Supplemental Information contents

Supplemental experimental procedures

To represent the diverse human populations worldwide or in The Gambia (depending on the analysis, as specified in the text and figure and table legends), epitope prediction analyses were performed using the most prevalent HLA-A, -B, -DR, -DP and -DQ alleles in the given population, as defined by the IEDB and the Allele Frequency Database (http://www.allelefrequencies.net/). The results of population coverage are shown in Table S4.

Epitope diversity analysis

M. tuberculosis T cell epitope encoding sequences were retrieved from the IEDB (http://www.iedb.org/) on the 24th of April, 2015. We selected 1,730 epitopes using the following criteria: linear peptides, *M. tuberculosis* complex (ID:77643, Mycobacterium complex), positive assays only, T cell assays, any MHC restriction, host: humans, any diseases, any reference type. Because of the technical limitations of Illumina sequencing in repetitive regions, we excluded 277 epitopes located in PE/PPE genes, phages-related genes and transposases. We excluded 227 additional epitopes due to the incapacity to assign accurate genomic coordinates in the reference strain H37Rv. This was the result of either a duplication of the epitope in different proteins or the absence of the epitope-encoding sequence in the H37Rv genome as determined by blastP (Altschul et al., 1990). Hence, we surveyed the genetic diversity of 1,226 epitopes encoded by 304 antigens. The genetic diversity was assessed by estimating dN/dS of each genome with respect to the MTBC ancestor using PAML (Yang, 2007).

To assess whether the genetic variation found in epitopes differed from nonepitope regions of the same antigens, we used Wilcoxon signed rank test with continuity correction implemented in R to compare dN/dS observed within concatenated epitopes to the same ratio in 1) concatenated nonepitope regions of the 1,226 epitopes-encoding proteins, 2) essential genes and 3) nonessential genes. Essential and nonessential genes were classified according to Zhang, et al (Zhang et al., 2012).

Epitope predictions

Complete amino acid sequences of the proteins of interest in the *M. tuberculosis* reference strain H37Rv were retrieved from GenBank (http://www.ncbi.nlm.nih.gov/Genbank/). The IEDB "recommended method" (IEDB, www.iedb.org) was used for predicting CD8 and CD4 T cell epitopes in *M. tuberculosis* protein sequences (Lundegaard et al., 2008; Wang et al., 2008). The IEDB "recommended method" uses the best possible method for a given HLA molecule among a selection of prediction methods including: ANN, SMM, CombLib, NetMHCpan, NN-align (netMHCII-2.2), SMM-align (netMHCII-1.1), Sturniolo, and NetMHCIIpan. The output of these

predictions is a table, showing a percentile rank for each selected method which is generated by comparing the peptide's score against the scores of five million random 15 mer peptides selected from SWISSPROT database. A low percentile rank indicates the highest affinity. Threshold cut-off values corresponding to a maximal percentile rank of 10 was used in this work. CD8 T cell peptide length was not included as a criterion in the analysis and multiple allele/length pairs were submitted at a time.

Statistical analysis

We set up a Bayesian model to compare the probability of responding to only one (ancestral or variant) peptide (that is, the mutation affects the response), to the probability that the response was the same for both peptides (the mutation does not affect the response). Peptides with one amino acid change present four different possibilities: response (SI>2) to the ancestral but not the variant peptide, response to the variant but not the ancestral peptides, to both, or to none. Two independent Bernoulli distributions with probability p_1 and p_2 were assumed to model the response Y_1, Y_2 to the different epitopes. Two Beta distributions centered around 0.5 were specified as a priori distributions for the response probabilities p_1 and p_2. Similarly, epitopes with two amino acid changes had eight different possibilities and three independent Bernoulli distributions where probability p_1, p_2 and p_3 were assumed to model the response Y_1, Y_2, Y_3 to them. The posterior distribution was evaluated and the hypothesis $P(Y_1=Y_2=1) > P(Y_1=1, Y_2=0)+ P(Y_2=1, Y_1=0)$ was tested for peptides with 2 mutations. Similarly, in case of 2 mutations, the hypothesis tested was $P(Y_1=Y_2=Y_3=1)+ P(Y_1=Y_2=1,Y_3=0)+ P(Y_1=Y_3=1,Y_2=0)+ P(Y_2=Y_3=1,Y_1=0) > P(Y_1=1, Y_2=0, Y_3=0)+ P(Y_1=0, Y_2=1, Y_3=0)+ P(Y_1=0, Y_2=0, Y_3=1)+ P(Y_1=0, Y_2=0, Y_3=0)$. The Bayesian p-values are reported in Table S8.

Supplemental References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403-410.

Arlehamn, C.S.L., Gerasimova, A., Mele, F., Henderson, R., Swann, J., Greenbaum, J.A., Kim, Y., Sidney, J., James, E.A., Taplitz, R.*, et al.* (2013). Memory T Cells in Latent Mycobacterium tuberculosis Infection Are Directed against Three Antigenic Islands and Largely Contained in a CXCR3(+)CCR6(+) Th1 Subset. Plos Pathog. *9*,e1003130

Comas, I., Coscolla, M., Luo, T., Borrell, S., Holt, K.E., Kato-Maeda, M., Parkhill, J., Malla, B., Berg, S., Thwaites, G.*, et al.* (2013). Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. Nat. Genet. *45*, 1176-U1311.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: An information aesthetic for comparative genomics. Genome Res. *19*, 1639-1645.

Lew, J.M., Kapopoulou, A., Jones, L.M., and Cole, S.T. (2011). TubercuList-10 years after. Tuberculosis *91*, 1-7.

Lundegaard, C., Lamberth, K., Harndahl, M., Buus, S., Lund, O., and Nielsen, M. (2008). NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. Nucleic Acids Res *36*, 509-512.

Vilella, A.J., Blanco-Garcia, A., Hutter, S., and Rozas, J. (2005). VariScan: analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. Bioinformatics, *21*,2791-3.

Vita, R., Zarebski, L., Greenbaum, J.A., Emami, H., Hoof, I., Salimi, N., Damle, R., Sette, A., and Peters, B. (2010). The Immune Epitope Database 2.0. Nucleic Acids Res. *38*, 854-862.

Wang, P., Sidney, J., Dow, C., Mothe, B., Sette, A., and Peters, B. (2008). A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. PloS Comput. Biol. *4*, e1000048.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. *24*, 1586-1591.

Zhang, Y.J., Ioerger, T.R., Huttenhower, C., Long, J.E., Sassetti, C.M., Sacchettini, J.C., and Rubin, E.J. (2012). Global assessment of genomic regions required for growth in Mycobacterium tuberculosis. Plos Pathog. *8*, e1002946.