Early progression to active tuberculosis is a highly heritable trait driven by 3q23 in Peruvians

Yang Luo 1,2,3,4,5, Sara Suliman 1, Samira Asgari 1,2,3,4,5, Tiffany Amariuta 1,2,3,4,5, Yuriy Baglaenko 1,2,3,4,5, Marta Martínez-Bonet 1, Kazuyoshi Ishigaki 1,2,3,4,5, Maria Gutierrez-Arcelus 1,2,3,4,5, Roger Calderon 7, Leonid Lecca 7, Segundo R. León 7, Judith Jimenez 7, Rosa Yataco 7, Carmen Contreras 7, Jerome T. Galea 8, Mercedes Becerra 9, Sergey Nejentsev 10,11, Peter A. Nigrovic 1,12, D. Branch Moody 1, Megan B. Murray 9 & Soumya Raychaudhuri 1,2,3,4,5,13

Of the 1.8 billion people worldwide infected with *Mycobacterium tuberculosis*, 5-15% will develop active tuberculosis (TB). Approximately half will progress to active TB within the first 18 months after infection, presumably because they fail to mount an effective initial immune response. Here, in a genome-wide genetic study of early TB progression, we genotype 4002 active TB cases and their household contacts in Peru. We quantify genetic heritability ($h^2_g$) of early TB progression to be 21.2% (standard error 0.08). This suggests TB progression has a strong genetic basis, and is comparable to traits with well-established genetic bases. We identify a novel association between early TB progression and variants located in a putative enhancer region on chromosome 3q23 (rs73226617, OR = 1.18; P = 3.93 × $10^{-8}$). With in silico and in vitro analyses we identify rs73226617 or rs148722713 as the likely functional variant and ATP1B3 as a potential causal target gene with monocyte specific function.
he infectious pathogen *Mycobacterium tuberculosis* (*M*. *tuberculosis*) infects about one-quarter of the world’s population. Approximately 5–15% of infected individuals progress to active TB while the vast majority remain infected with viable latent *M*. *tuberculosis* (*M*. *tb*) (Fig. 1a). In 2017, approximately 10 million new patients were diagnosed with active TB, and 1.6 million people died from TB-related diseases. Active TB can develop immediately (within the first 18 months) after recent *M*. *tb* infection or after many years of latency, presumably caused via distinct disease mechanisms. Late progression or TB reactivation is more likely the consequence of acquired immune compromise due to other diseases or ageing, whereas early progression is presumably due to failure in mounting the initial immune response that contains the bacterial spread. Previous studies have indicated a strong heritable component of population-wide TB susceptibility, that includes early disease progression, reactivation, and infection. But whether early progression has a different genetic architecture compared to population-wide susceptibility has yet to be defined.

Reported associations for TB and other infectious diseases have to be considered in the context of TB diagnostic criteria and selected control groups. To date genome-wide association studies (GWAS) of TB have compared mixed pools of TB patients with early progression or reactivation, to population controls, who may not have been exposed to *M*. *tb* at all. Hence, known human genetic loci associations with clinical outcomes might represent risk factors for *M*. *tb* infection, progression from recent *M*. *tb* exposure to active TB, or reactivation of TB after a period of latency. Infection, progression, and reactivation represent pathophysiology distinct disease transitions likely involving distinct mechanisms of transmission, early innate immune response, and control by adaptive immunity. Thus, the study of mixed TB populations using controls of unknown exposure status may underestimate or miss genetic associations for these separate stages of disease.

In this study, we perform a genome-wide association study of a large sample of early TB progression cases (2175 recently exposed cases and 1827 controls). We first establish early TB progression has a strong genetic basis that is comparable to other complex traits. We further identify a novel association with early TB progression, prioritize likely causal variants and functional genes, and propose new candidate mechanisms of host response in early TB progression.

**Results**

**Building an early progression to active tuberculosis cohort.** To identify host factors that drive pulmonary early TB progression, we conducted a large, longitudinal genetic study in Lima, Peru (Fig. 1b), where the TB incidence rate is one of the highest in the region. We enrolled patients with microbiologically confirmed pulmonary TB. Within 2 weeks of enrolling an index patient, we identified their household contacts (HHCs) and screened for infection as measured by a tuberculin skin test (TST) and for signs and symptoms of pulmonary and extra-pulmonary TB. HHCs were re-evaluated at 2, 6, and 12 months. We considered individuals to be early progressors if they are (1) index patients whose *M*. *tb* isolates shared a molecular fingerprint with isolates from other enrolled patients; (2) HHCs who developed TB disease within 1 year after exposure to an index patient and (3) index patients who were 40-years old or younger at time of diagnosis. We considered HHCs who were TST positive at baseline or any time during the 12 month follow up period, but who had no previous history of TB disease and remained disease free, as non-progressing controls (Methods, Fig. 1b). In total, we genotyped 2175 recently exposed pulmonary TB cases (early progressors) versus 1827 HHCs with latent tuberculosis infection, who had not progressed to active TB during 1 year of follow-up (non-progressors), as controls (Methods, Supplementary Table 1).

**Genomic analysis demonstrates the distinct genetic heritage of Peruvians.** Peru is a country with a complex demographic history and underexplored genomic variation. When Spanish conquistadors arrived in the region in the 16th century, Peru was the center of the vast Inca Empire and was inhabited by a large Native American population. During the colonial period, Europeans and Africans (brought in as slaves) arrived in large numbers to Peru. After Peru gained its independence in 1821, there was a flow of immigrants from southern China to all regions of Peru as a replacement for slaves. As a result, based on the analysis of our large genomic cohort, the genetic background of the current Peruvian population is shaped by different levels of admixture between Native Americans, Europeans, African and Asian immigrants that arrived in waves with specific and dated historical antecedents. When compared to individuals from other South American countries, Peruvians tend to share a greater genetic similarity with Andean indigenous people such as Quechua and Aymara (Fig. 2, Supplementary Fig. 1, Methods).

This unique genetic heritage provides both a challenge and an opportunity for biomedical research. To optimally capture genetic variation, and particularly rare variations in Peruvians, we designed a 712,000-SNP customized array (LIMAArray) with genome-wide coverage based on whole-exome sequencing data from 116 active TB cases (Methods, Supplementary Table 2, Supplementary Fig. 2). When compared to other more comprehensive genotyping platforms available at the time, LIMAArray showed an ~5% increase in imputation accuracy, particularly for population-specific and low-frequency variants (Supplementary Table 3). We derived estimated genotypes for ~8 million variants using the 1000 Genomes Project Phase 3 as the reference panel and tested single marker and rare-variant burden associations with linear mixed models that account for both population stratification and relatedness in the cohort (Supplementary Figs. 3–4, Methods). Genome-wide association results of 2160 cases and 1820 controls after quality control (Methods) are summarized in Supplementary Fig. 5. We observed no inflation of test statistics (λ*GCTA* = 1.03, λ*GAT* = 1.00 for common and rare association analyses respectively), which suggests potential biases were strictly controlled in our study. We observed no significant rare variant (minor allele frequency (MAF) <1%) association with TB progression after performing gene-based generalized linear mixed model (Methods).

**Progression of recent *Mycobacterium tuberculosis* exposure to active tuberculosis is a highly heritable complex trait.** To investigate the genetic basis of early TB progression, we first estimated its variant-based heritability (*h*^2^). Using GCTA we estimated *h*^2^ of TB progression to be 21.2% (standard error (s.e.) = 0.08, *F*~GAUSSIAN~ = 2.64 × 10^−3^) on the liability scale with assumed incidence rate of 0.05 in the cohort (Methods). To avoid biases introduced from calculating genetic relatedness matrices (GRMs) in admixed individuals, we calculated two different GRMs based on admixture-aware relatedness estimation methods and removed related individuals. Both admixture-aware methods reported similar *h*^2^ estimates (Supplementary Table 4), indicating our reported heritability estimation is robust under different model assumptions. We quantified *h*^2^ of TB progression and observed a surprisingly strong genetic basis. This degree of heritability is comparable to traits with a well-established genetic basis (Supplementary Table 5). For example,
GWAS have identified ~200 risk loci for Crohn’s disease\(^{24,25}\), which has a reported \(h^2\) of 28.4% (s.e. = 0.02, \(P_{\text{Gaussian}} = 8.62 \times 10^{-71}\))\(^{24}\).

To compare the genetic heritability between early TB progression and population-wide TB susceptibility, we subsequently obtained genotypes from a previous TB study conducted in Russia with 11,137 individuals\(^{11}\). Using GCTA, we estimated the \(h^2\) of population-wide TB susceptibility to be 17.8% (s.e. = 0.02, \(P_{\text{Gaussian}} = 2.85 \times 10^{-21}\)) with assumed prevalence of 0.04\(^{26}\). Even though the point estimate of \(h^2\) of TB progression is greater than that of population-wide TB risk in the Russian study, these estimates are not statistically different from each other (two-tailed \(t\)-test \(P = 0.68\), Supplementary Fig. 6). Regardless, the strong host genetic basis of TB progression suggests that larger progression studies may be well-powered to discover additional variants.

**Genome-wide association study identifies a novel association at 3q23.** We next identified a novel risk locus associated with TB progression on chromosome 3q23, which is comprised of 11 variants in non-coding regions downstream of RASA2 and upstream of RNF7 (\(P < 1 \times 10^{-5}\)) (Fig. 3a, Supplementary Table 6, Dataset 1). The strongest association with early TB progression was at a genotyped variant rs73226617 (OR = 1.18; \(P = 3.93 \times 10^{-8}\)). To test for artifacts and to identify stronger associations that might have been missed due to genotyping and imputation, we first checked the genotype intensity cluster plot of the top associated variant which showed a clear separation between genotypes AA, AG, and GG (Supplementary Fig. 7). We then designed individual TaqMan genotyping assays for four top associated variants (Methods, Supplementary Table 7). We genotyped these four SNPs in 4002 subjects and concluded that all four variants show a high concordance rate (>99%) with imputed genotypes (Supplementary Table 6, Dataset 1). Because all 11 variants in the risk locus are in high linkage disequilibrium (LD) with each other (Supplementary Fig. 8), the other imputed variants are also likely to have high imputation quality.

To determine whether the reported risk locus at 3q23 also has an independent association with TB progression from recent *M. tb* infection, we conducted a case-only analysis removing age from our case selection criteria. This approach is based on the premise that TB cases that share a DNA fingerprint for *M. tb* and HHCs who developed active TB are epidemiologically related while cases in which *M. tb* fingerprints are different might have resulted from remote infection that reactivated during the study assessment\(^{27}\). 1472 out of 2175 presumed early progressors shared molecular fingerprint of *M. tb* isolates with another case or developed active TB during the 1 year of follow-up (Supplementary Fig. 9). Other cases did not have a shared molecular fingerprint among *M. tb* isolates or did not come from the same household as the index case, leading to a lower degree of certainty in the early progression status of these cases. In this case-only analysis, the top associated signal rs73226617 was nominally associated with early progression (OR = 1.09, \(P = 0.016\)). A heritability analysis restricted to those that shared the same molecular fingerprint or from the same household estimated in a \(h^2\) of 22.1% (s.e. = 0.06, \(P_{\text{Gaussian}} = 1.32 \times 10^{-4}\)) despite the smaller number of samples. To assess the independence of the stratified cases compared to the overall case-control analysis, we first compared reported effect sizes in both analyses and observed a low Pearson correlation \((r = 0.014\), Supplementary Fig. 10). To test the significance of the reported association, we performed a permutation analysis, where we randomly permuted the case/control status in the stratified analysis. After permuting for 10,000 times, the observed OR (1.09) has a \(P\)-value of 0.017 (Supplementary Fig. 11). We next performed a Bayesian analysis to test whether the reported association is restricted to the early progressors after recent exposure to *M. tb* (Methods). The disease specific approximate Bayes Factor\(^{28}\) (i.e., the ratio of the marginal likelihood for a model where the variant is only associated with early progressor who has a shared molecular fingerprints and/or a secondary cases and for a model where is associated with all progressors) is 0.42. This suggested that the SNP is most likely to be associated with early progressors who have recent exposure to *M. tb* alone, but
almost equally likely to be associated with TB progression in general.

We examined the 11 most associated variants for early TB progression identified in the Peruvian cohort in previously published GWAS datasets9–11 (Supplementary Table 8, Dataset 2). These SNPs were less frequent (<1%) in the African populations than in the European and Peruvian populations, resulting in lower statistical power to detect association. We
Fig. 2 Global ancestry analysis of Peruvian populations. a ADMIXTURE plot of admixed individuals and continental reference panels. Each individual is represented as a thin vertical bar. The colors can be interpreted as different ancestries. Reference panels are either from the 1000 Genomes project (1000G) or Native American individuals collected from Reich et al. Nature. Han Chinese are from Beijing, China; Yoruba are from Ibadan, Nigeria; European individuals are Utah Residents (CEPH) with Northern and Western European Ancestry; Puerto Ricans are from Puerto Rico; Colombians are from Medellin, Colombia; Mexican individuals are from Los Angeles, California; Peruvians are from Lima, Peru. Northern Americanindian includes individuals from Maya, Mixe, and Kqchikel. Central American includes individuals from Pima, Zapotec, Mixtec, Yaqui, Chorotega, Tepehuano. Southern American includes individuals from Piapeco, Karitiana, Surui, Wayuu, Jamamadi, Parakana, Guarani, Kaingang, Ticuna, Palikur, Toba, Arara, Wichi, Chane and Guahibo. Andean population includes Quechua and Aymara. K = 6 models are shown above, K = 3 through K = 15 models are available in Supplementary Fig. 1. Source data are provided as a Source Data file. b Map of locations of sampled Native American populations.

therefore examined the SNPs in two previously published Russian (5530 TB cases and 5607 controls) and Icelandic (4049 TB cases and 6543 TST + controls) GWAS datasets. We observed that the effects in the Russian cohort were similar, as they shared comparable ORs of 1.10 (Russian) and 1.18 (Peru) for rs73226617 (PRussian = 0.065). In contrast, there was no signal observed in the Icelandic cohort (OR = 1.06, P Iceland = 0.437). Consistent with our previous case-only analysis, the weaker signals observed in both European cohorts indicate that 3q23 is associated with early TB progression. The lack of association observed in the two European cohorts could be due to the inclusion of reactivation TB cases and noninfected controls; differences in TB prevalence (Methods).

We next examined how previously published TB GWAS risk loci are associated with progression in this study. We detected evidence of association in a previously reported TB locus at rs9272785 in the HLA region (OR = 1.04, P = 4.49 × 10−3), but did not detect signals at other reported risk loci (Supplementary Table 9). Thus, previously reported loci may relate to infection or reactivation phenotypes, rather than early TB progression whereas HLA association may affect both early progression and reactivation. Next, we performed an HLA imputation using a multi-ethnic model framework (Methods, Supplementary Fig. 12), we tested associations between specific amino acid positions and TB progression which identified the most significant association at amino acid position 73 of HLA-A (OR = 1.12, P = 1.03 × 10−6). We noted several other amino acids of class I genes with suggestive associations (P < 1 × 10−3) including position 97 of HLA-B (OR = 1.05, P = 8.99 × 10−6). Notably, amino acid variability at this position affects the structure and flexibility of the peptide-binding groove and is associated with many infectious and autoimmune phenotypes, such as HIV-1 viral load and ankylosing spondylitis. These results suggest that HLA class I genes might play a role in TB progression.

To try to identify which of the variants in our reported risk locus is likely to be the functional polymorphism affecting the risk of pulmonary TB progression, we employed the FINE-MAP software (Methods). The 90% credible set includes seven genetic variants, with rs73226617 having the highest posterior probability (0.54), followed by rs58538713 (0.16) and the indel rs148722713 (0.05) among 713 variants in the region (Fig. 3b, Supplementary Table 6, Dataset 1).

A monocyte-specific regulatory element in 3q23 is implicated in TB progression. To identify likely functional variants and target genes, we employed a method called IMPACT (Inference and Modeling of phenotype-related ACTive Transcription) briefly, IMPACT identifies regions predicted to be involved in transcriptional regulatory processes related to a key transcription factor of a cell type (Methods) by leveraging information from approximately 400 chromatin and sequence annotations in public databases (Fig. 3c, Supplementary Table 10, Dataset 3). Each variant is assigned a probability between 0 (least likely to be a regulatory element) and 1 (most likely to be a regulatory element). We tiled through the 23,308 base pair region on a per-nucleotide basis, computing the probability of a cell-type regulatory element separately for 15 different cell types and cell states of which 10 are immune cell types with known roles in TB outcomes, including T cells, B cells, monocytes, macrophages, and peripheral blood cells (Fig. 3e). We observed monocyte-specific predicted regulatory elements at rs73226617 and rs148722713 (IMPACT score 0.79 and 0.41, respectively, Fig. 3d).

We next searched for other epigenomic evidence that may indicate changes in transcriptional enhancers and other cis-regulatory elements. Given the possible monocyte-specific activity of the identified risk locus, we actively sought datasets that include monocyte primary cells or cell lines. We used data presented in the BLUEPRINT project to search for chromQTLs. We observed significant chromQTL present in the region (characterized by the presence of H3K4me1) in monocytes (Supplementary Fig. 13) suggesting that this region indeed has an active enhancer. The rs73226617 SNP was included in this region, but did not itself have evidence of chromQTL activity; however, it is in high LD with the top associated chromQTL signal (rs1568171, D’ = 1.0).

Based on the IMPACT analysis and the suggested enhancer activity in monocytes, we studied mononuclear cells (THP1) as the most likely experimental model for locus-specific gene regulatory activity. We performed electrophoretic mobility shift assays (EMSA) to test whether the variants differentially bound nuclear complexes in an allele-specific manner among the seven variants that constitute the 90% credible set (Methods). We could detect differential protein binding that was competed out by unlabeled probes for three of the risk alleles (rs73226617, rs58538713, and rs148722713) (Supplementary Fig. 14), providing evidence that these alleles might confer differential transcription factor binding activity, and in the right context may lead to altered enhancer activity.

On the basis of posterior probabilities from the genetic data, EMSA binding assays demonstrating the capacity to alter binding of nuclear extract protein, and localization to an enhancer region with regulatory potential, we identified rs73226617, and rs148722713 as the most likely causal alleles.

Potential target genes implicated by the TB progression risk locus. Next, we searched public promoter Hi-C databases to identify any significant interactions between the monocyte-specific enhancer harboring our most likely causal allele, rs73226617 and rs148722713. We found that in monocytes, both of the risk variants (rs73226617, rs148722713) are in a region that interacts with the promoter of ATP1B3 (Supplementary Fig. 15a, b). Similar to the IMPACT results, we found the variant-gene interactions are strongest in monocytes compared to other cell types (Supplementary Fig. 15c, d), suggesting cell-type-specific activities in the
identified TB risk locus. ATP1B3 (ATPase Na+/K+ Transporting Subunit Beta 3) is a protein-coding gene, which belongs to the family of Na+/K+ and H+/K+ ATPases. Na+/K+–ATPases are composed of an alpha, beta, and FXYD subunits, are integral membrane proteins responsible for establishing and maintaining the electrochemical gradients of sodium and potassium ions across the plasma membrane through active transport against their osmotic gradients. A recent study demonstrated that the Na,
Fig. 3 Genome-wide association details of the 3q23 locus. a A regional association plot of the 3q23 locus including all genotyped and imputed variants. The horizontal line indicates the genome-wide significant threshold at 1.78 × 10^{-7} for Peruvian populations (Methods). b Fine-mapping posterior probability of all variants in the chr3:140221602-145217859 region. c Number of overlaps between all variants in the risk locus and ~400 epigenetic features. d Predicted posterior probability of cell-type-specific gene regulatory activity using IMPACT (Inference and Modeling of phenotype-related ACTive Transcription) based on the epigenetic chromatin signature of binding sites of the transcription factor CEBPB in monocytes. e Intersection of nucleotide-resolution of variant cell-state IMPACT annotations with potential causal variants in 3q23 locus. The γ-axis shows the posterior probability of predicted cell-state regulatory activity among each variant in 15 different cell types and cell states. The x-axis shows the genomic positions of all variants among the identified risk locus. The bolded variant (rs73226617) is the leading risk variant from the association study which shows the highest predicted cell-state regulatory activity in monocytes (masked by CEBPB transcription factor). Dashed lines highlight 11 top associated variants. Genotyped variant rs73226617 is highlighted in red. Source data are provided as a Source Data file.

K ATPase Beta 3 subunit in monocytes has an important function in mediating a normal T cell response. Indeed ligating it with an antibody resulted in a blunted T cell response after stimulation. This effect was specific to the monocytes population. Consistent with these findings, differential expression of ATP1B3 in whole blood, along with genes coding for other members of the Na+/K +-ATPases, was recently reported to be associated with TB progression in an African cohort of household contacts of TB patients. Collectively, the Hi-C analysis and reported association with TB progression point to ATP1B3 as a candidate gene of the risk locus in 3q23.

Since in silico evidence suggested that our identified TB risk locus harbors monocyte-specific regulatory elements, we used the CRISPR/Cas9 system to introduce insertions/deletions around the top associated variant rs73226617 (Methods, Supplementary Fig. 16a). Among 23 sorted and grown clones that had unchanged risk loci or harbored unique edits and deletions (Supplementary Fig. 16a), we did not observe differential gene expression between edited and unedited THP1 clones in the cis-genes around the 500 kb window of the leading rs73226617 variant (ANOVA, P-value > 0.05, Methods, Supplementary Fig. 17a). This CRISPR/Cas9 approach to disrupt the putative enhancer has multiple limitations. Firstly, while we observed no effect in THP1 cell lines, this might result from differences between primary monocytes and transformed THP1 cell lines, or failure to identify the relevant activation conditions and cell context to test enhancer activities, which are known to influence eQTL interactions. Secondly, although we chose guide RNA sequences optimized to target the 3q23 region, and did not identify other likely genomic targets by nucleotide homology, off-target effects are still possible (Supplementary Table 20). We analyzed independent edited THP1 clones to reduce the likelihood that we propagated additional off-target genomic edits. However, a genome-wide analysis of differential expression also did not detect any other differentially expressed targets outside the local neighborhood surrounding rs73226617 (Supplementary Fig. 17b), suggesting that off-target disruptions were unlikely. In particular, we noted the enhancer activity seen in primary monocytes, is not seen in THP1 cell lines.

Discussion

Overall, our results argue that rapid TB progression is a highly heritable trait, comparable to other human diseases with an established genetic origin. More generally, these results begin to address general questions about genomic approaches to infectious diseases, which have lagged behind other diseases and complex traits in terms of locus discovery (Supplementary Table 5). Infections, especially chronic infectious diseases, play out in highly distinct phases that involve exposure, crossing epithelial boundaries, pathogen expansion, locating a host niche, and in the case of TB, decades-long persistence, reactivation, and retransmission. Each of these stages can be controlled by distinct host factors. Our analysis indicates that progression from recent M.tb exposure to active TB has a different genetic basis compared to TB reactivation. Specific analysis of clinical progression as a distinct phase allows for a more powerful detection of risk factors for an equal number of samples, as compared to case-control studies, which are an amalgamation of different phenotypes. Thus, this work suggests that while detailed, stage-specific phenotypic profiling may be more costly, it may offer key advantages for infectious disease genetic studies. Specifically, it allows for precise phenotype definitions and identification of biological targets with specific implications. Therefore, detailed phenotypic profiling should become an additional valuable approach for future genetic studies of infectious diseases. Detailed phenotyping enables investigators to dissect pathogenic mechanisms at different stages of infection and disease progression.

Methods

Ethics statement. We recruited 4002 subjects from a large catchment area of Lima, Peru that included 20 urban districts and ~3.3 million residents to donate a blood sample for use in our study.

We obtained written informed consent from all the participants. The study protocol was approved by the Institutional Review Board of Harvard School of Public Health and by the Research Ethics Committee of the National Institute of Health of Peru.

Preparation of genome-wide genetic data. We enrolled index cases as adults (aged 15 and older) who presented with clinically suspected pulmonary TB at any of 106 participating health centers. We excluded patients who resided outside the catchment area, who had received treatment for TB before and those who were unable to give informed consent. Pulmonary TB patients have been diagnosed by the presence of acid fast bacilli in sputum smear or a positive M.tb culture at any time from enrollment to the end of treatment. All cultures of the index cases were genotyped using mycobacterial interspersed repetitive units-variable number of tandem repeats (MIRU-VNTR). Within 2 weeks of enrolling an index patient, we enrolled his or her household contacts (HHCs). The M.tb status was determined using the Tuberculin Skin Test (TST). All HHCs were evaluated for signs and symptoms of pulmonary and extra-pulmonary TB disease at 2, 6, and 12 months after enrollment. All cases were HIV-negative, culture-positive and drug-sensitive to pulmonary TB. We defined cases who were likely to have recently exposed TB, if a case satisfied at least one of the three criteria: (1) exposed HHCs who developed active TB during a 12 month follow up period; (2) index patients whose M.tb isolates shared a molecular fingerprint with isolates from other enrolled patients and (3) index patients who were 40 years old or younger at time of diagnosis. To maximize the likelihood that controls were exposed to M.tb but did not develop active disease, we chose them from among TST positive HHCS with no previous history of TB disease, and who remained disease free at the time of recruitment both by directly re-contacting individuals to inquire about their latest medical history and by checking their names against lists of notified TB patients at all of the 106 health clinics. Where possible, we chose controls who are less than second-degree related to the index cases.

Customized axiom array for Peruvian populations. We developed a custom array (LIMMArray based on whole-exome sequencing data from 116 active TB cases to optimize the capture of genome-wide genetic variation in Peruvians. Many markers were included because of known associations with, or possible roles in, phenotypic variation, particularly TB-related (Supplementary Table 12). The array also includes coding variants across a range of minor allele frequencies (MAFs), including rare markers (<1% MAF), and markers that provide good genome-wide coverage for imputation in Peruvian populations in the common (>5%), low frequency (1–5%) and rare (0.5–1%) MAF ranges (Supplementary Table 3). This
approach allowed the detection of rare population-specific coding variants and those which predisposed individuals to TB risk.

Genotyping and quality control. We extracted genomic DNA from whole blood of the participating subjects. Genotyping of all samples was performed using our customized Affymetrix LIMAAArray. Genotypes were called in a total of 4002 samples using the apt-genotype-axiom48. Individuals were excluded if they were missing more than 5% of the genotype data, had an excess of heterozygous genotypes (≥3.5 standard deviations, Supplementary Table 13), duplicated with identifying state of 50% or in index cases with an allelic score greater than 10 years old. After excluding these individuals, we excluded variants with a call rate less than 95%, with duplicated position markers, those with a batch effect (P < 10−5), Hardy–Weinberg (HWE) P-value below 10−5 in controls, and a missing rate per SNP difference in cases and controls greater than 10−5 (Supplemental Table 14). In total, there were 877,232 SNPs left for imputation and association analyses after quality control.

Imputation and association analyses. The genotyped data were pre-phased using SHAPEIT249, IMPUTE250 was then used to impute genotypes at untyped genetic variants using the 1000 Genomes Project Phase 3 dataset19 as a reference panel. For SNPs using GEMMA55 version 0.7 association analyses (Supplementary Table 6). This result supports that our reported strengths genome-wide (Supplementary Fig. 19) and in our reported top-association factor (i.e., the ratio of the marginal likelihood for a model where the variant is only associated with all progressors (log(ABP) = −0.02); if twice as high, 0.26 (s.e. 0.09). If the true prevalence was in fact half as high, our estimate would instead be 0.17 (i.e. 0.02); if twice as high, 0.06 (i.e. 0.09). H2 on the observed scale is 0.24 (i.e. 0.09).

Bayesian meta-analysis on GWAS summary statistics. Briefly, we adopted a Bayesian meta-analysis approach39 to test whether the reported top association is restricted to the early progressors only. We calculated the approximate Bayes factor (ABF)28 for the top associated variant (rs73226617), testing the hypothesis that the reported association is specific to early progressors with a shared molecular fingerprint, and assumed the variance σ2 of the true effect to be 0.04 as suggested by previous studies26,59. We assumed the predicted correlation of true effects (ρ) between two phenotypes to be 0.5. The disease-specific log(ABP)(i.e., the ratio of the marginal likelihood for a model where the variant is only associated with early progressors who has a shared molecular fingerprints and/or a secondary cases (log(ABP) = 0.81) and for a model where is associated with all progressors, (log(ABP) = 6.19) is −0.38. To test the robustness of the model using different priors (σ2 and ρ), we varied the values of σ = (0.1,0.2,0.3,0.4) and ρ = 0.01,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9) but did not detect a strong difference that would alter the conclusion above (Supplementary Table 17).

In silico functional annotation of candidate causal variants. We combined multiple sources of in silico genome-wide functional annotations from publicly available databases to help identify potential functional variants and target genes in the 3q23 novel risk locus. To investigate functional elements enriched across the region encompassing the strongest candidate causal variants, we aggregated ~400 epigenomic and sequence annotations including cell-type-specific annotation types such as ATAC-seq, DNase-seq, FAIRE-seq, HiChIP-HK372ac, HiChIP-CTCF, polyA-seq, and DNA methylation. As well as cell-type-specific nonspecific annotations, such as conservation scores and sequence annotation, such as coding, intronic, intergenic, etc. A list of all included resources is summarized in Supplementary Table 10, Dataset 3.

Using IMPACT, we built a model that predicts cell type gene regulatory elements by learning the epigenomic profiles of key TF binding sites in the cell type. Briefly, we trained IMPACT to distinguish regulatory elements from non-regulatory elements among 11 immune-related TFs and 4 others (Supplementary Table 18). To create the class of gold standard regulatory elements, we scanned the ChIP-seq peaks of the 10 most enriched TFs, mentioned above, for matches to the TF binding motif, using HOMER52 [4,8,3] and retained the genomic location of the highest scoring match for each ChIP-seq peak to the regulatory class. To create the class of non-regulatory elements, we scanned the entire genome for motif matches of each of the 14 master TFs, again using HOMER, and selected motif matches with no overlap with the ChIP-seq peaks. IMPACT learns an epigenomic profile representative of cell type regulatory elements in 10-fold cross validation (CV) using the complete sets of regulatory and non-regulatory elements. We scored regions of interest according to the learned feature profile from this CV.

Electrophoretic mobility shift assay (EMSA). Frozen cell pellets from the THP1 cell line (ATCC) were used for preparation of nuclear extracts using NE-PER Nuclear and Cytoplasmic Extraction reagent (ThermoFisher) according to the manufacturer’s instructions, then dialyzed overnight at 4 °C with gentle stirring in 10 mM Tris–HCl (pH 7.5), 100 mM NaCl, 10 mM MgCl2, 10% glycerol, 0.5 mM EDTA, 0.5 mM DTT, 0.5 mM PMSF, 0.5 mM benzamidine, 0.5 mM pepstatin, 0.5 mM aprotinin, 0.5 mM leupeptin, 0.5 mM ethylenediaminetetraacetic acid (EDTA), 0.5 mM phenylmethylsulfonyl fluoride (PMSF). Samples were quantified using BCA Protein Assay Kit (ThermoFisher, catalogue no. 23227) and stored in 1x Halt protease inhibitor cocktail (ThermoFisher, catalogue no. 78437) at −80 °C until use. We designed single-stranded oligonucleotides (15bp corresponding to the set of alleles, Supplementary Table 19), and biotinylated the forward and reverse sequences separately using the Biotin 3'End DNA Labeling Kit (ThermoFisher Scientific, catalogue no. 20160) following the manufacturer’s instructions. Single-stranded probes were annealed by incubation for 5 min at 95 °C followed by 1 h at room temperature. EMSA reactions were performed using the LightShift Chemoshift immuno- luminescent EMSA kit (ThermoFisher, catalogue no. 20148). Binding reactions were performed in a volume of 20 μL: 2 μL of 10x binding buffer, 16 μg nuclear extract, 2.5% glycerol, 5 mM MgCl2, 0.05% Nonidet P-40 and 50 μg Poly d(dC):d(dG) as a nonspecific DNA competitor, and 20 μl of biotinylated probes with or without unlabeled competitor probes at 200 fold molar excess. The binding reaction was resolved on 5% Tris-borate–EDTA (TBE) gels (BioRad, catalogue no. 3450049) at 110 volts. Gels were transferred for 1 h at 4 °C at 40 volts onto pre-cut zeta-probe
nylon membrane (Bio-Rad, catalogue no. 162-0165). Transferred DNA was UV
crosslinked for 10 min, then blocked and incubated with stabilize streptavidin-
horseradish peroxidase conjugate, at 1:300 dilution in ECSA blocking buffer, 
then washed and detected by chemiluminescence. Finally, exposed on CL-Xposur™
films (ThermoFisher Scientific, catalogue no. 34089).

CRISPR/Cas9 editing around rs73226617 in 3q23, and did not have additional targets in the genome. For
pyruvate, 0.05 2-mercaptoethanol, 1× penicillin-streptomycin). To disrupt the
202) were cultured in complete RPMI (RPMI-1640, Gibco,10% fetal bovine serum,
14, 17 and 19 are provided as a Source Data
scripts and data for generating

1. Houben, R. M. G. J. & Dodd, P. J. The global burden of latent tuberculosis infection: a re-estimation using mathematical modelling. PLoS Med. 13, e1002152 (2016).

2. Organization, W. H. & Others. Global tuberculosis report 2018. (World Health Organization, 2018).

3. Behr, M. A., Edelstein, P. H. & Ramakrishnan, L. Revisiting the timetable of tuberculosis. BMJ 362, k2738 (2018).

4. van der Eijk, E. A., van de Vosse, E., Vandenbroucke, J. P. & van Dissel, J. T. Heredity versus environment in tuberculosis in twins: the 1950s United
Kingdom Prophit Survey Simonds and Comstrock revisited. Am. J. Respir. Crit. Care Med. 176, 1281–1288 (2007).

5. Kallmann, F. J. & Reiner, D., Others. Twin studies on genetic variations in resistance to tuberculosis. J. Hered. 34, 269–276 (1943).

6. Cobat, A. et al. High heritability of antituberculosis immunity in an area of hyperendemicity for tuberculosis disease. J. Infect. Dis. 201, 15–19 (2010).

7. Stein, C. M. Genetic epidemiology of tuberculosis susceptibility: impact of study design. PLoS Pathog. 7, e1001189 (2011).

8. Abel, L., El-Baghdadi, J., Bousilha, A. A., Casanova, J.-L. & Schurr, E. Human genetics of tuberculosis: a long and winding road. Philos. Trans. R. Soc. Lond. B Biol. Sci. 369, 20130428 (2014).

9. Thye, T. et al. Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. Nat. Genet. 42, 739–741 (2010).

10. Thye, T. et al. Common variants at 11p13 are associated with susceptibility to tuberculosis. Nat. Genet. 44, 257–259 (2012).

11. Curtis, J. et al. Susceptibility to tuberculosis is associated with variants in the ASAP1 gene encoding a regulator of dendritic cell migration. Nat. Genet. 47, 523–527 (2015).

12. Mahasirimongkol, S. et al. Genome-wide association studies of tuberculosis in Asians identify distinct at-risk locus for young tuberculosis. J. Hum. Genet. 57, 363–367 (2012).

13. Chimusa, E. R. et al. Genome-wide association study of ancestry-specific TB risk in the South African Coloured population. Hum. Mol. Genet. 23, 796–809 (2014).

14. WHO Global tuberculosis report 2017. (2017).

15. Sandovall, J. R. et al. Tracing the genomic ancestry of Peruvians reveals a major legacy of pre-Columbian ancestors. J. Hum. Genet. 58, 627–634 (2013).

16. Wang, S. et al. Geographic patterns of genome admixture in Latin American Mestizos. PLoS Genet. 4, e1000307 (2008).

17. Hui Dehart, E. From slavery to freedom: Chinese coolies on the sugar plantations of nineteenth century Cuba. Labour Hist. 31:51–77 (2017).

18. Gonzales, M. J. Chinese plantation workers and social conflict in Peru in the late nineteenth century. J. Lat. Am. Stud. 21, 385–424 (1989).

19. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. Nature 526, 75–81 (2015).

20. Reich, D. et al. Reconstructing native American population history. Nature 488, 370–374 (2012).

21. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88, 76–82 (2011).

22. Conomos, M. P. et al. Genetic diversity and association studies in US plantations of nineteenth century Cuba. J. Hum. Genet. 57, 199–212 (2012).

23. Thornton, T. et al. Estimating kinship in admixed populations. Am. J. Hum. Genet. 91, 122–138 (2012).

24. Luo, Y. et al. Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7. Nat. Genet. 49, 186–192 (2017).

25. Liu, J. Z. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat. Genet. 47, 979–986 (2015).

26. Speed, D. et al. Reevaluation of SNV heritability in complex human traits. Nat. Genet. 49, 986–992 (2017).

27. Murray, M. Determinants of cluster distribution in the molecular epidemiology of tuberculosis. Proc. Natl Acad. Sci. USA. 99, 1538–1543 (2002).

28. Wakefield, J. Bayes factors for genome-wide association studies: comparison with P-values. Genet. Epidemiol. 33, 79–86 (2009).

29. Sveinbjornsson, G. et al. HLA Class II sequence variants influence tuberculosis risk in populations of European ancestry. Nat. Genet. 48, 318–322 (2016).

30. International HIV Controllers Study. et al. The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. Nat. Genet. 41, 1465–1470 (2013).

31. McLaren, P. J. et al. Polymorphisms of large effect explain the majority of the host genetic contribution to variation of HIV-1 virus load. Proc Natl Acad. Sci. USA 112, 14568–14566 (2013).

32. Cortes, A. et al. Major histocompatibility complex associations of ankylosing spondylitis are complex and involve further epistasis with ERAPI. Nat. Commun. 6, 7416 (2015).

33. Eigenmayer, C. et al. FINE-MAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics 32, 1493–1501 (2016).
34. Amariuta, T. et al. IMPACT: Genomic annotation of cell-state-specific regulatory elements inferred from the epigenome of bound transcription factors. Am. J. Hum. Genet. 104, 879–895 (2019).
35. Chen, L. et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. Cell 167, 1398–1414 (2016).
36. Schofield, E. C. et al. ChIPC: a web-based tool for the integrative and interactive visualization of promoter capture Hi-C datasets. Bioinformatics 32, 2511–2513 (2016).
37. Javiere, B. M. et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. Cell 167, 1369–1384 (2016).
38. Takheew, N. et al. Ligation of Na, K ATPase β3 subunit on monocytes by a specific monoclonal antibody mediates T cell hypofunction. PLoS ONE 13, e0203557 (2018).
39. Dufy, F. J. et al. Immunometabolic signatures predict risk of progression to active tuberculosis and disease outcome. Front. Immunol. 10, 527 (2019).
40. Dimas, A. S. et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. Science 325, 1246–1250 (2009).
41. The GTex Consortium. The genotype-tissue expression (GTEx) pilot analysis: multissue gene regulation in humans. Science 348, 648–660 (2015).
42. Gutierrez-Arcelus, M. et al. Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing. PLoS Genet. 11, e1004958 (2015).
43. Gutierrez-Arcelus, M., Baglaenko, Y., Arora, J. & Hannes, S. Allele-specific expression changes dynamically during T cell activation in HLA and other autoimmune loci. Preprint at https://www.biorxiv.org/content/10.1101/599449v1.full (2019).
44. Corces, M. R. et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat. Genet. 48, 1193–1203 (2016).
45. Phanstiel, D. H. et al. Static and dynamic DNA loops form AP-1-bound activation hubs during macrophage development. Mol. Cell 67, 1037–1048 (2017).
46. Mohaghegh, N. et al. NextPBM: a platform to study cell-specific transcription factor binding and cooperativity. Nucleic Acids Res. 47, e51 (2019).
47. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
48. Schillert, A. & Ziegler, A. Genotype calling for the Affymetrix platform. Methods Mol. Biol. 850, 513–523 (2012).
49. O’Connell, J. et al. A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet. 10, e1004234 (2014).
50. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. 44, 955–959 (2012).
51. Jia, X. et al. Imputing amino acid polymorphisms in human leukocyte antigens. PLoS ONE 8, e64683 (2013).
52. Luo, Y. et al. Novel high-resolution multi-ethnic HLA imputation reference panels constructed based on high-coverage whole-genome sequencing data. Poster presented at: American Society of Human Genetics Annual Meeting; Orlando, USA 2017.
53. Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat. Methods 11, 407–409 (2014).
54. Price, A. L. et al. Long-range LD can confound genome scans in admixed populations. Am. J. Hum. Genet. 83, 132–135 (2008). ; author reply 135–9.
55. Chen, H. et al. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. Am. J. Hum. Genet. 98, 653–666 (2016).
56. Kanai, M., Tanaka, T. & Okada, Y. Empirical estimation of genome-wide significance thresholds based on the 1000 Genomes Project data set. J. Hum. Genet. 61, 861–866 (2016).
57. Benner, C. et al. Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. Am. J. Hum. Genet. 101, 539–551 (2017).
58. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42, 565–569 (2010).
59. Crochet, H. et al. Bayesian meta-analysis across genome-wide association studies of diverse phenotypes. Genet. Epidemiol. 43, 532–547 (2019).
60. Jostins, L. & McVean, G. Trinculo: Bayesian and frequentist multinomial logistic regression for genome-wide association studies of multi-category phenotypes. Bioinformatics 32, 1989–1990 (2016).
61. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 38, 576–589 (2010).
62. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34, 525–527 (2016).

Acknowledgements
The study was supported by the National Institutes of Health (NIH) TB Research Unit Network, Grant U19 AI111224 and NHGRI U01 HG009379. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. S.N. was supported by MRC (MR/M012382/1), the ERC Starting grant (260477) and the National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre. T.A. was supported by NIH (NHGRI T32 HG00295). The authors thank Garðar Sveinbjörnsson, Patrick Sulem, Ingileif Jonsdottir, and Kari Stefansson at deCODE genetics, Reykjavik, Iceland, for validating the association of rs73226617 with TB progression in the Icelandic population.

Author contributions
Y.L. designed the genotyping array, performed statistical analysis of the GWAS data and wrote the first draft of the manuscript. S.S. performed the EMSA and CRISPR/Cas9 experiments. S.A. carried out the rare association studies of the GWAS data. T.A. implemented the IMPACT model. Y.B. helped to develop the protocols for the CRISPR/Cas9 experiments. K.I. performed the chromQTL analysis. M.G.A. helped the low-input RNA sequencing data analysis. R.C., L.L., S. R. L., J. J., R. Y., C.C., J. T. G., M.B. and M.B.M. participated in study design, protocol development, and sample collection. S.N. contributed the Russian data for the meta and heritability analysis. M.B.B. and P.A.N. helped to develop the protocols for EMSA experiments. D.B.M supervised the EMSA and CRISPR/Cas9 experiments. M.B.M. participated in study design, protocol development, and study conception. S.R. and M.B.M conceived and supervised the study. All authors contributed to the writing of the manuscript.

Additional information
Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-019-11664-1.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/.

Peer review information: Nature Communications thanks James Gilchrist and the other anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.