Genome-wide association study in Brazil identifies risk factor-adjusted genetic susceptibility to pulmonary tuberculosis with cell-specific gene expression effects

Short title: Brazil PTB GWAS

Kimberly A Dill-McFarland[1*†], Bruno B Andrade[2,3*†], Marina C Figueiredo[4], Alice M S Andrade[2,3], Francys Avendaño-Rangel[2,3], Marcelo Cordeiro-Santos[5,6], Afrânio L Kritski[7], Valeria C Rolla[8], Juan M Cubillos-Angulo[2], Spyros A Kalams[4,9], Josh D Simmons[4], Jared M Oakes[9], Jonathan Peña Avila[10], Helder I Nakaya[10,11], Rama D Gangula[4], Peter F Rebeiro[4], Gustavo Amorim[12], Simon A Mallal[4], Regional Prospective Observational Research in Tuberculosis (RePORT)-Brazil Consortium, Timothy R Sterling[4†], Thomas R Hawn[1†]

[1] Allergy and Infectious Diseases, Department of Medicine, University of Washington, Seattle, WA, USA;
[2] Laboratório de Pesquisa Clínica e Translacional, Instituto Gonçalo Moniz, Fundação Oswaldo Cruz (FIOCRUZ), Salvador, BA, Brazil;
[3] Instituto de Pesquisa Clínica e Translacional, Medicina Zarns, Clariens Educação, Salvador, BA, Brazil;
[4] Division of Infectious Diseases, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA;
[5] Fundação de Medicina Tropical Dr. Heitor Vieira Dourado do Amazonas, Manaus, AM, Brazil;
[6] Curso de Medicina, Universidade do Estado do Amazonas (UEA), Manaus, AM, Brazil;
[7] Programa Acadêmico de Tuberculose da Faculdade de Medicina, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brazil;
[8] Laboratório de Pesquisa Clínica em Micobacteriose, Instituto Nacional de Infectologia Evandro Chagas (FIOCRUZ), Rio de Janeiro, RJ, Brazil;
[9] Department of Pathology, Microbiology, and Immunology, Vanderbilt University Medical Center, Nashville, TN, USA;
[10] Department of Clinical and Toxicological Analyses, University of São Paulo, São Paulo, SP, Brazil;
[11] Hospital Israelita Albert Einstein, São Paulo, SP, Brazil;
[12] Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA

* Co-first authors
[†] Co-corresponding authors

Corresponding authors
Kimberly A Dill-McFarland, kadm@uw.edu
Bruno B Andrade, Bruno.andrade@fiocruz.br
Timothy R Sterling, timothy.sterling@vumc.org
Thomas R Hawn, thawn@uw.edu

ABSTRACT

Although genetic factors contribute to tuberculosis (TB) risk, no cross-population causal variants have been identified by genome-wide association studies (GWAS). Here, we utilized low-pass whole genome sequencing (lpWGS) with imputation plus detailed epidemiologic risk factors and single-cell expression quantitative loci (sceQTL) to address prior GWAS limitations. Using 947 pulmonary tuberculosis (PTB) cases and 1807 close contact controls in the Regional Prospective Observational Research in TB (RePORT) study in Brazil, we estimated PTB heritability to be 47.7%. We identified 19 SNPs associated with PTB (P<5E-8) after adjustment for major risk factors (HIV, diabetes, smoking). Seven of these SNPs were associated with peripheral blood cell-specific sceQTLs in controls. Specifically, SNPs cis to transcription factors ZNF717 and MAML3 were associated with PTB disease and gene expression in monocytes, T cells, or B cells. Overall, this study utilized lpWGS, in-depth epidemiology, and single-cell analyses to detect population-specific genetic risk factors for PTB in Brazil.

INTRODUCTION

Multiple studies over the past 50 years suggest that host genetics regulate susceptibility to tuberculosis (TB), yet the major causal genes remain largely unknown [1,2]. This gap in knowledge regarding "missing heritability" is exemplified by the lack of reproducible findings among genome-wide association studies (GWAS) of TB in several countries [3–12]. The reasons for missing heritability in TB GWAS are multifactorial. Differences in underlying population genetics, population-specific genetic risk factors, genotyping platforms, imputation reference panels, *Mycobacterium tuberculosis* (Mtb) strains, and a lack of adjustment for clinical TB risk factors

may hinder the reproducibility of TB GWAS results. In addition, prior GWAS often lack direct experimental connections of associated genetic variants to functional outcomes that may impact TB outcomes. Designing studies to address these limitations may foster the discovery of the missing genetic regulators of susceptibility to TB.

Genotyping methods for GWAS studies continue to evolve to better measure high-confidence genetic variants within and across populations. High-pass whole genome sequencing (hpWGS) with at least 30X coverage is the gold-standard to comprehensively define genetic variation that may be population- and/or disease-specific. However, due to budgetary constraints, alternative methods are more often employed. Single nucleotide polymorphism (SNP) arrays with or without imputation are the most common genotyping platforms. These platforms were used in all prior TB GWAS, but they rely on pre-selected SNPs based on previously genetically characterized populations. Low-pass WGS (lpWGS, ≤ 1X coverage), along with recent advances in imputation, offers a cost-effective alternative to capture more population- and disease-specific variation compared to SNP arrays [13]. In addition, while the discovery of statistically significant SNPs remains the primary goal of GWAS, evidence of the functional effects of these SNPs is needed to fully elucidate their role in TB pathogenesis. In some studies, database-derived expression quantitative trail loci (eQTL) provide evidence of SNP function, but current transcriptomic databases are limited by the small number of cell types and/or human populations available for comparison to GWAS studies. Single-cell eQTL data in the same population as a GWAS could address this gap and expand the functional assessment of candidate GWAS SNPs.

In addition, precise phenotyping of traits is an essential feature of genetic studies. For TB cases, genetic signals may be obscured by other strong TB risk factors including HIV, diabetes mellitus (DM), smoking, alcohol consumption, malnutrition, and others. In addition, control misclassification can occur if there is insufficient Mtb exposure and/or follow-up duration among controls. Previous TB GWAS mainly focused on pulmonary TB (PTB) in HIV-negative populations with adjustments for age, sex, and ancestry but not other risk factors such as DM and smoking [3–12]. The method to control for potential confounders is most often a matched case-control design while an alternative full cohort, covariate-adjusted method is underexplored despite providing more statistical power.

To address some of the methodologic limitations of prior TB GWAS, we prospectively enrolled PTB cases and their close contact in the Regional Prospective Observational Research in TB (RePORT) study in Brazil, genotyped SNPs by lpWGS with imputation, and performed both full cohort covariate-adjusted and case-control analyses. We also assessed the functional relevance of candidate SNPs through eQTL analysis in a peripheral blood mononuclear cell (PBMC) single-cell RNA-seq substudy in RePORT-Brazil.

METHODS

Cohort. From 2015 to 2019, culture-confirmed adult pulmonary tuberculosis (PTB) patients were recruited at study clinics in Rio de Janeiro, Manaus, and Salvador as part of the Regional Prospective Observational Research in TB (RePORT) study in Brazil. Full details were reported

previously [14]. Briefly, PTB cases were confirmed by positive *Mycobacterium tuberculosis* sputum culture at enrollment (N = 1,060, **Figure S1**). Close contacts of PTB patients were then recruited using cumulative density sampling under the criteria of > 4 hrs/wk of contact with the index PTB cases within the past 6 months (N = 1,839). Controls were defined as close contacts with normal chest radiograph at enrollment and no symptoms for 24 months of follow-up. Additional radiograph tests and sputum were only performed when contacts presented with symptoms during the course of the study. Close contacts who progressed to active PTB during this time were classified as progressors (N = 25) and excluded from the GWAS. This study was approved by the institutional review boards of the study sites in Rio de Janiero (Instituto Nacional de Infectologia Evandro Chagas, Fundação Oswaldo Cruz [FIOCRUZ], Faculdade de Medicina e Centro Municipal de Saúde de Duque de Caxias, Universidade Federal do Rio De Janeiro, Secretaria Municipal de Saúde do Rio De Janeiro e Clínica Da Família Rinaldo Delamare), Salvador (Instituto Brasileiro Para Investigação da Tuberculose [IBIT], Fundação José Silveira), and Manaus (Fundação de Medicina Tropical Dr. Heitor Vieira Dourado) as well as Vanderbilt University Medical Center. Written informed consent was obtained from all participants.

Case-control subset. Among the 2,754 pass-filter individuals (see below), 819 individuals (241 controls, 578 PTB cases) were included in a case-control subset in which all were HIV-seronegative and without diabetes mellitus (DM). DM was self-reported for controls and defined by HbA1C at enrollment < 6.5 for cases. In addition, controls were IGRA-positive at enrollment, received < 30 days of any treatment for tuberculosis infection during the course of

130 the study, and had ≥ 12 months of follow-up during which they did not develop active PTB.
131 Treatments included isoniazid or rifampin monotherapy.
132
133 **Sample collection and DNA library preparation.** At enrollment, whole blood was collected in
134 EDTA tubes, aliquoted, and immediately frozen at −80 °C. DNA was extracted from 100 µl of
135 whole blood using the DNeasy Blood and Tissue Kit (Qiagen) following the manufacturer's
136 protocol with an additional final centrifugation to maximize DNA recovery. DNA was quantified
137 using Qubit Fluorometer (Thermo Fisher) and aliquoted in 96 well plates to 10 ul at 100 ng/ul.
138 Libraries were prepared using the Twist Biosciences kit (P/N: 104207) following a miniaturized
139 version of the manufacturers protocol and quantified using Qubit. Pool quality was assessed
140 using the Agilent Bioanalyzer and quantified using a qPCR-based method with the KAPA Library
141 Quantification Kit (P/N: KK4873) and the QuantStudio 12K instrument. Prepared DNA library
142 pools were pooled for sequencing in equimolar ratios based on qPCR.
143
144 **Low-pass sequencing and genotype calls.** In total, 2791 participants were sequenced using
145 Illumina low-pass whole genome sequencing (lpWGS) at 1-2X coverage at Vanderbilt University
146 Medical Center Technologies for Advanced Genomics (VANTAGE). DNA library pools were
147 subjected to cluster generation using the NovaSeq 6000 System, following the manufacturer's
148 protocols. Sequencing was performing using 150 bp paired-end reads on the NovaSeq 6000
149 platform targeting 10 million reads per sample (1-2X genomic coverage). Raw sequencing data
150 (FASTQ) was subjected to quality control analysis, including read quality assessment. Real Time
151 Analysis Software (RTA) and NovaSeq Control Software (NCS) (1.8.0; Illumina) were used for

base calling. MultiQC (v1.7; Illumina) was used for data quality assessments. Single nucleotide polymorphisms (SNPs) were called using Illumina DRAGEN Germline v4.0.3 with default low-pass parameters in Illumina BaseSpace. Genotypes were then imputed using Illumina DRAGEN Imputation v4.0.3, which implemented GLIMPSE [13] against the 1000 Genomes Project Phase 3 [15] IRPv1 autosomal SNP reference panel (N = 2,489). Imputation resulted in 49.5 million autosomal SNPs, of which 31.1 million (63%) had an imputation score > 0.6. Genotypes were further processed using PLINK2 [16,17] and R v4.4.0 [18]. SNPs on X and Y chromosomes were not included in the analysis. In addition, 3 samples (2 cases, 1 contact) failed quality controls prior to imputation and were not included in analyses.

Ancestry and kinship. Imputed genotypes were filtered by imputation score > 0.6, minor allele frequency (MAF) > 1%, and sliding window linkage disequilibrium (LD) with a 500 kb window, 1 kb slide, and maximum r^2 of 0.2 using PLINK2. With the resulting 12.6 million genotypes, identity by descent was calculated using the Robust KING method [19] in SNPRelate [20] and ancestry principal components (PCs) were estimated using PC-AiR [19,21] in GENESIS [22]. Kinship was then calculated using PC-Relate [23] with 2 ancestry PCs and a training set of the 1,554 unrelated individuals defined by PC-AiR. Sample pairs with high relatedness (kinship > 0.9) were further assessed and determined to be twins (N = 5 pairs), siblings (N = 1), duplicate samples (N = 5), or unrelated/unknown (N = 2). The first chronological sample was retained for duplicates, and high-kinship unrelated/unknown pairs were removed from analysis. This resulted in 2,754 individuals for GWAS analyses (1,807 controls, 947 PTB cases).

Pass filter samples were compared to 1000 Genomes Project Phase 3 [15] (N = 2,548) using principal component analysis (PCA). RePORT imputed genotypes were filtered to imputation score > 0.6, MAF > 5%, Hardy-Weinberg Equilibrium (HWE) P > 1E-6, 0% missingness, and > 60% concordance with high-pass WGS (see *Genotype validation*). Overlapping filtered RePORT and 1000 Genomes genotypes were then filtered to remove AT and CG SNPs, which are more prone to sequencing error, and LD filtered using a sliding window of 100 kb, 1 kb slide, and r^2 threshold of 0.1. Finally, genotypes were filtered to MAF > 10% in the combined dataset. Using these 96,426 SNPs, PCs were calculated using PLINK2, and RePORT participant self-reported race was compared to 1000 Genomes Super Populations.

GWAS model fitting. From the 12.6 million genotypes used for kinship, 10,000 random positions were selected for model fitting. Mixed effects logistic regression was performed to test genotype associations with PTB cases (1) vs close-contact controls (0) (**Table S1**). Participants predominantly identified as brown (57%), black (22%), or white (20%), which most closely corresponded to 1000 Genomes [15] admixed American, African, and European Super Population ancestry groups, respectively, though significant admixing was apparent (**Figure S2A,B**). Ancestry PCs were progressively added to the base GWAS model and assessed for total variation explained as well as model fit by Akaike Information Criterion (AIC). A covariate was retained if it significantly improved model fit in its univariate model and its removal worsened model fit in its leave-one-out model ($|\Delta AIC| > 8$). The first two ancestry PCs were included in the final model as they explained the most variation (PC $\geq$ 1%) (**Figure S2C**) and improved model fit for the majority of SNPs as indicated by reduced AIC compared to models with one

fewer PCs (**Figure S2D**). Next, kinship was added to the 2 PC logistic model as a pairwise covariance random effect and tested with GMMAT [24]. While kinship did not improve model fit (**Figure S2E**), the cohort contained 371 first-degree relationships (kinship > 0.5), including several twin pairs (**Figure S2F**). Thus, kinship was retained in the model to account for this relatedness.

Finally, additional clinical covariates age and sex as well as PTB risk covariates smoking, HIV status, and diabetes mellitus (DM) (self-reported for controls, enrollment HbA1C > 6.5 for cases) were assessed. Each covariate was assessed in univariate models (e.g. 2PC kinship model + one covariate, **Figure S2G**) and leave-one-out models (e.g. 2 PC kinship model + all covariates except one, **Figure S2H**). Smoking was classified as past/current vs never as there were no significant differences between past and current smokers in a univariate model (**Table S1**, P > 0.3). All covariates except age significantly improved model fit when added in univariate models and worsened model fit when removed in leave-one-out models (**Table S1**). Age initially worsened model fit when added to the simpler model but had no significant effect when removed from the more complex model (|ΔAIC| < 8). Thus, age was retained in the final model to facilitate comparisons to previous GWAS studies that commonly corrected for age. This resulted in the full cohort model

$$(1)\ logit(\mathrm{PTB} = 1|X, u) = \alpha + \beta_1 \mathrm{SNP}_i + \beta_2 \mathrm{PC1} + \beta_3 \mathrm{PC2} + \beta_4 \mathrm{sex} + \beta_5 \mathrm{age} + \beta_6 \mathrm{smoke} + \beta_7 \mathrm{HIV} + \beta_8 \mathrm{DM} + u,$$

$$u \sim \mathcal{N}\left(0, \sigma_g^2 K\right)$$

where logit(P(TB=1|X, u)) is the log-odds of PTB disease given the covariates X and random effects u, α is the intercept, β are the coefficients of main effects (SNP) and covariates (ancestry PCs, clinical covariates, PTB risk covariates), and u is the random effects with covariance structure determined by genetic kinship (K). The case-control cohort was modeled for a comparable model removing PTB risk covariates (HIV, DM) that were absent in this subset.

GWAS statistics. The full dataset imputed 49.5 million genotypes were filtered for imputation score > 0.6, MAF > 5%, HWE P > 1E-6 (calculated on controls only), and 0% missingness. This resulted in 6,944,453 genotype positions for analysis. The full cohort was modeled with adjustment for ancestry, kinship, clinical, and PTB risk covariates as in equation (1), and the case-control subset with this model without HIV and DM as person living with HIV and/or DM were excluded from this analysis. Both cohorts were additionally analyzed with a simplified model without all three PTB risk covariates (smoking, HIV, DM) to compare with previous GWAS studies.

The rationale for the full cohort analysis was to maximize sample size and potential statistical power while adjusting for covariates that could decrease precision without fully addressing confounding bias. The rationale for the case-control subset was to optimize case and control selection to exclude variability introduced by risk factors known to be associated with PTB disease (e.g. HIV, DM). The case-control design helped improve statistical efficiency for estimating the causal effect within and across strata of balanced characteristics (relying on discordant pairs to be informative) and more powerfully limited the confounding bias for

important *a priori*-identified confounders. Though the case-control design lent itself to estimating a causal effect less prone to bias and more likely to be internally valid, statistical power was reduced relative to the full cohort sample size. The merits of the full cohort analysis (improved power to detect small effects) and case-control design (improved internal validity to detect unbiased causal effects) are, therefore, complementary instead of redundant.

All genotypes were scored using GMMAT (glmm.score) [24] in the full and case-control cohorts separately. Genotypes with a score P < 1E-4 were then fit in the full GLMMkin model. In addition, all full cohort selected genotypes were also run in the case-control subset. Significant SNPs were defined at P < 5E-8 and suggestive SNPs at P < 5E-5. Lower MAF genotypes (1 - 5%) were then selected within 10 kb of significant SNPs and tested with the same models described above as well as assessed for linkage disequilibrium (LD) with significant SNPs in the region. Significant SNPs were also tested for associations with clinical and PTB risk covariates by modeling sex, age, smoking, HIV, or DM as outcomes in place of PTB status while retaining all other covariates. Finally, significant SNPs were cis-mapped to genes within 50 kb. Gene enrichment was performed against Broad MSigDB [25] Hallmark, canonical pathways, and gene ontology using iterative hypergeometric enrichment in SEARchways [26] to account for SNPs with multiple gene annotations.

The overall heritability of PTB progression was calculated using Genome-wide Complex Trait Analysis (GCTA [27]) with a genetic relationship matrix (GRM) and restricted maximum likelihood (REML [28]) calculated from concordant SNPs in 1020 unrelated individuals from the full cohort

(594 controls, 426 cases, GRM < 0.05). Prevalence was estimated as the mean PTB prevalence during the recruitment period (2025-2019) in RePORT participating Brazilian states (Amazonas, Bahia, Rio de Janeiro) according to the Pan American Health Organization (57.07 cases per 100,000 population) [29]. Models were corrected for covariates as in the full cohort (ancestry PCs, sex, age, smoking, HIV, DM) as well as without PTB risk covariates for comparison.

Genotype validation by high-pass sequencing. To validate imputed genotypes, a random subset of individuals (25 controls, 25 cases) was re-sequenced with 30X genome coverage at VANTAGE. DNA library pools were subjected to cluster generation using the NovaSeq X Plus System, following the manufacturer's protocols. Sequencing was performed using 150 bp paired-end sequencing on the NovaSeq X Plus platform targeting 350 - 400 million reads per sample (30X genomic coverage). Raw sequencing data (FASTQ) was quality assessed, and bases were called as for low-pass sequencing. Sequences were then cleaned using SEAsnake [30] (maximum ambiguous bases = 1, minimum length = 15, minimum quality = 30, 5-prime trimming = 10 bp, adapters = Illumina universal). Sequences were aligned with STAR (2.7.9a)[31] against the human genome GRCh38, release 112 and SNPs were called with bcftools mpileup v1.14 [32] with a minimum base quality of 20 and a minimum mapping quality of 30.

Significant GWAS SNPs were also compared to SNP manifests for 19 arrays used in prior TB GWAS studies. UCSC LiftOver [33] was used to convert this study's hg38 genomic locations to older genome versions (hg37, hg36, hg35) for comparisons. Only 6 of the 19 significant

concordant SNPs were present on one or more arrays. The most overlap was found with the custom LIMAArray (4 SNPs), which was designed for a Peruvian population [3].

Single-cell RNA sequencing. Single-cell RNA-sequencing (scRNAseq) was performed on peripheral blood mononuclear cells (PBMC) from a subset of close contacts. This included 25 controls who progressed to PTB and 33 sex-, age- (+/- 5 yrs), and interferon gamma release assay (IGRA)-matched controls from RePORT-Brazil. PBMC were thawed, washed twice with phosphate buffered saline (PBS), and resuspended at 10E6 cells/100 ml in RPMI (MT10040CV from Mediatech/Corning) with 10% human serum. Cells were rested 2 hours, CD40 blocking antibody was added for 15 minutes (1:200 dilution, 0.5 ul), and then CD40L (diluted 1:10) was added. Cells were incubated overnight, washed with PBS, and then washed with cell staining buffer. Cells were resuspended in 50 ml of staining buffer and 5ml of the Fc blocking agent Human true stain FcX (cat#422302, Biolegend) and placed on ice for 10 minutes. Cells were hash-tagged using one microliter of TotalSeq-C Human Universal Cocktail (Biolegend) per sample and incubated on ice for a further 30 minutes. Cells were then washed three times and resuspended in staining buffer. A 137 Cite-seq C antibody (cat#399905, Biolegend) panel was prepared as per the manufacturer's guidelines. For pools of 4-5 samples, 50ml of the antibody panel was added, and cells were incubated on ice for 30 minutes. Cells were washed three times with PBS with 2% BSA and resuspended in PBS with 1% BSA. Approximately 8,000 cells per individual were provided to the VANTAGE sequencing core. Libraries were created following the 10X Genomics standard protocol (3-prime version 4), and sequencing was performed on an Illumina NovaSeq 6000 at VANTAGE.

Raw sequences (FASTQ) were processed using 10x Genomics Cell Ranger Multi v9.0.0 using 10x Genomics Cloud Analysis [34]. Poor-quality cells, doublets, and high mitochondrial cells (> 10%) were removed. Genes not expressed in at least 150 cells were removed. Imputation was applied to the concatenated count matrix using Adaptively-thresholded Low Rank Approximation (ALRA) [35] in R v4.3.3 [18]. Further data cleaning and annotation was performed using python v3.10.16. A consensus set of highly variable genes (HVG, n = 5000) were calculated on the imputed counts across project batches. Principal component analysis (PCA, n = 30) was done. The PCA result was submitted to the Harmony algorithm for integration (theta = 2) [36]. K-nearest neighbors (kNN, n = 30) were found using the Harmony components. Leiden clustering (resolution = 3) was used to partition the kNN graph. Cell clusters were split into three groups: T and natural killer (NK) cells, myeloid cells, and B cells. Each subset was then analyzed separately by HVG, PCA, Harmony, kNN, and Leiden clustering. Uniform Manifold Approximation and Projection (UMAP) was used for visualization. Leiden clusters were annotated according to canonical marker expression profiles. Subset-specific HVG n-genes were set as follows: T/NK (n = 4000), myeloid (n = 2500), and B (n = 2000). Harmony theta was set as follows: T/NK (n = 4), myeloid (n = 4), and B (n = 2). Leiden resolution was set as follows: T/NK = 2.5, myeloid = 1, and B = 1.

Single-cell expression quantitative loci (sceQTL). In total, 58 participants with both scRNAseq and GWAS data were selected. A subset of participants (N = 25) were progressors to PTB and not included in the original GWAS; thus, ancestry PCs and kinship were re-calculated for this

combined subset as described above. Kinship values were highly correlated across controls processed in this subset versus the full cohort (Pearson R = 0.998, P < 2.2E-16). sceQTL analysis was performed in R v4.4.2 [18]. Single-cell data (543,229 cells) were pseudo-bulked using Seurat v5.1.0 [37], and the 32 cell types were processed independently. Counts were filtered to protein-coding genes, those with expression in at least 3 samples, trimmed mean of means (TMM) normalized, and transformed to log2 counts per million (CPM) with gene-level quality weights using edgeR [38] and limma [39,40]. This resulted in 7,297 to 13,029 genes for analysis per cell type. SNP data were filtered to overlapping single-cell participants and GWAS-significant SNPs that were annotated to at least one protein-coding gene, concordant with high-pass data, and expressed in at least one cell type in single-cell data. This yielded 9 SNPs annotated to 8 protein-coding genes, which resulted in 12 unique gene-SNP pairs. These pairs were modeled against gene expression in the 32 single-cell RNAseq cell types using a kimma [41] additive model. Covariate inclusion was assessed in univariate and leave-one-out models as with GWAS model fitting (**Table S2**), which resulted in a final sceQTL model without clinical or PTB risk covariates (**Figure S3**).

$$(2)\ expression = \alpha + \beta_1 \mathrm{SNP}_i + \beta_2 \mathrm{PC1} + \beta_3 \mathrm{PC2} + u,$$

$$u \sim \mathcal{N}\left(0, \sigma_g^2 K\right)$$

where components are as described in equation (1), except expression is the pseudo-bulk gene expression per individual within cell type. FDR correction was applied within cell type across SNPs annotated to the same gene. Significance was determined at FDR < 0.05. Sensitivity analyses were performed including covariates relevant to GWAS models (sex, age, smoking, DM

as in equation (1)). Dominant models were also assessed for SNPs with a measured but rare homozygous uncommon genotype (RGS9).

RESULTS

GWAS study design.

To discover genetic determinants of pulmonary tuberculosis (PTB) susceptibility in Brazil, we examined genotypes by low-pass whole genome sequencing (lpWGS) with imputation (N = 6,944,453 SNPs with minor allele frequency [MAF] > 5%) in culture-confirmed PTB cases and close contact controls from the RePORT-Brazil cohort (**Figure 1A, Figure S1**). We performed two complementary analyses: a full cohort risk-adjusted design (N = 947 cases, 1807 controls) to maximize statistical power to detect smaller effects and a case-control strategy (N = 578 cases, 241 controls) to reduce PTB risk factor variance in order to detect less biased causal effects. When compared to controls in the full cohort (**Table 1**), PTB cases had a higher male frequency (67% cases vs 41% controls, P = 1E-37), were older (38 ± 14 vs 33 ± 19, P = 2E-14), had lower BMI (21 ± 3.5 vs 25 ± 6.5, P = 7E-71), and had higher prevalence of tobacco use (52% vs 27%, P = 8E-40), HIV (20% vs 2%, P = 1E-53), and DM (23% vs 5%, P = 5E-45). Among PLWH, PTB cases had higher usage of ART (41% vs 0%) and lower CD4 counts (218 ± 225 vs 467 ± 340, P = 1E-06) than controls. Also, 30% of controls received TB infection (TBI) treatment while on study (190 ± 84 days). The case-control subset fully controlled for HIV and DM (0%) and was more balanced for age (P = 0.07) and TBI treatment (maximum 30 days in controls) (**Table 1**).

GWAS SNP-level primary results with high-pass WGS confirmation.

366 Previous TB GWAS studies commonly corrected for age, sex, and genetic relatedness (ancestry
367 PCs and/or kinship). To assess whether to include additional covariates, we tested GWAS
368 models on a random sample of 10,000 SNPs in the full cohort for improved fit as measured by
369 reductions in the Akaike information criterion (AIC). We selected a full cohort model adjusting
370 for ancestry, sex, age, smoking, HIV, DM, and kinship (equation (1)). The case-control subset
371 was modeled for a comparable model removing PTB risk covariates (HIV, DM) that were absent
372 in this subset. The full cohort and case-control analyses had an expected quantile-quantile (QQ)
373 trend with the majority of SNPs following a one-to-one ratio and significant SNPs deviating from
374 this trend with low genomic inflation factors (λ = 0.985 and 0.994, respectively) (**Figure S4**). In
375 the full cohort model, 68 low-pass WGS genotyped SNPs were significantly associated with PTB
376 disease (P < 5E-08) and an additional 686 SNPs were suggestive (P < 5E-05) (**Table S1**). In the
377 case-control subset, no SNPs reached significance but 164 were suggestive (P < 5E-05) (**Table**
378 **S1**).
379
380 A random subset of 50 participants (25 PTB cases and 25 controls) were assessed using 30X
381 high-pass whole genome sequencing (hpWGS) to confirm lpWGS imputation results. Selected
382 individuals represented the overall genetic diversity of the cohort as seen in PCA (**Figure S5A**).
383 Of the 49.5 million imputed lpWGS SNPs, 31% were called in at least 1 participant in the high-
384 pass subset (**Figure S5B**). The majority of lpWGS SNPs that were called in hpWGS were high-
385 quality, including > 90% (28% of all SNPs) with imputation scores > 0.6. In contrast, nearly half
386 of lpWGS SNPs that were not called in hpWGS failed imputation quality filtering (score < 0.6),
387 indicating poor-quality in both data sets. Among the 15.5 million lpWGS SNPs called in all 50

high-pass participants, concordance was 74 ± 26% between high-pass and imputed low-pass data (**Figure S5C**, **Table S3**), which is slightly below the GLIMPSE2 reported values for 1X imputed data (r = 0.8) [42]. Among imputation and MAF pass-filter GWAS-significant SNPs (N = 68), there was lower coverage (50%) (**Figure S5D**) and lower concordance (57 ± 29%) compared to the high pass data (**Figure S5E**). Specifically among significant low-pass SNPs, 34 were not called in high-pass data, 19 had concordance ≥ 60%, and 15 had concordance < 60% (**Figure S5F**, **Table S3**). Significant SNPs with concordance ≥ 60% were used in further analyses. In the full cohort, 19 concordant SNPs were significant (P < 5E-08) and an additional 238 were suggestive (P < 5E-05) in GWAS models (**Figure 1B**, **Table 2**). The case-control subset had 64 suggestive concordant SNPs, none of which overlapped with full cohort significant or suggestive SNPs (**Figure 1B**, **Table 2**). Significant SNPs were spread across the genome, including 11 different chromosomes (**Figure 1C, D**).

Utilizing the 2,527,008 concordant SNPs across the entire lpWGS data set and a subset of 1020 unrelated individuals, we estimated heritability by Genome-wide Complex Trait Analysis (GCTA). GCTA found that 47.7 ± 23.8 s.e. % of the variance in PTB disease in this cohort can be explained by genetic factors when corrected for ancestry, age, sex, and PTB risk covariates (**Table S4**). Given the low rate of PTB in the population relative to participants used in the GCTA analysis (0.06 vs 42%), the scaled heritability was estimated to be 15.8 ± 7.9%. When PTB risk covariates were not incorporated (smoking, HIV, DM), heritability estimates were lower at 18.2 ± 22.5% overall and 7.3 ± 7.5% scaled for population prevalence. Together, these data suggest PTB progression in Brazil is partially genetically regulated with high heritability and that 19 concordant SNPs were associated with PTB in the full cohort analysis.

Impact of clinical covariate correction.

To assess the impact of PTB risk co-variate inclusion in our study design, we performed analyses with a "standard" GWAS model including age, sex, ancestry PCs, and kinship as well as targeted PTB models lacking one of the PTB risk covariates included in the final full cohort model (HIV, DM, smoking). While P-values were correlated between the model with all PTB risk covariates and the "standard" model with no PTB risk covariates (Pearson R = 0.83, P < 2.2E-16), the "standard" model identified only 3 significant SNPs, which fully overlapped with the full model's results (P < 5E-08, **Figure 3A**). Thus, accounting for PTB risk factors yielded 16 more significant SNPs without loss of any signal found with a "standard" GWAS model. Among the 19 concordant SNPs significant in the full cohort final model, the majority were detected when at least 2 out of 3 PTB risks were accounted for (**Figure 3B**). Notably, the addition of HIV and/or DM resulted in the greatest increases in significant SNPs detected. Smoking had less impact on significant SNP detection, though 2 SNPs were identified only when all three risk covariates were included. Thus, correction for PTB risk covariates improved model fit (**Figure S2G,H**) and retained statistical power to detect associations with PTB in the full cohort (**Figure 3**).

SNP functional and genomic annotation.

We next mapped the significant SNPs to assess their genomic and functional features. We first assessed LD patterns surrounding significant lead SNPs. We assessed lower frequency SNPs (> 1% MAF) within 10 kb of significant concordant GWAS SNPs for linkage disequilibrium (LD R^2) with the lead SNP (**Table S1**). ZNF717 was the only SNP with surrounding SNPs with high LD and

low P-values (**Figure 2**). In contrast, the remaining significant SNPs did not have clustering due to limited concordant calls and/or low LD in the region.

The majority of significant, concordant SNPs (17 of 19 SNPs) were annotated within genes or within 50 kb cis of genes, including protein-coding and RNA genes (lncRNA, snRNA, snoRNA) as well as pseudogenes (**Figure 1E**, **Table S5**). In total, 5 significant concordant SNPs mapped to 5 protein-coding gene transcripts and 7 SNPs were 50 kb cis to 6 protein-coding genes (**Table 2**). Among these 11 genes, there was minimal gene set enrichment within the Hallmark, canonical pathways, or gene ontology databases (FDR < 0.3, overlap > 1, **Table S6**). These included Reactome RNA polymerase II transcription (MAML3, ZNF717) and Wikipathways VEGFA-VEGFR2 signaling pathway (ATF4, CGNL1, CYP2C8). None of the 12 protein-coding concordant SNPs were significant predictors of a sex, age, smoking, HIV, or DM (**Table S1**), further supporting their association with PTB after correction for these clinical covariates. Further analyses focus on the 12 protein-coding region SNPs to enrich for potentially causative relationships to PTB disease.

Single-cell eQTL analysis.

To assess whether significant concordant SNPs were associated with cis gene expression, we examined a PBMC single-cell RNAseq dataset from the same cohort. SNPs significantly associated with PTB disease (P < 5e-08), concordant in high-pass data (> 60%), and annotated to protein-coding genes (N = 12) were modeled against expression of their annotated genes. In total, 58 participants were present in both datasets; this subset was 36 ± 18 years old, 34%

male (N = 20), and had similar PTB risk factors as the larger cohort including 25 ± 6.6 BMI, 10% DM (N = 6), 26% tobacco users (N = 15), and no persons living with HIV (PLWH). In total, 9 GWAS-significant concordant SNPs were mapped to 8 protein-coding genes expressed in at least one cell type in single-cell RNA-seq. CGNL1, CYP2C8, and MPPED1 were not expressed in any cell type; thus, their associated SNPs were not modeled in sceQTL analysis. Among the 32 annotated cell types (**Figure S6A**), all expressed at least one gene of interest, and genes of interest were generally expressed in the majority of cell types with a wide range of expression levels (**Figure S6B**). The exception to this was CACNA1I (calcium voltage-gated channel subunit alpha), which were expressed in only 12 cell types.

Overall, 9 unique SNP-gene-cell models were significant eQTLs (FDR < 0.05, **Table S2**). These represented 7 SNPs, 4 genes, and 7 cell types (**Figure 4A**). Specifically, SNPs associated with a higher risk of PTB in the clinical GWAS were associated with lower ZNF717 (zinc finger protein) expression in classical and monocytes and naïve B cells (**Figure 4B**). For monocytes, these effects were seen for the lead SNP (3:75816280:A:T) as well as two additional SNPs in LD in this region (**Figure S7A**). Co-localization was observed for ZNF717 GWAS SNPs and sceQTLs in monocytes but less so in B cells, which may indicate indirect effects or a different causal variant in this cell type (**Figure S7A**). Among the ZNF717-significant SNPs, only one (3:75816219:T:G) was present on any previously employed SNP arrays (Illumina MEGA^{EX} , **Table S7**). In addition, an intronic COX10 (cytochrome C oxidase assembly factor) SNP was significantly associated with higher gene expression in B cells expressing IFN stimulated genes (isg) (**Figure 4C**). Despite sparse genotyping in this region, a more highly significant SNP in LD with several suggestive

SNPs was found upstream (**Figure S7B**), thus indicating that the GWAS-significant SNP may not be causally associated with gene expression. An intronic RGS9 (regulator of G protein signaling) SNP was associated with higher gene expression in CD56-bright natural killer cells (**Figure 4D**) with some GWAS colocalization but no LD with the sceQTL SNP (**Figure S7C**). This effect was present in either an additive (FDR = 0.035) or dominant model (FDR = 0.025), thus ensuring that the effect was not driven by the single homozygous uncommon individual. Finally, an intronic MAML3 (mastermind-like transcriptional coactivator) SNP previously identified on the LIMAArray (**Table S7**) significantly associated with lower expression of this gene in stressed myeloid cells and higher expression in CD8+ and other T cells (**Figure 4E**). Minimal nearby SNPs were successfully genotyped in this region, which resulted in no apparent LD structure though some co-localization was apparent due to the closest surrounding SNPs not significantly associated with gene expression (**Figure S7D**). Together, these expression studies suggest that several SNPs were cell-specific eQTLs with directionality that informs biological models of PTB pathogenesis.

DISCUSSION

We used a genome wide association study (GWAS) to discover genetic variants associated with pulmonary tuberculosis (PTB) in Brazil with three important study design features: inclusion of extended epidemiologic risk factors in case-control definitions, genotyping with imputed low-pass whole genome sequencing (lpWGS) to capture more comprehensive population-specific variants, and the assessment of the functional significance of candidate genetic variants with single-cell eQTL (sceQTL) analysis in the same cohort. We discovered several genetic variants

associated with PTB with genome-wide level significance and associated function in cell-specific sceQTLs.

To gain insight into genetic regulation of susceptibility to PTB, we examined high-confidence imputed lpWGS SNPs that were associated with PTB disease, concordant in high-pass WGS (hpWGS), mapped to protein-coding genes, and expressed in a matched single-cell RNA-seq experiment. Notably, zinc finger protein 717 (ZNF717) SNPs were associated with gene expression in monocytes or B cells. This region was the only one to display a characteristic GWAS "build-up" of signal among SNPs with lower P-values and high linkage disequilibrium (LD). Importantly, only one of these SNPs (rs113392015) was present on any commonly used SNP array. ZNF717 encodes a Kruppel-associated box (KRAB) zinc-finger protein, which binds DNA and has predicted roles in regulation of transcription, but no published experimental studies of its function. Here, multiple ZNF717 SNPs were sceQTLs associated with decreased expression in classical monotypes while a single SNP was associated with increased expression in naïve and interferon stimulated gene (ISG) expressing B cells. Together, these data suggest that genetic regulation of ZNF717 across multiple loci impacts corresponding gene expression and is associated with increased susceptibility to PTB.

Additionally, an intronic mastermind-like transcriptional coactivator 3 (MAML3) SNP was associated with PTB risk and gene expression in multiple cell populations. The MAML3 SNP was associated with decreased expression of MAML3 expression in stressed myeloid cells but increased expression in CD8+ cytotoxic and other T cells. MAML3 is a transcription coactivator

520 involved in positive regulation of transcription, particularly Notch signaling. Notch pathways
521 have been implicated in PTB disease including T cell cytokine production [43], Th1/Th2 balance [44],
522 and apoptosis in response to *Mycobacterium tuberculosis* (Mtb) antigen [45]. Thus, genetic
523 regulation of Notch through MAML3 may contribute to PTB risk. Also, an intronic Regulator of G
524 protein signaling 9 (RGS9) SNP was associated with increased PTB risk and RGS9 gene
525 expression in CD56-bright natural killer (NK) cells. While the canonical role of RGS9 is regulation
526 of neuronal G protein signaling pathways [46,47], it has been linked to a tissue resident-like
527 phenotype in NK cells [48]. In particular, CD56-bright NK cells are characterized by higher
528 interferon-gamma (IFNG), tumor necrosis factor-beta (TNFB), interleukin-10 (IL10), and IL12
529 production in response to monocyte-derived cytokines when compared to their CD56-dim
530 counterparts [49]. Thus, these NK-derived cytokines may provide an important innate immune
531 response to PTB downstream of primary monocyte infection. These data suggest a possible
532 genetic regulator of the tissue resident-like NK phenotype through RGS9. Overall, the
533 combination of GWAS and sceQTL datasets in the same cohort enables rapid screening of
534 GWAS SNPs for those most likely to impact PTB disease through transcriptomic outcomes. The
535 single-cell approach further facilitates detailed identification of the involved cell types as well as
536 directionality and magnitude of effects.
537
538 Our study had several limitations. First, although one distinguishing feature of this study was its
539 use of unbiased genome-wide genotyping with lpWGS, imputation discrepancies did exist
540 between lpWGS and hpWGS, even among SNPs that passed lpWGS imputation quality and
541 minor allele frequency (MAF) filtering. This may indicate genomic regions prone to low quality

542 sequencing and/or imputation that should be excluded in future studies. However, testing with
543 additional populations and reference data sets is necessary to confirm these regions,
544 particularly given the sparsity of Brazilian individuals in current references. Nevertheless, this
545 method yielded millions of SNPs not represented by current or past SNP arrays, and overall,
546 imputed lpWGS genotype calls were highly concordant with hpWGS. Second, this study cohort
547 displayed diverse ethnic and geographic backgrounds [50] as well as variability in a number of
548 known PTB risk factors. Traditionally, GWAS correct for age, sex, and ancestry as well as often
549 exclude individuals with known risk factors. Here, we have shown through statistical
550 assessment that exclusion was not necessary, and many factors can be accounted for as model
551 covariates. This preserved sample size, and thus statistical power, which would otherwise be
552 reduced in case-control designs. Specifically, PTB risk factors including HIV, DM, and smoking
553 improved model fit as well as increased PTB signal despite reduced degrees of freedom in the
554 more complex statistical model. The more controlled, smaller case-control study failed to
555 identify these statistically significant SNPs, likely as a result of reduced power. This underscores
556 the importance of these factors in PTB risk assessment and may contribute to the lack of
557 reproducibility among TB GWAS that did not account for similar risk factors. Third, the sample
558 size of this PTB GWAS did not allow for the detection of rare variants (MAF < 5%) and was
559 under-powered to detect SNPs with small effect sizes. Finally, despite advances in design and
560 methodology, this GWAS failed to identify PTB-associated SNPs of significance described in
561 prior GWAS, even SNPs present on prior SNP arrays. This is a common feature among TB GWAS
562 and may indicate an abundance of population-specific genetic risk factors. Mtb has co-evolved
563 with humans for thousands of years, and there is considerable Mtb strain variability worldwide

[51]. Thus, population-specific SNPs relevant to PTB disease risk are likely present, and unbiased genotyping across multiple populations will be necessary to identify population-spanning SNPs.

In summary, this PTB GWAS identified 12 protein-coding loci associated with PTB disease and linked seven of these SNPs to potential functional consequences through sceQTL analysis. These findings provide new insights into genetic risks of PTB in the highly admixed Brazilian population. While no overlap was found between this and prior TB GWAS, future studies employing lpWGS may yield population-spanning SNPs not previously measured across populations due to the limitations of prior SNP arrays. Thus, this study provides important groundwork for both lpWGS methods and the assessment of genetic risks for PTB.

DATA AVAILABILITY

Sequence files for this project are available in the database of Genotypes and Phenotypes (dbGaP) [in progress]. All scripts associated with this project can be found in https://github.com/hawn-lab/ReporTB_GWAS_public.

ACKNOWLEDGEMENTS

We thank the individual study participants of the Regional Prospective Observational Research in Tuberculosis (RePORT)-Brazil consortium, study coordinators, and the clinical and research staff at all sites.

FUNDING

This work was supported by NIH R01 AI147765 (BBA, TRS, TRH), R01 AI20790 (VCR, TRS), U01AI172064 (BBA, TRS), U01 AI069923, and CRDF Global DAA9-19-65383-1 (TRH). ALK, BBA, MCS, and VCR are fellows from the National Council for Scientific and Technological Development (CNPq), Brazil.

FINANCIAL DISCLOSURES

KADM reports consulting fees for bioinformatic tool development from Seattle BioSoftware and EuropaDX. All other authors report no competing financial interests.

AUTHOR CONTRIBUTIONS

Study design (KADM, BBA, SAK, PFR, GA, SAM, TRS, TRH), study supervision (BBA, SAK, TRS, TRH), program management (MCF), study enrollment and data collection (AMSA, MCS, ALK, VCR), sample collection and processing (AMSA, FAR, JMCA, RDG), data cleaning (KADM, JDS, RDG), single cell annotation (JDS, JMO, JPA, HIN), GWAS and sceQTL statistics (KADM), and primary manuscript writing (KADM, TRH). All authors reviewed and approved the final manuscript.

The Regional Prospective Observational Research in Tuberculosis (RePORT)-Brazil Consortium

Salvador, BA, Brazil

Laboratório de Pesquisa Clínica e Translacional, Instituto Gonçalo Moniz, Fundação Oswaldo Cruz (FIOCRUZ): Mariana Araújo-Pereira

Instituto de Pesquisa Clínica e Translacional, Medicina Zarns, Clariens Educação: Mariana Araújo-Pereira, Artur T L Queiroz

Instituto Brasileiro para Investigação da Tuberculose, Fundação José Silveira: Michael Rocha, Vanessa Nascimento, Victor Porfirio

Rio de Janeiro, RJ, Brazil

Programa Acadêmico de Tuberculose da Faculdade de Medicina, Universidade Federal do Rio de Janeiro: Adriana Moreira, André Luiz Bezerra, Anna Cristina Carvalho

Laboratório de Pesquisa Clínica em Micobacteriose, Instituto Nacional de Infectologia Evandro Chagas (FIOCRUZ): Adriano Gomes Silva, Aline Benjamin, Quezia Medeiros, Flavia M Sant'Anna

Secretaria Municipal de Saúde do Rio De Janeiro e Clínica Da Família Rinaldo Delamare: Solange Cavalcante, Betina Durovni, Jamile Garcia, João Marine

Manaus, AM, Brazil

Fundação de Medicina Tropical Dr. Heitor Vieira Dourado do Amazonas: Allyson G Costa, Brenda Carvalho, Alexandra Brito, Leandro Garcia, Renata Spener

Nashville, TN, USA

Division of Infectious Diseases, Department of Medicine, Vanderbilt University Medical Center:

Angela Jones, Hilary V Riley, Cody Staats, Megan Turner

REFERENCES

1. Rosain, J. *et al.* Mendelian susceptibility to mycobacterial disease: 2014–2018 update. *Immunol Cell Biol* **97**, 360–367 (2019).
2. Hill, A. V. S. The Genomics and Genetics of Human Infectious Disease Susceptibility. *Annu Rev Genomics Hum Genet* **2**, 373–400 (2001).
3. Luo, Y. *et al.* Early progression to active tuberculosis is a highly heritable trait driven by 3q23 in Peruvians. *Nat Commun* **10**, 3765 (2019).
4. Sveinbjornsson, G. *et al.* HLA class II sequence variants influence tuberculosis risk in populations of European ancestry. *Nat Genet* **48**, 318–322 (2016).
5. Png, E. *et al.* A genome wide association study of pulmonary tuberculosis susceptibility in Indonesians. *BMC Med Genet* **13**, 5 (2012).
6. Thye, T. *et al.* Common variants at 11p13 are associated with susceptibility to tuberculosis. *Nat Genet* **44**, 257–259 (2012).
7. Thye, T. *et al.* Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. *Nat Genet* **42**, 739–741 (2010).
8. Grant, A. V *et al.* A genome-wide association study of pulmonary tuberculosis in Morocco. *Hum Genet* **135**, 299–307 (2016).
9. 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).
10. Bai, H. *et al.* Genome-wide association study of tuberculosis in the western Chinese Han and Tibetan population. *MedComm (Beijing)* **4**, e250 (2023).
11. Zheng, R. *et al.* Genome-wide association study identifies two risk loci for tuberculosis in Han Chinese. *Nat Commun* **9**, 4072 (2018).
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. Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J. & Delaneau, O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat Genet* **53**, 120–126 (2021).
14. Arriaga, M. B. *et al.* Novel stepwise approach to assess representativeness of a large multicenter observational cohort of tuberculosis patients: The example of RePORT Brazil. *International Journal of Infectious Diseases* **103**, 110–118 (2021).
15. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
16. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559–575 (2007).
17. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, (2015).
18. R Core Team. R: A language and environment for statistical computing. Preprint at https://www.r-project.org/ (2023).
19. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
20. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).

21. Conomos, M. P., Miller, M. B. & Thornton, T. A. Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness. *Genet Epidemiol* **39**, 276–293 (2015).
22. Gogarten, S. M. *et al.* Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics* **35**, 5346–5348 (2019).
23. Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. Model-free Estimation of Recent Genetic Relatedness. *The American Journal of Human Genetics* **98**, 127–148 (2016).
24. Chen, H., Conomos, M. & Pham, D. GMMAT: Generalized Linear Mixed Model Association Tests. Preprint at (2023).
25. Dolgalev, I. msigdbr: MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format. Preprint at https://cran.r-project.org/package=msigdbr (2019).
26. Benson, B., Dill-McFarland, K. A., Cox, M. S. & Janczyk, T. SEARchways: gene set and pathway enrichment. Preprint at https://doi.org/10.5281/zenodo.13145502 (2024).
27. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A Tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics* **88**, 76–82 (2011).
28. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565–569 (2010).
29. Silva Júnior, J. N. de B. *et al.* Trends in tuberculosis incidence and mortality coefficients in Brazil, 2011–2019: analysis by inflection points. *Revista Panamericana de Salud Pública* **47**, 1 (2023).
30. Dill-McFarland, K. A., Benson, B. & Segnitz, R. M. SEAsnake: a snakemake pipeline for processing RNA-seq fastq to counts. *Zenodo* (2024) doi:10.5281/zenodo.11646755.
31. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
32. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
33. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* **34**, D590–D598 (2006).
34. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049 (2017).
35. Linderman, G. C. *et al.* Zero-preserving imputation of single-cell RNA-seq data. *Nat Commun* **13**, 192 (2022).
36. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* **16**, 1289–1296 (2019).
37. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* **42**, 293–304 (2024).
38. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
39. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).
40. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**, R29 (2014).
41. Dill-McFarland, K. A. *et al.* Kimma: flexible linear mixed effects modeling with kinship covariance for RNA-seq data. *Bioinformatics* **39**, btad279 (2023).

42. Rubinacci, S., Hofmeister, R. J., Sousa da Mota, B. & Delaneau, O. Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes. *Nat Genet* **55**, 1088–1090 (2023).
43. Schaller, M. A., Allen, R. M., Kimura, S., Day, C. L. & Kunkel, S. L. Systemic Expression of Notch Ligand Delta-Like 4 during Mycobacterial Infection Alters the T Cell Immune Response. *Front Immunol* **7**, (2016).
44. Li, Q. *et al.* Down-regulation of Notch signaling pathway reverses the Th1/Th2 imbalance in tuberculosis patients. *Int Immunopharmacol* **54**, 24–32 (2018).
45. Palaga, T. *et al.* Notch signaling regulates expression of Mcl-1 and apoptosis in PPD-treated macrophages. *Cell Mol Immunol* **10**, 444–452 (2013).
46. Chen, C.-K. J. Chapter Three - RGS Protein Regulation of Phototransduction. in *Progress in Molecular Biology and Translational Science* (ed. Fisher, R. A.) vol. 133 31–45 (Academic Press, 2015).
47. Masuho, I., Wakasugi-Masuho, H., Posokhova, E. N., Patton, J. R. & Martemyanov, K. A. Type 5 G Protein β Subunit (Gβ5) Controls the Interaction of Regulator of G Protein Signaling 9 (RGS9) with Membrane Anchors*. *Journal of Biological Chemistry* **286**, 21806–21813 (2011).
48. Brownlie, D. *et al.* Expansions of adaptive-like NK cells with a tissue-resident phenotype in human lung and blood. *Proceedings of the National Academy of Sciences* **118**, e2016580118 (2021).
49. Cooper, M. A. *et al.* Human natural killer cells: a unique innate immunoregulatory role for the CD56bright subset. *Blood* **97**, 3146–3151 (2001).
50. Piekos, J. A. *et al.* Genetic ancestry proportion influences risk of adverse events from tuberculosis treatment in Brazil. *medRxiv* 2024.08.29.24312595 (2024) doi:10.1101/2024.08.29.24312595.
51. Brites, D. & Gagneux, S. The Nature and Evolution of Genomic Diversity in the Mycobacterium tuberculosis Complex. in *Strain Variation in the Mycobacterium tuberculosis Complex: Its Role in Biology, Epidemiology and Control* (ed. Gagneux, S.) 1–26 (Springer International Publishing, Cham, 2017). doi:10.1007/978-3-319-64371-7_1.

1 **Table 1. Cohort demographics of Brazil Regional Prospective Observational Research in TB**

2 **(RePORT) GWAS**

		Full cohort			Case-control cohort		
		Close contacts	**TB cases**	**P[a]**	**Close contacts**	**TB cases**	**P[a]**
	N	1807	947		241	578	
Clinical	Age	33 ± 19	38 ± 14	2E-14	38 ± 17	36 ± 14	7E-02
	< 5 yrs (N)	94	0		2	0	
	Sex (male)	738 (41%)	631 (67%)	1E-37	83 (34%)	367 (64%)	5E-14
	Race			3E-04			8E-03
	Brown	1078 (60%)	489 (52%)		130 (54%)	261 (45%)	
	Black	360 (20%)	243 (26%)		79 (33%)	177 (31%)	
	White	350 (20%)	195 (21%)		28 (12%)	129 (22%)	
	Other	19 (1%)	20 (2%)		4 (2%)	11 (2%)	
TB risk	BMI	25 ± 6.5	21 ± 3.5	7E-71	26 ± 5.8	20 ± 3.1	8E-49
	Tobacco use (current, past)	478 (27%)	491 (52%)	8E-40	75 (31%)	279 (48%)	9E-06
	Diabetes mellitus (DM)	96 (5%)	221 (23%)	5E-45	0	0	
	Hemoglobin A1C (HbA1c)[b]	*NA*	6.4 ± 2.0%				
	HIV	44 (2%)	186 (20%)	1E-53	0	0	
	Anti-retroviral therapy (ART)	0	77 (41%)				
	CD4 counts [min, max]	467 [108, 1561]	218 [6, 1092]	1E-06	*NA*	*NA*	
	LTBI treatment	437 (30%)	0		9 (3%)	0	
	Cumulative treatment (days)	190 ± 84			15 ± 10		

3 [a] Numeric variable compared using Wilcoxon test. Categorical variables compared using Chi-squared. [b] DM was self-

4 reported in controls and determined by HgA1c in cases. NA: not applicable

5 **Table 2. GWAS-level significant SNPs annotated to protein-coding genes [a]**

			Full cohort					Case-control cohort		Impute	hpWGS	Annotation		sceQTL[b]
	snpID	rsID	P	OR	MAF control	MAF case	P other models	P	OR	Score	Match (%)	Gene	Detail / kb to gene	Cell type FDR < 0.05
Intragenic	4:139801643:G:T	rs13114088	2.6E-09	2.2	5.5	8.8		0.036	1.6	0.65	72	MAML3	Intron 2	sMyeloid, o tCD8
	15:57418934:T:C	rs56380268	5.4E-09	2.2	4.8	7.6		0.35	1.2	0.66	86	CGNL1	Intron 1	*nd*
	17:14155330:A:G	rs34369729	3.5E-08	2.1	5.1	8.3		0.43	1.2	0.64	74	COX10	Intron 4	isgB
	17:65130128:A:T	rs9303479	1.0E-10	2.4	5.1	8.6		0.085	1.5	0.65	84	RGS9	Intron 1	NK-CD56+
	22:43432220:G:A	rs1458265905	2.9E-09	2.2	5.4	8.9		0.0021	2.1	0.69	76	MPPED1	Intron 2	*nd*
Cis up to 50 kb	3:75816211:A:G	rs111435847	3.4E-10	2.4	4.3	7.6		0.016	1.8	0.83	88	ZNF717	30.6	cMono-2
	3:75816219:T:G	rs113392015	1.4E-10	2.4	4.5	7.9		0.022	1.7	0.84	88	ZNF717	30.6	
	3:75816280:A:T	rs78962383	6.2E-11	2.4	4.7	8.3		0.025	1.7	0.81	80	ZNF717	30.7	
	3:75816310:C:A	rs113424442	3.4E-10	2.4	4.7	8.0		0.054	1.6	0.71	76	ZNF717	30.7	
	3:75816342:A:C	rs113386032	3.2E-08	2.2	4.4	6.9		0.088	1.5	0.80	84	ZNF717	30.8	nB
	10:95079377:G:A	rs200480013	1.8E-11	2.3	5.4	9.4		0.015	1.7	0.79	66	CYP2C8	9.9	
	22:39566026:A:T	rs199749312	2.5E-10	2.5	4.2	7.5		0.016	1.8	0.81	82	ATF4	43.3	
												CACNA1I	4.7	*nd*
												MIEF1	47.9	
												RPS19BP1	33.3	

6 [a] Details on all significant SNPs can be found in **Table S5**. SNPs in this table were derived from imputed low pass
7 WGS dataset, had > 60% concordance with high pass WGS, and were annotated to protein-coding genes. [b]
8 sMyeloid: stressed myeloid cell, oT: other T cell, tCD8: T cell CD8+ cytotoxic, isgB: B cell expressing IFN stimulated
9 genes, cMono: classical monocyte, nB: B cell naïve, *nd*: Not detected in single cell expression. MAF: minor allele
10 frequency, OR: odds ratio, hpWGS: high-pass whole genome sequencing, sceQTL: single-cell expression quantitative
11 trait loci

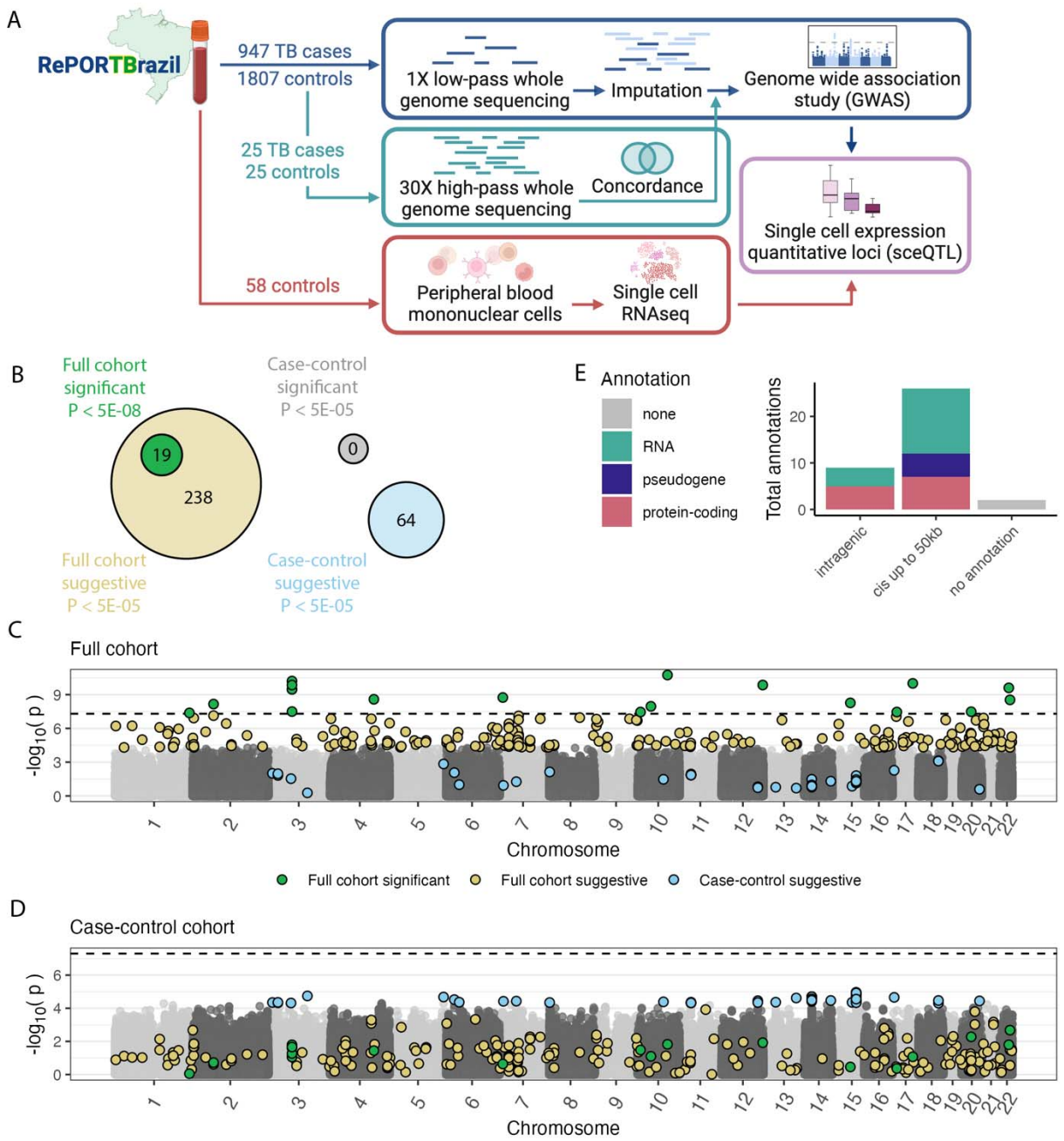
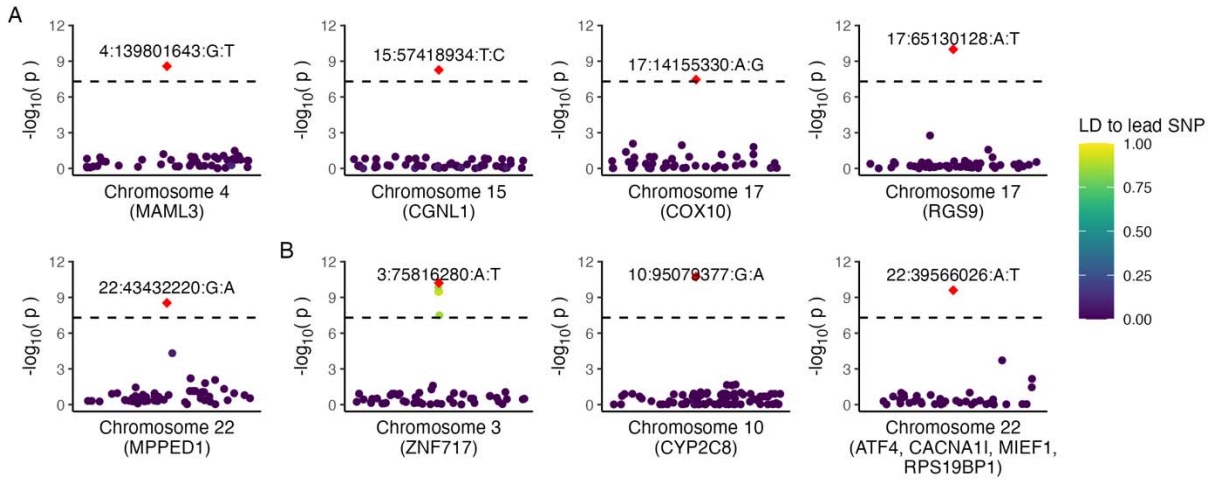


Figure 1. **SNPs associated with TB in Brazil GWAS.** SNPs at > 5% MAF and verified > 60% concordant in high-pass data are included. (A) Study design and analysis. Additional details can be found in Figure S1. Created with biorender.com. (B) In the full cohort, significant (green, P < 5E-8) and suggestive SNPs (yellow, P < 5E-5) did not overlap with any suggestive SNPs identified in the case-control cohort (blue, P < 5E-5). No SNPs were significant in the case-control cohort

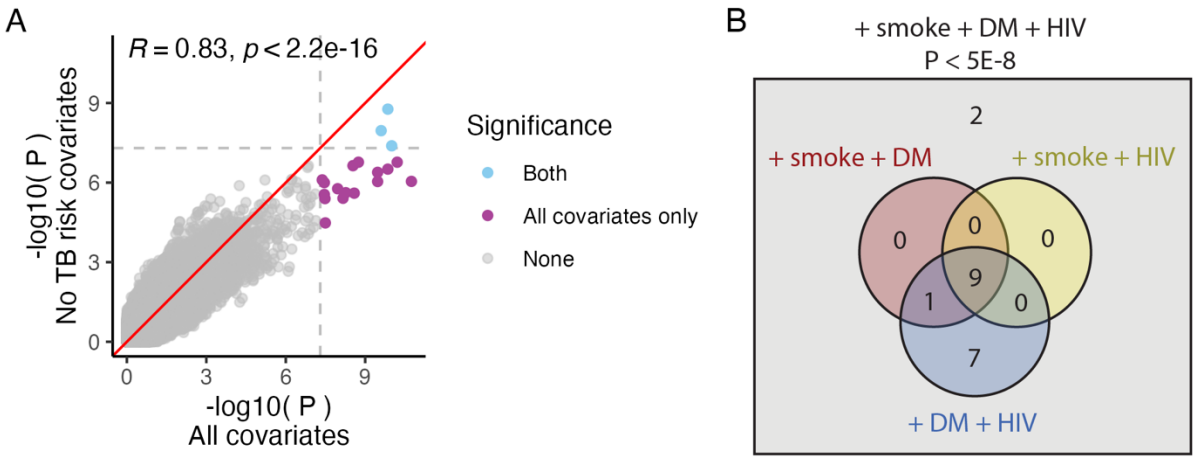
18 (grey, P < 5E-8). Manhattan plots for the (C) full cohort and (D) case-control subset. Concordant
19 genotypes are colored if significant or suggestive in either the full cohort or case-control with
20 colors as in (B). For nonsignificant SNPs, alternating odd and even numbered chromosomes are
21 shaded in greys. (E) Gene-level annotations of concordant significant SNPs. Genes are grouped
22 by their closest annotation within a gene or within 50 kb cis of a gene. SNP annotations are
23 colored by protein-coding, pseudogene, RNA (lncRNA, snRNA, snoRNA), and no annotation
24 groups.
25



26

27 **Figure 2. Genomic regions of significant protein-coding SNPs in the full cohort**. SNPs significant
28 for TB progression in the full cohort as well as > 60% concordant with high-pass sequencing
29 were selected. SNPs were annotated to protein-coding genes (A) intragenic and (B) cis up to 50
30 kb. Manhattan plots with concordant SNPs at > 1% MAF and +/- 10 kb from these SNPs are
31 shown with color indicating linkage disequilibrium (LD R^2) with lead SNPs, which are highlighted
32 as red diamonds. Gene annotations are noted on the x-axis.
33

34 -



35

36 **Figure 3. GWAS model impact of correcting for TB risk covariates.** The primary full cohort
37 model for TB progressions corrected for ancestry PCs, genetic kinship, age, sex, smoking, HIV,
38 and DM (all covariates). A second, simplified model without TB risk covariates (smoking, HIV,
39 DM) was then applied to the same data. Only SNPs > 5% MAF and > 60% concordance in the
40 high-pass data are shown. (A) Pearson correlation of P-values for all SNPs in the two models.
41 The red line indicates a 1:1 fit. Dashed lines and colored points indicate model significance (P <
42 5E-8), and there were no SNPs significant in only the no TB risk covariate model. (B) Impacts of
43 individual TB risk covariates. Leave-one-out models were applied to the 19 SNPs concordant in
44 high-pass and significant in the full cohort model with all covariates (grey box). Significant SNPs
45 in each model lacking one TB risk covariates are totaled in the Venn diagram with 2 SNPs
46 outside the Venn, indicating significance only when all three TB risk covariates are included in
47 the model.

48

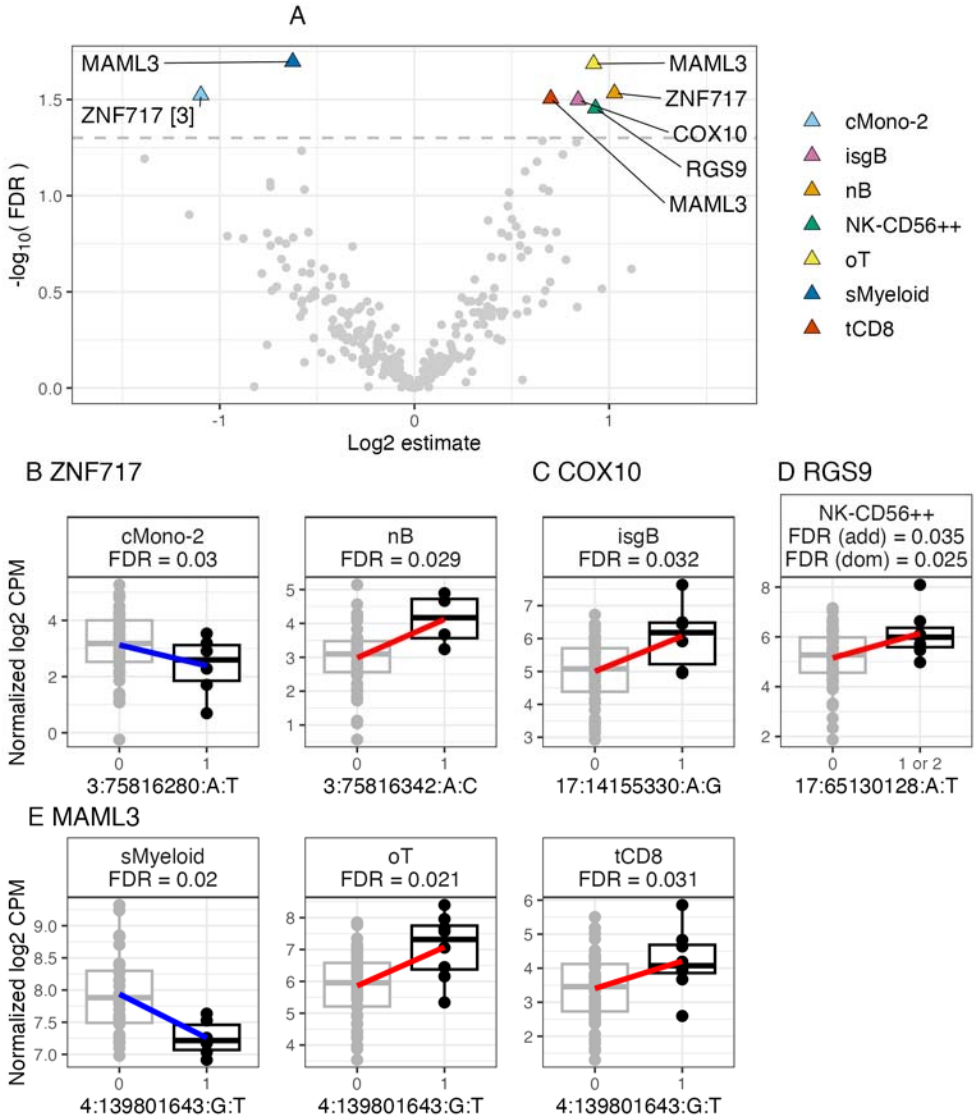


Figure 4. Single cell cis-expression quantitative loci (sceQTL) among TB GWAS significant SNPs. GWAS significant concordant SNPs were mapped 50 kb cis to protein-coding genes that were expressed in single-cell data. Gene expression was modeled against corresponding mapped SNPs. (A) Volcano plot of sceQTLs. Color and shape indicate significant sceQTL cell populations (FDR < 0.05). Labels note the gene annotation for significant SNPs with bracketed numbers indicating the total number of significant SNPs when > 1. (B) ZNF717, (C) COX10, (D) RGS9, and (E) MAML3 sceQTLs. Pseudobulk normalized counts per million (CPM) per donor in sceQTLs

57 with SNPs significant in GWAS (P < 5E-8). Genotypes are represented as the number of alternate
58 alleles (0: grey, 1 or 2: black) with SNPs identified by chromosome:position:reference:alternate.
59 Trend lines represent linear models with positive (red) and negative (blue) effects of the
60 alternate allele. cMono-2: Monocyte classical cluster 2, isgB: B cell expressing IFN stimulated
61 genes, nB: B cell naïve, NK-CD56++: natural killer cell CD56 bright, oT: T cell other, sMyeloid:
62 myeloid cell stressed, tCD8: T cell CD8+ cytotoxic