy* axis) is represented on the left *yy* axis. Genomic positions (*xx* axis) are according to the reference genome H37Rv (GenBank accession number NC00962.3). *Pe_pgrs16*, *pe_pgrs17* and *pe_pgrs18* were excluded from the analysis and are not represented (stroke line). ** $p < 0.01$; *** $p < 0.01$.

Considering that SNPs under diversifying selection may reflect targets of molecular adaptation driven by host-related pressures, the iVEGI genes were screened for diversifying selection with PAML and FUBAR [105,107]. The site models M2a and M8 included in PAML package were applied and Bayes Empirical Bayes (BEB) analysis revealed significant evidence for diversifying selection (M2a pp > 0.95 or M8 pp > 0.95) in seven codon sites corresponding to amino acid positions 186 of LprP (M2a pp = 0.928 and M8 pp = 0.977), 666 of AccA2 (M2a pp = 0.999 and M8 pp = 1.000), 51 and 233 of AccD2 (M2a pp = 0.901 and M8 pp = 0.953; M2a pp = 0.997 and M8 pp = 0.999, respectively), 54 and 68 of Rv0990c (M2a pp = 0.957 and M8 pp = 0.982; M2a pp = 0.960 and M8 pp = 0.983, respectively) and 81 of RimJ

(M2a pp = 0.950 and M8 pp = 0.981) (Table 1). FUBAR identified four additional putative targets of diversifying selection (pp ($\beta>\alpha$) > 0.9) in amino acid residues 169 of Rv0987 (pp $\beta>\alpha$ = 0.941); 191 of Rv0988 (pp $\beta>\alpha$ = 0.967); 141 of GrcC2 (pp $\beta>\alpha$ = 0.942) and 303 of Rv0996 (pp $\beta>\alpha$ = 0.918) (Table 1). An overlap between the predictions of PAML and FUBAR was found for the amino acid 233 of AccD2. To test a possible influence of recombination, nucleotide sequences were analysed with RDP, GENECONV and Bootscan, with none of the methods detecting signals supportive of recombination (data not shown). Overall, evidences for diversifying selection were detected by PAML or FUBAR in 11 codons of nine distinct iVEGI genes (Table 1). These data support high levels of nucleotide diversity and the occurrence of targeted events of diversifying selection in the iVEGI.

Table 1. Amino acid substitutions under diversifying selection in iVEGI.

Gene name	Genomic position	Gene position	Amino acid substitution	PAML				FUBAR	
				M2a		M8			
				BEB ppω>1	ω ratio ± SE	BEB ppω>1	ω ratio ± SE	β-α	pp $\beta>\alpha$
lprP (*Rv0962c*)	1074996	557	L186P	0.928	8.138 ± 2.683	0.977*	8.561 ± 2.145	4.798	0.876
accA2 (*Rv0973c*)	1083755	1996	S666P	0.999**	8.705 ± 1.835	1.000**	7.992 ± 1.954	2.658	0.788
accD2 (*Rv0974c*)	1087193	153	K51N	0.901	7.778 ± 2.910	0.953*	8.006 ± 2.516	3.300	0.819
	1086648	698	R233L	0.997**	8.482 ± 1.973	0.999**	8.341 ± 2.015	6.601	0.930[a]
Rv0987	1103046	505	V169L	0.289	1.231 ± 1.276	0.361	1.242 ± 1.220	7.789	0.941[a]
Rv0988	1105686; 1105687	571;572	L191A	0.704	3.977 ± 3.117	0.815	4.316 ± 3.069	7.887	0.967[a]
grcC2 (*Rv0989c*)	1106961;1106962	421;422	V141L/A	0.831	5.613 ± 3.174	0.925	5.749 ± 2.902	7.140	0.942[a]
Rv0990c	1107940	160	A54S	0.957*	7.555 ± 2.542	0.982*	7.333 ± 2.484	4.183	0.861
	1107897	203	A68V	0.960*	7.571 ± 2.524	0.983*	7.340 ± 2.477	4.422	0.881
rimJ (*Rv0995*)	1111852	241	D81Y	0.951*	6.780 ± 2.667	0.981*	6.428 ± 2.570	3.516	0.840
Rv0996	1113290;1113292	907;909	Q303E/H	0.696	4.076 ± 3.210	0.827	4.595 ± 3.204	6.693	0.912[a]

BEB, Bayes Empirical Bayes; pp, posterior probability; β, nonsynonymous substitution rate; α, synonymous substitution rate; *$p < 0.05$; **$p < 0.01$

[a] pp above 0.900 considered as highly supportive of diversifying selection by FUBAR.

Specific iVEGI variants under diversifying selection impact T cell epitope prediction

The sustained high *in vivo* expression of iVEGI in mouse models of *M. tuberculosis* infection [157,158] makes its encoded proteins good candidates for antigen presentation. Taking this into account and considering its genetic diversity, the genomic island was next tested for the presence of unknown T cell epitopes that could be the underlying cause of diversifying selection. For this purpose, the binding affinity of the peptides overlapping the iVEGI sites under diversifying selection was predicted to a set of 72 class I and 41 class II HLA molecules (Supplementary Table 2). The analysis was performed with NetMHCpan [144] and NetMHCIIpan [145]. A total of 1530 putative high binding affinity peptides were identified, including 244 for class I HLA and 1286 for class II HLA. To allow class I versus class II comparisons, the results were normalized for the number of peptide lengths and HLA molecules tested and expressed as the percentage of the maximum possible number of high binding affinity peptides (Figure 7 and Supplementary Figure 1). The majority of the sites with significant evidences for being under diversifying selection (eight out of 11) were found to be included in regions encoding peptides predicted to bind with high affinity to at least one HLA of both classes (Figure 7A). The exceptions were the amino acids AccA2 S666, with no predicted high binding affinity peptides, and AccD2 K51 and Rv0990c A68, with high binding affinity peptides detected only to class I HLAs. HLA class II predictions identified in mean five times more immunogenic peptides than HLA class I predictions (Figure 7A). Interestingly, about 90% of the putative $CD4^+$ T cell epitopes were concentrated in the overlapping regions of three sites (Rv0987 V169, Rv0988 L191 and Rv0996 Q303).

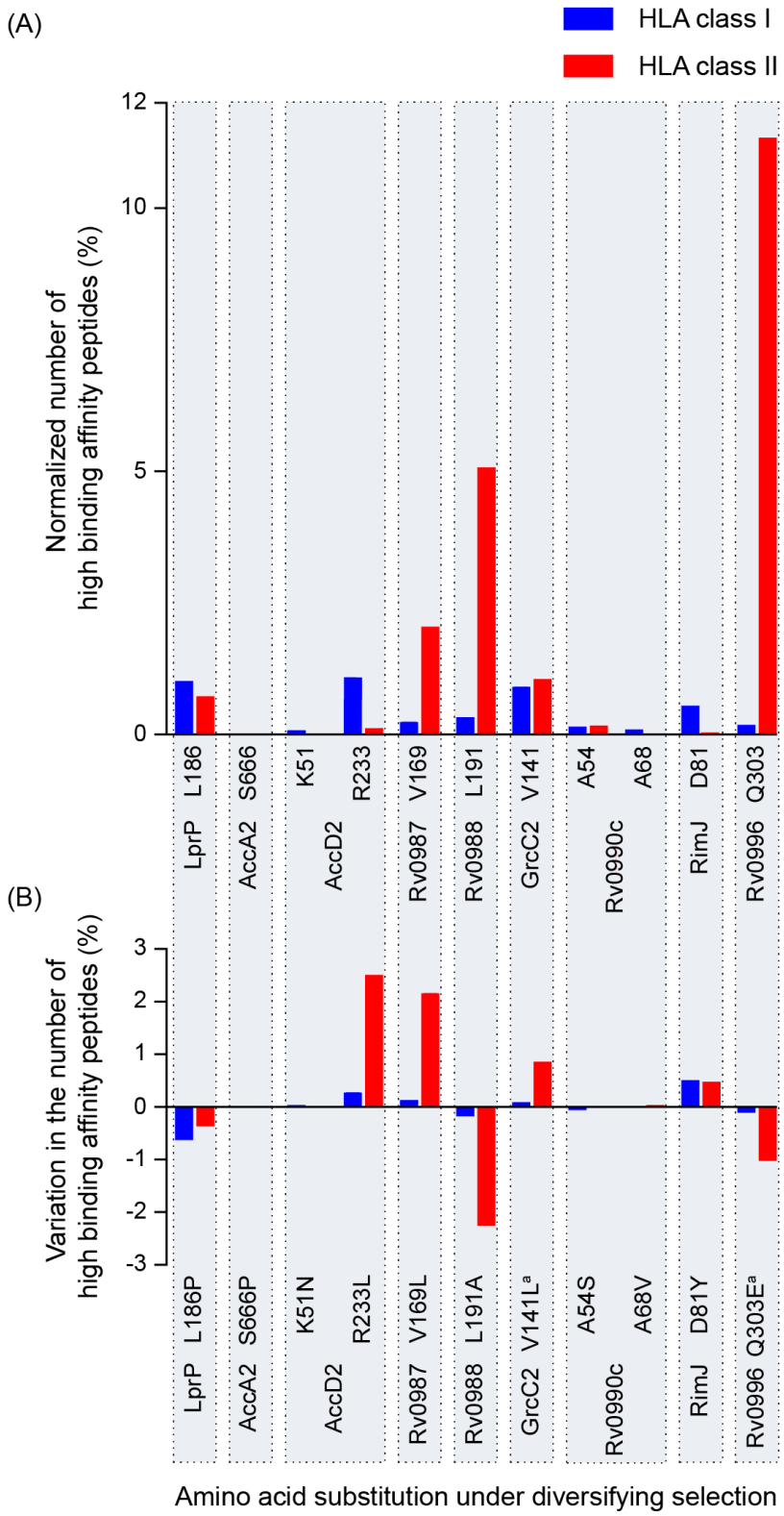


Figure 7. Immunoinformatics analysis of the sites under diversifying selection in iVEGI. (A) Number of putative $CD4^+$ and $CD8^+$ T cell epitopes in amino acid sequences including sites under diversifying selection. High binding affinity peptides were predicted by NetMHCpan and NetMHCIIpan to a representative set of HLA class I (in blue) and class II (in red) alleles (Supplementary Table 2). The number of predicted immunogenic peptides is represented as percentage of the total possibilities for each HLA class (Supplementary Figure 1). (B) Variation in the number of high binding affinity peptides due to diversifying selection in the iVEGI. [a] More than one amino acid substitution in the same codon; only the substitution resulting in the highest variation was represented.

To evaluate the functional impact of sequence diversity on T cell epitope prediction, the HLA-binding affinity of wild-type peptide sequences was compared to that of the variant peptides under selection (Figure 7B). Variations in class II HLA were highly predominant and responsible for 84% of the total

differences registered. Strikingly, three substitutions (AccD2 R233L, Rv0987 V169L and Rv0988 L191A) contributed to 72% of all the predicted alterations in class II HLA-binding (Figure 7B). Indeed, variants AccD2 R233L and Rv0987 V169L increased the number of potential $CD4^+$ T epitopes by 2.5% and 2.1%, respectively. Contrarily, Rv0988 L191A decreased this number by 2.2% when compared to the wild-type sequence. Overall, the data revealed three amino acid substitutions (AccD2 R233L, Rv0987 V169L and Rv0988 L191A) with high probability to be under $CD4^+$ T cell driven diversifying selection.

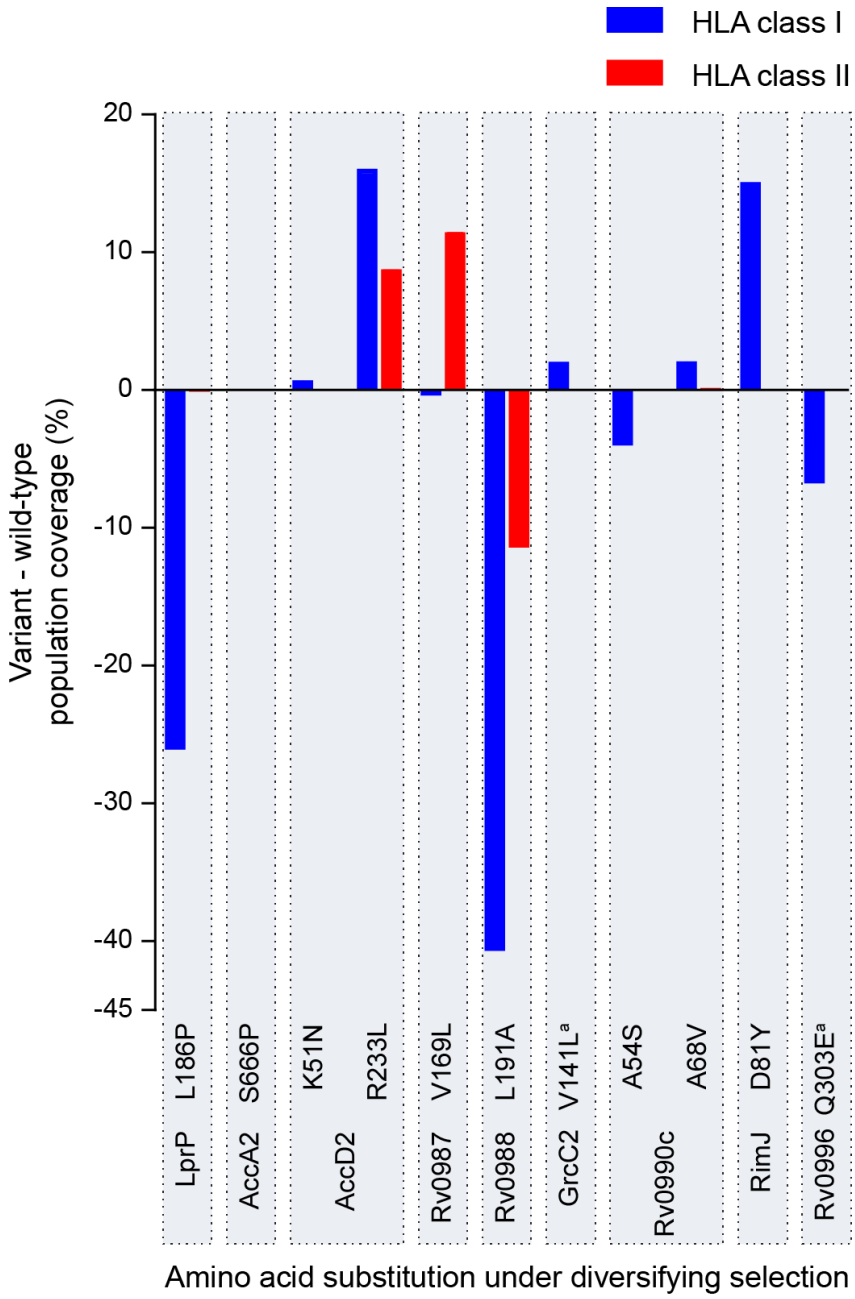


Figure 8. Variations in the worldwide population coverage in sites under diversifying selection. The percentage of the human population with potential to recognize predicted T cell epitopes (worldwide population coverage) was estimated for HLA class I (in blue) and HLA class II (in red). The difference between the population coverage of the variants under diversifying selection relative to the wild-type is on the *yy* axis. [a] More than one amino acid substitution in the same codon; only the substitution that resulted in the highest variation was represented.

Geographic correlations between population coverage and MTBC lineage frequency

T cell epitopes might be differentially recognized by diverse individuals due to the variable frequencies of HLA alleles across the human population, thus resulting in distinct population coverage. For this reason, the fraction of the human population with potential to recognize the putative epitopes under study (worldwide population coverage) was predicted. Next, the variation resulting from the presence of amino acids under diversifying selection was calculated (Figure 8 and Supplementary Table 3). Despite the low number of high binding affinity peptides to HLA class I (Figure 7), the highest alterations in population coverage imposed by the variant epitopes were found for HLA class I (Figure 8). Of particular note were four amino acid substitutions (LprP L186P, AccD2 R233L, Rv0988 L191A and RimJ D81Y) which showed a high impact on the HLA class I worldwide population coverage (-26%, +16%, -41% and +15%, respectively, as compared to the wild-type sequence) (Figure 8).

The largest variations in class II HLAs worldwide population coverage were attributed to AccD2 R233L and Rv0987 V169L (which were expected to increase coverage by 9% and 11%, respectively) and to Rv0988 L191A (expected to decrease the coverage by 11%). The analysis was expanded by investigating if the regional variations in population coverage were associated with the geographic distribution of the MTBC lineages harbouring the selected variants (Figure 9). For that purpose, the frequency of the MTBC lineages and the alterations in regional population coverage induced by the variants under diversifying selection were compared across 15 world geographic regions (Supplementary Figure 2). No correlations were found for the HLA class I. In contrast, three statistically significant correlations were found between the variations in HLA class II population coverage for AccD2 R233L (r_τ = 0.39, $p < 0.05$), Rv0987 V169L (r_τ = 0.47, $p < 0.05$) and Rv0988 L191A (r_τ = -0.41, $p < 0.05$) and the frequency of lineage 1 and lineage 2 strains (Figure 9 and Supplementary Table 4). AccD2 R233L occurred in all the lineage 1 strains analysed (Figure 6C and Supplementary Table 1) and the variant amino acid induced a calculated increase in regional population coverage ranging from 1% in Central America to nearly 30% in Southeast Asia (Figure 9A). This increase in population coverage significant and positively correlated with the percentage of lineage 1 frequency in the 15 tested regions (graph in Figure 9A). Rv0987 V169L was found in 10% of the lineage 1 strains and 8% of the lineage 2 strains analysed (Figure 6C and Supplementary Table 1) and resulted in a predicted increase of the regional population coverage ranging from 3% in Central America to 29% in Southeast Asia (Figure 9B). The estimated increase in coverage was also significant and positively correlated with the regional prevalence of lineage 1 and 2 strains (graph in Figure 9B). In contrast, Rv0988 L191A, that was found in 69% of the lineage 2 strains (Figure 6C and

Supplementary Table 1), was estimated to decrease the regional population coverage from 3% in Central America to 29% in Southeast Asia (Figure 9C). This was significant and negatively correlated with the regional frequency of the lineage 2 strains (graph in Figure 9C). Overall, the results suggest that the amino acid substitutions AccD2 R233L, Rv0987 V169L and Rv0988 L191A were likely relevant to the sympatric adaptation of MTBC strains to certain host populations.

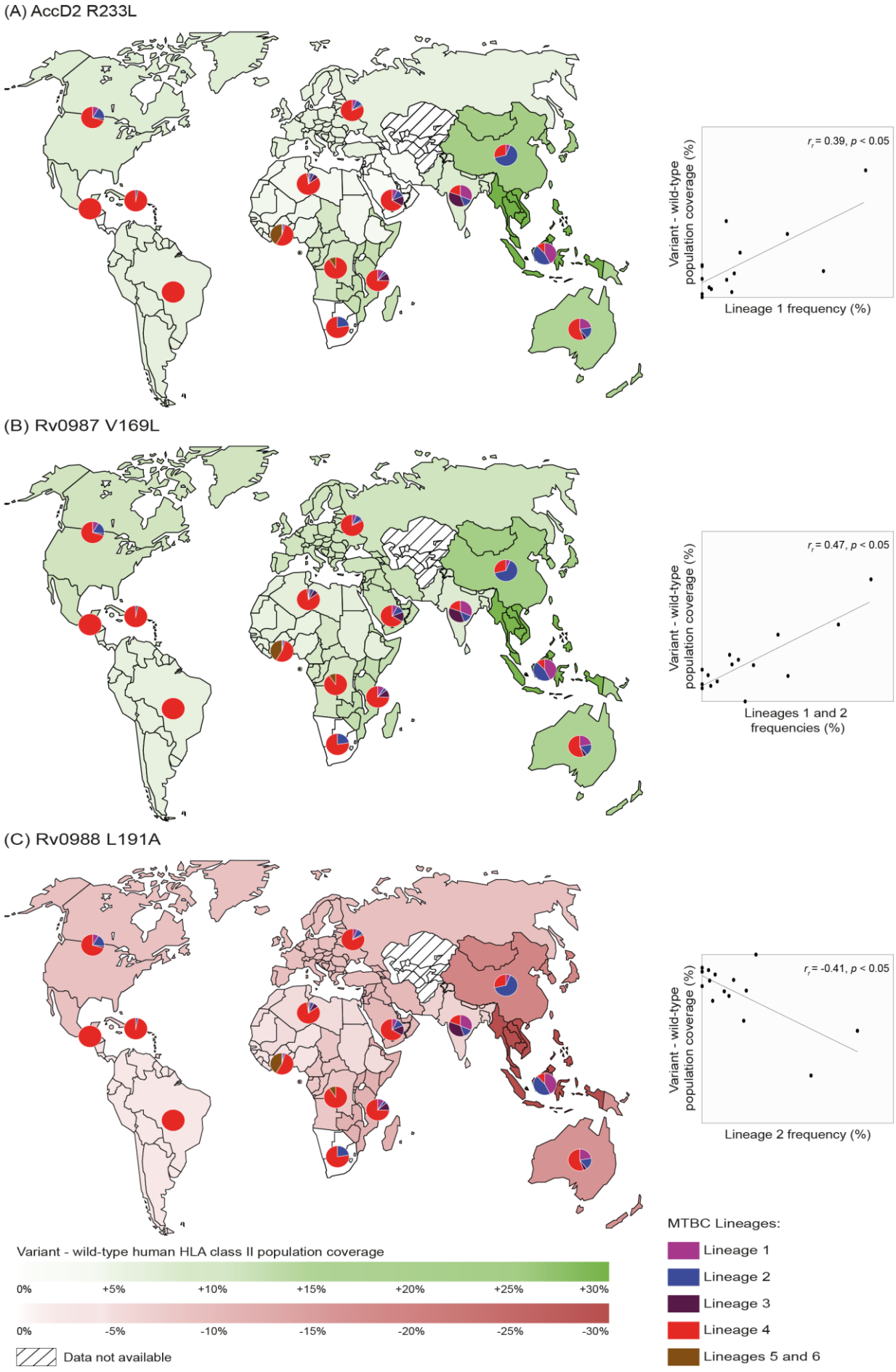


Figure 9. Association between differential population coverage of predicted epitopes and the geographic distribution of the MTBC. The frequency (in pie charts) of the MTBC lineages harboring sites under diversifying selection (in *xx* axis of each graph) correlated with the variations in population coverage due to differential binding to human HLA class II (green and red color gradients) of peptides harbouring three amino acid substitutions: (A) AccD2 R233L (r_τ = 0.39, p < 0.05); (B) Rv0987 V169L (r_τ = 0.47, p < 0.05) and (C) Rv0988 L191A (r_τ = -0.41, p < 0.05).

In vitro validation of the predicted CD4+ T cell epitopes under diversifying selection

Typical T cell epitopes are expected to bind with high affinity and stability to an HLA molecule [123]. Thus, large alterations in the HLA-binding properties are anticipated in epitopes under T cell driven diversifying selection. Taking this into account, the *in vitro* HLA-binding affinity and stability were compared for the set of wild-type and variant peptides that i) were predicted to be under diversifying selection; ii) showed large variation in the predicted HLA-binding affinity and iii) impacted the population coverage and correlated with the geographical distribution of the MTBC lineages. This set was formed by six peptides with the amino acid substitutions AccD2 R233L, Rv0987 V169L and Rv0988 L191A. The HLA class II DRA1*01:01 DRB1*01:01 was chosen for the *in vitro* assays since this was the HLA with the broadest population coverage [160] among the ones with variable predicted binding affinity. The binding affinity and stability of each peptide was compared to a validated DRB1*01:01 CD4+ T cell epitope of the *M. tuberculosis* ESAT-6 protein [155]. The results confirmed the variant forms of AccD2$_{228\text{-}241}$ and Rv0987$_{164\text{-}177}$, but not of Rv0988$_{185\text{-}202}$, as being epitopes with strong binding affinity to DRB1*01:01. In agreement with the *in silico* analysis, radical alterations (> 58%) in the binding affinity between wild-type and variant peptides were observed for AccD2$_{228\text{-}241}$ and Rv0987$_{164\text{-}177}$ (Figure 10). Most importantly, the variant AccD2$_{228\text{-}241}$ peptide, in addition to displaying a binding affinity as high as ESAT-6$_{3\text{-}17}$, also formed a highly stable complex with DRA1*01:01 DRB1*01:01. This suggests that the variant AccD2$_{228\text{-}241}$, but notthe wild-type peptide, has the potential to function as a previously unidentified immunodominant CD4+ T cell epitope in individuals expressing the HLA DRA1*0101 DRB1*01:01.

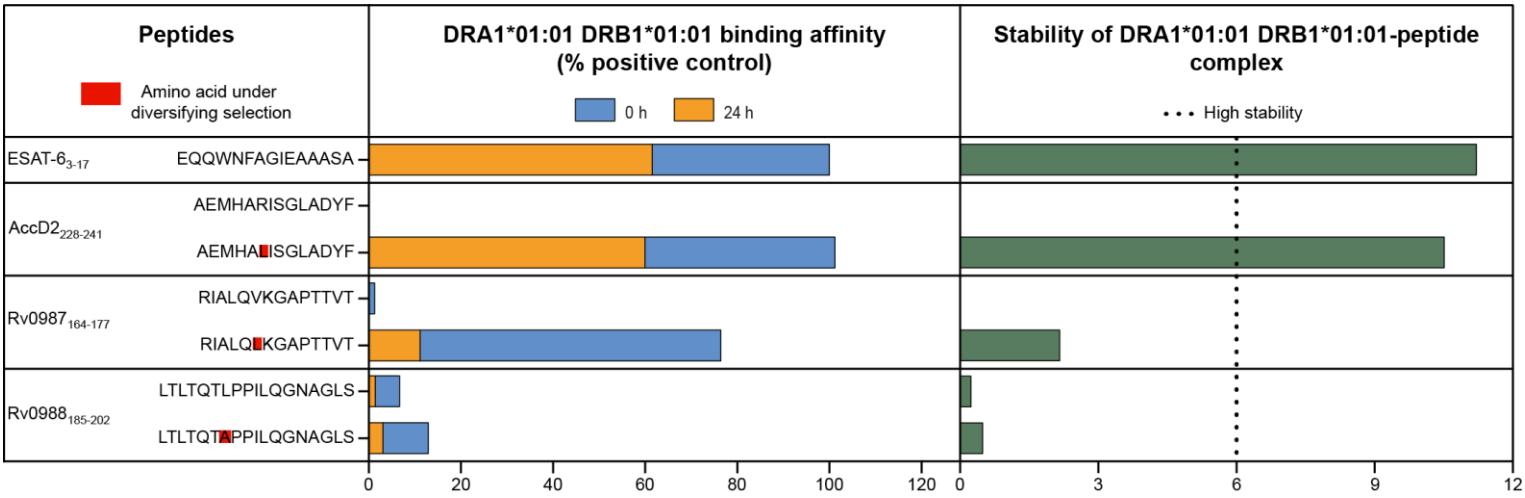


Figure 10. *In vitro* validation of the candidate CD4+ T cell epitopes under diversifying selection. Amino acid substitutions with the highest impact on HLA class II binding predictions (AccD2 R233L, Rv0987 V169L and Rv0988 L191A) were subjected to *in vitro* confirmation of the estimated HLA-binding affinity. The HLA class II molecule with the broadest population coverage among the ones with variable predicted binding affinity (DRA1*01:01 DRB1*01:01) was used. The binding affinity of wild-type and variant peptide sequences was measured at 0 h and 24 h after incubation with the DRA1*01:01 DRB1*01:01 protein and compared with a validated *M. tuberculosis* CD4+ T epitope (ESAT-6$_{3\text{-}17}$). The stability index of the different DRA1*01:01 DRB1*01:01-peptide epitope complexes was also evaluated.

DISCUSSION

Control of TB, a global disease that affects millions of people every year, highly relies on the development of better preventive strategies as the only vaccine in use (BCG) is of limited efficiency [5,6]. In fact, much effort has been put on the development of new vaccines but so far the results have been at best similar to BCG [161,162]. To achieve better outcomes, a deep understanding of the characteristics of protective immune responses and of how they are triggered and modulated is critical [162,163]. In this context, and considering that the efficiency of host T cell-mediated immunity is determinant for infection control [1,113], it is greatly relevant to obtain a full picture of the factors that might influence T cell responses in TB. Among these factors, the genetic and phenotypic diversity associated with the MTBC has been neglected. Indeed, several studies showed that strains of *M. tuberculosis* differently interact with innate immune cells [64,65,69,164,165], thus potentially being able to modulate the outcome of the acquired immune response [166]. Another factor with a potential impact on the type and quality of T cells responses is the existence of varying epitopes within the MTBC. It has been previously described that several T cell epitopes are evolutionarily hyperconserved [138], suggesting that the T cell responses induced in the host are similar across the MTBC diversity. However, the sympatric associations between MTBC lineages and human populations [48] and its disruption in HIV-1 infected individuals [114] suggests that at least an unknown set of relevant and variable MTBC $CD4^+$ T cell epitopes may exist. Furthermore, considering the long co-evolution of MTBC with its human host [51,88], it is likely that the host immune system represents a strong pressure shaping the evolution of the MTBC complex and namely at the level of the epitope-HLA interaction. In fact, different HLA haplotypes have been associated with resistance or susceptibility to TB caused by uncharacterized strains [167–169]) or even associated to TB caused by specific MTBC phylogenetic taxa [170]. In this thesis, a pipeline of molecular evolution, immunoinformatics and *in vitro* techniques was applied to reveal novel T cell epitopes in MTBC, taking into consideration the genetic variation within MTBC and its phylogeographic adaptation to the human host. Herein presented data constitute the first evidences for T cell driven diversifying selection in MTBC, namely within the genomic island iVEGI.

The iVEGI is a good candidate region within the *M. tuberculosis* genome for the discovery of MTBC-varying epitopes, due to its *in vivo* expression profile [157] and high genetic diversity as shown herein. Screening *in vivo-expressed* regions for the discovery of novel *M. tuberculosis* antigens has previously allowed the successful identification of a novel antigen encoded by *Rv2034* [171]. The iVEGI contains several genes implicated in the survival and pathogenesis [158] of *M. tuberculosis* and with prolonged high expression

levels in immunocompetent animal models of TB infection but not in immunosuppressed animals or during *in vitro* growth [157]. Data for high genetic diversity in the iVEGI were firstly reported in a study comparing 20 genomes across four MTBC lineages, which highlighted a three genes operon (*Rv0986-Rv0988*) in this region with high SNP density [89]. Using 270 genomes from all the known seven MTBC lineages, it was revealed that a total of 21 genes in the iVEGI had a level of nucleotide diversity above the whole-genome average. The results also showed that SNPs in the iVEGI were evenly distributed across the MTBC lineages, which is in contrast with the genome-wide data in support for higher genetic conservation among the 'modern' (lineages 2, 3 and 4) when compared to 'ancient' lineages (lineages 1, 5, 6 and 7). This finding points to differences in the evolutionary process of this particular region. In addition, evidence supporting the action of diversifying selection was found in 11 iVEGI loci. Thus, genetic diversity and molecular evolution analyses strongly support the involvement of this genomic island in the long-term evolution of the MTBC population. Subsequent immunoinformatics data highlighted T cells as one of the driving forces for this evolution. Some variant iVEGI peptides were predicted to have differential binding affinity to class I or class II HLAs, with the more pronounced alterations found in HLA class II. This is congruent with previous studies showing that the contribution of $CD4^+$ T cells for the immunity against TB is more relevant than the one of $CD8^+$ T cells [115,172]. Distinct host-pathogen interactions could be the cause of failure in producing a universal vaccine with high worldwide population coverage. Indeed, three distinct TB vaccine candidates were found to have considerably different overall population coverage, thus threatening their efficacy in certain regions of the globe [173]. In this context, much of the existing T cell epitope discovery studies in *M. tuberculosis* field have been using HLA molecules with high frequency in white individuals, which is likely to have a negative impact in the efficacy of the vaccine in several world regions with high TB burden [141]. The *in silico*-based approach followed in this thesis addresses this bias by including in the predictions HLA alleles that are predominant in TB endemic regions. The largest variations in HLA-binding affinity were concentrated in three specific codons of three distinct iVEGI-encoded proteins and were estimated to have a large impact in the world population coverage. Interestingly, for these three amino acid substitutions, there were significant correlations between the alterations in HLA class II population coverage and the frequency of the MTBC strains harbouring the selected variants. Therefore, the referred substitutions may represent evolutionary markers in MTBC of sympatric adaptation driven by $CD4^+$ T cells. The binding affinity and the stability of wild-type and variant forms of the peptides containing these markers was measured *in vitro*. The *in vitro* assays showed that the single amino acid substitutions predicted to be under diversifying selection in both $AccD2_{228-241}$ and $Rv0987_{164-177}$ induced large alterations (> 58%) in the binding affinity between wild-

type and variant peptides, which was not observed for Rv0988$_{185-202}$. However, the most striking finding was that although the wild-type sequence of AccD2$_{228-241}$ did not bind to DRB1*01:01, the lineage 1-associated variant peptide displayed a level of binding affinity and stability as high as the experimentally confirmed immunodominant epitope ESAT-6$_{3-17}$ [155]. Since the stability of the HLA-peptide complex was associated with immunodominance [150–153], the variant AccD2$_{228-241}$ peptide may constitute a relevant lineage-specific CD4^{+} T cell epitope.

CONCLUSION AND FUTURE PERSPECTIVES

The workflow described in this thesis allowed the identification of predicted T cell epitopes under diversifying selection in iVEGI, their study in terms of envisaged HLA-binding, impact in worldwide population coverage and geographic association between the human population and MTBC and, finally, the functional validation of the most promising candidates. The use of these approaches provided evidence for the existence of HLA-restricted T cell epitopes evolving under diversifying selection across the MTBC in response to immune pressure. From the 11 amino acid substitutions predicted to be under diversifying selection in iVEGI, three (AccD2 R233L, Rv0987 V169L and Rv0988 L191A) were associated with immune system pressure *in silico*. Of this, strong evidences to be a variable T cell epitope were confirmed *in vitro* for a peptide encompassing AccD2 R233L. However, T cells cannot be discarded as drivers of Rv0987 V169L and Rv0988 L191A, as performing the *in vitro* binding assays with other HLA predicted to have variable binding affinity (or even with other peptide) could yield different results. For the eight remaining sites, the responsible selective pressure has yet to be uncovered. Protein signature recognition tools such as InterProScan [174] could be the starting point to understand the cause and importance of these substitutions to *M. tuberculosis*. These tools perform homology search for conserved motifs, enzymatic active sites and many other features [174], and are thus a vital aid in the functional characterization of a protein.

The variant form of $AccD2_{228-241}$ was confirmed to have different binding affinity and stabilities comparatively to the wild-type. Nevertheless, further experiments are needed to ensure that the peptide is presented to T cells and is differentially immunogenic in humans infected with lineage 1 strains. To achieve this, wild-type and mutant peptides could be synthesised and used to stimulated blood collected from TB patients that i) express the HLA class II of interest and ii) were infected with lineage 1 strains. Assuming that memory T cells are created upon the recognition of the $AccD2_{228-241}$-DRB1*01:01 complex, differences in the levels and type of T cell responses are expected upon stimulation. For instance, this could be assessed by the quantification of cytokines known to be associated with TB disease, such as IFN-γ and interleukin-10 (IL-10) [1]. The implications of this study can also be extended to vaccination field, i.e. the BCG vaccine strains harbour the wild-type $AccD2_{228-241}$ peptide that was found not to bind to DRB1*01:01. It is thus tempting to speculate that boosting the BCG-induced immune response with the variant $AccD2_{228-241}$ could influence the protection to infections caused by lineage 1 MTBC strains.

Overall, the results presented in this thesis contributed to highlight T cells as one of the drivers in the evolutionary process of highly expressed *M. tuberculosis* genes, with consequences in the host-pathogen sympatric associations. The findings further support the impact of MTBC diversity in host immune-responses and pave the way to the discovery of novel MTBC lineage-specific epitopes. A better understanding of the molecular targets under T cell driven evolution will offer unprecedented tools in the development of TB preventive and therapeutic strategies with broader effectiveness.

MATERIALS AND METHODS

Sequence retrieval

To comprehensively represent all the seven known phylogenetic lineages of the MTBC, two previously described *M. tuberculosis* whole-genome sequence sets [51,111] were combined with five additional *M. tuberculosis* genomes publically available in the databases (Beijing/NITR203, CAS/NITR204, EAI5/NITR206, Haarlem/NITR202 and NCGM2209; GenBank accession numbers NC_021054, NC_021193, NC_021194, NC_021192 and NZ_DF126614, respectively), resulting in a total of 270 MTBC genomes (48 from lineage 1, 49 from lineage 2, 40 from lineage 3, 95 from lineage 4, 17 from lineages 5, 17 from lineages 6 and 4 from lineage 7). The region of 44,849 nucleotides known as iVEGI [157] was extracted with MegaBlast [176] from the 270 *M. tuberculosis* genomes using H37Rv (GenBank accession number NC_000962.3) as the reference. Three *pe_pgrs* genes present in the iVEGI region (*pe_pgrs16*, *pe_pgrs17* and *pe_pgrs18*) displayed high homology between its coding and intergenic repetitive regions [177,178] and were thus excluded from the analysis due to the technical difficulties in obtaining reliable nucleotide sequences. Multiple sequence alignments were performed by MUSCLE [179], with subsequent manual curation.

Nucleotide diversity and recombination tests

Whole genome and iVEGI nucleotide sequences were divided into 100 equal parts and its nucleotide diversity (π) was calculated using MEGA-CC [180]. Statistical significance in the comparisons among groups was assessed with *t*-test in SPSS v22. Nucleotide sequences of iVEGI were screened for recombination events through the automated RDP, GENECONV and Bootscan algorithms implemented in RDP v4.36 software [181].

Analysis of diversifying selection

Signatures of diversifying selection were detected by the CODEML program included in PAML package [105] and by FUBAR [107] implementation available on the Datamonkey web server (http://www.datamonkey.org) [182]. Inference of diversifying selection in CODEML was performed through pairwise comparison of site models that do not allow ratios of nonsynonymous to synonymous substitutions greater than one (ω>1) with an alternative model that allow ω>1 (M1-M2a and M7-M8), by a likelihood ratio test (LRT). As recommended [105], a Bayes empirical Bayes (BEB) approach [106] was used to calculate the posterior probability (BEB pp) of each codon to be under diversifying selection only when LRT values > 5.99 (meaning that the model admitting selection fitted better to data at a 5% significance level). The minimum requirement for a given site to be considered as evolving under diversifying selection was BEB pp > 0.95 in one of the M2a or M8 CODEML models. In FUBAR, the posterior probability of the nonsynonymous substitution rate (β) to be greater than the synonymous substitution rate (pp β>α) was considered as indicative of diversifying selection only for pp β>α values above 0.9.

T cell epitope prediction

iVEGI regions comprising amino acid residues under diversifying selection were investigated for the presence of $CD4^+$ T and $CD8^+$ T cell epitopes with NetMHCpan 2.8 [144] and NetMHCIIpan 3.0 [145], respectively. These methods were chosen since they were reported to be the best overall predictors [146–148]. A comprehensive set of HLA alleles (Supplementary Table 2) was selected based on its high prevalence in the human population worldwide [183,184] or in TB endemic regions [173]. HLA-C alleles were not selected for the HLA class I predictions due to its limited polymorphism and lower expression levels (about 90% less) comparatively to HLA-A and HLA-B [185]. Regarding HLA class II, HLA-DP and HLA-DQ loci were not considered since the predictions are thought to be less reliable due to few experimental binding data available [145]. On the –DR locus, only DRB1 alleles were used, as they bind the majority of the *M. tuberculosis* epitopes and were described to be five times more expressed than other DR alleles [186]. Half maximal inhibitory concentrations (IC_{50}) < 50 nM or rank scores < 0.5% were used as cutoff for high binding affinity peptides. All peptide lengths available were used, for both HLA class I (peptides of 8-14 amino acids) and HLA class II (peptides of 9-19 amino acids). To allow comparison, a normalization to the number of peptide lengths and HLA molecules tested was applied and the results were expressed as the percentage of the maximum possible number of high binding

affinity peptides (Supplementary Figure 1). In cases of more than one amino acid substitution per codon, only the substitution with the highest impact on the predictions was represented in the figures.

Assessment of the worldwide population coverage and the geographic distribution of the MTBC

The fraction of the human population with the ability to recognize the predicted epitopes (worldwide population coverage) was estimated using the implementation of a previously described algorithm [135] at the Immune Epitope Database (http://tools.immuneepitope.org/tools/population/iedb_input) [160]. The input data were the HLA alleles with at least one high binding affinity peptide overlapping the selected sites (Supplementary Table 3). The frequency of the MTBC lineages (n = 41974 strains) across 15 geographic regions (Supplementary Figure 2) was obtained from the SITVITWEB database [44]. Lineages were assigned to spoligotype data in accordance with the literature [187]. Unknown spoligotypes or the ones with regional frequencies below 1% were excluded from the analysis. The correlation between variations in the population coverage and the frequency of MTBC lineages was evaluated with the non-parametric Kendall's Tau (τ) test in SPSS Statistics v22. A p-value < 0.05 was applied as threshold for statistical significance.

In vitro HLA-binding assays

Candidate epitopes were tested for their binding affinity to DRA*01:01 DRB1*01:01 with the ProImmune REVEAL® HLA-peptide binding assay. All the peptides were synthetized with high purity (Supplementary Table 5). Binding affinity was determined as a function of the signal generated by a fluorescently labelled antibody against the native conformation of the HLA-peptide complex. A binding score was calculated for each peptide by comparison with a previously validated (both *in vitro* and in TB patients [155]) $CD4^+$ T cell epitope (ESAT-6_{3-17}). Measurements were conducted at 0 and 24 h following incubation at 37°C. To obtain the stability index of the HLA-peptide complex, approximated half-life times were derived through a one-phase dissociation equation and subsequent multiplication by the binding score of each peptide. HLA-peptide complexes with resulting values ≥ six were considered to be highly stable.

REFERENCES

1. O'Garra A, Redford PS, McNab FW, Bloom CI, Wilkinson RJ, Berry MPR. The immune response in tuberculosis. Annu Rev Immunol. 2013;31: 475–527. doi:10.1146/annurev-immunol-032712-095939

2. Davies PDO, Gordon SB, Davies G, editors. Clinical Tuberculosis. 5th ed. Boca Raton, FL: CRC Press; 2014.

3. Yang D, Kong Y. The bacterial and host factors associated with extrapulmonary dissemination of *Mycobacterium tuberculosis*. Front Biol. 2015;10(3): 252–261. doi:10.1007/s11515-015-1358-y

4. Ernst JD. The immunological life cycle of tuberculosis. Nat Rev Immunol. 2012;12(8): 581–591. doi:10.1038/nri3259

5. Fine PE. Variation in protection by BCG: implications of and for heterologous immunity. Lancet. 1995;346(8986): 1339–1345. doi:10.1016/S0140-6736(95)92348-9

6. Kaufmann SHE. Tuberculosis vaccines: time to think about the next generation. Semin Immunol. 2013;25(2): 172–181. doi:10.1016/j.smim.2013.04.006

7. World Health Organization. Global tuberculosis report 2014. Geneva; 2014.

8. Bychkov A V., Dorosevich AE, D'Souza JW. Postmortem investigations following human immunodeficiency virus infection. Int J Collab Res Intern Med Public Heal. 2009;1(2): 28–46.

9. Roberts CA, Buikstra JE. The bioarchaeology of tuberculosis: a global view on a reemerging disease. Gainesville, FL: University Press of Florida; 2003.

10. Baker O, Lee OY-C, Wu HHT, Besra GS, Minnikin DE, Llewellyn G, et al. Human tuberculosis predates domestication in ancient Syria. Tuberculosis (Edinb). 2015;95 Suppl 1: S4–S12. doi:10.1016/j.tube.2015.02.001

11. Hershkovitz I, Donoghue HD, Minnikin DE, Besra GS, Lee OY-C, Gernaey AM, et al. Detection and molecular characterization of 9,000-year-old *Mycobacterium tuberculosis* from a Neolithic settlement in the Eastern Mediterranean. PLoS One. 2008;3(10): e3426. doi:10.1371/journal.pone.0003426

12. Nicklisch N, Maixner F, Ganslmeier R, Friederich S, Dresely V, Meller H, et al. Rib lesions in skeletons from early neolithic sites in Central Germany: on the trail of tuberculosis at the onset of agriculture. Am J Phys Anthropol. 2012;149(3): 391–404. doi:10.1002/ajpa.22137

13. Zink AR, Sola C, Reischl U, Grabner W, Rastogi N, Wolf H, et al. Characterization of *Mycobacterium tuberculosis* complex DNAs from Egyptian mummies by spoligotyping. J Clin Microbiol. 2003;41(1): 359–367.

14. Mays S, Taylor GM. A first prehistoric case of tuberculosis from Britain. Int J Osteoarchaeol. 2003;13(4): 189–196. doi:10.1002/oa.671

15. Fletcher HA, Donoghue HD, Holton J, Pap I, Spigelman M. Widespread occurrence of *Mycobacterium tuberculosis* DNA from 18th-19th century Hungarians. Am J Phys Anthropol. 2003;120(2): 144–152. doi:10.1002/ajpa.10114

16. Nerlich AG, Lösch S. Paleopathology of human tuberculosis and the potential role of climate. Interdiscip Perspect Infect Dis. 2009;2009: 437187. doi:10.1155/2009/437187

17. Spiro SG, Silvestri GA, Agusti A. Clinical Respiratory Medicine. 4th ed. Philadelphia, PA: Saunders; 2012.

18. Sakula A. Robert Koch: centenary of the discovery of the tubercle bacillus, 1882. Can Vet J. 1983;24(4): 127–31.

19. Gangadharam PRJ, Jenkins PA, editors. Mycobacteria: Basic Aspects. 1st ed. New York, NY: Chapman & Hall; 1998.

20. Smith T. A comparative study of bovine tubercle bacilli and of human bacilli from sputum. J Exp Med. 1898;3: 451–511.

21. Aranaz A, Liébana E, Gómez-Mampaso E, Galán JC, Cousins D, Ortega A, et al. *Mycobacterium tuberculosis* subsp. *caprae* subsp. nov.: a taxonomic study of a new member of the *Mycobacterium tuberculosis* complex isolated from goats in Spain. Int J Syst Bacteriol. 1999;49(3): 1263–1273. doi:10.1099/00207713-49-3-1263

22. Bastida R, Loureiro J, Quse V, Bernardelli A, Rodríguez D, Costa E. Tuberculosis in a wild subantarctic fur seal from Argentina. J Wildl Dis. 1999;35(4): 796–798. doi:10.7589/0090-3558-35.4.796

23. Frota CC, Hunt DM, Buxton RS, Rickman L, Hinds J, Kremer K, et al. Genome structure in the vole bacillus, *Mycobacterium microti*, a member of the *Mycobacterium tuberculosis* complex with a low virulence for humans. Microbiology. 2004;150(5): 1519–1527.

24. Brosch R, Gordon S V., Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, et al. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. Proc Natl Acad Sci U S A. 2002;99(6): 3684–3689. doi:10.1073/pnas.052548299

25. Smith NH, Kremer K, Inwald J, Dale J, Driscoll JR, Gordon S V., et al. Ecotypes of the *Mycobacterium tuberculosis* complex. J Theor Biol. 2006;239(2): 220–225. doi:10.1016/j.jtbi.2005.08.036

26. Stead WW. The origin and erratic global spread of tuberculosis: How the past explains the present and is the key to the future. Clin Chest Med. 1997;18(1): 65–77. doi:10.1016/S0272-5231(05)70356-7

27. Tortoli E, Tortoli E. Impact of genotypic studies on mycobacterial taxonomy: the new mycobacteria of the 1990s. 2003;16(2): 319–354. doi:10.1128/CMR.16.2.319-354.2003

28. O'Reilly LM, Daborn CJ. The epidemiology of *Mycobacterium bovis* infections in animals and man: a review. Tuber Lung Dis. 1995;76 Suppl 1: 1–46. doi:10.1016/0962-8479(95)90591-X

29. Gordon S V, Eiglmeier K, Garnier T, Brosch R, Parkhill J, Barrell B, et al. Genomics of *Mycobacterium bovis*. Tuberculosis (Edinb). 2001;81(1/2): 157–163. doi:10.1054/tube.2000.0269

30. Supply P, Mazars E, Lesjean S, Vincent V, Gicquel B, Locht C. Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. Mol Microbiol. 2000;36(3): 762–771. doi:10.1046/j.1365-2958.2000.01905.x

31. Supply P, Warren RM, Bañuls A-L, Lesjean S, Van Der Spuy GD, Lewis L-A, et al. Linkage disequilibrium between minisatellite loci supports clonal evolution of *Mycobacterium tuberculosis* in a high tuberculosis incidence area. Mol Microbiol. 2003;47(2): 529–538. doi:10.1046/j.1365-2958.2003.03315.x

32. Cvetnic Z, Katalinic-Jankovic V, Sostaric B, Spicic S, Obrovac M, Marjanovic S, et al. *Mycobacterium caprae* in cattle and humans in Croatia. Int J Tuberc Lung Dis. 2007;11(6): 652–658.

33. Park D, Qin H, Jain S, Preziosi M, Minuto JJ, Mathews WC, et al. Tuberculosis due to *Mycobacterium bovis* in patients coinfected with human immunodeficiency virus. Clin Infect Dis. 2010;51(11): 1343–1346. doi:10.1086/657118

34. Berg S, Smith NH. Why doesn't bovine tuberculosis transmit between humans? Trends Microbiol. 2014;22(10): 552–553. doi:10.1016/j.tim.2014.08.007

35. Blouin Y, Cazajous G, Dehan C, Soler C, Vong R, Hassan MO, et al. Progenitor "*Mycobacterium canettii*" clone responsible for lymph node tuberculosis epidemic, Djibouti. Emerg Infect Dis. 2014;20(1): 21–28. doi:10.3201/eid2001.130652

36. Supply P, Marceau M, Mangenot S, Roche D, Rouanet C, Khanna V, et al. Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of *Mycobacterium tuberculosis*. Nat Genet. 2013;45(2): 172–179. doi:10.1038/ng.2517

37. Gutierrez MC, Brisse S, Brosch R, Fabre M, Omaïs B, Marmiesse M, et al. Ancient origin and gene mosaicism of the progenitor of *Mycobacterium tuberculosis*. PLoS Pathog. 2005;1(1): e5. doi:10.1371/journal.ppat.0010005

38. Krzywinska E, Krzywinski J, Schorey JS. Naturally occurring horizontal gene transfer and homologous recombination in *Mycobacterium*. Microbiology. 2004;150: 1707–1712. doi:10.1099/mic.0.27088-0

39. Derbyshire KM, Gray TA. Distributive conjugal transfer: new insights into horizontal gene transfer and genetic exchange in mycobacteria. Microbiol Spectr. 2014;2(1): MGM2–0022–2013. doi:10.1128/microbiolspec.MGM2-0022-2013

40. Bottai D, Stinear TP, Supply P, Brosch R. Mycobacterial pathogenomics and evolution. Microbiol Spectr. 2014;2(1): MGM2–0025–2013. doi:10.1128/microbiolspec.MGM2-0025-2013

41. Jagielski T, Van Ingen J, Rastogi N, Dziadek J, Mazur PK, Bielecki J. Current methods in the molecular typing of *Mycobacterium tuberculosis* and other mycobacteria. Biomed Res Int. 2014;2014: 645802. doi:10.1155/2014/645802

42. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, et al. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J Clin Microbiol. 1997;35(4): 907–914.

43. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rüsch-Gerdes S, Willery E, et al. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. J Clin Microbiol. 2006;44(12): 4498–4510. doi:10.1128/JCM.01392-06

44. Demay C, Liens B, Burguière T, Hill V, Couvin D, Millet J, et al. SITVITWEB - A publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology. Infect Genet Evol. 2012;12(4): 755–766. doi:10.1016/j.meegid.2012.02.004

45. Brudey K, Driscoll JR, Rigouts L, Prodinger WM, Gori A, Al-Hajoj SA, et al. *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol. 2006;6: 23. doi:10.1186/1471-2180-6-23

46. Hirsh AE, Tsolaki AG, DeRiemer K, Feldman MW, Small PM. Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. Proc Natl Acad Sci U S A. 2004;101(14): 4871–4876. doi:10.1073/pnas.0305627101

47. Tsolaki AG, Gagneux S, Pym AS, de la Salmoniere Y-OLG, Kreiswirth BN, Soolingen D Van, et al. Genomic deletions classify the Beijing/W strains as a distinct genetic lineage of *Mycobacterium tuberculosis*. J Clin Microbiol. 2005;43(7): 3185–3191. doi:10.1128/JCM.43.7.3185-3191.2005

48. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, et al. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. Proc Natl Acad Sci U S A. 2006;103(8): 2869–2873. doi:10.1073/pnas.0511240103

49. Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, Homolka S, et al. High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. PLoS Biol. 2008;6(12): e311. doi:10.1371/journal.pbio.0060311

50. Roetzer A, Diel R, Kohl T a., Rückert C, Nübel U, Blom J, et al. Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. PLoS Med. 2013;10(2): e1001387. doi:10.1371/journal.pmed.1001387

51. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, et al. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat Genet. 2013;45(10): 1176–1182. doi:10.1038/ng.2744

52. Firdessa R, Berg S, Hailu E, Schelling E, Gumi B, Erenso G, et al. Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis, Ethiopia. Emerg Infect Dis. 2013;19(3): 460–463. doi:10.3201/eid1903.120256

53. Wirth T, Hildebrand F, Allix-Béguec C, Wölbeling F, Kubica T, Kremer K, et al. Origin, spread and demography of the *Mycobacterium tuberculosis* complex. PLoS Pathog. 2008;4(9): e1000160. doi:10.1371/journal.ppat.1000160

54. Rasmussen M, Guo X, Wang Y, Lohmueller K, Rasmussen S, Albrechtsen A, et al. An Aboriginal Australian genome reveals separate human dispersals into Asia. Science. 2011;334(6052): 94–98. doi:10.1126/science.1211177

55. Oppenheimer S. Out-of-Africa, the peopling of continents and islands: tracing uniparental gene trees across the map. Philos Trans R Soc B Biol Sci. 2012;367(1590): 770–784. doi:10.1098/rstb.2011.0306

56. Comas I, Gagneux S. The past and future of tuberculosis research. PLoS Pathog. 2009;5(10): e1000600. doi:10.1371/journal.ppat.1000600

57. Yimer SA, Norheim G, Namouchi A, Zegeye ED, Kinander W, Tønjum T, et al. *Mycobacterium tuberculosis* lineage 7 strains are associated with prolonged patient delay in seeking treatment for pulmonary tuberculosis in Amhara Region, Ethiopia. J Clin Microbiol. 2015;53(4): 1301–1309. doi:10.1128/JCM.03566-14

58. Parwati I, Alisjahbana B, Apriani L, Soetikno RD, Ottenhoff TH, van der Zanden AGM, et al. *Mycobacterium tuberculosis* Beijing genotype is an independent risk factor for tuberculosis treatment failure in Indonesia. J Infect Dis. 2010;201(4): 553–557. doi:10.1086/650311

59. Stavrum R, PrayGod G, Range N, Faurholt-Jepsen D, Jeremiah K, Faurholt-Jepsen M, et al. Increased level of acute phase reactants in patients infected with modern *Mycobacterium tuberculosis* genotypes in Mwanza, Tanzania. BMC Infect Dis. 2014;14: 309. doi:10.1186/1471-2334-14-309

60. Albanna AS, Reed MB, Kotar K V, Fallow A, McIntosh FA, Behr MA, et al. Reduced transmissibility of East African Indian strains of *Mycobacterium tuberculosis*. PLoS One. 2011;6(9): e25075. doi:10.1371/journal.pone.0025075

61. Cowley D, Govender D, February B, Wolfe M, Steyn L, Evans J, et al. Recent and rapid emergence of W-Beijing strains of *Mycobacterium tuberculosis* in Cape Town, South Africa. Clin Infect Dis. 2008;47(10): 1252–1259. doi:10.1086/592575

62. Kato-Maeda M, Kim EY, Flores L, Jarlsberg LG, Osmond D, Hopewell PC. Differences among sublineages of the East-Asian lineage of *Mycobacterium tuberculosis* in genotypic clustering. Int J Tuberc Lung Dis. 2010;14(5): 538–544.

63. Gehre F, Antonio M, Faïhun F, Odoun M, Uwizeye C, de Rijk P, et al. The first phylogeographic population structure and analysis of transmission dynamics of *M. africanum* West African 1– combining molecular data from Benin, Nigeria and Sierra Leone. PLoS One. 2013;8(10): e77000. doi:10.1371/journal.pone.0077000

64. Carmona J, Cruz A, Moreira-Teixeira L, Sousa C, Sousa J, Osorio NS, et al. *Mycobacterium tuberculosis* strains are differentially recognized by TLRs with an impact on the immune response. PLoS One. 2013;8(6): e67277. doi:10.1371/journal.pone.0067277

65. Portevin D, Gagneux S, Comas I, Young D. Human macrophage responses to clinical isolates from the *Mycobacterium tuberculosis* complex discriminate between ancient and modern lineages. PLoS Pathog. 2011;7(3): e1001307. doi:10.1371/journal.ppat.1001307

66. Chen Y-Y, Chang J-R, Huang W-F, Hsu S-C, Kuo S-C, Sun J-R, et al. The pattern of cytokine production *in vitro* induced by ancient and modern Beijing *Mycobacterium tuberculosis* strains. PLoS One. 2014;9(4): e94296. doi:10.1371/journal.pone.0094296

67. Reiling N, Homolka S, Walter K, Brandenburg J, Niwinski L, Ernst M, et al. Clade-specific virulence patterns of *Mycobacterium tuberculosis* complex strains in human primary macrophages and aerogenically infected mice. MBio. 2013;4(4): e00250–13. doi:10.1128/mBio.00250-13

68. Marais BJ, Hesseling a. C, Schaaf HS, Gie RP, Van Helden PD, Warren RM. *Mycobacterium tuberculosis* transmission is not related to household genotype in a setting of high endemicity. J Clin Microbiol. 2009;47(5): 1338–1343. doi:10.1128/JCM.02490-08

69. Krishnan N, Malaga W, Constant P, Caws M, Thi Hoang Chau T, Salmons J, et al. *Mycobacterium tuberculosis* lineage influences innate immune response and virulence and is associated with distinct cell envelope lipid profiles. PLoS One. 2011;6(9): e23870. doi:10.1371/journal.pone.0023870

70. Jereb J, Etkind SC, Joglar OT, Moore M, Taylor Z. Tuberculosis contact investigations: outcomes in selected areas of the United States, 1999. Int J Tuberc Lung Dis. 2003;7(Suppl 3): S384–390.

71. Marks SM, Taylor Z, Qualls NL, Shrestha-Kuwahara RJ, Wilce M a., Nguyen CH. Outcomes of contact investigations of infectious tuberculosis patients. Am J Respir Crit Care Med. 2000;162(6): 2033–2038. doi:10.1164/ajrccm.162.6.2004022

72. Thye T, Owusu-Dabo E, Vannberg FO, van Crevel R, Curtis J, Sahiratmadja E, et al. Common variants at 11p13 are associated with susceptibility to tuberculosis. Nat Genet. 2012;44(3): 257–259. doi:10.1038/ng.1080

73. Al-Muhsen S, Casanova JL. The genetic heterogeneity of mendelian susceptibility to mycobacterial diseases. J Allergy Clin Immunol. 2008;122(6): 1043–1051. doi:10.1016/j.jaci.2008.10.037

74. Bustamante J, Boisson-Dupuis S, Abel L, Casanova J-L. Mendelian susceptibility to mycobacterial disease: genetic, immunological, and clinical features of inborn errors of IFN-γ immunity. Semin Immunol. 2014;26(6): 454–470. doi:10.1016/j.smim.2014.09.008

75. Casanova JL. Mendelian susceptibility to mycobacterial infection in man. Swiss Med Wkly. 2001;131(31): 445–454.

76. Stead W, Senner J, Reddick W, Lofgren J. Racial differences in susceptibility to infection by *Mycobacterium tuberculosis*. N Engl J Med. 1990;322(7): 422–427. doi:10.1056/NEJM199002153220702

77. Coussens AK, Wilkinson RJ, Nikolayevskyy V, Elkington PT, Hanifa Y, Islam K, et al. Ethnic variation in inflammatory profile in tuberculosis. PLoS Pathog. 2013;9(7): e1003468. doi:10.1371/journal.ppat.1003468

78. Möller M, De Wit E, Hoal EG. Past, present and future directions in human genetic susceptibility to tuberculosis. FEMS Immunol Med Microbiol. 2010;58(1): 3–26. doi:10.1111/j.1574-695X.2009.00600.x

79. Abel L, El-Baghdadi J, Bousfiha AA, Casanova J-L, Schurr E. Human genetics of tuberculosis: a long and winding road. Philos Trans R Soc Lond B Biol Sci. 2014;369(1645): 20130428. doi:10.1098/rstb.2013.0428

80. Stein CM. Genetic epidemiology of tuberculosis susceptibility: impact of study design. PLoS Pathog. 2011;7(1): e1001189. doi:10.1371/journal.ppat.1001189

81. Reed MB, Pichler VK, McIntosh F, Mattia A, Fallow A, Masala S, et al. Major *Mycobacterium tuberculosis* lineages associate with patient country of origin. J Clin Microbiol. 2009;47(4): 1119–1128. doi:10.1128/JCM.02142-08

82. Gagneux S. Host-pathogen coevolution in human tuberculosis. Philos Trans R Soc Lond B Biol Sci. 2012;367(1590): 850–859. doi:10.1098/rstb.2011.0316

83. De Jong BC, Hill PC, Aiken A, Awine T, Antonio M, Adetifa IM, et al. Progression to active tuberculosis, but not transmission, varies by *Mycobacterium tuberculosis* lineage in the Gambia. J Infect Dis. 2008;198(7): 1037–1043. doi:10.1086/591504

84. Caws M, Thwaites G, Dunstan S, Hawn TR, Lan NTN, Thuong NTT, et al. The influence of host and bacterial genotype on the development of disseminated disease with *Mycobacterium tuberculosis*. PLoS Pathog. 2008;4(3): e1000034. doi:10.1371/journal.ppat.1000034

85. Herb F, Thye T, Niemann S, Browne ENL, Chinbuah M a., Gyapong J, et al. *ALOX5* variants associated with susceptibility to human pulmonary tuberculosis. Hum Mol Genet. 2008;17(7): 1052–1060. doi:10.1093/hmg/ddm378

86. Intemann CD, Thye T, Niemann S, Browne ENL, Amanua Chinbuah M, Enimil A, et al. Autophagy gene variant *IRGM* -261T contributes to protection from tuberculosis caused by *Mycobacterium tuberculosis* but not by *M. africanum* strains. PLoS Pathog. 2009;5(9): e1000577. doi:10.1371/journal.ppat.1000577

87. Thye T, Niemann S, Walter K, Homolka S, Intemann CD, Chinbuah MA, et al. Variant G57E of mannose binding lectin associated with protection against tuberculosis caused by *Mycobacterium africanum* but not by *M. tuberculosis*. PLoS One. 2011;6(6): e20908. doi:10.1371/journal.pone.0020908

88. Brites D, Gagneux S. Co-evolution of *Mycobacterium tuberculosis* and Homo sapiens. Immunol Rev. 2015;264(1): 6–24. doi:10.1111/imr.12264

89. Namouchi A, Didelot X, Schöck U, Gicquel B, Rocha EP. After the bottleneck: Genome-wide diversification of the *Mycobacterium tuberculosis* complex by mutation, recombination, and natural selection. Genome Res. 2012;22(4): 721–734. doi:10.1101/gr.129544.111

90. Yang Z, Bielawski JP. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 2000;15(12): 496–503. doi:10.1016/S0169-5347(00)01994-7

91. Nielsen R. Molecular signatures of natural selection. Annu Rev Genet. 2005;39: 197–218. doi:10.1146/annurev.genet.39.073003.112420

92. Woolhouse MEJ, Webster JP, Domingo E, Charlesworth B, Levin BR. Biological and biomedical implications of the co-evolution of pathogens and their hosts. Nat Genet. 2002;32(4): 569–577. doi:10.1038/ng1202-569

93. Allison AC. Protection afforded by sickle-cell trait against subtertian malareal infection. Br Med J. 1954;1(4857): 290–294.

94. Boucher C a, O'Sullivan E, Mulder JW, Ramautarsing C, Kellam P, Darby G, et al. Ordered appearance of zidovudine resistance mutations during treatment of 18 human immunodeficiency virus-positive subjects. J Infect Dis. 1992;165(1): 105–110. doi:10.1093/infdis/165.1.105

95. Biswas S, Akey JM. Genomic insights into positive selection. Trends Genet. 2006;22(8): 437–446. doi:10.1016/j.tig.2006.06.005

96. Hanchard NA, Rockett KA, Spencer C, Coop G, Pinder M, Jallow M, et al. Screening for recently selected alleles by analysis of human haplotype similarity. Am J Hum Genet. 2006;78(1): 153–159. doi:10.1086/499252

97. Fay JC. Weighing the evidence for adaptation at the molecular level. Trends Genet. 2011;27(9): 343–349. doi:10.1016/j.tig.2011.06.003

98. Zhai W, Nielsen R, Slatkin M. An investigation of the statistical power of neutrality tests based on comparative and population genetic data. Mol Biol Evol. 2009;26(2): 273–283. doi:10.1093/molbev/msn231

99. Wright SI, Charlesworth B. The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model. Genetics. 2004;168(2): 1071–1076. doi:10.1534/genetics.104.026500

100. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 1986;3(5): 418–426.

101. Yang Z, Nielsen R, Goldman N, Pedersen AM. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 2000;155(1): 431–449.

102. Crandall K a, Kelsey CR, Imamichi H, Lane HC, Salzman NP. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. Mol Biol Evol. 1999;16(3): 372–382. doi:10.1093/oxfordjournals.molbev.a026118

103. Bofkin L, Goldman N. Variation in evolutionary processes at different codon positions. Mol Biol Evol. 2007;24(2): 513–521. doi:10.1093/molbev/msl178

104. Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 1994;11(5): 725–736.

105. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24(8): 1586–1591. doi:10.1093/molbev/msm088

106. Yang Z, Wong WSW, Nielsen R. Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol. 2005;22(4): 1107–1118. doi:10.1093/molbev/msi097

107. Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Pond SLK, et al. FUBAR: a Fast, Unconstrained Bayesian Approximation for inferring selection. Mol Biol Evol. 2013;30(5): 1196–1205. doi:10.1093/molbev/mst030

108. Anisimova M, Nielsen R, Yang Z. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics. 2003;164(3): 1229–1236.

109. Smith GJD, Naipospos TSP, Nguyen TD, de Jong MD, Vijaykrishna D, Usman TB, et al. Evolution and adaptation of H5N1 influenza virus in avian and human hosts in Indonesia and Vietnam. Virology. 2006;350(2): 258–268. doi:10.1016/j.virol.2006.03.048

110. Smith EE, Sims EH, Spencer DH, Kaul R, Olson M V. Evidence for diversifying selection at the pyoverdine locus of *Pseudomonas aeruginosa*. J Bacteriol. 2005;187(6): 2138–2147. doi:10.1128/JB.187.6.2138-2147.2005

111. Osório NS, Rodrigues F, Gagneux S, Pedrosa J, Pinto-Carbó M, Castro AG, et al. Evidence for diversifying selection in a set of *Mycobacterium tuberculosis* genes in response to antibiotic- and nonantibiotic-related pressure. Mol Biol Evol. 2013;30(6): 1326–1336. doi:10.1093/molbev/mst038

112. Farhat M, Shapiro B, Kieser K, Sultana R, Jacobson K, Victor T, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. Nat Genet. 2013;45(10): 1183–1189. doi:10.1038/ng.2747

113. Cooper AM. Cell-mediated immune responses in tuberculosis. Annu Rev Immunol. 2009;27: 393–422. doi:10.1146/annurev.immunol.021908.132703

114. Fenner L, Egger M, Bodmer T, Furrer H, Ballif M, Battegay M, et al. HIV infection disrupts the sympatric host-pathogen relationship in human tuberculosis. PLoS Genet. 2013;9(3): e1003318. doi:10.1371/journal.pgen.1003318

115. North R, Jung Y. Immunity to tuberculosis. Annu Rev Immunol. 2004;22: 599–623. doi:10.1146/annurev.immunol.22.012703.104635

116. Lancioni C, Nyendak M, Kiguli S, Zalwango S, Mori T, Mayanja-Kizza H, et al. $CD8^+$ T cells provide an immunologic signature of tuberculosis in young children. Am J Respir Crit Care Med. 2012;185(2): 206–212. doi:10.1164/rccm.201107-1355OC

117. Chen CY, Huang D, Wang RC, Shen L, Zeng G, Yao S, et al. A critical role for CD8 T cells in a nonhuman primate model of tuberculosis. PLoS Pathog. 2009;5(4): e1000392. doi:10.1371/journal.ppat.1000392

118. Neefjes J, Jongsma MLM, Paul P, Bakke O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. Nat Rev Immunol. 2011;11(12): 823–836. doi:10.1038/nri3084

119. Rudolph MG, Stanfield RL, Wilson I a. How TCRs bind MHCs, peptides, and coreceptors. Annu Rev Immunol. 2006;24: 419–466. doi:10.1146/annurev.immunol.23.021704.115658

120. Huang J, Meyer C, Zhu C. T cell antigen recognition at the cell membrane. Mol Immunol. 2012;52(3-4): 155–164. doi:10.1016/j.molimm.2012.05.004

121. Zuñiga J, Torres-García D, Santos-Mendoza T, Rodriguez-Reyna TS, Granados J, Yunis EJ. Cellular and humoral mechanisms involved in the control of tuberculosis. Clin Dev Immunol. 2012;2012: 193923. doi:10.1155/2012/193923

122. Paul S, Kolla R V, Sidney J, Weiskopf D, Fleri W, Kim Y, et al. Evaluating the immunogenicity of protein drugs by applying *in vitro* MHC binding data and the immune epitope database and analysis resource. Clin Dev Immunol. 2013;2013: 467852. doi:10.1155/2013/467852

123. Patronov A, Doytchinova I. T-cell epitope vaccine design by immunoinformatics. Open Biol. 2013;3(1): 120139. doi:10.1098/rsob.120139

124. De Groot AS, Sbai H, Aubin C Saint, McMurry J, Martin W. Immuno-informatics: Mining genomes for vaccine components. Immunol Cell Biol. 2002;80(3): 255–269. doi:10.1046/j.1440-1711.2002.01092.x

125. Davies MN, Flower DR. Harnessing bioinformatics to discover new vaccines. Drug Discov Today. 2007;12(9-10): 389–395. doi:10.1016/j.drudis.2007.03.010

126. Mustafa AS, Al-Attiyah R, Hanif SNM, Shaban FA. Efficient testing of large pools of *Mycobacterium tuberculosis* RD1 peptides and identification of major antigens and immunodominant peptides recognized by human Th1 cells. Clin Vaccine Immunol. 2008;15(6): 916–924. doi:10.1128/CVI.00056-08

127. Mustafa AS, Shaban F. Mapping of Th1-cell epitope regions of *Mycobacterium tuberculosis* protein MPT64 (*Rv1980c*) using synthetic peptides and T-cell lines from *M. tuberculosis*-infected healthy humans. Med Princ Pract. 2010;19(2): 122–128. doi:10.1159/000273073

128. Brusic V, Petrovsky N. Immunoinformatics and its relevance to understanding human immune disease. Expert Rev Clin Immunol. 2005;1(1): 145–157. doi:10.1586/1744666X.1.1.145

129. Sette A, Rappuoli R. Reverse vaccinology: developing vaccines in the era of genomics. Immunity. 2010;33(4): 530–541. doi:10.1016/j.immuni.2010.09.017

130. Keogh E, Fikes J, Southwood S, Celis E, Chesnut R, Sette a. Identification of new epitopes from four different tumor-associated antigens: recognition of naturally processed epitopes correlates with HLA-A*0201-binding affinity. J Immunol. 2001;167(2): 787–796. doi:10.4049/jimmunol.167.2.787

131. Rosa DS, Ribeiro SP, Cunha-Neto E. $CD4^+$ T cell epitope discovery and rational vaccine design. Arch Immunol Ther Exp (Warsz). 2010;58(2): 121–130. doi:10.1007/s00005-010-0067-0

132. Trowsdale J, Knight JC. Major histocompatibility complex genomics and human disease. Annu Rev Genomics Hum Genet. 2013;14: 301–323. doi:10.1146/annurev-genom-091212-153455

133. Gonzalez-Galarza FF, Takeshita LYC, Santos EJM, Kempson F, Maia MHT, Silva a. LSD, et al. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. Nucleic Acids Res. 2015;43(D1): D784–D788. doi:10.1093/nar/gku1166

134. Schubert B, Lund O, Nielsen M. Evaluation of peptide selection approaches for epitope-based vaccine design. Tissue Antigens. 2013;82(4): 243–251. doi:10.1111/tan.12199

135. Bui H-H, Sidney J, Dinh K, Southwood S, Newman MJ, Sette A. Predicting population coverage of T-cell epitope-based diagnostics and vaccines. BMC Bioinformatics. 2006;7: 153. doi:10.1186/1471-2105-7-153

136. Caccamo N, Meraviglia S, La Mendola C, Bosze S, Hudecz F, Ivanyi J, et al. Characterization of HLA-DR- and TCR-binding residues of an immunodominant and genetically permissive peptide of the 16-kDa protein of *Mycobacterium tuberculosis*. Eur J Immunol. 2004;34(8): 2220–2229. doi:10.1002/eji.200425090

137. Höhn H, Kortsik C, Tully G, Nilges K, Necker A, Freitag K, et al. Longitudinal analysis of *Mycobacterium tuberculosis* 19-kDa antigen-specific T cells in patients with pulmonary

tuberculosis: association with disease activity and cross-reactivity to a peptide from HIV_{env} gp120. Eur J Immunol. 2003;33(6): 1613–1623. doi:10.1002/eji.200323480

138. Comas I, Chakravartti J, Small PM, Galagan J, Niemann S, Kremer K, et al. Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. Nat Genet. 2010;42(6): 498–503. doi:10.1038/ng.590

139. Ivanyi J. Function and potentials of *M. tuberculosis* epitopes. Front Immunol. 2014;5: 107. doi:10.3389/fimmu.2014.00107

140. Kunnath-Velayudhan S, Porcelli SA. Recent advances in defining the immunoproteome of *Mycobacterium tuberculosis*. Front Immunol. 2013;4: 335. doi:10.3389/fimmu.2013.00335

141. Axelsson-Robertson R, Magalhaes I, Parida SK, Zumla A, Maeurer M. The immunological footprint of *Mycobacterium tuberculosis* T-cell epitope recognition. J Infect Dis. 2012;205(Suppl 2): S301–S315. doi:10.1093/infdis/jis198

142. Tomar N, De RK, editors. Immunoinformatics: a brief review. Immunoinformatics. 1st ed. New York, NY: Springer Science & Business Media; 2014. pp. 23–55. doi:10.1007/978-1-4939-1115-8_3

143. Zhang L, Udaka K, Mamitsuka H, Zhu S. Toward more accurate pan-specific MHC-peptide binding prediction: a review of current methods and tools. Brief Bioinform. 2012;13(3): 350–364. doi:10.1093/bib/bbr060

144. Hoof I, Peters B, Sidney J, Pedersen LE, Sette A, Lund O, et al. NetMHCpan, a method for MHC class I binding prediction beyond humans. Immunogenetics. 2009;61(1): 1–13. doi:10.1007/s00251-008-0341-z

145. Karosiene E, Rasmussen M, Blicher T, Lund O, Buus S, Nielsen M. NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. Immunogenetics. 2013;65(10): 711–724. doi:10.1007/s00251-013-0720-y

146. Chaves FA, Lee AH, Nayak JL, Richards KA, Sant AJ. The utility and limitations of current Web-available algorithms to predict peptides recognized by CD4 T cells in response to pathogen infection. J Immunol. 2012;188(9): 4235–4248. doi:10.4049/jimmunol.1103640

147. Zhang L, Chen Y, Wong H-S, Zhou S, Mamitsuka H, Zhu S. TEPITOPEpan: extending TEPITOPE for peptide binding prediction covering over 700 HLA-DR molecules. PLoS One. 2012;7(2): e30483. doi:10.1371/journal.pone.0030483

148. Trolle T, Metushi IG, Greenbaum JA, Kim Y, Sidney J, Lund O, et al. Automated benchmarking of peptide-MHC class I binding predictions. 2015;31(13): 2174–2181. doi:10.1093/bioinformatics/btv123

149. Berzofsky JA. Immunodominance in T lymphocyte recognition. Immunol Lett. 1988;18(2): 83–92. doi:10.1016/0165-2478(88)90046-6

150. Harndahl M, Rasmussen M, Roder G, Dalgaard Pedersen I, Sørensen M, Nielsen M, et al. Peptide-MHC class I stability is a better predictor than peptide affinity of CTL immunogenicity. Eur J Immunol. 2012;42(6): 1405–1416. doi:10.1002/eji.201141774

151. Lazarski CA, Chaves FA, Jenks SA, Wu S, Richards KA, Weaver JM, et al. The kinetic stability of MHC Class II : peptide complexes is a key parameter that dictates immunodominance. Immunity. 2005;23(1): 29–40. doi:10.1016/j.immuni.2005.05.009

152. Yin L, Trenh P, Guce A, Wieczorek M, Lange S, Sticht J, et al. Susceptibility to HLA-DM is determined by a dynamic conformation of major histocompatibility complex class II molecule bound with peptide. J Biol Chem. 2014;289(34): 23449–23464. doi:10.1074/jbc.M114.585539

153. Lazarski CA, Chaves FA, Sant AJ. The impact of DM on MHC class II-restricted antigen presentation can be altered by manipulation of MHC-peptide kinetic stability. J Exp Med. 2006;203(5): 1319–1328. doi:10.1084/jem.20060058

154. Brodin P, Rosenkrands I, Andersen P, Cole ST, Brosch R. ESAT-6 proteins: protective antigens and virulence factors? Trends Microbiol. 2004;12(11): 500–508. doi:10.1016/j.tim.2004.09.007

155. Arlehamn CSL, Sidney J, Henderson R, Greenbaum J a, James E a, Moutaftsi M, et al. Dissecting mechanisms of immunodominance to the common tuberculosis antigens ESAT-6, CFP10, Rv2031c (hspX), Rv2654c (TB7.7), and Rv1038c (EsxJ). J Immunol. 2012;188(10): 5020–5031. doi:10.4049/jimmunol.1103556

156. Kim A, Sadegh-Nasseri S. Determinants of immunodominance for CD4 T cells. Curr Opin Immunol. 2015;34: 9–15. doi:10.1016/j.coi.2014.12.005

157. Talaat AM, Lyons R, Howard ST, Johnston SA. The temporal expression profile of *Mycobacterium tuberculosis* infection in mice. Proc Natl Acad Sci U S A. 2004;101(13): 4602–4607. doi:10.1073/pnas.0306023101

158. Ward SK, Abomoelak B, Marcus S a, Talaat AM. Transcriptional profiling of *Mycobacterium tuberculosis* during infection: lessons learned. Front Microbiol. 2010;1: 121. doi:10.3389/fmicb.2010.00121

159. Coscolla M, Gagneux S. Consequences of genomic diversity in *Mycobacterium tuberculosis*. Semin Immunol. 2014;26(6): 431–444. doi:10.1016/j.smim.2014.09.012

160. Vita R, Overton J a., Greenbaum J a., Ponomarenko J, Clark JD, Cantrell JR, et al. The immune epitope database (IEDB) 3.0. Nucleic Acids Res. 2014;43(D1): D405–D412. doi:10.1093/nar/gku938

161. Andersen P, Kaufmann SHE. Novel vaccination strategies against tuberculosis. Cold Spring Harb Perspect Med. 2014;4(6): a018523. doi:10.1101/cshperspect.a018523

162. Pitt JM, Blankley S, McShane H, O'Garra A. Vaccination against tuberculosis: How can we better BCG? Microb Pathog. 2013;58: 2–16. doi:10.1016/j.micpath.2012.12.002

163. Robinson RT, Orme IM, Cooper AM, Cooper AM. The onset of adaptive immunity in the mouse model of tuberculosis and the factors that compromise its expression. Immunol Rev. 2015;264(1): 46–59. doi:10.1111/imr.12259

164. Wang C, Peyron P, Mestre O, Kaplan G, van Soolingen D, Gao Q, et al. Innate immune response to *Mycobacterium tuberculosis* Beijing and other genotypes. PLoS One. 2010;5(10): e13594. doi:10.1371/journal.pone.0013594

165. Sarkar R, Lenders L, Wilkinson KA, Wilkinson RJ, Nicol MP. Modern lineages of *Mycobacterium tuberculosis* exhibit lineage-specific patterns of growth and cytokine induction in human monocyte-derived macrophages. PLoS One. 2012;7(8): e43170. doi:10.1371/journal.pone.0043170

166. Iwasaki A, Medzhitov R. Control of adaptive immunity by the innate immune system. Nat Immunol. 2015;16(4): 343–353. doi:10.1038/ni.3123

167. Yuliwulandari R, Sachrowardi Q, Nakajima H, Kashiwase K, Hirayasu K, Mabuchi A, et al. Association of HLA-A, -B, and -DRB1 with pulmonary tuberculosis in western Javanese Indonesia. Hum Immunol. 2010;71(7): 697–701. doi:10.1016/j.humimm.2010.04.005

168. Wu F, Zhang W, Zhang L, Wu J, Li C, Meng X, et al. NRAMP1, VDR, *HLA-DRB1*, and *HLA-DQB1* gene polymorphisms in susceptibility to tuberculosis among the Chinese Kazakh population: a case-control study. Biomed Res Int. 2013;2013: 484535. doi:10.1155/2013/484535

169. Tong X, Chen L, Liu S, Yan Z, Peng S, Zhang Y, et al. Polymorphisms in *HLA-DRB1* Gene and the risk of tuberculosis: a meta-analysis of 31 studies. Lung. 2015;193(2): 309–318. doi:10.1007/s00408-015-9692-z

170. Salie M, Van Der Merwe L, Möller M, Daya M, Van Der Spuy GD, Van Helden PD, et al. Associations between human leukocyte antigen class I variants and the *Mycobacterium tuberculosis* subtypes causing disease. J Infect Dis. 2014;209(2): 216–223. doi:10.1093/infdis/jit443

171. Commandeur S, van den Eeden SJF, Dijkman K, Clark SO, van Meijgaarden KE, Wilson L, et al. The in vivo expressed *Mycobacterium tuberculosis* (IVE-TB) antigen Rv2034 induces $CD4^+$ T-cells that protect against pulmonary infection in HLA-DR transgenic mice and guinea pigs. Vaccine. 2014;32(29): 3580–3588. doi:10.1016/j.vaccine.2014.05.005

172. Mogues BT, Goodrich ME, Ryan L, Lacourse R, North RJ. The relative importance of T cell subsets in immunity and immunopathology of airborne *Mycobacterium tuberculosis* infection in mice. J Exp Med. 2001;193(3): 271–280. doi:10.1084/jem.193.3.271

173. Davila J, McNamara LA, Yang Z. Comparison of the predicted population coverage of tuberculosis vaccine candidates Ag85B-ESAT-6, Ag85B-TB10.4, and Mtb72f via a bioinformatics approach. PLoS One. 2012;7(7): e40882. doi:10.1371/journal.pone.0040882

174. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9): 1236–1240. doi:10.1093/bioinformatics/btu031

175. Billerbeck E, Horwitz JA, Labitt RN, Donovan BM, Vega K, Budell WC, et al. Characterization of human antiviral adaptive immune responses during hepatotropic virus infection in HLA-transgenic human immune system mice. J Immunol. 2013;191(4): 1753–1764. doi:10.4049/jimmunol.1201518

176. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. J Comput Biol. 2000;7(1-2): 203–214. doi:10.1089/10665270050081478

177. Karboul A, Gey van Pittius NC, Namouchi A, Vincent V, Sola C, Rastogi N, et al. Insights into the evolutionary history of tubercle bacilli as disclosed by genetic rearrangements within a PE_PGRS duplicated gene pair. BMC Evol Biol. 2006;6: 107. doi:10.1186/1471-2148-6-107

178. McEvoy CRE, Cloete R, Müller B, Schürch AC, van Helden PD, Gagneux S, et al. Comparative analysis of *Mycobacterium tuberculosis pe* and *ppe* genes reveals high sequence variation and an apparent absence of selective constraints. PLoS One. 2012;7(4): e30593. doi:10.1371/journal.pone.0030593

179. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5): 1792–1797. doi:10.1093/nar/gkh340

180. Kumar S, Stecher G, Peterson D, Tamura K. MEGA-CC: Computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. Bioinformatics. 2012;28(20): 2685–2686. doi:10.1093/bioinformatics/bts507

181. Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evol. 2015;1(1): ve–vev003. doi:10.1093/ve/vev003

182. Delport W, Poon AFY, Frost SDW, Kosakovsky Pond SL. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics. 2010;26(19): 2455–2457. doi:10.1093/bioinformatics/btq429

183. McKinney DM, Southwood S, Hinz D, Oseroff C, Arlehamn CSL, Schulten V, et al. A strategy to determine HLA class II restriction broadly covering the DR, DP, and DQ allelic variants most commonly expressed in the general population. Immunogenetics. 2013;65(5): 357–370. doi:10.1007/s00251-013-0684-y

184. Copin R, Coscollá M, Seiffert S. Sequence diversity in the *pe_pgrs* genes of *Mycobacterium tuberculosis* is independent of human T cell recognition. MBio. 2014;5(1): e00960–13. doi:10.1128/mBio.00960-13

185. Zemmour J, Parham P. Distinctive polymorphism at the HLA-C locus: implications for the expression of HLA-C. J Exp Med. 1992;176(4): 937–950. doi:10.1084/jem.176.4.937

186. Blythe MJ, Zhang Q, Vaughan K, de Castro R, Salimi N, Bui H-H, et al. An analysis of the epitope knowledge related to Mycobacteria. Immunome Res. 2007;3: 10. doi:10.1186/1745-7580-3-10

187. Stucki D, Malla B, Hostettler S, Huna T, Feldmann J, Yeboah-Manu D, et al. Two new rapid SNP-typing methods for classifying *Mycobacterium tuberculosis* complex into the main phylogenetic lineages. PLoS One. 2012;7(7): e41253. doi:10.1371/journal.pone.0041253

SUPPLEMENTARY DATA

Single-nucleotide substitutions found in iVEGI

Supplementary Table 1. Single-nucleotide substitutions found in iVEGI across 270 MTBC strains from all known lineages.

Gene/intergenic region	Common name	Genome position	Nucleotide substitution	Type	Amino acid substitution	SNP frequency (%)							dbSNP[a]	TBDB[b]	tbvar[c]
						Lineage 1	Lineage 2	Lineage 3	Lineage 4	Lineage 5	Lineage 6	Lineage 7			
Rv0960	-	1073704	G -> A	Nonsynonymous	A54T	0.0	2.0	0.0	0.0	0.0	0.0	0.0			
Rv0960	-	1073716	C -> T	Nonsynonymous	R58W	31.3	0.0	0.0	0.0	0.0	0.0	0.0	rs157705827	x	x
Intergenic_*Rv0960-Rv0961*	-	1074039	C -> T	-	-	0.0	2.0	0.0	0.0	0.0	0.0	0.0			
Rv0961	-	1074152	G -> T	Nonsynonymous	A27S	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0962c	*lprP*	1074558	A -> G	Nonsynonymous	L186P	0.0	0.0	0.0	3.2	0.0	0.0	0.0	rs157727929	x	x
Intergenic_*Rv0962c-Rv0963c*	-	1075269	C -> T	-	-	0.0	0.0	0.0	4.6	0.0	0.0	0.0	rs1577058883; rs157727947	x	x
Intergenic_*Rv0962c-Rv0963c*	-	1075279	C -> T	-	-	0.0	0.0	0.0	18.3	0.0	0.0	0.0	rs157705884	x	x
Rv0963c	-	1075584	C -> T	Nonsynonymous	G172R	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0963c	-	1075618	G -> A	Synonymous	S160S	22.9	0.0	0.0	0.0	0.0	0.0	0.0	rs157705894	x	x
Rv0963c	-	1075766	G -> A	Nonsynonymous	P111L	0.0	2.0	0.0	0.0	0.0	0.0	0.0			
Rv0963c	-	1075904	T -> C	Nonsynonymous	D65G	0.0	8.2	0.0	0.0	0.0	0.0	0.0			x
Rv0963c	-	1075986	C -> G	Nonsynonymous	A38P	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0963c	-	1075997	C -> G	Nonsynonymous	R34P	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0963c	-	1076020	C -> A	Nonsynonymous	K26N	0.0	0.0	0.0	1.1	0.0	0.0	0.0	rs157705937	x	x
Rv0963c	-	1076052	T -> G	Synonymous	R16R	0.0	0.0	2.5	0.0	0.0	0.0	0.0			

Rv0964c	-	1076196	T -> C	Stoplost	X161W	0.0	0.0	0.0	0.0	0.0	5.9	0.0			
Rv0964c	-	1076309	T -> G	Nonsynonymous	T124P	0.0	0.0	0.0	6.3	0.0	0.0	0.0	rs157705967	x	x
Rv0964c	-	1076532	G -> A	Synonymous	I49I	0.0	0.0	2.5	0.0	0.0	0.0	0.0			
Intergenic_*Rv0964c-Rv0965c*	-	1076689	A -> C	-	-	0.0	75.0	0.0	0.0	0.0	0.0	0.0	rs157705991	x	x
Rv0965c	-	1076880	C -> T	Synonymous	A106A	0.0	89.8	0.0	0.0	0.0	0.0	0.0	rs157706002	x	x
Rv0965c	-	1077102	C -> T	Synonymous	A32A	0.0	0.0	0.0	0.0	100.0	0.0	0.0	rs157706012; rs157727994	x	x
Rv0965c	-	1077106	C -> A	Nonsynonymous	R31L	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Intergenic_*Rv0965c-Rv0966c*	-	1077211	C -> T	-		0.0	0.0	0.0	0.0	0.0	13.3	0.0			
Rv0966c	-	1077280	A -> C	Nonsynonymous	W186G	0.0	0.0	0.0	0.0	100.0	100.0	0.0	rs157706020; rs157728002	x	x
Rv0966c	-	1077312	G -> A	Nonsynonymous	A175V	0.0	0.0	0.0	1.1	0.0	0.0	0.0	rs157706021	x	x
Rv0966c	-	1077365	G -> A	Synonymous	H157H	6.3	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0966c	-	1077521	C -> T	Synonymous	P105P	0.0	0.0	2.5	0.0	0.0	0.0	0.0			
Rv0966c	-	1077713	C -> T	Synonymous	Q41Q	0.0	0.0	0.0	0.0	0.0	0.0	100.0			x
Rv0966c	-	1077754	C -> T	Nonsynonymous	A28T	100.0	0.0	0.0	0.0	0.0	0.0	0.0	rs157706026	x	x
Intergenic_*Rv0966c-Rv0967c*	-	1077911	C -> T	-	-	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Intergenic_*Rv0966c-Rv0967c*	-	1077917	G -> A	-	-	0.0	0.0	0.0	0.0	0.0	41.2	0.0	rs157706031		
Intergenic_*Rv0966c-Rv0967c*	-	1077921	C -> T	-	-	0.0	0.0	0.0	1.1	0.0	5.9	0.0			
Rv0967	-	1077981	A -> G	Nonsynonymous	K3E	2.1	0.0	0.0	0.0	0.0	0.0	0.0	rs157706035	x	x
Rv0967	-	1078166	G -> A	Synonymous	T64T	0.0	2.0	0.0	0.0	0.0	0.0	0.0	rs157706036	x	x
Rv0968	-	1078398	G -> A	Stogain	W3X	17.5	0.0	0.0	0.0	0.0	0.0	0.0	rs157706037	x	x
Rv0969	*ctpV*	1078777	C -> A	Nonsynonymous	A12E	0.0	0.0	0.0	3.2	0.0	0.0	0.0			
Rv0969	*ctpV*	1079274	G -> A	Nonsynonymous	V178M	0.0	0.0	0.0	2.2	0.0	0.0	0.0			
Rv0969	*ctpV*	1079457	C -> T	Nonsynonymous	R239C	21.3	0.0	0.0	0.0	0.0	0.0	0.0	rs15770648	x	x
Rv0969	*ctpV*	1079473	G -> A	Nonsynonymous	R244K	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0969	*ctpV*	1079558	G -> C	Nonsynonymous	E272D	0.0	2.0	0.0	0.0	0.0	0.0	0.0			
Rv0969	*ctpV*	1079830	G -> T	Nonsynonymous	G363V	0.0	0.0	0.0	0.0	5.9	0.0	0.0			
Rv0969	*ctpV*	1079927	A -> C	Synonymous	T395T	2.1	0.0	0.0	7.5	0.0	0.0	0.0	rs157706050	x	x

Rv0969	*ctpV*	1080013	C -> T	Nonsynonymous	T424I	0.0	8.2	0.0	0.0	0.0	0.0	0.0			x
Rv0969	*ctpV*	1080191	T -> C	Synonymous	P483P	0.0	0.0	0.0	1.1	0.0	0.0	0.0	rs157731275		x
Rv0969	*ctpV*	1080192	A -> G	Nonsynonymous	N484D	2.1	0.0	2.7	100.0	0.0	0.0	0.0	rs157706051	x	x
Rv0969	*ctpV*	1080290	A -> G	Nonsynonymous	I516M	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0969	*ctpV*	1080366	G -> A	Nonsynonymous	G542R	0.0	0.0	0.0	0.0	0.0	11.8	0.0			
Rv0969	*ctpV*	1080376	A -> C	Nonsynonymous	K545T	0.0	0.0	2.7	0.0	0.0	0.0	0.0			
Rv0969	*ctpV*	1080493	T -> C	Nonsynonymous	V584A	0.0	2.0	0.0	0.0	0.0	0.0	0.0			
Rv0969	*ctpV*	1080528	G -> A	Nonsynonymous	A596T	0.0	0.0	0.0	9.7	0.0	0.0	0.0	rs157731276		x
Rv0969	*ctpV*	1080542	T -> C	Synonymous	G600G	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0969	*ctpV*	1080914	C -> T	Synonymous	Y724Y	0.0	0.0	0.0	0.0	0.0	11.8	0.0			
Rv0970	-	1081062	A -> G	Nonsynonymous	D4G	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0970	-	1081200	C -> T	Nonsynonymous	A50V	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0970	-	1081601	T -> G	Nonsynonymous	S184A	0.0	2.0	0.0	0.0	0.0	0.0	0.0	rs157706055; rs157728012	x	x
Rv0970	-	1081681	C -> T	Synonymous	V210V	0.0	0.0	0.0	3.2	0.0	0.0	0.0	rs157728013	x	x
Rv0971c	*echA7*	1081831	G -> A	Synonymous	L252L	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0971c	*echA7*	1081882	A -> G	Nonsynonymous	S235P	0.0	0.0	0.0	0.0	0.0	0.0	100.0			x
Rv0971c	*echA7*	1082433	G -> A	Nonsynonymous	A51V	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0971c	*echA7*	1082445	T -> A	Nonsynonymous	E47V	0.0	0.0	0.0	9.7	0.0	0.0	0.0			x
Rv0971c	*echA7*	1082578	T -> G	Nonsynonymous	S3R	0.0	0.0	2.6	0.0	0.0	0.0	0.0			
Rv0972c	*fadE12*	1082891	A -> C	Nonsynonymous	L287W	0.0	2.0	0.0	0.0	0.0	0.0	0.0			
Rv0972c	*fadE12*	1082983	G -> A	Synonymous	G256G	0.0	0.0	2.5	0.0	0.0	0.0	0.0			
Rv0972c	*fadE12*	1083179	G -> A	Nonsynonymous	T191I	0.0	0.0	2.5	0.0	0.0	0.0	0.0			
Rv0972c	*fadE12*	1083576	G -> T	Nonsynonymous	P59T	0.0	0.0	0.0	5.3	0.0	0.0	0.0			
Rv0972c	*fadE12*	1083684	C -> T	Nonsynonymous	V23M	0.0	0.0	0.0	0.0	0.0	5.9	0.0			
Rv0973c	*accA2*	1083755	A -> G	Nonsynonymous	S666P	100.0	0.0	0.0	0.0	0.0	0.0	0.0	rs157706062	x	x
Rv0973c	*accA2*	1083944	C -> T	Nonsynonymous	E603K	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0973c	*accA2*	1084048	G -> C	Nonsynonymous	P568R	0.0	0.0	5.0	0.0	0.0	0.0	0.0			
Rv0973c	*accA2*	1084091	C -> A	Nonsynonymous	A554S	0.0	4.2	0.0	0.0	0.0	0.0	0.0			x

Rv0973c	*accA2*	1084239	A -> G	Synonymous	R504R	0.0	0.0	2.5	0.0	0.0	0.0	0.0			
Rv0973c	*accA2*	1084280	A -> G	Nonsynonymous	W491R	0.0	0.0	5.0	0.0	0.0	0.0	0.0			
Rv0973c	*accA2*	1084409	C -> G	Nonsynonymous	G448R	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0973c	*accA2*	1084476	C -> T	Synonymous	L425L	0.0	0.0	0.0	1.1	0.0	0.0	0.0	rs157706065	x	x
Rv0973c	*accA2*	1084567	T -> C	Nonsynonymous	Y395C	0.0	0.0	5.0	0.0	0.0	0.0	0.0			
Rv0973c	*accA2*	1084812	G -> A	Synonymous	C313C	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0973c	*accA2*	1084831	T -> C	Nonsynonymous	Q307R	0.0	2.1	0.0	0.0	0.0	0.0	0.0	rs157706066	x	x
Rv0973c	*accA2*	1084863	T -> C	Synonymous	E296E	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0973c	*accA2*	1084911	G -> A	Synonymous	Y280Y	0.0	0.0	15.0	0.0	0.0	0.0	0.0	rs157706067	x	x
Rv0973c	*accA2*	1084967	C -> T	Nonsynonymous	A262T	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0973c	*accA2*	1085489	C -> A	Stogain	E88X	0.0	2.1	0.0	0.0	0.0	0.0	0.0			
Rv0973c	*accA2*	1085516	C -> T	Nonsynonymous	G79S	0.0	0.0	0.0	2.1	0.0	0.0	0.0			
Rv0974c	*accD2*	1085919	C -> T	Nonsynonymous	G476D	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0974c	*accD2*	1085983	C -> A	Nonsynonymous	A455S	0.0	0.0	0.0	0.0	0.0	11.8	0.0			
Rv0974c	*accD2*	1086076	G -> C	Nonsynonymous	P424A	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0974c	*accD2*	1086367	G -> A	Nonsynonymous	H327Y	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0974c	*accD2*	1086525	G -> T	Nonsynonymous	P274Q	0.0	0.0	0.0	4.3	0.0	0.0	0.0	rs157706073	x	x
Rv0974c	*accD2*	1086635	C -> A	Nonsynonymous	L237F	0.0	2.0	0.0	0.0	0.0	0.0	0.0			
Rv0974c	*accD2*	1086648	C -> A	Nonsynonymous	R233L	100.0	0.0	0.0	0.0	0.0	0.0	0.0	rs157706074	x	x
Rv0974c	*accD2*	1086687	G -> A	Nonsynonymous	S220F	0.0	0.0	2.5	0.0	0.0	0.0	0.0			
Rv0974c	*accD2*	1086778	C -> T	Nonsynonymous	G190S	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0974c	*accD2*	1086896	T -> A	Nonsynonymous	K150N	0.0	0.0	0.0	0.0	0.0	5.9	0.0			
Rv0974c	*accD2*	1086902	G -> T	Synonymous	T148T	0.0	0.0	0.0	0.0	0.0	11.8	0.0			
Rv0974c	*accD2*	1087193	C -> G	Nonsynonymous	K51N	2.1	0.0	2.5	18.1	0.0	0.0	0.0	rs157706078	x	x
Rv0974c	*accD2*	1087210	G -> T	Nonsynonymous	H46N	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0974c	*accD2*	1087279	T -> C	Nonsynonymous	K23E	0.0	0.0	0.0	11.7	0.0	0.0	0.0	rs157731280; rs157733291		x
Rv0975c	*fadE13*	1087478	C -> T	Nonsynonymous	G340D	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0975c	*fadE13*	1087534	C -> A	Nonsynonymous	K321N	2.1	0.0	0.0	0.0	0.0	0.0	0.0			

Rv0975c	*fadE13*	1087699	C -> A	Synonymous	R266R	0.0	0.0	0.0	2.1	0.0	0.0	0.0			x
Rv0975c	*fadE13*	1087760	A -> G	Nonsynonymous	L246P	0.0	0.0	0.0	0.0	11.8	0.0	0.0	rs157706079	x	x
Rv0975c	*fadE13*	1087812	T -> G	Nonsynonymous	N229H	0.0	0.0	0.0	2.1	0.0	0.0	0.0	rs157731281		x
Rv0975c	*fadE13*	1087824	C -> A	Nonsynonymous	V225L	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0975c	*fadE13*	1087852	G -> A	Synonymous	T215T	16.7	0.0	0.0	0.0	0.0	0.0	0.0			x
Rv0975c	*fadE13*	1088097	C -> T	Nonsynonymous	D134N	0.0	0.0	0.0	0.0	11.8	0.0	0.0	rs157706080	x	x
Rv0975c	*fadE13*	1088159	C -> T	Nonsynonymous	R113Q	0.0	2.0	0.0	0.0	0.0	0.0	0.0			
Rv0975c	*fadE13*	1088189	C -> T	Nonsynonymous	G103D	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0976c	-	1088514	C -> T	Synonymous	E554E	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0976c	-	1088600	T -> C	Nonsynonymous	I526V	0.0	2.0	2.5	0.0	0.0	0.0	0.0			
Rv0976c	-	1088676	C -> T	Synonymous	L500L	0.0	0.0	0.0	3.2	0.0	0.0	0.0			
Rv0976c	-	1088909	G -> A	Nonsynonymous	R423C	0.0	0.0	2.5	0.0	0.0	0.0	0.0			
Rv0976c	-	1089078	G -> A	Synonymous	A366A	0.0	0.0	5.0	0.0	0.0	0.0	0.0			
Rv0976c	-	1089421	C -> T	Nonsynonymous	R252Q	0.0	0.0	2.5	0.0	0.0	0.0	0.0			
Rv0976c	-	1089554	T -> C	Nonsynonymous	T208A	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0976c	-	1089922	G -> T	Nonsynonymous	P85H	0.0	0.0	0.0	1.1	0.0	0.0	0.0	rs157731282; rs157733292		x
Rv0976c	-	1090077	G -> A	Synonymous	Y33Y	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0976c	-	1090133	A -> G	Nonsynonymous	S15P	0.0	0.0	0.0	0.0	0.0	41.2	0.0			
Intergenic_*Rv0979c-Rv0979A*	-	1094885	C -> T	-	-	0.0	0.0	0.0	1.1	0.0	0.0	0.0	rs157731290		x
Rv0979A	*rpmF*	1095053	G -> T	Nonsynonymous	K56N	0.0	0.0	0.0	0.0	0.0	100.0	0.0	rs157706106	x	x
Rv0981	*mprA*	1096905	A -> C	Nonsynonymous	E28D	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0981	*mprA*	1097023	G -> A	Nonsynonymous	G68S	0.0	100.0	100.0	100.0	0.0	0.0	0.0	rs157706114	x	x
Rv0981	*mprA*	1097220	C -> T	Synonymous	S133S	0.0	100.0	100.0	0.0	0.0	0.0	0.0	rs157706115	x	x
Rv0981	*mprA*	1097328	G -> C	Synonymous	R169R	4.2	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0981	*mprA*	1097442	C -> T	Synonymous	D207D	0.0	89.8	0.0	0.0	0.0	0.0	0.0	rs157706116	x	x
Rv0982	*mprB*	1097633	C -> T	Synonymous	A42A	0.0	0.0	0.0	0.0	100.0	0.0	0.0	rs157706117	x	x
Rv0982	*mprB*	1097894	C -> T	Synonymous	T129T	0.0	0.0	2.5	0.0	0.0	0.0	0.0			
Rv0982	*mprB*	1098141	C -> T	Nonsynonymous	P212S	2.1	0.0	0.0	0.0	0.0	0.0	0.0			

Rv0982	*mprB*	1098263	T -> C	Synonymous	R252R	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0982	*mprB*	1098266	C -> T	Synonymous	T253T	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0982	*mprB*	1098459	G -> A	Nonsynonymous	V318I	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0982	*mprB*	1098523	A -> T	Nonsynonymous	H339L	0.0	0.0	0.0	100.0	0.0	0.0	0.0	rs1577016118	x	x
Rv0982	*mprB*	1098698	C -> G	Synonymous	G397G	0.0	0.0	0.0	5.3	0.0	0.0	0.0			x
Rv0982	*mprB*	1098710	C -> A	Synonymous	P401P	0.0	0.0	5.0	0.0	0.0	0.0	0.0			
Rv0982	*mprB*	1098711	G -> A	Nonsynonymous	V402M	0.0	0.0	0.0	0.0	0.0	5.9	0.0			
Rv0982	*mprB*	1098984	G -> A	Nonsynonymous	V493I	0.0	0.0	7.5	0.0	0.0	0.0	0.0	rs157706119	x	x
Intergenic_*Rv0982-Rv0983*	-	1099058	G -> A	-	-	0.0	100.0	100.0	0.0	0.0	0.0	0.0	rs157706120	x	x
Rv0983	*pepD*	1099092	C -> T	Synonymous	G9G	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0983	*pepD*	1099212	C -> T	Synonymous	F49F	18.8	0.0	0.0	0.0	0.0	0.0	0.0	rs157706122	x	x
Rv0983	*pepD*	1099237	C -> A	Nonsynonymous	P58T	0.0	0.0	0.0	0.0	0.0	11.8	0.0			
Rv0983	*pepD*	1099573	T -> C	Synonymous	L170L	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0983	*pepD*	1099678	C -> A	Nonsynonymous	P205T	0.0	2.0	0.0	0.0	0.0	0.0	0.0	rs157706125		
Rv0983	*pepD*	1099797	C -> A	Synonymous	G244G	0.0	8.2	0.0	0.0	0.0	0.0	0.0			x
Rv0983	*pepD*	1100234	C -> T	Nonsynonymous	P390L	0.0	0.0	0.0	3.2	0.0	0.0	0.0	rs157706128	x	x
Rv0983	*pepD*	1100414	G -> A	Nonsynonymous	G450D	0.0	0.0	27.5	0.0	0.0	0.0	0.0	rs157735240		x
Rv0984	*moaB2*	1100589	C -> T	Nonsynonymous	H44Y	4.2	0.0	0.0	0.0	0.0	0.0	0.0	rs157706130	x	x
Rv0985c	*mscL*	1101092	C -> T	Nonsynonymous	G130E	0.0	0.0	0.0	4.3	0.0	0.0	0.0	rs157706132	x	x
Rv0985c	*mscL*	1101174	C -> T	Nonsynonymous	V103I	0.0	0.0	2.5	0.0	0.0	0.0	0.0			
Rv0985c	*mscL*	1101179	C -> T	Nonsynonymous	G101E	0.0	0.0	2.5	0.0	0.0	0.0	0.0			
Rv0985c	*mscL*	1101317	C -> G	Nonsynonymous	G55A	0.0	4.2	0.0	0.0	0.0	0.0	0.0			
Rv0985c	*mscL*	1101353	A -> C	Nonsynonymous	I43S	0.0	0.0	0.0	2.1	0.0	0.0	0.0			x
Intergenic_*Rv0985c-Rv0986*	-	1101516	G -> T	-	-	2.1	0.0	0.0	0.0	0.0	0.0	0.0	rs157706135	x	x
Intergenic_*Rv0985c-Rv0986*	-	1101578	C -> G	-	-	0.0	0.0	2.5	0.0	0.0	0.0	0.0			
Rv0986	-	1101814	A -> C	Nonsynonymous	Q4H	0.0	2.0	0.0	0.0	0.0	0.0	0.0			
Rv0986	-	1102117	G -> A	Synonymous	V105V	0.0	89.8	0.0	0.0	0.0	0.0	0.0	rs157706139	x	x
Rv0986	-	1102175	G -> A	Nonsynonymous	V125M	0.0	0.0	0.0	0.0	0.0	11.8	0.0			

Rv0986	-	1102273	T -> G	Nonsynonymous	I157M	0.0	0.0	0.0	0.0	0.0	0.0	100.0			x
Rv0986	-	1102468	C -> A	Synonymous	G222G	0.0	100.0	0.0	0.0	0.0	0.0	0.0	rs157706141	x	x
Rv0986	-	1102484	G -> T	Nonsynonymous	V228L	0.0	0.0	0.0	0.0	100.0	100.0	0.0	rs157706142	x	x
Rv0986	-	1102520	A -> G	Nonsynonymous	T240A	0.0	0.0	2.5	1.1	0.0	0.0	0.0			
Rv0987	-	1102646	G -> A	Synonymous	A35A	0.0	89.8	0.0	0.0	0.0	0.0	0.0	rs157706143	x	x
Rv0987	-	1102788	G -> T	Nonsynonymous	V83F	0.0	0.0	0.0	7.4	0.0	0.0	0.0	rs157731291		x
Rv0987	-	1102805	C -> T	Synonymous	H88H	0.0	0.0	0.0	0.0	0.0	17.6	0.0			
Rv0987	-	1102924	T -> C	Nonsynonymous	L128S	4.2	0.0	0.0	0.0	0.0	0.0	0.0			x
Rv0987	-	1102973	C -> G	Synonymous	P144P	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0987	-	1103046	G -> C	Nonsynonymous	V169L	10.4	8.2	0.0	0.0	0.0	0.0	0.0			x
Rv0987	-	1103126	C -> A	Nonsynonymous	D195E	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0987	-	1103206	A -> C	Nonsynonymous	K222T	0.0	2.0	0.0	0.0	0.0	0.0	0.0			
Rv0987	-	1103249	C -> T	Synonymous	A236A	0.0	0.0	0.0	21.3	0.0	0.0	0.0	rs157706148	x	x
Rv0987	-	1103289	G -> C	Nonsynonymous	V250L	0.0	0.0	0.0	0.0	5.9	0.0	0.0			
Rv0987	-	1103487	G -> A	Nonsynonymous	G316R	12.5	0.0	0.0	0.0	0.0	0.0	0.0	rs157706151; rs157728037	x	x
Rv0987	-	1103656	C -> T	Nonsynonymous	A372V	0.0	73.5	0.0	0.0	0.0	0.0	0.0	rs157706152	x	x
Rv0987	-	1103786	G -> T	Nonsynonymous	L415F	0.0	0.0	0.0	12.8	0.0	0.0	0.0			
Rv0987	-	1103908	T -> C	Nonsynonymous	M456T	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0987	-	1103955	C -> T	Stogain	R472X	0.0	2.0	0.0	0.0	0.0	0.0	0.0	rs157706153	x	x
Rv0987	-	1103995	C -> T	Nonsynonymous	T485I	22.9	0.0	0.0	0.0	0.0	0.0	0.0	rs157706154	x	x
Rv0987	-	1104044	C -> T	Synonymous	G501G	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0987	-	1104140	A -> G	Synonymous	T533T	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0987	-	1104308	A -> G	Synonymous	R589R	0.0	0.0	0.0	2.1	0.0	0.0	0.0			
Rv0987	-	1104499	A -> G	Nonsynonymous	D653G	0.0	0.0	0.0	0.0	100.0	0.0	0.0	rs157706158	x	x
Rv0987	-	1104518	G -> A	Nonsynonymous	M659I	0.0	4.1	0.0	0.0	0.0	0.0	0.0			
Rv0987	-	1104626	C -> G	Synonymous	A695A	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0987	-	1104634	C -> T	Nonsynonymous	A698V	16.7	0.0	0.0	0.0	0.0	0.0	0.0	rs157706160	x	x
Rv0987	-	1104654	A -> G	Nonsynonymous	I705V	0.0	0.0	2.5	0.0	0.0	0.0	0.0			

Rv0987	-	1104690	G -> T	Nonsynonymous	V717F	0.0	0.0	0.0	100.0	0.0	0.0	0.0	rs157706161	x	x
Rv0987	-	1104740	C -> A	Synonymous	A733A	0.0	2.0	0.0	0.0	0.0	0.0	0.0			
Rv0987	-	1104928	C -> T	Nonsynonymous	A796V	16.7	0.0	0.0	0.0	0.0	0.0	0.0			x
Rv0987	-	1104944	C -> T	Synonymous	R801R	0.0	8.2	0.0	0.0	0.0	0.0	0.0			x
Rv0987	-	1104956	C -> T	Synonymous	V805V	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0987	-	1104967	G -> A	Nonsynonymous	G809D	0.0	2.0	0.0	0.0	0.0	0.0	0.0			
Rv0987	-	1104996	G -> C	Nonsynonymous	A819P	0.0	2.0	0.0	0.0	0.0	0.0	0.0			
Rv0987	-	1105056	C -> T	Nonsynonymous	P839S	0.0	2.0	0.0	0.0	0.0	0.0	0.0			
Rv0987	-	1105102	A -> C	Nonsynonymous	E854A	0.0	0.0	0.0	12.8	0.0	0.0	0.0			x
Intergenic_*Rv0987-Rv0988*	-	1105111	G -> T	-	-	0.0	0.0	0.0	1.1	0.0	0.0	0.0			x
Rv0988	-	1105216	A -> G	Nonsynonymous	D34G	0.0	0.0	0.0	0.0	5.9	0.0	0.0			
Rv0988	-	1105284	A -> G	Nonsynonymous	V57I	0.0	100.0	100.0	100.0	0.0	0.0	0.0	rs157706165	x	x
Rv0988	-	1105342	A -> G	Nonsynonymous	Y76C	0.0	0.0	2.5	0.0	0.0	0.0	0.0			
Rv0988	-	1105346	C -> A	Synonymous	T77T	0.0	8.3	0.0	0.0	0.0	0.0	0.0	rs157706166	x	x
Rv0988	-	1105367	C -> T	Synonymous	D84D	8.3	0.0	0.0	0.0	0.0	0.0	0.0			x
Rv0988	-	1105424	C -> T	Synonymous	G103G	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0988	-	1105505	G -> T	Synonymous	S130S	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0988	-	1105525	C -> T	Nonsynonymous	A137V	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0988	-	1105557	G -> T	Nonsynonymous	A148S	0.0	0.0	0.0	0.0	100.0	0.0	0.0	rs157706167	x	x
Rv0988	-	1105587	T -> C	Nonsynonymous	W158R	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0988	-	1105686	C -> G	Nonsynonymous	L191A	0.0	68.8	0.0	0.0	0.0	0.0	0.0	rs157706168	x	x
Rv0988	-	1105687	T -> C	Nonsynonymous	L191A	0.0	68.8	0.0	0.0	0.0	0.0	0.0	rs157706169	x	x
Rv0988	-	1105739	G -> C	Synonymous	P208P	0.0	0.0	0.0	0.0	0.0	0.0	100.0			x
Rv0988	-	1105832	C -> G	Nonsynonymous	S239R	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0988	-	1105877	T -> C	Synonymous	D254D	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0988	-	1106099	C -> T	Synonymous	I328I	0.0	0.0	100.0	0.0	0.0	0.0	0.0	rs157706173	x	x
Rv0988	-	1106219	G -> T	Nonsynonymous	M368I	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Intergenic_*Rv0988-Rv0989c*	-	1106289	G -> T	-	-	0.0	0.0	0.0	2.1	0.0	0.0	0.0			x

Intergenic_*Rv0988-Rv0989c*	-	1106311	G -> A	-	-	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0989c	*grcC2*	1106409	T -> C	Nonsynonymous	D325G	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0989c	*grcC2*	1106422	C -> T	Nonsynonymous	V321I	0.0	0.0	0.0	17.9	0.0	0.0	0.0	rs157706176	x	x
Rv0989c	*grcC2*	1106464	G -> A	Nonsynonymous	R307C	0.0	8.2	0.0	0.0	0.0	0.0	0.0			x
Rv0989c	*grcC2*	1106492	C -> T	Synonymous	A297A	0.0	0.0	5.0	0.0	0.0	0.0	0.0			
Rv0989c	*grcC2*	1106598	G -> C	Nonsynonymous	A262G	0.0	0.0	0.0	0.0	0.0	23.5	0.0	rs157706177; rs157728040	x	x
Rv0989c	*grcC2*	1106614	G -> T	Nonsynonymous	L257M	0.0	12.2	0.0	0.0	0.0	0.0	0.0	rs157706178; rs157728041	x	x
Rv0989c	*grcC2*	1106652	A -> G	Nonsynonymous	L244P	0.0	0.0	0.0	9.5	0.0	0.0	0.0			
Rv0989c	*grcC2*	1106838	G -> C	Nonsynonymous	A182G	0.0	10.2	0.0	0.0	0.0	0.0	0.0	rs157734969		x
Rv0989c	*grcC2*	1106842	T -> C	Nonsynonymous	I181V	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0989c	*grcC2*	1106875	G -> C	Nonsynonymous	L170V	0.0	0.0	5.0	0.0	0.0	0.0	0.0			
Rv0989c	*grcC2*	1106877	T -> C	Nonsynonymous	Y169C	0.0	0.0	0.0	7.4	0.0	0.0	0.0	rs157731295		x
Rv0989c	*grcC2*	1106959	C -> T	Nonsynonymous	A142T	0.0	0.0	0.0	2.1	0.0	0.0	0.0			
Rv0989c	*grcC2*	1106961	A -> G	Nonsynonymous	V141A	0.0	0.0	2.5	0.0	0.0	0.0	0.0	rs157706180		
Rv0989c	*grcC2*	1106962	C -> G	Nonsynonymous	V141L	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0989c	*grcC2*	1107024	T -> C	Nonsynonymous	D120G	0.0	0.0	0.0	0.0	100.0	0.0	0.0	rs157706181	x	x
Rv0989c	*grcC2*	1107077	C -> T	Synonymous	K102K	0.0	0.0	2.5	0.0	0.0	0.0	0.0			
Rv0989c	*grcC2*	1107134	C -> T	Nonsynonymous	M83I	4.2	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0989c	*grcC2*	1107320	G -> A	Synonymous	A21A	0.0	0.0	0.0	0.0	5.9	0.0	0.0			
Intergenic_*Rv0989c-Rv0990c*	-	1107434	A -> T	-	-	0.0	0.0	0.0	20.0	0.0	0.0	0.0	rs157706182	x	x
Rv0990c	-	1107897	G -> A	Nonsynonymous	A68V	0.0	0.0	0.0	0.0	0.0	100.0	0.0	rs157706183; rs157728042	x	x
Rv0990c	-	1107917	G -> T	Nonsynonymous	H61Q	0.0	2.0	0.0	21.1	0.0	0.0	0.0	rs157706184	x	x
Rv0990c	-	1107940	C -> A	Nonsynonymous	A54S	2.1	2.0	0.0	100.0	0.0	0.0	0.0	rs157706185; rs157728043	x	x
Rv0990c	-	1107986	C -> T	Nonsynonymous	M38I	4.2	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0990c	-	1108072	G -> A	Synonymous	L10L	0.0	0.0	0.0	0.0	5.9	0.0	0.0			
Rv0990c	-	1108077	G -> A	Nonsynonymous	P8L	0.0	0.0	0.0	0.0	0.0	5.9	0.0			
Intergenic_*Rv0990c-Rv0991c*	-	1108113	G -> T	-	-	2.1	0.0	0.0	0.0	0.0	0.0	0.0			

Rv0991c	-	1108199	G -> A	Synonymous	S102S	0.0	0.0	5.0	0.0	0.0	0.0	0.0			
Rv0991c	-	1108325	G -> A	Synonymous	-	0.0	0.0	0.0	0.0	0.0	5.9	0.0			
Intergenic_*Rv0991c-Rv0992c*	-	1108521	G -> A	-	-	0.0	69.4	0.0	0.0	0.0	0.0	0.0	rs157706187	x	x
Rv0992c	-	1108599	G -> C	Nonsynonymous	I191M	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0992c	-	1108788	C -> T	Synonymous	A128A	0.0	2.0	0.0	0.0	0.0	0.0	0.0			
Rv0992c	-	1108939	C -> T	Nonsynonymous	R78H	2.1	0.0	0.0	0.0	0.0	0.0	0.0	rs157706188	x	x
Rv0992c	-	1109163	G -> C	Nonsynonymous	I3M	0.0	100.0	100.0	100.0	0.0	0.0	0.0	rs157706189	x	x
Rv0993	*galU*	1109278	C -> T	Nonsynonymous	R3C	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0993	*galU*	1109296	C -> T	Nonsynonymous	P9S	0.0	2.0	0.0	0.0	0.0	0.0	0.0			
Rv0993	*galU*	1109313	C -> G	Synonymous	V14V	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0993	*galU*	1109391	T -> G	Synonymous	T40T	0.0	0.0	0.0	0.0	58.8	0.0	0.0	rs157706190	x	x
Rv0993	*galU*	1109535	G -> A	Synonymous	K88K	0.0	0.0	0.0	2.1	0.0	0.0	0.0			x
Rv0993	*galU*	1109975	G -> A	Nonsynonymous	R235Q	0.0	0.0	0.0	6.3	0.0	0.0	0.0	rs157706191	x	x
Rv0993	*galU*	1110177	T -> G	Synonymous	G302G	0.0	0.0	0.0	0.0	11.8	0.0	0.0	rs157706192; rs157728045	x	x
Intergenic_*Rv0993-Rv0994*	-	1110225	G -> A	-	-	0.0	0.0	0.0	0.0	0.0	5.9	0.0			
Rv0994	*moeA1*	1110295	T -> G	Synonymous	A9A	0.0	0.0	2.6	0.0	0.0	0.0	0.0			
Rv0994	*moeA1*	1110574	G -> T	Nonsynonymous	R102S	0.0	0.0	0.0	0.0	0.0	5.9	0.0	rs157706193		
Rv0994	*moeA1*	1110721	T -> C	Synonymous	R151R	0.0	0.0	69.2	0.0	0.0	0.0	0.0	rs157735264		x
Rv0994	*moeA1*	1110750	T -> C	Nonsynonymous	V161A	6.3	0.0	0.0	0.0	0.0	0.0	0.0	rs157706194	x	x
Rv0994	*moeA1*	1110916	C -> A	Nonsynonymous	D216E	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0994	*moeA1*	1110956	G -> T	Nonsynonymous	G230C	0.0	100.0	100.0	100.0	0.0	0.0	0.0	rs157706195	x	x
Rv0994	*moeA1*	1111228	G -> A	Synonymous	L320L	14.6	0.0	0.0	0.0	0.0	0.0	0.0			x
Rv0994	*moeA1*	1111382	G -> A	Nonsynonymous	D372N	0.0	0.0	0.0	0.0	5.9	0.0	0.0			
Rv0994	*moeA1*	1111397	C -> T	Synonymous	L377L	0.0	0.0	0.0	0.0	0.0	5.9	0.0	rs157706196	x	x
Rv0994	*moeA1*	1111448	A -> G	Nonsynonymous	T394A	0.0	8.2	0.0	0.0	0.0	0.0	0.0			x
Rv0994	*moeA1*	1111462	G -> T	Synonymous	A398A	0.0	0.0	2.6	0.0	0.0	0.0	0.0			
Rv0994	*moeA1*	1111469	C -> T	Synonymous	L401L	14.6	0.0	0.0	0.0	0.0	0.0	0.0	rs157706197	x	x
Rv0994	*moeA1*	1111518	T -> C	Nonsynonymous	V417A	97.9	0.0	0.0	0.0	0.0	0.0	0.0	rs1577061698	x	x

Rv0994	*moeA1*	1111528	C -> T	Synonymous	A420A	4.2	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0995	*rimJ*	1111678	G -> A	Nonsynonymous	G23S	0.0	69.4	0.0	0.0	0.0	0.0	0.0	rs157706199; rs157733669	x	x
Rv0995	*rimJ*	1111784	C -> T	Nonsynonymous	P58L	0.0	0.0	0.0	2.1	0.0	0.0	0.0	rs157731297; rs157733670		
Rv0995	*rimJ*	1111826	G -> T	Nonsynonymous	R72L	0.0	0.0	0.0	10.5	0.0	0.0	0.0	rs157731298; rs157733671		x
Rv0995	*rimJ*	1111852	G -> T	Nonsynonymous	D81Y	0.0	0.0	100.0	0.0	0.0	0.0	0.0	rs157706200	x	x
Rv0995	*rimJ*	1111921	G -> A	Nonsynonymous	G104S	0.0	2.0	0.0	0.0	0.0	0.0	0.0	rs157706201	x	x
Rv0995	*rimJ*	1111925	A -> G	Nonsynonymous	Y105C	0.0	0.0	0.0	10.5	0.0	0.0	0.0	rs157731299; rs157733672		x
Rv0995	*rimJ*	1111993	T -> G	Nonsynonymous	C128G	6.3	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0995	*rimJ*	1112027	C -> T	Nonsynonymous	A139V	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Intergenic_*Rv0995-Rv0996*	-	1112226	C -> A	-	-	0.0	0.0	2.5	0.0	0.0	0.0	0.0			
Intergenic_*Rv0995-Rv0996*	-	1112327	C -> A	-	-	0.0	0.0	0.0	0.0	0.0	0.0	25.0			
Intergenic_*Rv0995-Rv0996*	-	1112333	T -> A	-	-	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0996	-	1112530	T -> C	Synonymous	A49A	0.0	0.0	0.0	1.1	0.0	0.0	0.0	rs157706202	x	x
Rv0996	-	1112630	G -> C	Nonsynonymous	E83Q	0.0	0.0	0.0	0.0	5.9	0.0	0.0			
Rv0996	-	1112752	G -> A	Synonymous	P123P	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0996	-	1112880	C -> T	Nonsynonymous	A166V	8.3	0.0	0.0	0.0	0.0	0.0	0.0			x
Rv0996	-	1113029	C -> A	Nonsynonymous	R216S	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0996	-	1113226	C -> G	Synonymous	T281T	0.0	0.0	0.0	1.1	0.0	0.0	0.0	rs157731301		
Rv0996	-	1113290	C -> G	Nonsynonymous	Q303E	0.0	100.0	100.0	100.0	0.0	0.0	0.0	rs157706205	x	x
Rv0996	-	1113292	G -> C	Nonsynonymous	Q303H	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0996	-	1113326	T -> C	Nonsynonymous	S315P	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0996	-	1113403	G -> A	Synonymous	A340A	12.5	0.0	0.0	0.0	0.0	0.0	0.0	rs157706207	x	x
Intergenic_*Rv0996-Rv0997*	-	1113733	C -> A	-	-	0.0	0.0	70.0	0.0	0.0	0.0	0.0	rs157706209	x	x
Intergenic_*Rv0996-Rv0997*	-	1113750	C -> T	-	-	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Intergenic_*Rv0996-Rv0997*	-	1113783	C -> T	-	-	4.2	0.0	0.0	0.0	0.0	0.0	0.0			
Intergenic_*Rv0996-Rv0997*	-	1113840	A -> G	-	-	0.0	0.0	0.0	2.1	0.0	0.0	0.0	rs157706210	x	x

Intergenic_*Rv0996-Rv0997*	-	1114129	T -> G	-	-	0.0	0.0	0.0	0.0	100.0	0.0	0.0	rs157706211	x	x
Intergenic_*Rv0996-Rv0997*	-	1114261	G -> C	-	-	0.0	0.0	2.5	0.0	0.0	0.0	0.0			
Intergenic_*Rv0996-Rv0997*	-	1114289	G -> A	-	-	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0997	-	1114361	C -> T	Synonymous	G23G	0.0	0.0	2.5	0.0	0.0	0.0	0.0			
Rv0997	-	1114494	G -> A	Nonsynonymous	G68S	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0997	-	1114503	C -> T	Stogain	Q71X	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0997	-	1114539	A -> C	Nonsynonymous	T83P	0.0	2.0	0.0	0.0	0.0	0.0	0.0			
Rv0998	-	1114860	C -> G	Nonsynonymous	P38R	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0998	-	1114981	C -> T	Synonymous	I78I	0.0	0.0	0.0	0.0	5.9	0.0	0.0			
Rv0998	-	1115075	G -> A	Nonsynonymous	G110S	0.0	0.0	0.0	0.0	0.0	29.4	0.0			
Rv0998	-	1115198	G -> T	Nonsynonymous	A151S	2.1	2.0	2.5	0.0	0.0	0.0	0.0			
Rv0998	-	1115509	C -> T	Synonymous	I254I	0.0	0.0	27.5	0.0	0.0	0.0	0.0	rs157735268		x
Rv0999	-	1115796	C -> A	Synonymous	A10A	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv0999	-	1115857	T -> G	Nonsynonymous	L31V	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv0999	-	1115880	A -> C	Synonymous	V38V	0.0	0.0	15.0	0.0	0.0	0.0	0.0	rs157706216		
Rv0999	-	1115884	G -> A	Nonsynonymous	A40T	0.0	0.0	2.5	0.0	0.0	0.0	0.0			
Rv1000c	-	1116592	T -> G	Nonsynonymous	K186T	0.0	0.0	0.0	2.1	0.0	0.0	0.0			x
Rv1000c	-	1116838	T -> A	Nonsynonymous	E104V	2.1	0.0	0.0	0.0	0.0	0.0	0.0			
Rv1001	*arcA*	1117308	C -> T	Synonymous	L42L	0.0	0.0	0.0	0.0	0.0	100.0	0.0	rs157706220	x	x
Rv1001	*arcA*	1117405	C -> T	Nonsynonymous	T74I	97.9	0.0	0.0	0.0	0.0	0.0	0.0	rs157706221	x	x
Rv1001	*arcA*	1117568	C -> T	Synonymous	N128N	0.0	0.0	2.5	0.0	0.0	0.0	0.0			
Rv1001	*arcA*	1117656	C -> T	Synonymous	L158L	0.0	0.0	0.0	0.0	5.9	0.0	0.0			
Rv1001	*arcA*	1117758	A -> G	Nonsynonymous	I192V	0.0	0.0	0.0	1.1	0.0	0.0	0.0			
Rv1001	*arcA*	1117773	C -> G	Nonsynonymous	P197A	0.0	0.0	0.0	3.2	0.0	0.0	0.0			x
Rv1001	*arcA*	1117990	T -> G	Nonsynonymous	L269R	0.0	0.0	2.5	0.0	0.0	0.0	0.0			
Rv1001	*arcA*	1118006	G -> C	Synonymous	T274T	0.0	0.0	0.0	3.2	0.0	0.0	0.0	rs157731303; rs157733297		x
Rv1001	*arcA*	1118020	A -> G	Nonsynonymous	D279G	0.0	0.0	0.0	1.1	0.0	0.0	0.0	rs157706223; rs157728046	x	x

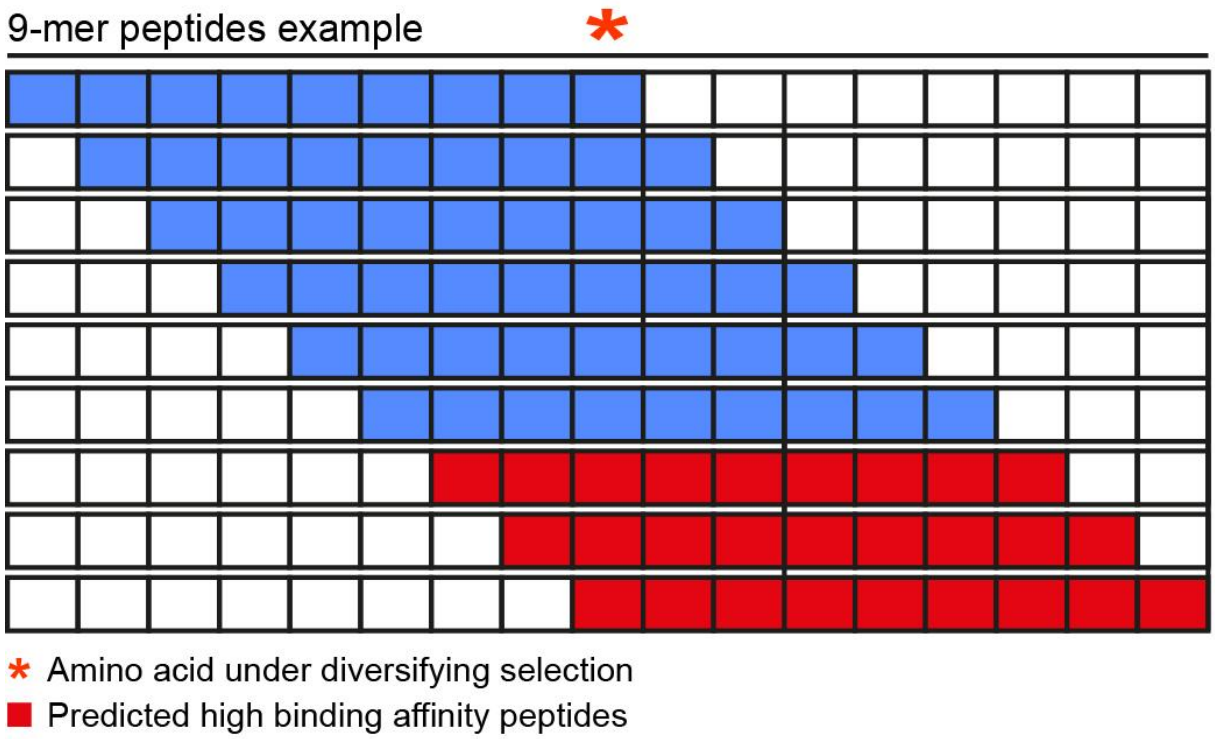
Rv1001	*arcA*	1118270	A -> G	Synonymous	V362V	0.0	0.0	0.0	0.0	0.0	100.0	0.0	rs157706228; rs157728050	x	x
Rv1001	*arcA*	1118293	A -> G	Nonsynonymous	D370G	0.0	0.0	0.0	1.1	0.0	0.0	0.0			

[a]Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29: 308–311. doi:10.1093/nar/29.1.308
[b]Reddy TBK, Riley R, Wymore F, Montgomery P, Decaprio D, Engels R, et al. TB database: An integrated platform for tuberculosis research. Nucleic Acids Res. 2009;37: 499–508. doi:10.1093/nar/gkn652
[c]Joshi KR, Dhiman H, Scaria V. tbvar: A comprehensive genome variation resource for *Mycobacterium tuberculosis*. Database (Oxford). England; 2014;2014: bat083. doi:10.1093/database/bat083

HLA alleles used in the immunoinformatics analysis

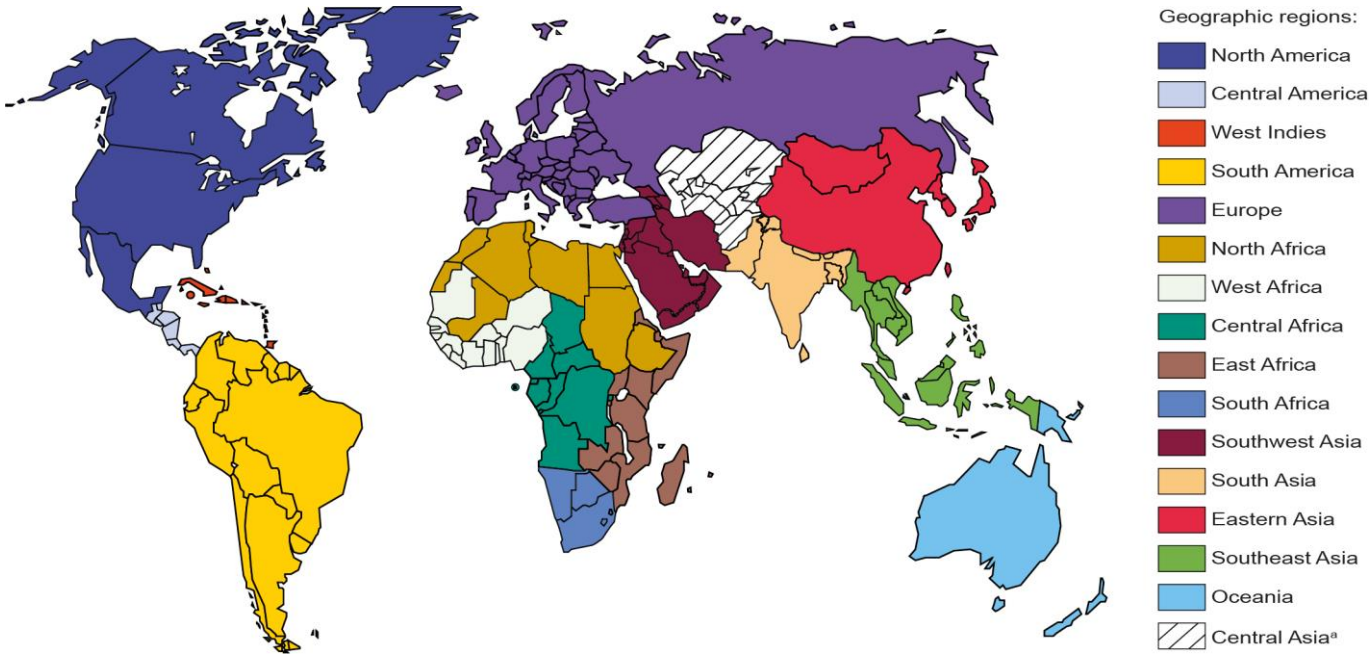
Supplementary Table 2. HLA class I and class II alleles used in the T cell epitope prediction.

HLA class I	HLA-A*01:01, HLA-A*02:01, HLA-A*02:02, HLA-A*02:03, HLA-A*02:06, HLA-A*02:07, HLA-A*02:22, HLA-A*03:01, HLA-A*11:01, HLA-A*23:01, HLA-A*24:02, HLA-A*24:07, HLA-A*24:20, HLA-A*26:01, HLA-A*29:02, HLA-A*30:01, HLA-A*31:01, HLA-A*31:08, HLA-A*32:01, HLA-A*33:01, HLA-A*33:03, HLA-A*34:01, HLA-A*68:01, HLA-A*74:01, HLA-B*07:02, HLA-B*07:05, HLA-B*08:01, HLA-B*13:01, HLA-B*13:02, HLA-B*14:01, HLA-B*14:02, HLA-B*14:05, HLA-B*15:01, HLA-B*15:02, HLA-B*15:03, HLA-B*15:04, HLA-B*15:07, HLA-B*15:10, HLA-B*15:11, HLA-B*15:13, HLA-B*15:21, HLA-B*15:32, HLA-B*18:01, HLA-B*18:02, HLA-B*27:04, HLA-B*27:05, HLA-B*35:01, HLA-B*35:03, HLA-B*35:05, HLA-B*35:06, HLA-B*39:01, HLA-B*39:06, HLA-B*40:01, HLA-B*40:02, HLA-B*40:06, HLA-B*42:01, HLA-B*44:03, HLA-B*44:04, HLA-B*44:06, HLA-B*45:01, HLA-B*46:01, HLA-B*48:01, HLA-B*48:03, HLA-B*51:01, HLA-B*52:01, HLA-B*53:01, HLA-B*53:03, HLA-B*55:02, HLA-B*56:01, HLA-B*57:01, HLA-B*58:01, HLA-B*58:02
HLA class II	DRB1*01:01, DRB1*01:02, DRB1*03:01, DRB1*03:02, DRB1*04:01, DRB1*04:02, DRB1*04:03, DRB1*04:04, DRB1*04:05, DRB1*04:07, DRB1*04:11, DRB1*07:01, DRB1*08:01, DRB1*08:02, DRB1*08:03, DRB1*08:04, DRB1*08:07, DRB1*09:01, DRB1*10:01, DRB1*11:01, DRB1*11:02, DRB1*11:03, DRB1*11:04, DRB1*12:01, DRB1*12:02, DRB1*13:01, DRB1*13:02, DRB1*13:03, DRB1*13:04, DRB1*14:01, DRB1*14:02, DRB1*14:03, DRB1*14:04, DRB1*14:05, DRB1*14:13, DRB1*15:01, DRB1*15:02, DRB1*15:03, DRB1*15:04, DRB1*16:01, DRB1*16:02



Supplementary Figure 1. Example of the calculation of the normalized percentage of predicted peptides with high binding affinity to HLA molecules.

Geographic regions used for the estimation of the regional population coverage and MTBC lineages distribution



Supplementary Figure 2. Geographic regions used for the estimation of the regional population coverage and MTBC lineages distribution. The world map was divided in 15 geographic regions (in different colors), according to the population coverage tool (http://tools.iedb.org/tools/population/iedb_input) of The Allele Frequency Net Database [133]. The distribution of the countries was further complemented on the basis of the United Nations guidelines (http://unstats.un.org/unsd/methods/m49/m49regin.htm). Northeast Asia and East Asia were grouped into Eastern Asia. [a]No datasets of HLA frequencies were available for Central Asia at the time of the analysis.

HLA alleles with predicted high binding affinity peptides encompassing wild-type and variant amino acid residues under diversifying selection

Supplementary Table 3. HLA alleles with predicted high binding affinity peptides encompassing wild-type and variant amino acid residues under diversifying selection.

Amino acid substitution	HLA class I				HLA class II			
	Wild-type		Variant		Wild-type		Variant	
	Number	Type	Number	Type	Number	Type	Number	Type
LprP L186P	29	HLA-A*24:02, HLA-A*24:07, HLA-A*24:20, HLA-A*31:01, HLA-A*31:08, HLA-A*33:03, HLA-A*68:01, HLA-B*07:02, HLA-B*07:05, HLA-B*15:03, HLA-B*15:04, HLA-B*15:10, HLA-B*15:13, HLA-B*35:01, HLA-B*35:03, HLA-B*35:05, HLA-B*35:06, HLA-B*39:01, HLA-B*39:03, HLA-B*42:01, HLA-B*44:06, HLA-B*46:01, HLA-B*51:01, HLA-B*53:01, HLA-B*53:03, HLA-B*55:02, HLA-B*56:01, HLA-B*57:01, HLA-B*58:01, HLA-B*58:02	17	HLA-B*07:02, HLA-B*07:05, HLA-B*35:03, HLA-B*35:06, HLA-A*31:01, HLA-A*31:08, HLA-A*33:03, HLA-A*68:01, HLA-B*15:03, HLA-B*15:10, HLA-B*15:13, HLA-B*44:06, HLA-B*46:01, HLA-B*53:01, HLA-B*57:01, HLA-B*58:01, HLA-B*58:02	3	DRB1*01:01, DRB1*01:02, DRB1*14:13	2	DRB1*01:01, DRB1*01:02
AccA2 S666P	0	-	0	-	0	-	0	-
AccD2 K51N	3	HLA-A*31:01, HLA-A*33:01, HLA-A*33:03	4	HLA-A*31:01, HLA-A*33:01, HLA-A*33:03, HLA-B*15:10	0	-	0	-

AccD2 R233L	29	HLA-A*26:01, HLA-A*30:01, HLA-A*30:02, HLA-A*68:01, HLA-B*07:02, HLA-B*07:05, HLA-B*08:01, HLA-B*13:01, HLA-B*14:02, HLA-B*14:05, HLA-B*15:02, HLA-B*15:03, HLA-B*15:04, HLA-B*15:10, HLA-B*15:11, HLA-B*15:13, HLA-B*15:21, HLA-B*18:01, HLA-B*39:01, HLA-B*39:03, HLA-B*39:06, HLA-B*40:01, HLA-B*40:02, HLA-B*40:06, HLA-B*44:03, HLA-B*44:04, HLA-B*45:01, HLA-B*48:01, HLA-B*48:03	36	HLA-A*01:01,HLA-A*02:01,HLA-A*02:03,HLA-A*02:06,HLA-A*02:07,HLA-A*02:22,HLA-A*29:02,HLA-A*30:02,HLA-B*13:01,HLA-B*15:04,HLA-B*15:10,HLA-B*18:01,HLA-B*35:01,HLA-B*35:03,HLA-B*35:05,HLA-B*35:06,HLA-B*39:01,HLA-B*39:03,HLA-B*39:06,HLA-B*44:03,HLA-B*44:04,HLA-B*45:01,HLA-B*53:03,HLA-A*26:01,HLA-B*14:02,HLA-B*14:05,HLA-B*15:02,HLA-B*15:03,HLA-B*15:11,HLA-B*15:13,HLA-B*15:21,HLA-B*40:01,HLA-B*40:02,HLA-B*40:06,HLA-B*48:01,HLA-B*48:03	2	DRB1*01:01, DRB1*01:02	5	DRB1*01:01, DRB1*01:02, DRB1*12:01, DRB1*12:02, DRB1*14:13
Rv0987 V169L	8	HLA-A*02:03,HLA-A*02:06,HLA-A*02:22,HLA-B*13:01,HLA-B*13:02,HLA-B*15:03,HLA-B*48:01,HLA-B*48:03	9	HLA-A*02:03,HLA-A*02:06,HLA-A*02:22,HLA-A*24:07,HLA-B*13:01,HLA-B*13:02,HLA-B*15:03,HLA-B*48:01,HLA-B*48:03	6	DRB1*01:01, DRB1*01:02, DRB1*07:01, DRB1*13:03, DRB1*13:04, DRB1*14:13	9	DRB1*01:01, DRB1*01:02, DRB1*07:01, DRB1*11:04, DRB1*12:01, DRB1*12:02, DRB1*13:03, DRB1*13:04, DRB1*14:13
Rv0988 L191A	10	HLA-A*02:01, HLA-A*02:03, HLA-A*02:06, HLA-A*02:22, HLA-B*13:02, HLA-B*48:01, HLA-B*48:03, HLA-B*51:01, HLA-B*52:01, HLA-B*56:01	5	HLA-A*02:06, HLA-B*13:02, HLA-B*48:01, HLA-B*48:03, HLA-B*52:01	9	DRB1*01:01, DRB1*01:02, DRB1*07:01, DRB1*11:04, DRB1*12:01, DRB1*12:02, DRB1*13:03, DRB1*13:04, DRB1*14:13	6	DRB1*01:01, DRB1*01:02, DRB1*07:01, DRB1*13:03, DRB1*13:04, DRB1*14:13
GrcC2 V141L	22	HLA-A*02:01, HLA-A*02:03, HLA-A*02:06, HLA-A*02:22, HLA-A*26:01, HLA-A*34:01, HLA-A*68:02, HLA-B*15:02, HLA-B*15:03, HLA-B*15:11, HLA-B*15:13, HLA-B*15:21, HLA-B*35:01, HLA-B*35:03, HLA-B*35:05, HLA-B*35:06,	25	HLA-A*02:01, HLA-A*02:03, HLA-A*02:06, HLA-A*02:07, HLA-A*02:22, HLA-A*34:01, HLA-B*35:06, HLA-B*39:01, HLA-B*39:03, HLA-A*26:01, HLA-A*68:02, HLA-B*15:02, HLA-B*15:03, HLA-B*15:11, HLA-B*15:13, HLA-B*15:21,	4	DRB1*01:01, DRB1*01:02, DRB1*11:04, DRB1*14:13	4	DRB1*01:01, DRB1*01:02, DRB1*11:04, DRB1*14:13

		HLA-B*44:06, HLA-B*46:01, HLA-B*51:01, HLA-B*52:01, HLA-B*53:01, HLA-B*53:03		HLA-B*35:01, HLA-B*35:03, HLA-B*35:05, HLA-B*44:06, HLA-B*46:01, HLA-B*51:01, HLA-B*52:01, HLA-B*53:01, HLA-B*53:03				
Rv0990c A54S	7	HLA-A*02:03, HLA-B*39:01, HLA-B*39:06, HLA-B*40:01, HLA-B*40:02, HLA-B*44:03, HLA-B*44:04	5	HLA-A*02:03, HLA-B*39:01, HLA-B*40:01, HLA-B*44:03, HLA-B*44:04	4	DRB1*01:01, DRB1*01:02, DRB1*11:04, DRB1*14:13	4	DRB1*01:01, DRB1*01:02, DRB1*11:04, DRB1*14:13
Rv0990c A68V	4	HLA-A*11:01, HLA-B*07:02, HLA-B*07:05, HLA-B*15:10	5	HLA-A*11:01, HLA-A*68:02, HLA-B*07:02, HLA-B*07:05, HLA-B*15:10	0	-	1	DRB1*14:13
RimJ D81Y	26	HLA-A*23:01, HLA-A*26:01, HLA-A*31:08, HLA-A*32:01, HLA-A*34:01, HLA-B*13:01, HLA-B*13:02, HLA-B*15:01, HLA-B*15:03, HLA-B*15:07, HLA-B*15:11, HLA-B*35:01, HLA-B*35:05, HLA-B*40:01, HLA-B*40:02, HLA-B*40:06, HLA-B*44:03, HLA-B*44:04, HLA-B*44:06, HLA-B*45:01, HLA-B*48:01, HLA-B*48:03, HLA-B*51:01, HLA-B*52:01, HLA-B*53:01, HLA-B*53:03	36	HLA-A*23:01, HLA-A*24:02, HLA-A*24:07, HLA-A*24:20, HLA-A*26:01, HLA-A*29:02, HLA-A*30:02, HLA-A*31:08, HLA-A*32:01, HLA-A*34:01, HLA-B*15:03, HLA-B*15:04, HLA-B*15:13, HLA-B*27:04, HLA-B*27:05, HLA-B*35:01, HLA-B*35:05, HLA-B*44:06, HLA-B*46:01, HLA-B*52:01, HLA-B*53:01, HLA-B*53:03, HLA-A*23:01, HLA-B*13:01, HLA-B*13:02, HLA-B*15:01, HLA-B*15:07, HLA-B*15:11, HLA-B*40:01, HLA-B*40:02, HLA-B*40:06, HLA-B*44:03, HLA-B*44:04, HLA-B*48:01, HLA-B*48:03, HLA-B*51:01	1	DRB1*01:02	1	DRB1*01:02
Rv0996 Q303E	6	HLA-A*30:01, HLA-A*31:01, HLA-A*74:01, HLA-B*15:03, HLA-B*27:04, HLA-B*27:05	3	HLA-A*31:01, HLA-B*27:04, HLA-B*27:05	29	DRB1*01:01, DRB1*01:02, DRB1*03:01, DRB1*03:02, DRB1*07:01, DRB1*08:01, DRB1*08:02, DRB1*08:03, DRB1*08:04, DRB1*08:07, DRB1*11:01, DRB1*11:02, DRB1*11:03, DRB1*11:04,	29	DRB1*01:01, DRB1*01:02, DRB1*03:01, DRB1*03:02, DRB1*07:01, DRB1*08:01, DRB1*08:02, DRB1*08:03, DRB1*08:04, DRB1*08:07, DRB1*11:01, DRB1*11:02, DRB1*11:03, DRB1*11:04,

						DRB1*12:01, DRB1*12:02, DRB1*13:01, DRB1*13:03, DRB1*13:04, DRB1*14:01, DRB1*14:02, DRB1*14:03, DRB1*14:04, DRB1*14:05, DRB1*14:13, DRB1*15:01, DRB1*15:03, DRB1*15:04, DRB1*16:02		DRB1*12:01, DRB1*12:02, DRB1*13:01, DRB1*13:03, DRB1*13:04, DRB1*14:01, DRB1*14:02, DRB1*14:03, DRB1*14:04, DRB1*14:05, DRB1*14:13, DRB1*15:01, DRB1*15:03, DRB1*15:04, DRB1*16:02

Regional population coverage for peptides encompassing residues under diversifying selection

Supplementary Table 4. Predicted population coverage in 15 geographic regions for peptides encompassing wild-type and variant amino acid residues under diversifying selection.

	Amino acid substitution under diversifying selection											
	AccD2 R233L				Rv0987 V169L				Rv0988 L191A			
	Variant MTBC strains		HLA class II		Variant MTBC strains		HLA class II		Variant MTBC strains		HLA class II	
Region	Lineage	Frequency (%)	Wild-type population coverage (%)	Variant population coverage (%)	Lineage	Frequency (%)	Wild-type population coverage (%)	Variant population coverage (%)	Lineage	Frequency (%)	Wild-type population coverage (%)	Variant population coverage (%)
North America	1	8.40	17.32	22.93	1 and 2	26.89	36.67	45.20	2	18.49	45.20	36.67
Central America	1	0.00	3.44	4.19	1 and 2	0.00	15.88	19.01	2	0.00	19.01	15.88
West Indies	1	0.00	14.95	22.26	1 and 2	3.17	38.43	44.62	2	3.17	44.62	38.43
South America	1	0.00	6.81	11.13	1 and 2	0.00	19.33	23.38	2	0.00	23.38	19.33
Europe	1	6.43	17.14	21.24	1 and 2	15.66	41.30	50.00	2	9.29	50.00	41.30
North Africa	1	2.43	11.00	12.92	1 and 2	7.90	40.32	45.05	2	5.47	45.05	40.32
West Africa	1	1.85	10.73	13.06	1 and 2	4.55	43.43	47.09	2	2.70	47.09	43.43
Central Africa	1	0.00	11.64	19.17	1 and 2	0.00	32.30	39.79	2	0.00	39.79	32.30
East Africa	1	9.93	15.54	26.04	1 and 2	14.37	25.35	36.32	2	4.44	36.32	25.35
South Africa	1	0.00	0.00	0.00	1 and 2	22.53	0.00	0.00	2	22.53	0.00	0.00
Southwest Asia	1	7.75	6.64	7.85	1 and 2	18.98	20.64	30.45	2	11.23	30.45	20.64
South Asia	1	31.74	9.25	15.41	1 and 2	44.94	40.14	46.14	2	13.20	46.14	40.14
Eastern Asia	1	6.32	6.71	24.60	1 and 2	71.29	15.07	33.15	2	64.97	33.15	15.07
Southeast Asia	1	42.76	0.78	30.62	1 and 2	88.31	10.12	38.84	2	45.55	38.84	10.12
Oceania	1	22.30	0.81	15.68	1 and 2	39.72	2.67	18.36	2	17.42	18.36	2.67

In vitro HLA-binding assays

Supplementary Table 5. Purity of synthesized peptides for the HLA-binding *in vitro* assays.

Peptide	Type	Peptide sequence	Purity (%)
ESAT-6$_{3-17}$	Positive control	EQQWNFAGIEAAASA	84.81
AccD2$_{228-241}$	Wild-type	AEMHARISGLADYF	83.35
AccD2$_{228-241}$	Variant	AEMHALISGLADYF	90.11
Rv0987$_{164-177}$	Wild-type	RIALQVKGAPTTVT	81.27
Rv0987$_{164-177}$	Variant	RIALQLKGAPTTVT	93.62
Rv0988$_{185-202}$	Wild-type	LTLTQTAPPILQGNAGLS	82.03
Rv0988$_{185-202}$	Variant	LTLTQTLPPILQGNAGLS	84.11