Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Genetic control of the transcriptional response to active tuberculosis disease and treatment

David MacHugh

david.machugh@ucd.ie

University College Dublin https://orcid.org/0000-0002-8112-4704

John Francis O'Grady

University College Dublin

Alexander Leonard

ETH Zurich https://orcid.org/0000-0001-8425-5630

Hou-Cheng Li

Aarhus University

Lingzhao Fang

Aarhus University https://orcid.org/0000-0003-1103-3679

Hubert Pausch

ETH Zürich https://orcid.org/0000-0002-0501-6760

Isobel Gormley

University College Dublin

Stephen Gordon

University College Dublin https://orcid.org/0000-0002-4833-5542

Article

Keywords:

Posted Date: August 29th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7311967/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Version of Record: A version of this preprint was published at Communications Medicine on July 29th, 2026. See the published version at https://doi.org/10.1038/s43856-026-01804-2.

Genetic control of the transcriptional response to active tuberculosis disease and treatment

John F. O'Grady[1*], Alexander S. Leonard[2], Houcheng Li[3], Lingzhao Fang[3], Hubert Pausch[2], Isobel C. Gormley[4], Stephen. V. Gordon[5,6,7], David E. MacHugh[1,6,7*]

[1] UCD School of Agriculture and Food Science, University College Dublin, Belfield, Dublin, D04 V1W8, Ireland.

[2] Animal Genomics, ETH Zurich, Universitaetstrasse 2, 8006, Zurich, Switzerland.

[3] Centre for Quantitative Genetics and Genomics, Aarhus University, C.F. Mollers Allé 3, 8000, Aarhus, Denmark

[4] UCD School of Mathematics and Statistics, University College Dublin, Belfield, Dublin, D04 V1W8, Ireland.

[5] UCD School of Veterinary Medicine, University College Dublin, Belfield, Dublin, D04V1W8, Ireland.

[6] UCD Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Belfield, Dublin, D04 V1W8, Ireland.

[7] UCD One Health Centre, University College Dublin, Belfield, Dublin, D04 V1W8, Ireland.

[*] Corresponding authors

E-mail addresses:

JFOG: john.ogrady@ucdconnect.ie

ASL: alexander.leonard@usys.ethz.ch

HL: houcheng.li@qgg.au.dk

LF: lingzhao.fang@qgg.au.dk

HP: hubert.pausch@usys.ethz.ch

ICG: claire.gormley@ucd.ie

SVG: stephen.gordon@ucd.ie

DEM: david.machugh@ucd.ie

Abstract

The functional impact of genomic regulatory variation during active TB disease and anti-TB treatment (ATT) remains poorly understood. Here, using 240 paired whole-blood RNA-seq samples from n = 48 active TB patients who underwent ATT, we call and impute 1,506,948 genome-wide variants from these transcriptomes to map *cis*-expression quantitative trait loci (eQTL), response-eQTL, and cell type interaction eQTL (ieQTL) by computationally deconvolving the bulk RNA-seq data. We show that the whole-blood transcriptome is substantially perturbed during ATT by identifying 7,246 differentially expressed genes. We characterise 5,129 *cis*-eQTL and 587 response-eQTL and show that genes associated with reQTL are significantly enriched in pathways impacting the host response to mycobacterial challenge and xenobiotic metabolism. We highlight significant changes in cell type proportions due to ATT, notably for NK cells and classical monocytes. We characterise 921 ieQTL, including an active-TB classical monocyte-specific ieQTL for the gene *NOD2*. Our work sheds light on the immunogenetics of TB disease and treatment, while providing a framework for integrative genomics studies using only RNA-seq data.

Introduction

Mycobacterium tuberculosis is the principal causative agent of human tuberculosis (TB), an infectious disease that caused the deaths of approximately 1.25 million people in 2023 and with more than 10.8 million newly diagnosed cases during the same reporting period[1]. Worldwide, it is estimated that approximately 25% or circa two billion individuals are latently infected with *M. tuberculosis*[2]. Of these latent TB infected individuals, between 5–10%, will develop active TB disease that will require treatment[3]. Without treatment, the death rate from human TB disease is high, with estimates ranging from 53-86%[4]; however, close to 85% of cases can be cured with treatment[1]. Anti-TB treatment (ATT) is a protracted process, characterised by the administration of strong antibiotics during two phases: 1) an intensive phase including rifampicin, isoniazid and pyrazinamide in combination with ethambutol for a period of two months, and 2) a continuation phase encompassing four months of rifampicin or isoniazid[5]. Unfortunately, ATT is associated with unwanted adverse effects including fever, hyperuricemia and most notably, hepatotoxicity[6].

Genetic linkage analyses and more recently, genome-wide association studies (GWAS) have been conducted for numerous tuberculosis-related phenotypic traits reviewed by [7,8]. These have included studies focusing on anti-TB drug-induced liver injury/hepatotoxicity, albeit with very small case numbers[9,10]. However, as is common to many GWAS studies, genome-wide significant single-nucleotide polymorphisms (SNPs) are largely located in non-coding genomic regions and inherited as dense, highly correlated haplotype blocks making identification of the causal variant challenging[11-13].

Expression quantitative trait loci (eQTL) are genomic polymorphisms associated with variability in mRNA transcript abundance and can be classified arbitrarily as *cis* if proximal to the associated gene (generally ≤ 1Mb) or *trans* if located further away from the associated gene (> 1Mb, including locations on different chromosomes)[14-16]. Expression QTL are functionally important for polygenic traits such as disease resilience because they are often enriched in GWAS signals[17] and provide a route for assigning regulatory polymorphisms to complex trait variation[18]. In this regard, previous studies have characterised the regulatory architecture of mammalian genomes, primarily focusing on gene expression across tissues in a range of species (e.g., humans, domestic cattle, and pigs)[19-21], at key developmental stages[22], and in particular cell types during specific disease contexts[23].

In the context of TB disease, several studies have integrated genotype and expression data to explore the genomic architecture of the transcriptional response during TB disease; however, these studies have either focused on a modest number of candidate genes[24], or were conducted *in vitro* using dendritic cells (DCs) or monocyte-derived macrophages (MDMs)[25-28]. To the best of our

knowledge, no high-resolution genome-wide eQTL study has been conducted for TB disease or to profile the genomic architecture *in vivo* of the ATT transcriptional response. However, the study by Barreiro, et al. [25] is noteworthy, because genomic variants were identified that alter mRNA expression only upon *M. tuberculosis* infection and likely represent functionally relevant polymorphisms involved in the host response during TB disease. These variants are known as response eQTL (reQTL) and represent a bridge between disease/trait-relevant phenotypes and genetic variation[29].

Most of the consortia-based eQTL studies outlined above were focused on mapping regulatory loci in bulk tissues. Although this is informative, and given the *cis*-genetic correlation of gene expression across tissues is high at approximately 0.7–0.8[30], the cross cell type *cis*-genetic correlation of gene expression is much lower at circa 0.25–0.75[31] indicating cell type specific effects. Recent large-scale studies have therefore focused on mapping such *cis*-eQTL at single-cell resolution (sc-eQTL)[32]. While this is costly and computationally challenging, a method to circumvent these obstacles centres on identifying proxy sc-eQTL known as cell type interacting eQTL (ieQTL) through *in-silico* deconvolution of bulk RNA-seq based on a scRNA-seq reference[33], and fitting an interaction term in the linear model between genotype and inferred cell type proportions[34,35]. Identifying such ieQTL can characterise additional loci important in complex trait variation that are otherwise masked in bulk RNA-seq[34]. However, hitherto, mapping of sc-eQTL or cell type ieQTL has not been performed in the context of active TB disease or ATT *in vivo*.

An eQTL study requires both genotype and gene expression information from the same individual; however, for some studies this is often not the case. In this regard, it has been shown that calling variants directly from high-throughput RNA-seq data can be used to address this problem[36,37]. For example, this approach has been recently used to conduct multi-tissue eQTL analyses in cattle and pigs[20,21]. Moreover, various tools that call variants from short-read sequence data have been developed, including the genome analysis tool kit (GATK)[38] and the DeepVariant tool[39], with the latter being specifically modified to call variants from RNA-seq data[40].

A previous study conducted by Tabone, et al. [41] examined transcriptional differences between a variety of TB phenotypes ranging from healthy controls to TB patients post anti-TB therapy in peripheral blood (PB) derived from the Leicester TB cohort. The aim of this current study was to leverage the RNA-seq data from Tabone and colleagues[41] and determine if genetic variability exists that influences the inter-individual transcriptional response during active TB disease and during ATT *in vivo*. This hypothesis was tested using PB RNA-seq data, derived from matched individuals with active TB who underwent anti-TB therapy[41], and by performing RNA-seq variant calling and *in silico* deconvolution to map *cis*-eQTL and ieQTL associated with active TB disease, and to characterise reQTL that alter their effect due to ATT *in vivo*.

Methods

RNA-seq sample download and preprocessing

A total of 240 high-throughput paired-end (2 × 100 bp) peripheral blood (PB) RNA-seq fastq files generated by Tabone, et al. [41] derived from a total of $n = 48$ individuals across five timepoints (a single untreated timepoint (T0) and four timepoints post commencement of ATT (T1–T4)) were downloaded from the Sequence Read Archive[42] (GEO accession number: GSE157657) using fastq-dump (v. 2.10.5) in addition to the *--split-files* parameter. The quality of the RNA-seq reads was assessed using FastQC (v. 0.11.9)[43] and MultiQC (v. 1.9)[44]. Trimmomatic (v. 0.38)[45] was used to sequentially remove Illumina adapters, filter raw reads < 36 bp in length and remove trailing bases with a phred-scale quality score < 30. Filtered RNA-seq reads were aligned to the human reference genome (GRCh38_full_analysis_set_plus_decoy_hla) using STAR (v. 2.7.10b)[46]. The reference genome index was generated using the *genomeGenerate* command with the gencode.v43.annotation.gtf annotation file specified in addition to the *--sjdbOverhang* 99 parameter. Read counts for each gene were quantified using featureCounts (v. 2.0.3)[47] by specifying the *-p -B -C -s2* flags to indicate that paired-end reads were being analysed, to only count pairs with both ends aligned, to exclude chimeric fragments and to denote that reversely stranded RNA-seq reads were being analysed. We performed expression quantification for each timepoint separately resulting in a total of five count matrices.

RNA-seq variant calling and genotype concordance

Single-nucleotide polymorphisms (SNPs) and indels were called from the aligned reads for each timepoint separately using the DeepVariant (v. 1.5)[39] RNA-seq variant calling pipeline[40], including the RNA-seq model from DeepVariant and by specifying the *--split_skip_reads* parameter. Individual sample genomic VCF files derived from DeepVariant were merged and harmonized using GLnexus[48] generating timepoint-specific VCF files. Across each timepoint separately, we retained variants with a minor allele frequency (MAF) > 0.1, which did not deviate significantly from Hardy-Weinberg Equilibrium (HWE) ($P < 1 \times 10^{-6}$), and that possessed a genotype call rate > 95%. The genotype concordance of samples across timepoints at matched genomic loci was compared to ensure that no sample swaps occurred. Genotype concordance was defined as the proportion of genotypes (0/0, 0/1, 1/1) at identical loci that matched between any two samples across two timepoints. We compared the genotype concordance for all timepoint pairwise comparisons (20 in total).

To maximise genome coverage and depth, and to generate a final set of variants for each of the $n = 48$ samples, we used SAMtools (v 1.15.1)[49] to merge individual specific aligned bam files across timepoints together. The two exceptions to this method included samples for patient 3 and patient

157, which possessed a moderately lower genotype concordance at timepoint 1 and timepoint 0, respectively, which was hypothesised to be due to reduced RNA quality. Therefore, for these two samples, we excluded these specific timepoints from the merging process. We then reimplemented the DeepVariant/GLnexus RNA-seq variant calling pipeline on the merged bam files. Read coverage at coding sequences was assessed using mosdepth (v. 0.3.3)[50]. To ensure variants (reference and alternative alleles) were being called correctly, we compared locus-specific reference and alternative allele pairs called in our dataset that were present in benchmarked regions across four Genome in a Bottle (GIAB) data sets (HG001, HG005, HG006 and HG007; https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/) (v. 4.2.1)[51].

Phasing and imputation

Prior to phasing and imputation, we excluded rare variants (MAF < 0.03), which deviated significantly from HWE ($P < 1 \times 10^{-5}$) or that displayed a poor Phred quality score (QUAL < 20) and that were missing in more than 80% of samples. Phasing was performed on each chromosome separately using SHAPEIT (v 5.1)[52] with individual chromosome recombination maps specified (https://github.com/odelaneau/shapeit4/blob/master/maps/genetic_maps.b38.tar.gz) in addition to a phased multi-ancestry reference panel derived from Byrska-Bishop, et al. [53] composed of 3,202 WGS (30× sequencing depth) individuals. Imputation was performed on each chromosome separately using Beagle (v. 5.4)[54] with default settings implemented and specifying the same reference panel as used for the phasing procedure. We visualised the relationships between gene expression, variant calling, and imputation performance using the karyoplotR package[55]. Briefly we first plotted the $\log_2$ transcripts per million (TPM) + 1[56,57] expression values of genes on chromosome 14 (HSA14) and visualised the dosage R^2 (DR2) values of directly called and imputed variants on the same chromosome. Following imputation, we retained variants with a DR2 value > 0.6 and that had a MAF > 0.05. Functional annotation of called and imputed variants was performed using Ensembl's Variant Effect Predictor (VEP) (v. 114.1)[58].

Population genetic structure was assessed using the *--pca* parameter from PLINK (v. 1.9)[59]. The *--keep-allele-order*, and *--indep-pairwise 50 10 0.1* flags were used to remove variants in high linkage disequilibrium (LD). The orthogonal eigenvectors of principal components (PCs) 1 and 2 were plotted and data points were coloured according to the self-reported ancestry for each individual.

Differential expression and functional enrichment analyses

A paired DEA was conducted between the naïve, untreated group (T0) and each treated group (T1, T2, T3, and T4) separately using DESeq2 (v.1.40.2)[60] and a design matrix incorporating the patient ID as a covariate. Genes with raw expression counts ≥ 6 in at least 20% of samples were

retained prior to the DEA. For each DEA, the null hypothesis tested was that the absolute logarithmic fold change (LFC) between the untreated timepoint and each of the treated timepoints for the expression of a particular gene was greater than 0.2. For these contrasts, raw *P*-values were inferred using the Wald test and corrected using the Benjamini-Hochberg (BH) method[61] to account for multiple tests. Genes with a BH false discovery rate (FDR) adjusted *P*-value ($P_{adj.}$) < 0.05 and an absolute LFC > 0.2 were considered significantly differently expressed (DE).

To identify pathways/ontologies enriched by significant DE genes, for each contrast, we took the top 300 genes showing increased and decreased expression based on $P_{adj.}$ and performed a gene-set enrichment analysis (GSEA) using the g:GOSt function implemented in the g:Profiler2 R package (v. 0.2.3)[62,63]. Of note, for DE genes identified from the T1 versus T0 comparison, we did not perform this analysis as the number of input DE genes was too small (n = 153) Following best practice recommendations from Wijesooriya, et al. [64], the background set of genes consisted of all genes that were tested in the DE analysis. Raw *P*-values were corrected using the BH method and pathways with a $P_{adj.}$ value < 0.05 were considered significantly enriched with DE genes.

***Cis*-eQTL mapping**

For the mapping of *cis*-eQTL, we used the human Genotype Tissue Expression (GTEx) Consortium[19] eQTL analysis pipeline with some minor modifications. We conducted the *cis*-eQTL analysis on each experimental group (T0, T1, T2, T3, and T4) separately. Raw RNA-seq read counts were normalised using the trimmed mean of the M values (TMM) method[65] and the expression values for each gene were then inverse normally transformed across samples to ensure the molecular phenotypes followed a normal distribution. Genes with raw expression counts ≥ 6 and a TPM[56,57] normalised expression count ≥ 0.1 in at least 20% of samples were retained for the eQTL analysis. For each group, we used the PCAForQTL R package (v.0.1.0)[66] to identify hidden confounders in the normalised and filtered expression matrices. We used the *runElbow* function in PCAForQTL to select the number of latent variables using the elbow method. We then merged these inferred/latent covariates with known covariates (the top three genotype PCs of the imputed genotype data set, age in years, and days since ATT commencement) and removed highly correlated known covariates captured well by the inferred covariates (unadjusted R^2 ≥ 0.9) using the PCAForQTL *filterKnownCovariates* function.

For the *cis*-eQTL mapping procedure, we used TensorQTL (v.1.0.8)[67]. To the best of our knowledge, no previous published study has investigated the optimal genomic interval at which to conduct a *cis*-eQTL analysis using imputed variants originally derived from RNA-seq data. Therefore, we conducted the *cis*-eQTL analysis considering distance windows of ±10–1000 kb from the TSS and selected the window size that resulted in the maximum number of unique *cis*-eGenes

(i.e., genes with at least one significant *cis*-eQTL) across all groups for downstream analyses. To identify significant *cis*-eQTL, we invoked the permutation strategy in TensorQTL[68] to estimate variant-phenotype associated empirical *P*-values with the parameter --mode *cis* to account for multiple variants being tested per molecular phenotype. We then used the Storey and Tibshirani FDR procedure[69] to correct the *beta* distribution-extrapolated empirical *P*-values to account for multiple phenotypes being tested genome-wide. A gene with at least one significant associated *cis*-eQTL (FDR $P_{adj.} < 0.1$) was considered a TensorQTL *cis*-eGene.

To identify significant *cis*-eVariants associated with detected *cis*-eGenes, we followed the procedure implemented by the Pig GTEx Consortium[21]. Briefly, we first obtained nominal *P*-values of association for each variant-gene pair using the parameter --mode *cis_nominal*. We then defined the empirical *P*-value of a gene which was closest to an FDR of 0.1 as the genome wide empirical *P*-value threshold (*pt*). Next, we calculated the gene-level threshold for each gene from the beta distribution by using the *qbeta(pt*, *beta_shape1*, *beta_shape2*) command in R (v. 4.3.2)[70] with beta_shape1 and beta_shape2 being derived from TensorQTL. Variants with a nominal *P*-value of association below the gene-level threshold were included in the final list of variant-gene pairs and were considered as significant TensorQTL *cis*-eVariants.

Power calculation for *cis*-eQTL mapping analysis

We estimated the effect size required to achieve an 80% power of detection using the *pwr.f2.test* function from the pwr (v. 1.3-0) R package[71]. We set the parameters $u = 1$, $v = 32–35$ depending on the final number of covariates used in each group, respectively, *sig.level* $= 1 \times 10^{-4}$ and *power* $= 0.8$. Based on these parameters, effect sizes of 0.761 in T0, 0.816 in T1, 0.788 in T2 and T3, respectively, and 0.847 T4 were estimated to be detectable at 80% power.

Replication of *cis*-eQTL results

To assess the replicability of *cis*-eQTLs identified in each timepoint, we downloaded summary statistics from the INTERVAL study, which performed *cis*-eQTL mapping in $n = 4{,}732$ individuals of European ancestry[72]. We used three different measurements of agreement of eQTL effects when comparing eQTLs across the two studies: allele concordance (AC), Story's π_1 statistic, and Spearman correlation ($\rho_{Spearman}$) of effect size estimates between the discovery and replication cohorts, respectively. To calculate AC, we matched significant eQTLs in the discovery cohort to significant *cis*-eVariants in the replication cohort. We then calculated the proportion of these cis-eQTLs/*cis*-eVariants that showed the same direction of effect. To calculate π_1, we obtained the *P*-values in the replication cohort of significant associations identified in the discovery cohort and used the *qvalue* function in R to estimate π_0. We then calculated π_1 as $1 - \pi_0$. To obtain the $\rho_{Spearman}$

statistics, we calculated the $\rho_{Spearman}$ between significant eQTLs identified in this study to matched variant:gene pairs in the replication cohort, regardless of significance.

Response eQTL (reQTL) mapping

To identify variants that significantly alter their effect due to ATT, we followed the procedure implemented by Lin, et al. [73]. Briefly, we used mashr (v. 0.2.79)[74], which maximises power through jointly analysing genetic effects on gene expression across timepoints to improve effect size and reduce standard error estimates, thereby facilitating the identification of both shared and condition specific *cis*-eQTL. Following the mashr eQTL analysis outline pipeline (https://stephenslab.github.io/mashr/articles/eQTL_outline.html), we first extracted a "strong set" of effect size estimates derived from the top eQTL for each gene across timepoints to prevent double counting of eQTL that are in LD and impact the same gene. We then extracted a subset of 200,000 randomly selected variants and considered a variant as a *cis*-eQTL and as having a significant effect on gene expression if the local false sign rate (LFSR) was < 0.05.

We next compared the effect sizes of putative eQTL between T0 and each of the treated timepoints (T1–T4). We classified an eQTL as a response eQTL if the effect was significant in at least one group (LFSR < 0.05) and if the absolute $\log_2$ fold change in the posterior effect size was greater than 2[73].

Enrichment analysis of reQTL

We used the *g:GOSt* function implemented in the g:Profiler2 R package (v. 0.2.3)[62,63] to determine the biological processes impacted by response-eGenes. To do this, we selected all response-eGenes and ordered them based on their absolute magnitude fold change relative to T0 and removed duplicate genes. The background set of genes consisted of all genes that were tested in the reQTL analysis. Raw *P*-values were corrected using the BH method and pathways with a $P_{adj.}$ value < 0.05 were considered significantly enriched for response-eGenes genes.

Single cell RNA-seq data analysis

Alignment and quality control

A total of $n = 5$ peripheral blood mononuclear cell (PBMC) scRNA-seq data sets derived from $n = 3$ individuals with active tuberculosis and $n = 2$ individuals with latent tuberculosis infection (LTBI) were obtained from Cai, et al. [75] and downloaded from the SRA (accession no. SRP247583; Sequence Read Run (SRR) IDs SRR11038989–SRR11038993). For each sample, the cleaned data filtered for low-quality reads and unrelated sequences were imported to CellRanger (v. 7.2.0)[76] and aligned to the human reference genome (hg38, GRCh38) specifying the count command and setting

the *–expect-cells* parameter to 10,000. The resulting count matrix files were loaded into R using the *Read10x* function from the Seurat (v 5.0.2) R package[77]. Ambient RNA was first removed in each sample using the *autoEstCont* function in SoupX (v.1.6.2)[78]. To remove possible doublets, we used scDblFinder (v. 1.17.1)[79] setting the dbr parameter equal to 0.01. Finally, we filtered and removed low-quality cells with < 500 or > 3000 expressed genes or that that had >7% of UMIs mapping to mitochondrial genes.

Batch effect correction and dimension reduction

Using the Seurat R package, data were log-transformed using the *NormalizeData* function and the top 2000 variable genes were selected using the *FindVariableFeatures* function setting the selection method to "vst". To remove batch effects, we first obtained genes that were variable across samples using the *SelectIntegrationFeatures* function and identified anchors using the *FindIntegrationAnchors* function setting the reduction parameter = "cca". Finally, we used the *IntegrateData* function to merge different samples together. The top 12 eigenvectors derived from the *RunPCA* function of the integrated data were used for downstream uniform manifold approximation and projection (UMAP) visualization and clustering.

Cell clustering, marker gene identification, and cell type annotation

The Louvain algorithm was used to cluster cells setting the resolution parameter to 0.4 in the *FindClusters* function resulting in 16 distinct cell clusters. We annotated cells using a comprehensive procedure involving automatic reference-based annotation, marker-based automatic annotation and marker gene-based annotation using marker genes curated from the literature. For the reference-based annotation, we used SingleR (v. 2.2.0)[80], specifying the Human Primary Cell Atlas (HPCA)[81], Blueprint/ENCODE (BPEN)[82,83], Immunological Genome Project (ImmGen)[84] and Monaco immune (Monaco)[85] data sets as references from the celldex (v. 1.10.1)[80] package and extracted both the main and high-resolution (fine) cell type annotation. For the marker-assisted annotation, we used SCINA[86] specifying markers derived from PBMCs collated by Diaz-Mejia, et al. [87] in .gmt format. Lastly, we used marker genes that have been systematically curated[88,89].

To identify marker genes capable of demarcating each annotated cluster, we used the *FindConservedMarkers* for each cluster, grouping on disease status and only considering positively DE genes. A gene was considered significantly DE if it was expressed in at least 25% of the cells within the cluster, the difference in the fraction of detection to the other clusters was at least 0.25 and the LFC was > 0.75. The one exception to this was for *Undefined T cells* where we relaxed the logfc.threshold parameter to 0.5 and the min.diff.pct parameter to 0.15. We then grouped DE genes together based on their cluster, ordered by LFC and selected up to a maximum of 100 genes per

cluster and removed duplicated genes. The resulting signature matrix was used in the downstream deconvolution analysis where specified.

Benchmarking of bulk RNA-seq deconvolution strategies

To identify a suitable tool to deconvolve the bulk RNA-seq data using the scRNA-seq data from Cai, et al. [75] as a reference, we performed a deconvolution benchmarking assessment of four algorithms: CIBERSORT[33], MuSiC (v 1.0.0) and non-negative least squares (NNLS) (v 1.0.0)[90], and Bisque (v. 1.0.5)[91]. We first generated pseudo-bulk RNA-seq data representing n = 100 samples derived from the Cai, et al. [75] PBMC scRNA-seq data using the SimBu R package (v. 1.2.0)[92] setting the *scenario* parameter to mirror_db and the *balance_even_mirror_scenario* parameter to 0.5 to introduce variability. We then visualised the cell type proportions of the 100 pseudo-bulk samples using the *plot_simulation* function. Given that MuSiC and Bisque perform internal normalisation, but CIBERSORT does not (unless quantile normalisation is specified that is recommended for microarray data), we separately TPM- and counts per million (CPM-) normalised the pseudo-bulk RNA-seq samples used for the CIBERSORT deconvolution procedure. Additionally, we also separately TPM-normalised the scRNA-seq data prior to deconvolution benchmarking using CIBERSORT.

In total, across all algorithms and normalisation strategies, we performed 12 benchmarking assessments. This included MuSiC, NNLS, and Bisque with, and without specifying marker genes from the signature matrix, and CIBERSORT with the raw, TPM- or CPM-normalised bulk RNA-seq data and the raw or TPM-normalised scRNA-seq data. Comparing the observed cell type proportions post-deconvolution to the expected cell type proportions, we assessed deconvolution performance using two metrics: 1) the average Pearson correlation (ρ) and 2) the average inverse of the root mean squared error (RMSE) across cell types where the RMSE is calculated as:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Where $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ = are the predicted values and $y_1, y_2, \ldots, y_n$ are the true simulated values in the pseudo-bulk samples for a particular cell type.

Deconvolution of bulk RNA-seq data

To perform the actual deconvolution of the bulk RNA-seq data, we used CIBERSORT[33] with the CPM-normalised bulk data and the raw counts from the signature matrix generated in this study. We disabled batch correction, quantile normalisation and conducted the analysis in relative mode

specifying 100 permutations. Deconvolved cell type proportions were visualised in R using ggplot2 (v. 3.4.4)[93].

To test for significant differences in cell type proportions throughout treatment, we first normalised the deconvolved cell type proportions using the ascrin method as part of the propeller function derived from the speckle R package[94]. We then used the *lmFit* function in limma (v. 3.56.2)[95] to fit a linear model accounting for individual ID as a random variable. To test for significant differences across groups, we performed an ANOVA using the *topTable* function and corrected for multiple testing using the BH method. The null hypothesis tested here was that difference in cell type proportions for a particular cell between all the timepoints is equal to 0. Cell types with an FDR $P_{adj.} < 0.05$ were considered to significantly change their cell type proportions over time.

Interaction eQTL (ieQTL) mapping

To identify putative variants that modulate RNA transcript abundance within a particular cell type, we used TensorQTL to fit a linear model that included an interaction term between genotype and cell type proportions that accounted for the same covariates used in the standard *cis*-eQTL mapping procedure. Considering only well-defined cell types with a mean proportion > 0.01 across all timepoints (NK cells, classical monocytes, non-classical monocytes, naïve B cells and memory $CD8^+$ T cells), we first normalised the proportions of each cell type across samples using the inverse normal transformation procedure. Within the regression framework, to mitigate against potential regression outlier effects, we restricted ieQTL mapping to variants with MAF ≥ 0.05 in the samples belonging to each of the top and bottom halves of the cell type proportion distribution, for each timepoint-cell type combination using the *--maf_threshold_interaction* 0.05 option in TensorQTL[34]. To identify genes in each group with at least one significant ieQTL, the top nominal *P*-value for each gene within each cell type analysis was selected and corrected for multiple testing at the gene level using eigenMT[96]. Genome-wide significance was then determined by computing BH FDR on the eigenMT-corrected *P*-values. Following this approach, there were no interaction terms significant after multiple testing correction ($P_{adj.} < 0.05$).

To leverage the longitudinal nature of this study and to increase power, for each cell type, we used mashr[74] to combine information across timepoints and improve effect size estimates of putative ieQTL. To do this we first filtered for common variant-gene pairs tested in the standard analysis using TensorQTL across all timepoints for a particular cell type. Following the mashr eQTL analysis outline pipeline (https://stephenslab.github.io/mashr/articles/eQTL_outline.html), for each cell type, we selected the lead variant-gene pair across timepoints as the "strong subset" of associations. Next, we selected a random set of 200,000 variant-gene associations to act as the "random set" and invoked a

threshold of LFSR < 0.01 to establish if an interaction effect was significant within a particular timepoint.

Visualisation of eQTL, reQTL, and ieQTL results

To plot results of the eQTL, reQTL, and ieQTL analyses, we extracted the residualised expression levels of genes using the QTLtools (v. 1.3.1)[97] *correct* function specifying the phenotype .bed file with the parameter *--bed* and the covariates .txt file with the *--cov* parameter. For genes of interest, we plotted the residualised expression values for each individual separated by their corresponding genotype as a boxplot for the eQTL, and reQTL analyses. For the ieQTL analysis, we plotted the residualised expression values and the inverse normally transformed cell type proportions as a scatter plot with data points coloured according to an individual's corresponding genotype for a particular locus.

Results

Study design overview and data description

A total of 240 PB RNA-seq samples published by Tabone, et al. [41] from n = 48 active TB patients longitudinally sampled across five timepoints (a single untreated timepoint (T0) and four timepoints post commencement of ATT (T1–T4)) were downloaded from the Sequence Read Archive (SRA)[42] (**Fig. 1**). **Supplementary Table S1** illustrates the Sequence Read Run (SRR) IDs for each of the patients across all five timepoints. **Supplementary Table S2** details the metadata information for each of the n = 48 patients including, age, sex, and day of sampling relative to anti-TB therapy commencement. The median day (D) of sampling relative to anti-TB commencement (± SD) was −1D ± 1.5, +7D ± 2.8, +55D ± 4.2, +166D ± 16, and +286D ± 41 for T0–T4, respectively (**Supplementary Fig. S1a**; **Supplementary Table S2**). Based on the self-reported ethnicity of participants, there were 22 South Asian, 14 European, 10 East African, one British-Indian, and one South African individual(s) present (**Supplementary Fig. S1b**). Of the 48 individuals, 30 were male and 18 were female. The mean age of study participants was 37 ± 11 years; for males and females separately, the mean age was 39 ± 11 years and 32 ± 10 years, respectively (**Supplementary Fig. S1c**; **Supplementary Table S2**).

Peripheral blood RNA-seq post-QC and filtering yielded a mean number of reads (± SD) per individual sample library of 27,659,544 ± 3,408,460 available for alignment. Of these, 16,872,495 ± 3,323,385 reads mapped uniquely to the human reference genome; 6,967,655 ± 3,189,931 reads mapped to multiple loci ($1 < N \leq 10$); 173,901 ± 50,990 reads mapped to too many loci ($N > 10$); 3,622,640 ± 1,218510 reads could not be mapped as they were too short; and 22,852 ± 7,672 reads could not be mapped to any genomic locus (**Supplementary Table S3**). Following quantification of gene expression levels, none of the libraries exhibited an abnormal distribution of counts (**Supplementary Fig. S2a**) and principal component analysis (PCA) based on the top 1,500 most variable genes following variance stabilising transformation of the raw counts in DESeq2 demonstrated that no sample, both within and across time points appeared to be an outlier. However, we observed that principal component 2 (PC2) separated individuals based on their reported sex at timepoints T0–T2 whereas PC1 separated individuals based on their reported sex at timepoints T3 and T4, respectively (**Supplementary Fig. S2b**). Evaluating the expression of the *RPS4Y1* and *XIST* genes across all n = 240 individuals/timepoints demonstrated that samples were assigned the correct sex ID (**Supplementary Fig. S2c**).

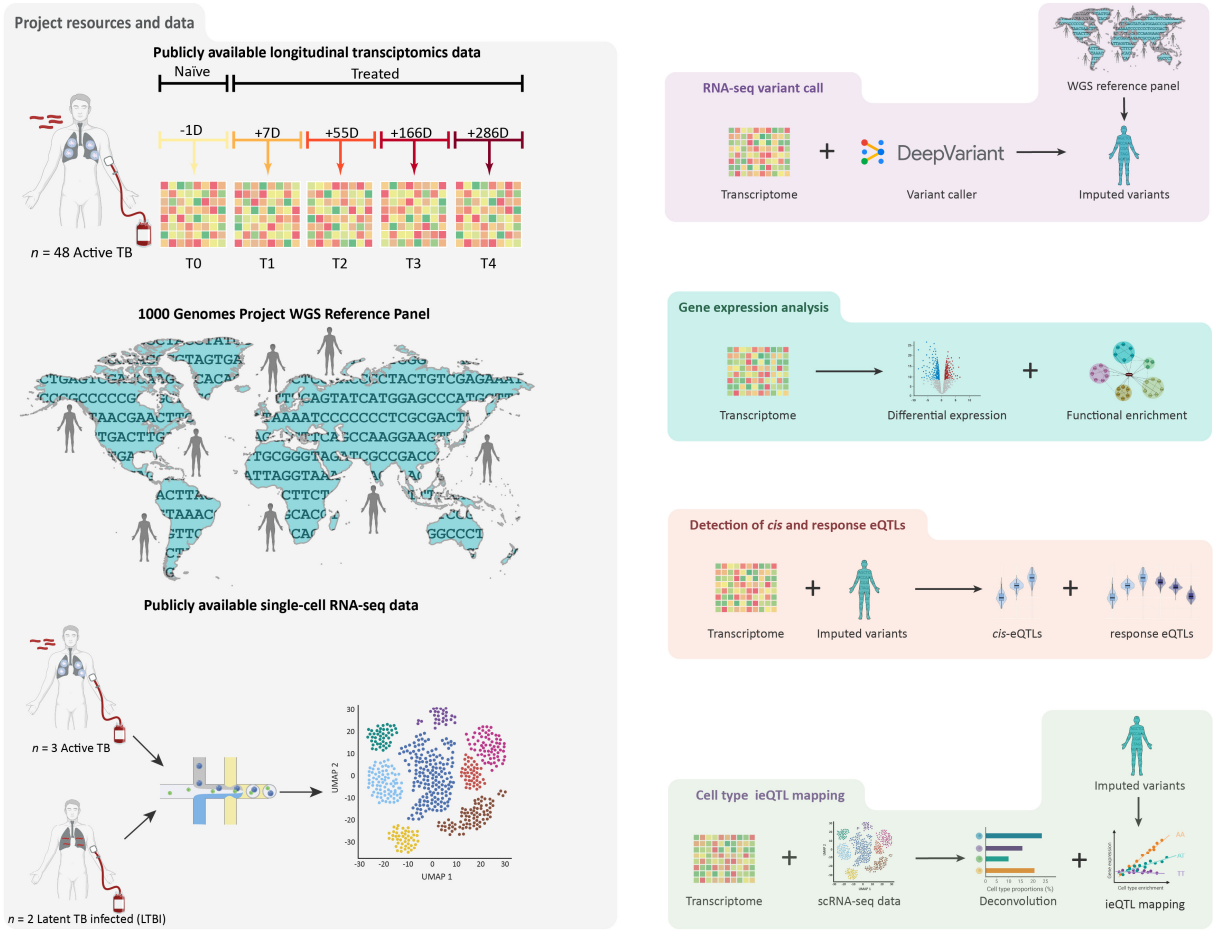


Fig. 1: Overview of study design, project resources, and computational analyses. We downloaded publicly available longitudinal peripheral blood (PB) RNA-seq data from Tabone, et al. [41] consisting of *n* = 48 patients with active tuberculosis (TB) who were sampled in a median sense −1D (T0), +7D (T1), +55D (T2), +166D (T3), and +286D (T4), before and after commencing ATT, respectively. We also obtained whole-genome sequence (WGS) data from 3,202 individuals to act as an imputation reference panel from Byrska-Bishop, et al. [53] and peripheral blood mononuclear single-cell RNA-seq data from *n* = 3 active TB patients and *n* = 2 latent TB infected (LTBI) patients published by Cai, et al. [75] (grey). The computational genomics analyses included: (1) variant calling with DeepVariant directly from the RNA-seq data and using these for imputation with the 1000 Genomes WGS reference panel (violet); (2) differential expression and functional enrichment analysis between each of the treated timepoints (T1–T4) and the untreated timepoint (T0) (neon blue); (3) integration of RNA-seq data and imputed variants for *cis*-eQTL and response-eQTL mapping (peach); and (4) integration of bulk RNA-seq, scRNA-seq data and imputed variants for *in silico* deconvolution and cell type interaction eQTL (ieQTL) mapping (mint). Some figure components created with a Biorender.com license.

Variant calling from RNA-seq data and imputation

Using DeepVariant, for each individual at each timepoint, we called variants directly from the RNA-seq data and merged timepoint specific samples together using GLnexus. Using this approach,

we characterised 1,523,539; 1,602,815; 1,783,843; 1,675,180; and 1,709,479 raw variants for the T0–T4 groups, respectively. We observed high genotype concordance between paired samples at matched genomic loci, with a mean (± SD) genotype concordance of 0.9899 ± 0.0125, confirming there were no sample swaps (**Supplementary Fig. S3**; **Supplementary Table S4**).

To maximise sequence coverage for variant calling, we merged patient-specific BAM files for each of the 48 individuals across timepoints (**Supplementary Table S5**). The short-read sequence data was generated using RNA sequencing; therefore, the depth at coding sequences (CDS) was high, with about 25% of protein-coding loci covered at a depth of approximately 50× (**Supplementary Fig. S4a**), whereas the coverage across the genome was relatively poor with roughly 5% of the genome covered by approximately two reads (**Supplementary Fig. S4b**). Using the merged BAM files, we called 4,034,779 raw variants across all 48 patients. Of these variants, a total of 791,555, 767,033, 764,644 and 763,894 were present in benchmarked regions of the Genome In A Bottle (GIAB) samples: HG001, HG005, HG006, and HG007, respectively[51]. On average, across all four samples, DeepVariant identified the correct reference and alternative allele pairs for 97.898% of called variants (**Supplementary Fig. S5a**). Removing putative triallelic sites increased the mean percentage of correctly called reference and alternative allele pairs to 99.392% (**Supplementary Fig. S5b**), which demonstrated that DeepVariant can produce highly accurate raw variant calls from PB RNA-seq data.

Following imputation and filtering of variants, a total of 1,506,948 genome-wide variants remained. Of these, 428,027 were directly called from the RNA-seq data and 1,078,921 were imputed from the WGS reference panel. The directly genotyped variants were enriched for variants residing in coding sequence (CDS) regions and in 5' or 3' untranslated regions (UTRs) whereas the imputed variants were enriched for variants residing upstream or downstream of a gene's CDS (≤ 5 kbps) and in intergenic regions (**Supplementary Fig. S6**). These results suggest that the imputation procedure captured a substantial number of variants residing in non-coding regions of the genome that would be important for mapping of *cis*-eQTLs. A principal component analysis (PCA) generated form 22,746 pruned genome-wide variants demonstrated that PC1 and PC2 accounted for 9.33% and 7.47% of the variation observed across the top 20 genotype PCs, respectively, with PC1 separating individuals based on their self-reported African ancestry and PC2 capturing European-Asian differences (**Fig. 2a**). Although the imputation procedure captured a substantial amount of non-coding variation (**Supplementary Fig. S6**), we observed that our ability to successfully call and impute variants is largely restricted to genomic loci within or proximal to expressed regions of the genome as imputation performance decreased substantially in regions of the genome not captured in the RNA-seq alignment (e.g., gene deserts or regions void of expression in these samples) (**Fig. 2b**), which is consistent with previous studies[37].

Taken together, based on these results, we were confident that 1) all samples were labelled correctly, 2) DeepVariant successfully called CDS variants with the correct reference and alternative allele, 3) the genotype data successfully differentiated individuals based on their self-reported ethnicity, and 4) imputation performance in non-coding and non-expressed regions of the genome was suboptimal but non-coding variation is still captured and retained.

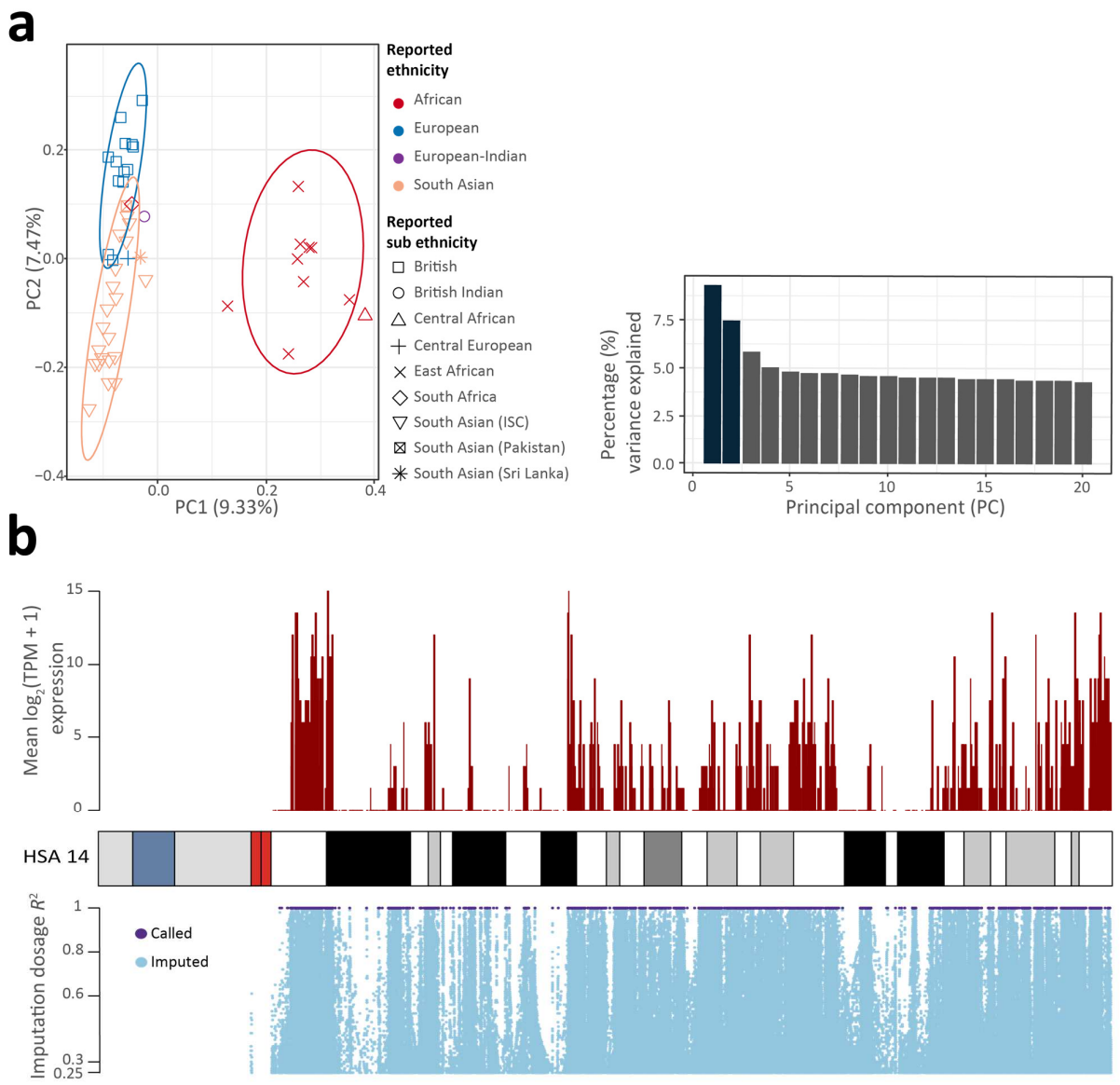


Fig. 2: Variant calling from RNA-seq data reveals global geographic structure but high-quality imputation restricted to areas within or proximal to expressed regions of the genome. (**a**) Principal component analysis (PCA) plot of principal components 1 and 2 (PC1 and PC2) derived from 22,746 pruned imputed variants for the *n* = 48 individuals in this study. The data points are coloured based on the self-reported primary ethnicity of the participants and the shapes represent self-reported sub-ethnicity. A bar plot of the relative variance contributions in the top 20 PCs is also shown with PC1 and PC2 accounting for 9.33% and 7.47% of the variation in the top 20 PCs, respectively. (**b**) A karyoplot of human chromosome 14 (HSA14) showing the relationship between $\log_2$ (TPM + 1) normalised gene expression from the pooled RNA-seq samples across

all n = 48 individuals and imputation performance for 267,985 HSA14 variants (Dosage $R^2 \geq 0.25$) originally called by DeepVariant (purple) or imputed with Beagle (steel blue).

Differential expression analysis highlights profound transcriptional changes that occur during anti-TB treatment

The original study by Tabone, et al. [41] compared transcriptional perturbations between treated individuals and healthy controls. Here, using a paired design, we tested a related but distinct hypothesis: do significant transcriptional changes occur in PB following ATT compared to the same individuals with active TB? We observed that substantial changes occur in the transcriptional profiles of PB during ATT with 154 genes significantly differentially expressed (DE) ($P_{adj.} < 0.05$, $|LFC| > 0.2$) at T1 versus T0 (**Fig. 3a**). Of these, 77 genes exhibited increased expression and 77 exhibited decreased expression in the T1 group. The number of significant DE genes increased during ATT up to a maximum of 1,367 and 1,847 genes that exhibited increased and decreased expression, respectively, in the T3 group versus the T0 group. However, in the comparison between T4 and T0, the number of significant DE genes reduced to 1,212 and 1,592 genes exhibiting increased and decreased expression, respectively, in the T4 group (**Fig. 3a**; **Supplementary Table 6**).

In all groups, gene-set enrichment analysis (GSEA) revealed a larger number of significantly impacted pathways (SIPs) resulting from genes with increased expression (3025) than from genes with decreased expression (53) (**Fig. 3b**; **Supplementary Table 7**). For genes with decreased expression in the treated groups, the SIPs included the *Immunoglobulin complex* GO:CC ontology ($P_{adj.} = 1.17 \times 10^{-77}$) in the T2 vs. T0 set of genes; the *Immune response to tuberculosis* WikiPathways term ($P_{adj.} = 1.32 \times 10^{-8}$) in the T3 vs. T0 gene set; and the *AIM2 inflammasome complex assembly* GO:BP ontology ($P_{adj.} = 1.23 \times 10^{-6}$) in the T4 vs. T0 gene set. For genes with increased expression, the SIPs included the *LTC4-CYSLTR mediated IL4 production* Reactome term ($P_{adj.} = 2.15 \times 10^{-2}$) in the T2 vs. T0 gene set; the *B cell proliferation* GO:BP ontology ($P_{adj.} = 2.09 \times 10^{-2}$) in the T3 vs. T0 gene set; and the *B cell receptor signaling pathway* KEGG term ($P_{adj.} = 3.07 \times 10^{-2}$) in the T4 vs. T0 gene set. The complete results for each GSEA are detailed in **Supplementary Table 7.**

In total, across all contrasts, we identified 7,246 significantly DE genes of which 2,942 exhibited increased expression and 4,334 exhibited decreased expression in the treated groups (T1–T4) compared to the naïve untreated group (T0). **Fig. 3c** illustrates the overlap of significant DE genes detected within each contrast separated by their direction of change relative to T0. We observed a total of 103 DE genes that were consistently DE in the same direction across all contrasts; 72 were consistently decreased and 31 were consistently increased in the treated groups versus the untreated group (**Fig. 3c**, **Fig. 3d**). Of these consistently DE genes, 15 genes (*FAM20A*, *CD274*, *CASP5*, *TIFA*, *FCGR1A*, *C1QC*, *SDC3*, *C1QB*, *P2RY14*, *GRAMD1C*, *GBP6*, *PDCD1LG2*, *GRIN3A*, *GBP1P1* and

FCGR1CP) overlapped with the 27 TREAT-TB gene signature developed by Tabone, et al. [41]. The number of overlapping genes observed was significantly more than what would be expected by chance ($P < 2.2 \times 10^{-16}$; Fisher's exact test), supporting the hypothesis that the expressed transcripts of these genes may serve as useful biomarkers for profiling the PB transcriptional response during anti-TB therapy.

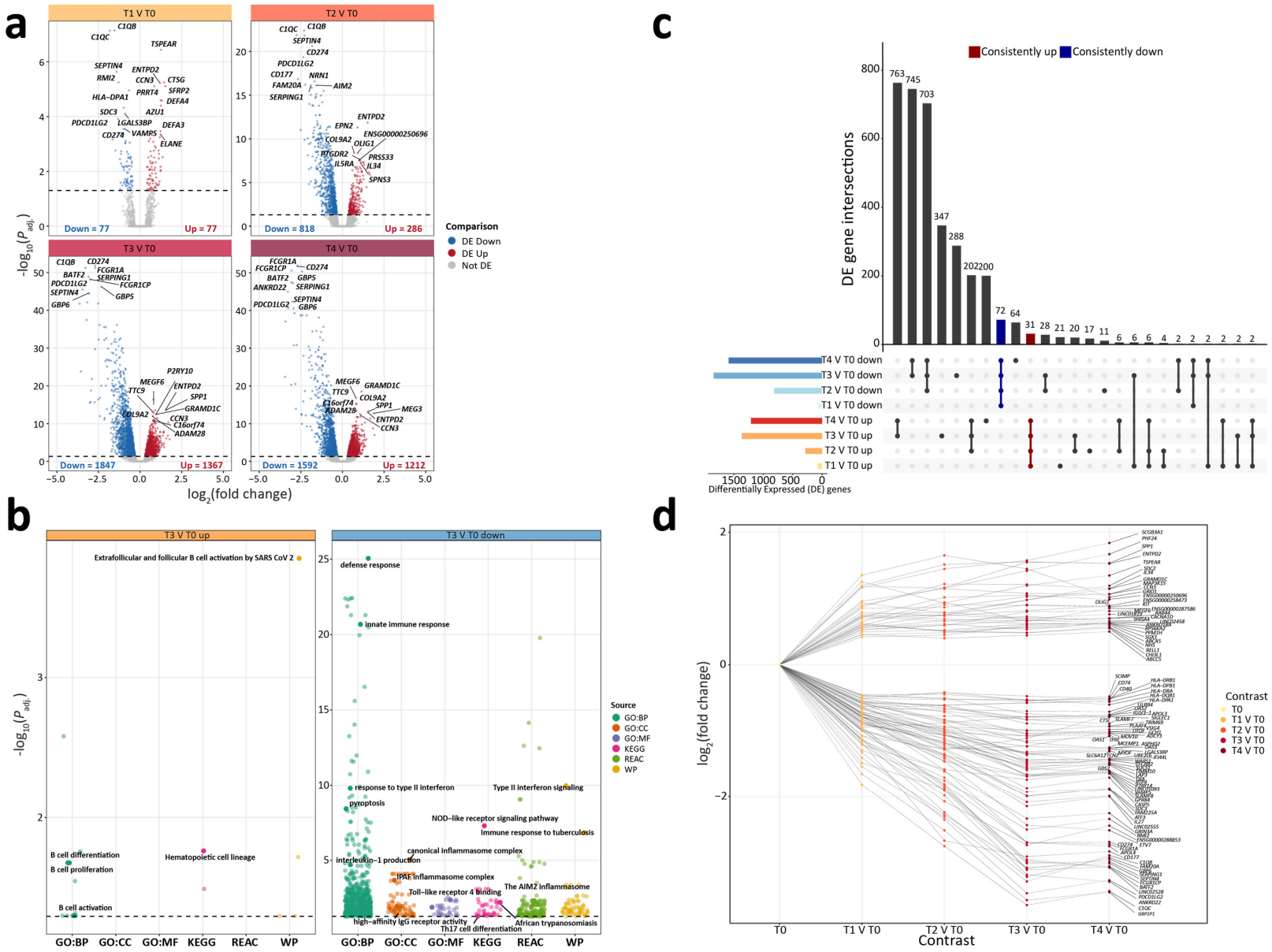

a
T1 V T0
T2 V T0
T3 V T0
T4 V T0
Down = 77
Up = 77
Down = 818
Up = 286
Down = 1847
Up = 1367
Down = 1592
Up = 1212
log2(fold change)
Comparison
DE Down
DE Up
Not DE
b
T3 V T0 up
T3 V T0 down
Extrafollicular and follicular B cell activation by SARS CoV 2
B cell differentiation
B cell proliferation
B cell activation
Hematopoietic cell lineage
defense response
innate immune response
response to type II interferon
pyroptosis
Type II interferon signaling
NOD-like receptor signaling pathway
Immune response to tuberculosis
canonical inflammasome complex
interleukin-1 production
IPAF inflammasome complex
Toll-like receptor 4 binding
The AIM2 inflammasome
high-affinity IgG receptor activity
Th17 cell differentiation
African trypanosomiasis
GO:BP
GO:CC
GO:MF
KEGG
REAC
WP
Source
c
Consistently up
Consistently down
DE gene intersections
T4 V T0 down
T3 V T0 down
T2 V T0 down
T1 V T0 down
T4 V T0 up
T3 V T0 up
T2 V T0 up
T1 V T0 up
Differentially Expressed (DE) genes
d
log2(fold change)
Contrast
T0
T1 V T0
T2 V T0
T3 V T0
T4 V T0

Fig. 3: Paired differential expression analysis highlights substantial perturbation of the transcriptome implicating immunologically relevant pathways. (**a**) Volcano plot for each of the differential expression contrasts conducted. The *y*-axis indicates the $-\log_{10}(P_{adj.})$ value and the *x*-axis shows the $\log_2$ fold change (LFC) for each gene between treated timepoints(T1–T4) compared to the untreated timepoint (T0). Genes significantly increased in expression ($P_{adj.} < 0.05$ and LFC > 0.2) are coloured red whereas genes significantly decreased in expression ($P_{adj.} < 0.05$ and LFC < 0.2) are coloured blue. Non-significant genes ($P_{adj.} > 0.05$) are coloured grey. (**b**) Jitter plot of significantly enriched pathways from g:Profiler for the top 300 differentially expressed increased (orange) and decreased (blue) genes from the analysis of T3 versus T0. The *y*-axis shows the $-\log_{10}(P_{adj.})$, the *x*-axis and the data point colours illustrates the respective databases from which significant pathways/ontologies originate: GO:BP (Gene Ontology Biological Processes), GO:CC (Gene Ontology Cellular Compartment), GO:MF (Gene Ontology Molecular Function), KEGG (Kyoto Encyclopaedia of Genes and Genomes), REAC (Reactome), and WP (WikiPathways). (**c**) Upset plot showing the overlap of significantly differentially expressed increased (yellow to red) and decreased (light blue to blue) genes across all contrasts. A total of 72 consistently decreased and 31 consistently increased genes are highlighted in dark blue and red, respectively. (**d**) Time-series line plot showing the LFC relative to T0 displayed on the *y*-axis at each contrast displayed on the *x*-axis for 103 consistently significant DE genes ($P_{adj.} < 0.05$, $|LFC| \geq 0.2$).

Mapping *cis*-expression QTL reveals the genomic architecture of the transcriptional response during TB disease and anti-TB treatment

Using TensorQTL, we mapped variants proximal to each gene to identify *cis*-eQTL. As *cis*-eQTL generally are arbitrarily defined as being within a ±1 Mb interval relative to the associated gene, our imputed variants are largely restricted clustered within or proximal to expressed regions of the genome (**Fig. 2b**; **Supplementary Fig. S6**), and we surmised that this may have implications for selecting an optimum window size for mapping *cis*-eQTL. We therefore selected various *cis* window sizes around the transcriptional start site (TSS) ranging from 0.01 to 1 Mb and mapped corresponding *cis* variants in each group accounting for the top three genotype PCs, age, days since ATT commencement, and between six and nine transcriptomic PCs derived from PCA4QTL to account for hidden confounders (**Supplementary Fig. S7**; **Supplementary Tables S8-S12**).

Assessing the number of unique *cis*-eGenes (i.e., a gene classified as having a significant variant ($P_{adj.} < 0.1$) in *cis*, a window size of ± 50 kb away from the TSS was determined to be optimal for the data set as it resulted in the largest number of unique *cis*-Genes being identified across all five timepoints **(Supplementary Fig. S8**). In total for this window size, we identified 725, 700, 767, 718, and 748 *cis*-eGenes in the T0, T1, T2, T3, and T4 groups, respectively ($P_{adj.} < 0.1$) (**Fig. 4a**; **Supplementary Table S13–S17**). Of these *cis*-eGenes, between 26.8–33.7% were timepoint-specific, and 208 were considered significant across all five groups, which accounted for between 27.1–29.7% of all *cis*-eGenes detected in each group. (**Fig. 4a**). We characterised 6,888, 6,551, 7,828, 6,985, and 7,213 significant *cis*-eVariants with a nominal *P*-value below the genome-wide significance threshold in the T0, T1, T2, T3, and T4 groups, respectively (**Supplementary Fig. S9**). Pleiotropic effects were evident because of these, between 4.0–7.6% were significantly associated with more than one *cis*-eGene.

The median effect size of top *cis*-eQTLs identified across all five timepoints was consistent with power calculations and ranged from 0.749–0.794. Additionally, we observed no significant difference in the distribution of absolute effect size estimates of top *cis*-eQTLs between timepoints (Wilcoxon rank-sum test; $P_{adj.} > 0.05$, **Supplementary Fig. S10**). These observations indicated that our *cis*-eQTL results are biased towards common variants with large effects on gene expression levels. We further compared our *cis*-eQTL results to those reported by the INTERVAL cohort[72] to assesses external replication of our findings. We observed consistently high allelic concordance (AC) estimates across timepoints ranging from 0.966–0.981. We also noted high π_1 statistics across timepoints ($T0\pi_1 = 0.944$ $T1\pi_1 = 0.946$, $T2\pi_1 = 0.964$, $T3\pi_1 = 0.969$, $T4\pi_1 = 0.919$) indicating good external replication rates. Lastly, we observed highly significant and strongly positive Spearman correlation coefficients between effect size estimates of *cis*-eQTLs identified in each timepoint and

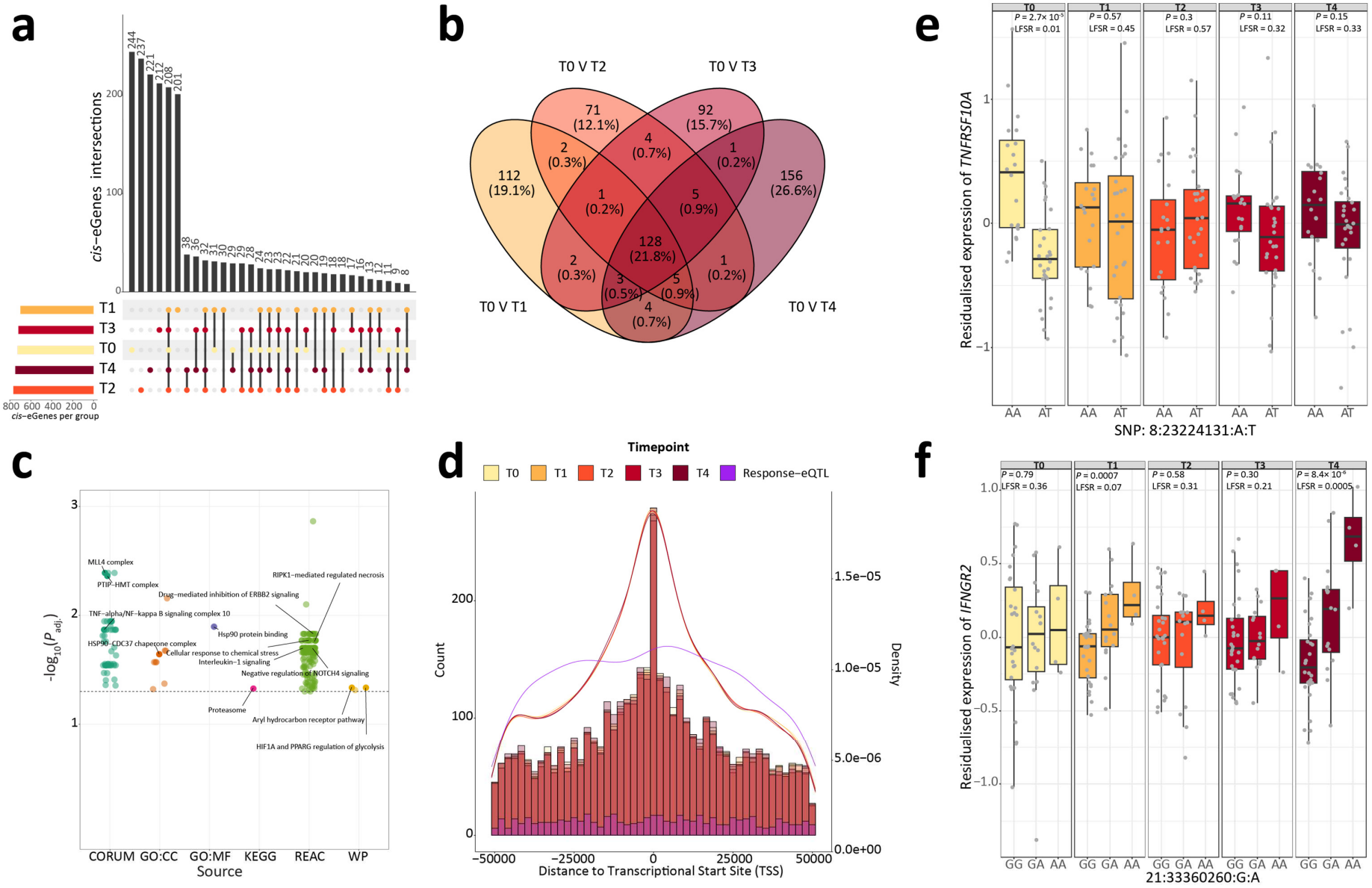
effect size estimates of matched variant-gene pairs reported in the INTERVAL cohort ($\rho_{spearman} \geq 0.798$; $P \leq 3.43 \times 10^{-107}$) (**Supplementary Fig. S11**)

The power in our study's experimental design lies in its longitudinal nature; therefore, we combined information across groups to maximise the statistical power for detecting *cis*-eQTL. To do this, we followed the procedure implemented by Lin, et al. [73] and used mashr[74] to recompute effect size and standard error estimates of *cis*-eQTL using inherent correlations shared between the summary statistics from each of the groups. Using this approach, we identified 4,061, 4,049, 3,954, 3,982, and 4,137 significant *cis*-eQTL (local false sign rate (LFSR) < 0.05) in the T0, T1, T2, T3, and T4 groups, respectively (**Supplementary Table S18**). In total, across all five groups, we identified 5,129 significant *cis*-eQTL of which 3,626 (70.7%) were shared across the five groups. The variance of the posterior mean effect sizes estimated using mashr was significantly lower than the variance of the effect size estimates obtained with TensorQTL (two-sided *F*-test; $P < 2.2 \times 10^{-16}$) across all groups. Similarly, the posterior standard deviation estimates from mashr were significantly lower than the TensorQTL standard error estimates for each group (two-sided paired *t*-test; $P < 2.2 \times 10^{-16}$) (**Supplementary Fig. S12**). We observed a strongly positive spearman correlation ($\rho_{spearman} \geq 0.808$; $P \leq 2.2 \times 10^{-16}$) between the posterior mean effect size estimates of significant *cis*-eQTLs (LFSR < 0.05) and the reported effect size of matched variant-gene pairs from the INTERVAL cohort (**Supplementary Fig. S13**).

We next wanted to characterise variants that change their effect due to the commencement of ATT (i.e., reQTL). To do this, we performed pairwise comparisons between the naïve untreated group (T0) and each of the treated groups (T1–T4) and defined an reQTL as one that was significant in at least one group (LFSR < 0.05), and where the magnitude of the absolute fold difference in the posterior mean effect size of the variant was ≥ 2 in a respective comparison. Using this approach, from each of the four comparisons, we characterised 587 unique reQTL. Of these, 431 (73.4%) were timepoint-specific and 156 (26.6%) were shared across more than one comparison indicating temporal and persistent effects of genetic variation on gene expression due to the ATT (**Fig. 4b**; **Supplementary Table S19–S22**). To assess if any of the genes associated with these reQTL were involved in biologically important pathways, we performed a GSEA using g:Profiler and ordered genes based on their absolute magnitude fold difference relative to T0. In total, we identified 217 SIPs ($P_{adj} < 0.05$) that encompassed, for example, transcriptional regulation, epigenetic reprogramming, the immune response during *M. tuberculosis* infection, and host response to xenobiotic treatment. These SIPs included the *TNF-alpha/NF-kappa B signaling complexes 6*, *7*, *8*, *9* and *10* ($P_{adj.} \leq 1.54 \times 10^{-2}$); the *PTIP-HMT complex* ($P_{adj.} = 4.3 \times 10^{-3}$) from the CURAM database; the *HSP90-CDC37 chaperone complex* ($P_{adj.} = 2.1 \times 10^{-2}$) from GO:CC; and the *Aryl hydrocarbon*

receptor ($P_{adj.}$ = 4.6 × 10^{-2}), *IPAF inflammasome* ($P_{adj.}$ = 3.0 × 10^{-2}), and *Cellular response to chemical stress* ($P_{adj.}$ = 1.7 × 10^{-2}) Reactome pathways (**Fig. 4c**). The complete results from this GSEA are detailed in **Supplementary Table S23**.

Similar to Lin, et al. [73], we observed that significant *cis*-eQTL (LFSR < 0.05) were enriched around the TSS with no particular trend in the 5' or 3' direction, whereas the reQTL tended to be more uniformly distributed across the genomic interval window (**Fig. 4d**). However, none of the *cis*-eQTL or reQTL, from any of the groups, were normally distributed around the TSS (Shapiro-Wilk normality test; $P \leq 7.8 \times 10^{-11}$ for all QTL sets). Our classification of 587 reQTL allowed us to characterise 137 active TB-specific reQTL that were significant (LFSR < 0.05) only in the T0 group and 438 treatment-induced reQTL that were significant in at least one of the treated groups but not T0. There were no significant reQTL present in all five groups. The active TB-specific response-eGenes included *TNFRSF10A* (**Fig. 4e**), *GPX4*, and *NAALADL1*. Response eGenes associated with ATT commencement included *RAB27A* at T1, *FBXO3* at T2, *PSMD3* at T3, and *IFNGR2* at T4 (**Fig. 4f**).


a
cis–eGenes intersections
244
237
221
212
208
201
38
36
32
31
30
29
29
28
24
23
23
22
21
20
20
19
18
18
17
16
13
12
11
9
8
200
100
0
T1
T3
T0
T4
T2
800 600 400 200 0
cis–eGenes per group
b
T0 V T2
T0 V T3
71 (12.1%)
92 (15.7%)
4 (0.7%)
2 (0.3%)
1 (0.2%)
112 (19.1%)
156 (26.6%)
1 (0.2%)
5 (0.9%)
128 (21.8%)
2 (0.3%)
1 (0.2%)
3 (0.5%)
5 (0.9%)
4 (0.7%)
T0 V T1
T0 V T4
e
T0
T1
T2
T3
T4
P = 2.7× 10-5
LFSR = 0.01
P = 0.57
LFSR = 0.45
P = 0.3
LFSR = 0.57
P = 0.11
LFSR = 0.32
P = 0.15
LFSR = 0.33
Residualised expression of TNFRSF10A
1
0
−1
AA AT AA AT AA AT AA AT AA AT
SNP: 8:23224131:A:T
c
-log10(Padj.)
3
2
1
MLL4 complex
PTIP–HMT complex
Drug–mediated inhibition of ERBB2 signaling
TNF–alpha/NF–kappa B signaling complex 10
HSP90–CDC37 chaperone complex
Hsp90 protein binding
Cellular response to chemical stress
Interleukin–1 signaling
RIPK1–mediated regulated necrosis
Negative regulation of NOTCH4 signaling
Proteasome
Aryl hydrocarbon receptor pathway
HIF1A and PPARG regulation of glycolysis
CORUM GO:CC GO:MF KEGG REAC WP
Source
d
Timepoint
T0
T1
T2
T3
T4
Response–eQTL
Count
200
100
0
Density
1.5e−05
1.0e−05
5.0e−06
0.0e+00
−50000 −25000 0 25000 50000
Distance to Transcriptional Start Site (TSS)
f
T0
T1
T2
T3
T4
P = 0.79
LFSR = 0.36
P = 0.0007
LFSR = 0.07
P = 0.58
LFSR = 0.31
P = 0.30
LFSR = 0.21
P = 8.4× 10-6
LFSR = 0.0005
Residualised expression of IFNGR2
1.0
0.5
0.0
−0.5
−1.0
GG GA AA GG GA AA GG GA AA GG GA AA GG GA AA
21:33360260:G:A

610

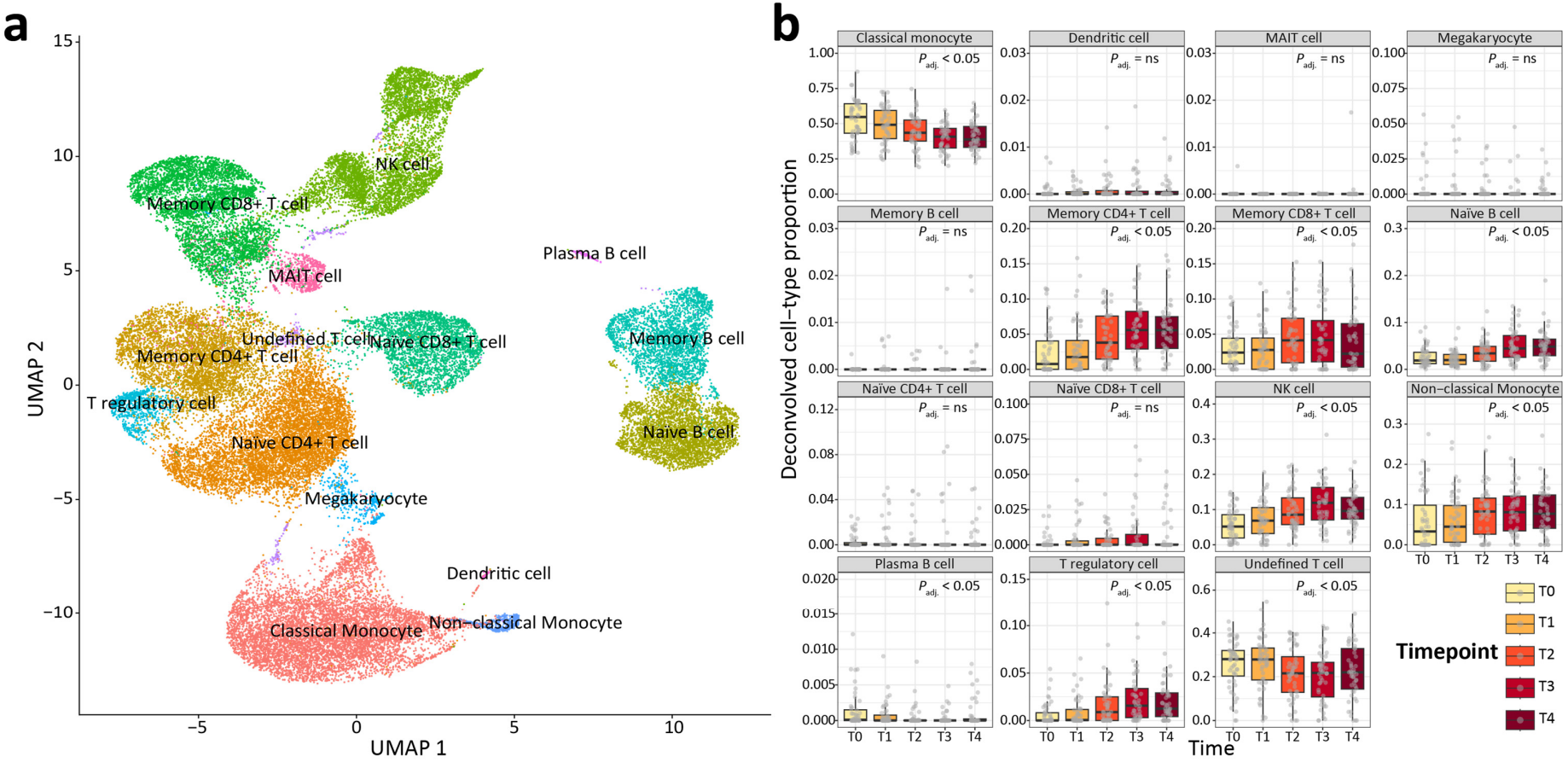
Fig. 4: Mapping of *cis* and response-eQTL illustrates genetic control of the transcriptional response to anti-TB treatment in peripheral blood. (**a**) Upset plot showing the overlap of significant *cis*-eGenes ($P_{adj.} < 0.1$) among the five timepoints (T0–T4). The coloured bar plots show the number of *cis*-eGenes detected at each timepoint. (**b**) Venn diagram illustrating the overlap of significant response-eQTL (reQTL) determined from each of the four treatment timepoints (T1–T4) versus the untreated group (T0). An reQTL was determined as being significant in at least one group (mashr local false sign rate (LFSR) < 0.05) and if the absolute magnitude of fold difference between the two groups in each contrast was ≥ 2. (**c**) Jitter plot of significantly enriched pathways from g:Profiler for all response-eGenes identified. The *y*-axis shows the $-\log_{10}(P_{adj.})$ and the *x*-axis and data point colours indicate the pathway/gene ontology databases. (**d**) Distance of all reQTL (purple) and significant *cis*-eQTL (LFSR < 0.05) to the transcriptional start site (TSS) of the target gene identified in each group separately. The overlaid histograms show the distribution in each group of eQTL. The curves show the kernel density estimate for each group of eQTL. (**e**) An example of an active-TB specific response *cis*-eQTL. The bar plot shows the residualised expression for *TNFRSF10A* with individuals plotted based on their genotype at the 8:23224131 locus. Raw *P*-values from the association analysis with TensorQTL and mashr LFSRs are also detailed for each timepoint. (**f**) An example of a treatment-induced response *cis*-eQTL. The bar plot shows the residualised expression for *IFNGR2* with individuals plotted based on their genotype at the 21:33360260 locus. Raw *P*-values from the association analysis with TensorQTL and mashr LFSRs are also detailed for each timepoint.

Single-cell RNA-seq analysis and computational deconvolution of bulk RNA-seq data

To computationally deconvolve our bulk RNA-seq samples and assess changes in cell type proportions during ATT, we first downloaded scRNA-seq data from peripheral blood mononuclear cells (PBMCs) obtained from $n = 2$ latent TB infection (LTBI) patients and $n = 3$ active TB cases. Following filtering of low-quality cells, removal of ambient RNA, and discarding putative doublets, a total of 44,009 cells (6,894–10,378 cells per sample) remained that were clustered into 15 major annotated cell subtypes (**Fig. 5a**). These cells were annotated using a comprehensive procedure involving automatic reference-based annotation with SingleR, marker-assisted annotation using SCINA, and marker-based annotation using published information on cell type gene markers[88,89] (**Supplementary Table S24-S25**). **Supplementary Table S26** details the annotation for each of the 44,009 cells based on the annotation procedure and the final annotation illustrated in **Fig. 5a**. From these clusters, we derived a signature matrix of 760 genes to use as reference for the deconvolution of the bulk RNA-seq data (**Supplementary Table S27)**.

Prior to deconvolving the PB RNA-seq data, we performed a benchmarking assessment on pseudo-bulk RNA-seq data for four deconvolution algorithms: Non-negative least squares (NNLS), MuSic, Bisque, and CIBERSORT with various parameter modifications totalling 12 different analyses (See Methods; **Supplementary Note 1**). From this analysis, based on the average Pearson correlation (ρ) and average root mean square error (RMSE) estimates across all 15 cell types, CIBERSORT, when specifying the raw scRNA-seq signature matrix and the CPM-normalised pseudo-bulk RNA-seq data, was determined to be the best performing algorithm ($\rho = 0.998$, RMSE = 0.003) (**Supplementary Note 1**).

Based on the results of the deconvolution benchmarking assessment, we deconvolved our CPM-normalised bulk RNA-seq data using CIBERSORT, specifying the signature matrix derived from the scRNA-seq data as the reference. The results of the deconvolution for each of the $n = 240$ bulk RNA-seq samples are presented in **Fig. 5b** and **Supplementary Table S30**. We observed significant changes (ANOVA F-test; $P_{adj.} < 0.05$) in cell type proportions during ATT. For example, natural killer (NK) cells significantly increased over time ($P_{adj.} = 4.92 \times 10^{-13}$), whereas classical monocytes ($CD14^{+}CD16^{-}$) significantly decreased over time ($P_{adj.} = 7.98 \times 10^{-12}$). Memory $CD4^{+}$ T cells significantly increased over time ($P_{adj.} = 4.42 \times 10^{-13}$), as did naïve B cells ($P_{adj.} = 7.92 \times 10^{-13}$) and T regulatory cells ($P_{adj.} = 1.39 \times 10^{-8}$). Other cell types that displayed significant changes included plasma B cells ($P_{adj.} = 4.49 \times 10^{-3}$), memory $CD8^{+}$ T cells ($P_{adj.} = 1.47 \times 10^{-2}$) and undefined T cells ($P_{adj.} = 2.23 \times 10^{-2}$). The complete results from the PB cell type deconvolution analysis are provided in **Supplementary Table S31**.


a
UMAP 2
UMAP 1
NK cell
Memory CD8+ T cell
MAIT cell
Plasma B cell
Undefined T cell
Naive CD8+ T cell
Memory B cell
Memory CD4+ T cell
T regulatory cell
Naive B cell
Naive CD4+ T cell
Megakaryocyte
Dendritic cell
Classical Monocyte
Non−classical Monocyte
b
Deconvolved cell−type proportion
Classical monocyte
Dendritic cell
MAIT cell
Megakaryocyte
Memory B cell
Memory CD4+ T cell
Memory CD8+ T cell
Naïve B cell
Naïve CD4+ T cell
Naïve CD8+ T cell
NK cell
Non−classical Monocyte
Plasma B cell
T regulatory cell
Undefined T cell
P adj. < 0.05
P adj. = ns
T0 T1 T2 T3 T4
Time
Timepoint

657

Fig. 5: Digital flow cytometry analysis using a scRNA-seq reference shows profound changes in cell type abundance due to anti-TB treatment. (**a**) Uniform manifold approximation and projection (UMAP) of 44,009 cells derived from PBMC scRNA-seq data from *n* = 2 LTBI individuals and *n* = 3 active TB patients. Cells are clustered and coloured into 15 major cell types that are denoted on each cluster. (**b**) Deconvolution results from CIBERSORT for all *n* = 240 bulk tissue samples based on the scRNA-seq reference matrix derived from (**a**) and consisting of 15 major cell types. Significant difference in cell type abundance was inferred based on an ANOVA *F*-test implemented in limma, where $P_{adj.}$ value < 0.05 indicate significant changes over time and "ns" indicates non-significance. Boxplots are coloured based on sampling timepoint. Horizontal lines inside the boxplots show the medians. Box bounds show the lower quartile (Q1, the 25th percentile) and the upper quartile (Q3, the 75th percentile). Whiskers are minima (Q1 – 1.5 × IQR) and maxima (Q3 + 1.5 × IQR) where IQR is the interquartile range (Q3–Q1). NK, natural killer; MAIT, mucosal-associated invariant T-cell; and DC, dendritic cell.

Cell type interaction eQTL (ieQTL) mapping

To interrogate the putative cell type specificity of *cis*-eQTL, and to characterise interaction *cis*-eQTL (ieQTL), we leveraged the deconvolved computational abundance estimates of well-characterised and common (median proportion ≥ 0.01 across all timepoints) cell types from CIBERSORT (classical monocytes ($CD14^+CD16^-$), non-classical monocytes ($CD14^+CD16^+$), memory $CD8^+$ T cells, naïve B cells, and NK cells). Fitting an interaction term between genotype and inverse normally transformed cell type abundance estimates in the linear model using TensorQTL accounting for the same covariates used in the standard *cis*-eQTL analysis, we identified no significant interaction terms after multiple testing correction ($P_{adj.} > 0.05$). We therefore employed a similar approach to increasing our power in the mapping of *cis*-eQTL in bulk tissue and used mashr to recompute effect size and standard error estimates of this interaction term across timepoints for each cell type separately. We defined a threshold of LFSR < 0.01 as evidence for a significant interaction effect between genotype and cell type abundance.

From this analysis, across all timepoints, we identified a total of 185 significant ieQTL (LFSR < 0.01) in classical monocyte cells. Similarly, across all timepoints, 220, 186, 220, and 110 ieQTL were observed in memory $CD8^+$ T cells, naïve B cells, NK cells, and non-classical monocytes, respectively (**Fig. 6a**). In total, across all cell types and timepoints, we identified significant interaction terms for 921 variant-gene-cell pairs (**Fig. 6a**; **Supplementary Table S32)**. We observed that ieQTL had a more uniform distribution around the *cis* window, similar to reQTL (**Fig. 4d**) across all cell types (**Fig. 6b**). Characterisation of significant ieQTL allowed us to identify a total of 115 active TB-specific ieQTL (**Supplementary Table S32)** across all five cell types tested. For example, we observed a significant classical monocyte-specific ieQTL for *NOD2* (variant = 16:50652646:A:C, $P_{g\times c} = 6.4 \times 10^{-5}$, LFSR = 0.0071) that was only significant at T0 (**Fig. 6c**). Other examples of active TB-specific ieQTL included ieQTL for *NLRC4* in classical monocytes (variant = 2:32236868:A:AAAAG, P_{GxC} = 0.0003, LFSR = 0.007), *NLRC3* in NK cells (variant = 16:3560641:T:C, $P_{g\times c}$ = 0.0002, LFSR = 0.007), and *IFITM2* in memory $CD8^+$ T cells (variant = 11:296203:A:G, $P_{g\times c}$ = 0.0003, LFSR = 0.003). We also identified treatment-induced ieQTL effects that were not significant in active TB (T0), including a significant ieQTL identified for *ATG4B* in non-classical monocytes early in ATT at T1 (variant = 2:241678377:A:G, $P_{g\times c}$ = 0.013, LFSR = 0.008) (**Fig. 6d**). Other treatment-induced interaction *cis*-Genes included *TNFAIP3*, *ABCA7*, and *S100A6*. The complete results of the ieQTL analysis are detailed **Supplementary Table S32**. Taken together, our results suggest that cell type-specific *cis*-eQTL are a feature of active TB disease and during ATT.

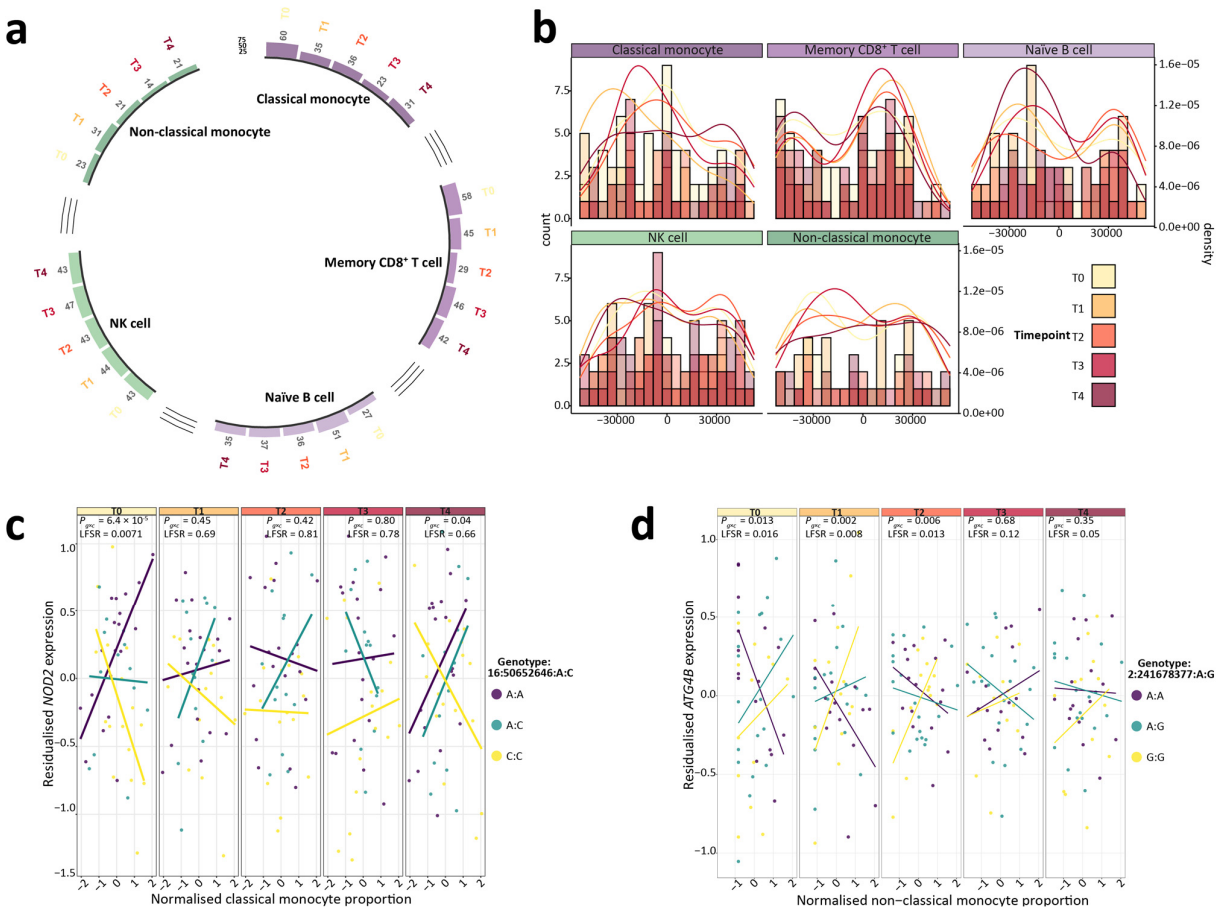


Fig. 6: Unearthing the cell type specificity of *cis*-eQTL during active TB and anti-TB treatment through interaction eQTL mapping. (**a**) Circular bar plot showing the number of statistically significant (local false sign rate (LFSR) < 0.01) cell type interaction *cis*-eQTL (ieQTL) identified in each cell type (classical monocytes, memory CD8$^+$ T cells, naïve B cells, and NK cells) across each of the five timepoints (T0–T4) separately. (**b**) Histogram organised into 5 kb bins showing distance of the significant ieQTL (LFSR < 0.01) to the transcriptional start site (TSS) of the target gene that were identified in each cell type. The overlaid histograms show the distribution for each timepoint. The curves show the kernel density estimate for each group of eQTL with scales represented on the second *y*-axis. (**c**) Example of an active TB classical monocyte-specific ieQTL identified for *NOD2* with the SNP 16:50652646:A:C. Scatter plot shows the residualised expression values for *NOD2* on the *y*-axis and the inverse normally transformed classical monocyte proportions on the *x*-axis with data points coloured based on corresponding genotype. Solid lines indicate the line of best fit for each genotype derived from the *geom_smooth* function in ggplot2. $P_{g\times c}$-values for the interaction term from TensorQTL are also shown as are the LFSR values obtained from mashr. (**d**) Example of a treatment-induced ieQTL identified in non-classical monocytes for the gene *ATG4*B with the SNP 2:241678377A:G in the T1 group. Scatter plot shows the residualised expression values for *ATG4B* on the *y*-axis and the inverse normally transformed non-classical monocyte proportions on the *x*-axis with data points coloured based on corresponding genotype. Solid lines indicate the line of best fit for each genotype derived from the *geom_smooth* function in ggplot2. $P_{g\times c}$-values for the interaction term from TensorQTL are also shown, as are the LFSR values obtained from mashr.

Discussion

Here, we have performed the first *in vivo* high-resolution genome-wide *cis*-eQTL, reQTL, digital flow cytometry, and cell type *cis*-ieQTL mapping analysis in the context of active TB disease and ATT using publicly available RNA-seq and scRNA-seq data. Our work illustrates the profound transcriptional and cell type abundance changes that occur during ATT and sheds light on how genomic variation influences this response, while also providing a framework for other studies that wish to exploit PB transcriptomics data to its full potential.

Variant calling from high-resolution transcriptomics data

An abundance of public PB-derived transcriptomics data now exists. For example, in case of human TB, this includes the Berry cohort[98], the Adolescent Cohort Study (ACS) and the Grand Challenges 6-74 Cohort Study (GC6-74) analysed by Zak, et al. [99], and the Leicester cohort used for the present study and previously for studying the transcriptional landscape during exposure to *M. tuberculosis*, active TB disease or LTBI, and during or post-ATT[41,100]. These data sets have been used primarily for identifying transcriptional biosignatures. Here, however, we show that calling variants directly from these RNA-seq data for the purposes of molecular QTL mapping is a powerful approach to decipher the genomic architecture underpinning the shared and divergent transcriptional responses within each of these phenotypic contexts.

We characterised 1,506,948 high-quality imputed and common variants through pooling of longitudinal RNA-seq data for individual TB patients across timepoints and showed that LD pruning and analyses of these variants recovered known population genetic structure[101] (**Fig. 2a**). However, our ability to impute distal variants located far away from expressed regions of the genome is significantly curtailed, likely because the imputation algorithm requires an even distribution of variants across the genome, which enables identification of long shared haplotypes in the reference set that can be imputed in the target set with high confidence[102]. In this study, because we are limited to RNA-seq data, our genotyped variants are largely restricted to, and primarily clustered within expressed coding regions that may not conform to default assumptions underpinning genome-wide imputation. We showed that this can impact a *cis*-eQTL analysis, primarily in terms of selecting an optimum *cis* window size (**Supplementary Fig. S8**). Future work to assess imputation strategies that capture extragenic polymorphisms using variants called from RNA-seq data will be important in this regard.

Expression QTL mapping reveals the genetic architecture of the transcriptional response to anti-TB treatment

Our mapping of reQTL identified hundreds of genomic variants that change their effect due to ATT (**Fig. 4b**) and we showed that the genes putatively regulated by these reQTL are involved in key immune and drug response pathways such as the *IL-1 signalling pathway*, *cellular response to chemical stress ontology*, and the *aryl hydrocarbon receptor* Wikipathway (**Fig. 4c**). The latter pathway is noteworthy, specifically in the context of ATT, because in zebrafish (*Danio rerio*) infected *in vivo* with *Mycobacterium marinum* and treated with rifabutin, inhibition of the aryl hydrocarbon receptor pathway with a small molecule inhibitor significantly enhanced bacterial killing compared to controls[103]. Therefore, targeting this pathway in humans, particularly those with variants associated with increased expression of the genes involved, may be a valuable strategy for host-directed therapy and for maximising the effectiveness of current anti-TB therapeutics and development of personalised treatment regimes. Interestingly, similar to Lin, et al. [73], we observed that reQTL tended to be uniformly distributed around the TSS with no particular enrichment at any locus, which is in contrast to standard *cis*-eQTL mapping where these eQTL showed clear enrichment around the TSS (**Fig. 4d**). These results suggest that functionally relevant polymorphisms implicated in TB disease and ATT response, at least in PB, may be located relatively far from the TSS. This is consistent with previous observations from analysis of DNase I hypersensitive sites, which show a large proportion of trait-associated polymorphisms from human GWAS studies reside in distal gene regulatory elements[104].

Computational deconvolution of bulk RNA-seq data is a robust approach for evaluating changes in cell type proportions during anti-TB treatment

Our deconvolution analysis highlighted significant changes in cell type abundance estimates before, and after commencing ATT that is consistent with previous flow cytometry analyses (**Fig. 5a**). For example, and most notably, the abundance of classical monocytes significantly decreased during ATT ($P_{adj.} < 0.05$). An increase in monocyte proportions has been documented in patients with pulmonary TB disease[105] and a high monocyte to lymphocyte ratio (MLR) has been proposed as a potential biomarker for active TB disease[106]. In addition, a meta-analysis conducted by Adane, et al. [107] demonstrated a significant decrease in the MLR following anti-TB therapy. Previous flow cytometry analysis of PBMCs has also shown that the proportion of circulating classical monocytes ($CD14^+CD16^-$) was significantly higher in patients with active TB compared to those sampled during ATT and in healthy controls[108]. Additionally, the scRNA-seq data analysed in this study reported that myeloid subsets of $CD14^+CD16^-$ monocytes were enriched in active TB patients compared to LTBI patients and healthy controls, respectively[75]. For the present study, we also observed that NK cells significantly increased their proportion during ATT ($P_{adj.} < 0.05$), which aligns with previous reports

that documented a depletion of NK cells during active TB disease[109]. Similarly, the proportion of memory $CD4^+$ T cells significantly increased following ATT ($P_{adj.} < 0.05$). This observation is in agreement with previous reports that show that the majority (~ 95%) of antigen-specific effector T cells die following pathogen clearance allowing a pool of memory $CD4^+$ T cells to remain (thereby increasing their relative proportion) and provide protection against reinfection[110]. This is further supported by our DE gene results, which showed that many of the genes decreased in expression across all contrasts are involved in proinflammatory immune response pathways (**Fig. 3a**; **Fig. 3c**) characteristic of active TB, indicating amelioration of the disease[111]. Naïve B cells also significantly increased their proportion ($P_{adj.} < 0.05$) following ATT, supporting previous observations that highlighted a significant decrease in the frequency of naïve B cells in patients with active TB compared to healthy controls[112]. Again, this is supported by our DE gene results where many of the genes exhibiting increased expression were significantly overrepresented in pathways associated with B cell proliferation and activation **(Fig. 3c**; **Supplementary Table S7**). Taken together, our results demonstrate that *in silico* deconvolution of bulk PB RNA-seq data using CIBERSORT is a powerful and robust approach for evaluating cell type composition changes during ATT.

Interaction eQTL mapping reveals putative cell type specificity of *cis*-eQTL

The deconvolved cell type proportions we obtained facilitated identification of 921 variant-gene pairs significantly interacting with the cell type abundance estimates for the five cell types tested. Of note, we identified a private, active TB specific ieQTL (16:50652646:A:C) in classical monocytes for the nucleotide binding oligomerization domain containing 2 gene (*NOD2*) (**Fig. 6c, Supplementary Table S32**) that was not considered a *cis*-eVariant in the standard *cis*-eQTL mapping. Specifically, we observed that individuals with the alternative allele displayed lower expression of *NOD2* in classical monocytes compared to heterozygotes or individuals homozygous for the alternative allele. The NOD2 protein is a member of the NLR (NOD, leucine-rich repeat (LRR)-containing protein) family of intracellular pattern recognition receptors through its ability to recognise bacterial cell wall components known as muramyl dipeptides[113]. In TB disease, NOD2 plays a central role in recognising *M. tuberculosis* pathogen-associated molecular patterns (PAMPs), and is essential for the activation of innate and adaptive immune responses to mycobacterial challenge through crosstalk with Toll-like receptors (TLRs)[114]. In this regard, numerous studies across diverse populations have suggested that polymorphisms in the *NOD2* gene are associated with susceptibility to TB[115]. Also, *NOD2*$^{-/-}$ mice experimentally infected with *M. tuberculosis* displayed higher bacterial burdens versus wild type mice, had impaired inflammatory responses and reduced concentrations of TNF and IFN-γ cytokines[116]. Considered together, our results suggest that individuals who are homozygous for the alternative allele of the SNP 16:50652646:A:C have reduced expression of

NOD2 in classical monocytes during active TB disease, which may impede their ability to recognise mycobacterial PAMPs upon phagocytosis and contribute to *M. tuberculosis* spread and pathogenesis.

Our results highlighted several treatment-induced ieQTL including the SNP 2:241678377:A:G associated with the expression of the autophagy related 4B cysteine peptidase gene (*ATG4B*) in non-classical monocytes for which the interaction term was significant at T1 (LFSR = 0.007) and that also tended towards significance at T0 (LFSR = 0.014) and T2 (LFSR = 0.011) (**Fig. 6d**). The *ATG4B* gene is a member of the ATG4 gene family encoding proteases that enable efficient autophagosome maturation and cellular autophagy, primarily through delipidation of ATG8 family members that attach to a developing phagophore membrane[117,118]. Considered primarily to be a homeostatic process, autophagy principally involves the breakdown of damaged cellular material through encasement in a double layered membrane organelle known as an autophagosome, which is then degraded further through fusion with an acidic lysosome[119]. Notwithstanding the role of autophagy in homeostasis, it has also emerged, over the past two decades, as a key host defence strategy against invading microbes, including tuberculous mycobacteria[120,121]. For example, in murine macrophages infected with *M. tuberculosis*, inhibition of autophagy through depletion of ATG5 resulted in substantial necrosis and increased host susceptibility to *M. tuberculosis* infection[122]. In the context of *ATG4B*, reduction of *ATG4B* expression levels has been shown to lead to changes in autophagy flux[117] and in murine macrophages infected with the *M. bovis* Bacillus Calmette-Guérin (BCG) strain, *Atg4b* expression and autophagy were shown to be inhibited by a micro-RNA (miR-129-3p) leading to enhanced intracellular survival of *M. bovis* BCG[123]. Collectively, our results indicate that individuals homozygous for the reference allele at the SNP 2:241678377:A:G have reduced expression of *ATG4B* in non-classical monocytes, which may hinder the autophagic capacity of these cells in the early stages of ATT thereby stunting the elimination of *M. tuberculosis*.

Limitations of the study design

There are necessary and inherent limitations associated with our analyses that should be acknowledged. Firstly, given that we have $n = 48$ samples per timepoint, even after using mashr to improve power, our ability to identify *cis*-eQTL is limited to those with large effects. Secondly, our *cis*-eQTL analysis is restricted to ±50 kb from the TSS. As a result, there may be independent upstream or downstream *cis*-eQTL/reQTL/ieQTL located further than 50 kb from the TSS, which influence the expression of the associated gene that we were unable to identify. Thirdly, while GWAS for anti-TB drug induced liver injury have been conducted, these were generated using relatively small sample sizes and summary statistics are not publicly available. As further GWAS are conducted for this trait and made publicly available, our results will form an important resource to colocalise[124] putative *cis*-eQTL or reQTL or to perform an integrative transcriptome-wide association study

(TWAS)[125] to characterise the functional genomic mechanisms underpinning this trait. Fourthly, *cis*-eQTLs explain a small proportion of the heritability in gene expression, whereas the combination of weak *trans*-eQTL effects are hypothesised to contribute a substantial proportion to gene expression variance[126], and represent an integral component of the omnigenic model of complex trait variability[127]. The relatively small sample size in the present study meant that we were unable to characterise distal genomic variants associated with variability in mRNA transcript abundance. However, exploiting the full set of individuals in the Berry[98] and Leicester cohorts[41,100] using our approach may facilitate the identification of *trans*-eQTL within an active TB disease context. Lastly, our deconvolution analysis focused on broad high-level cell types (e.g., classical monocytes and NK cells). As more active TB-specific and ATT-specific scRNA-seq data becomes available coupled with advancements in cell type annotation, *in silico* deconvolution can be performed at higher resolution, which would enable characterisation of significant changes in the fine-grained proportions of cell types due to anti-TB xenobiotic treatment.

Conclusion

Our high-resolution *cis*-eQTL, reQTL, and cell type interaction eQTL analysis has identified factors systematically linking host genomic variation, cell type abundance, and the inter-individual transcriptomic response during active TB disease and ATT. Our results form an important resource for researchers investigating how genetic variability impacts the inter-individual response to *M. tuberculosis* infection and ATT while also providing a computational framework for other studies that are limited to RNA-seq data alone. Further integrative genomics work, linking GWAS and molecular quantitative trait loci data together is required to fully elucidate the genomic architecture of active TB disease and the responses to anti-TB therapeutics.

References

1 World Health Organization. *Global Tuberculosis Report 2024*. (World Health Organization, 2024).

2 Houben, R. M. & Dodd, P. J. The global burden of latent tuberculosis infection: a re-estimation using mathematical modelling. *PLoS Med* **13**, e1002152 (2016). https://doi.org/10.1371/journal.pmed.1002152

3 Comstock, G. W., Livesay, V. T. & Woolpert, S. F. The prognosis of a positive tuberculin reaction in childhood and adolescence. *Am J Epidemiol* **99**, 131-138 (1974). https://doi.org/10.1093/oxfordjournals.aje.a121593

4 Tiemersma, E. W., van der Werf, M. J., Borgdorff, M. W., Williams, B. G. & Nagelkerke, N. J. Natural history of tuberculosis: duration and fatality of untreated pulmonary tuberculosis in HIV negative patients: a systematic review. *PLoS ONE* **6**, e17601 (2011). https://doi.org/10.1371/journal.pone.0017601

5 Nahid, P. *et al.* Official American Thoracic Society/Centers for Disease Control and Prevention/Infectious Diseases Society of America Clinical Practice Guidelines: Treatment of Drug-Susceptible Tuberculosis. *Clin Infect Dis* **63**, e147-e195 (2016). https://doi.org/10.1093/cid/ciw376

6 Gülbay, B. E. *et al.* Side effects due to primary antituberculosis drugs during the initial phase of therapy in 1149 hospitalized patients for tuberculosis. *Respir. Med.* **100**, 1834-1842 (2006). https://doi.org/10.1016/j.rmed.2006.01.014

7 Jiao, L. *et al.* Genetic architecture of tuberculosis susceptibility: a comprehensive research synopsis, meta-analyses, and epidemiological evidence. *Infect Genet Evol* **104**, 105352 (2022). https://doi.org/10.1016/j.meegid.2022.105352

8 Stein, C. M. Genetic epidemiology of resistance to *M. tuberculosis* infection: importance of study design and recent findings. *Genes Immun* **24**, 117-123 (2023). https://doi.org/10.1038/s41435-023-00204-z

9 Petros, Z. *et al.* Genome-wide association and replication study of hepatotoxicity induced by antiretrovirals alone or with concomitant anti-tuberculosis drugs. *OMICS* **21**, 207-216 (2017). https://doi.org/10.1089/omi.2017.0019

10 Nicoletti, P. *et al.* Genetic risk factors in drug-induced liver injury due to isoniazid-containing antituberculosis drug regimens. *Clin Pharmacol Ther* **109**, 1125-1135 (2021). https://doi.org/10.1002/cpt.2100

11 International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299-1320 (2005). https://doi.org/10.1038/nature04226

12 McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* **9**, 356-369 (2008). https://doi.org/10.1038/nrg2344

13 Uffelmann, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers* **1**, 59 (2021). https://doi.org/10.1038/s43586-021-00056-9

14 Morley, M. *et al.* Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743-747 (2004). https://doi.org/10.1038/nature02797

15 Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci* **368**, 20120362 (2013). https://doi.org/10.1098/rstb.2012.0362

16 Aguet, F. *et al.* Molecular quantitative trait loci. *Nature Reviews Methods Primers* **3**, 4 (2023). https://doi.org/10.1038/s43586-022-00188-6

17 Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* **6**, e1000888 (2010). https://doi.org/10.1371/journal.pgen.1000888

18 Nandakumar, S. K., Liao, X. & Sankaran, V. G. In the blood: connecting variant to function In human hematopoiesis. *Trends Genet* **36**, 563-576 (2020). https://doi.org/10.1016/j.tig.2020.05.006

19 GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020). https://doi.org/10.1126/science.aaz1776

20 Liu, S. *et al.* A multi-tissue atlas of regulatory variants in cattle. *Nat Genet* **54**, 1438-1447 (2022). https://doi.org/10.1038/s41588-022-01153-5

21 Teng, J. *et al.* A compendium of genetic regulatory effects across pig tissues. *Nat Genet* **56**, 112-123 (2024). https://doi.org/10.1038/s41588-023-01585-7

22 Jasinska, A. J. *et al.* Genetic variation and gene expression across multiple tissues and developmental stages in a nonhuman primate. *Nat Genet* **49**, 1714-1721 (2017). https://doi.org/10.1038/ng.3959

23 Natri, H. M. *et al.* Cell-type-specific and disease-associated expression quantitative trait loci in the human lung. *Nat Genet* **56**, 595-604 (2024). https://doi.org/10.1038/s41588-024-01702-0

24 Ai, J. W. *et al.* Gene expression pattern analysis using dual-color RT-MLPA and integrative genome-wide association studies of eQTL for tuberculosis suscepitibility. *Respir Res* **22**, 23 (2021). https://doi.org/10.1186/s12931-020-01612-9

25 Barreiro, L. B. *et al.* Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. *Proc Natl Acad Sci U S A* **109**, 1204-1209 (2012). https://doi.org/10.1073/pnas.1115761109

26 Blischak, J. D. *et al.* Predicting susceptibility to tuberculosis based on gene expression profiling in dendritic cells. *Sci Rep* **7**, 5702 (2017). https://doi.org/10.1038/s41598-017-05878-w

27 McHenry, M. L. *et al.* Tuberculosis severity associates with variants and eQTLs related to vascular biology and infection-induced inflammation. *PLoS Genet* **19**, e1010387 (2023). https://doi.org/10.1371/journal.pgen.1010387

28 Nieto-Caballero, V. E. *et al.* History of tuberculosis disease is associated with genetic regulatory variation in Peruvians. *PLoS Genet* **20**, e1011313 (2024). https://doi.org/10.1371/journal.pgen.1011313

29 Umans, B. D., Battle, A. & Gilad, Y. Where are the disease-associated eQTLs? *Trends Genet* **37**, 109-124 (2021). https://doi.org/10.1016/j.tig.2020.08.009

30 Liu, X. *et al.* Functional architectures of local and distal regulation of gene expression in multiple human tissues. *Am J Hum Genet* **100**, 605-616 (2017). https://doi.org/10.1016/j.ajhg.2017.03.002

31 Perez, R. K. *et al.* Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. *Science* **376**, eabf1970 (2022). https://doi.org/10.1126/science.abf1970

32 Yazar, S. *et al.* Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* **376**, eabf3041 (2022). https://doi.org/10.1126/science.abf3041

33 Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* **12**, 453-457 (2015). https://doi.org/10.1038/nmeth.3337

34 Kim-Hellmuth, S. *et al.* Cell type-specific genetic regulation of gene expression across human tissues. *Science* **369** (2020). https://doi.org/10.1126/science.aaz8528

35 Kasela, S. *et al.* Interaction molecular QTL mapping discovers cellular and environmental modifiers of genetic regulatory effects. *Am J Hum Genet* **111**, 133-149 (2024). https://doi.org/10.1016/j.ajhg.2023.11.013

36 Piskol, R., Ramaswami, G. & Li, J. B. Reliable identification of genomic variants from RNA-seq data. *Am J Hum Genet* **93**, 641-651 (2013). https://doi.org/10.1016/j.ajhg.2013.08.008

37 Deelen, P. *et al.* Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels. *Genome Med* **7**, 30 (2015). https://doi.org/10.1186/s13073-015-0152-4

38 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010). https://doi.org/10.1101/gr.107524.110

39 Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* **36**, 983-987 (2018). https://doi.org/10.1038/nbt.4235

40 Cook, D. E. *et al.* A deep-learning-based RNA-seq germline variant caller. *Bioinform Adv* **3**, vbad062 (2023). https://doi.org/10.1093/bioadv/vbad062

41 Tabone, O. *et al.* Blood transcriptomics reveal the evolution and resolution of the immune response in tuberculosis. *J Exp Med* **218** (2021). https://doi.org/10.1084/jem.20210915

42 Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Res* **53**, D20-d29 (2025). https://doi.org/10.1093/nar/gkae979

43 Andrews, S. FastQC: a quality control tool for high throughput sequence data. (2019). <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.

44 Ewels, P., Magnusson, M., Lundin, S. & Kaller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047-3048 (2016). https://doi.org/10.1093/bioinformatics/btw354

45 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014). https://doi.org/10.1093/bioinformatics/btu170

46 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013). https://doi.org/10.1093/bioinformatics/bts635

47 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930 (2014). https://doi.org/10.1093/bioinformatics/btt656

48 Yun, T. *et al.* Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* **36**, 5582-5589 (2021). https://doi.org/10.1093/bioinformatics/btaa1081

49 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009). https://doi.org/10.1093/bioinformatics/btp352

50 Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867-868 (2018). https://doi.org/10.1093/bioinformatics/btx699

51 Wagner, J. *et al.* Benchmarking challenging small variants with linked and long reads. *Cell Genom* **2**, 100128 (2022). https://doi.org/10.1016/j.xgen.2022.100128

52 Hofmeister, R. J., Ribeiro, D. M., Rubinacci, S. & Delaneau, O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *Nat Genet* **55**, 1243-1249 (2023). https://doi.org/10.1038/s41588-023-01415-w

53 Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426-3440 e3419 (2022). https://doi.org/10.1016/j.cell.2022.08.004

54 Browning, B. L., Zhou, Y. & Browning, S. R. A one-penny imputed genome from next-generation reference panels. *Am J Hum Genet* **103**, 338-348 (2018). https://doi.org/10.1016/j.ajhg.2018.07.015

55 Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088-3090 (2017). https://doi.org/10.1093/bioinformatics/btx346

56 Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci* **131**, 281-285 (2012). https://doi.org/10.1007/s12064-012-0162-3

57 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011). https://doi.org/10.1186/1471-2105-12-323

58 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016). https://doi.org/10.1186/s13059-016-0974-4

59 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). https://doi.org/10.1086/519795

60 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014). https://doi.org/10.1186/s13059-014-0550-8
61 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B, Methodological* **57**, 289-300 (1995). https://doi.org/DOI 10.1111/j.2517-6161.1995.tb02031.x
62 Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res* **47**, W191-W198 (2019). https://doi.org/10.1093/nar/gkz369
63 Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J. & Peterson, H. gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. *F1000Res* **9** (2020). https://doi.org/10.12688/f1000research.24956.2
64 Wijesooriya, K., Jadaan, S. A., Perera, K. L., Kaur, T. & Ziemann, M. Urgent need for consistent standards in functional enrichment analysis. *PLoS Comput Biol* **18**, e1009935 (2022). https://doi.org/10.1371/journal.pcbi.1009935
65 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). https://doi.org/10.1093/bioinformatics/btp616
66 Zhou, H. J., Li, L., Li, Y., Li, W. & Li, J. J. PCA outperforms popular hidden variable inference methods for molecular QTL mapping. *Genome Biol* **23**, 210 (2022). https://doi.org/10.1186/s13059-022-02761-4
67 Taylor-Weiner, A. *et al.* Scaling computational genomics to millions of individuals with GPUs. *Genome Biol* **20**, 228 (2019). https://doi.org/10.1186/s13059-019-1836-7
68 Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479-1485 (2016). https://doi.org/10.1093/bioinformatics/btv722
69 Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440-9445 (2003). https://doi.org/10.1073/pnas.1530509100
70 R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria, 2024).
71 pwr: Basic Functionf for Power Analysis (2020).
72 Tokolyi, A. *et al.* The contribution of genetic determinants of blood gene expression and splicing to molecular phenotypes and health outcomes. *Nat Genet* **57**, 616-625 (2025). https://doi.org/10.1038/s41588-025-02096-3
73 Lin, W. *et al.* Genetic regulatory effects in response to a high-cholesterol, high-fat diet in baboons. *Cell Genom* **4**, 100509 (2024). https://doi.org/10.1016/j.xgen.2024.100509
74 Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat Genet* **51**, 187-195 (2019). https://doi.org/10.1038/s41588-018-0268-8
75 Cai, Y. *et al.* Single-cell transcriptomics of blood reveals a natural killer cell subset depletion in tuberculosis. *EBioMedicine* **53**, 102686 (2020). https://doi.org/10.1016/j.ebiom.2020.102686
76 Zheng, G. X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049 (2017). https://doi.org/10.1038/ncomms14049
77 Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* **42**, 293-304 (2024). https://doi.org/10.1038/s41587-023-01767-y
78 Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* **9** (2020). https://doi.org/10.1093/gigascience/giaa151
79 Germain, P. L., Lun, A., Garcia Meixide, C., Macnair, W. & Robinson, M. D. Doublet identification in single-cell sequencing data using scDblFinder. *F1000Res* **10**, 979 (2021). https://doi.org/10.12688/f1000research.73600.2

80 Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* **20**, 163-172 (2019). https://doi.org/10.1038/s41590-018-0276-y

81 Mabbott, N. A., Baillie, J. K., Brown, H., Freeman, T. C. & Hume, D. A. An expression atlas of human primary cells: inference of gene function from coexpression networks. *BMC Genomics* **14**, 632 (2013). https://doi.org/10.1186/1471-2164-14-632

82 Martens, J. H. & Stunnenberg, H. G. BLUEPRINT: mapping human blood cell epigenomes. *Haematologica* **98**, 1487-1489 (2013). https://doi.org/10.3324/haematol.2013.094243

83 ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012). https://doi.org/10.1038/nature11247

84 Heng, T. S., Painter, M. W. & Immunological Genome Project Consortium. The Immunological Genome Project: networks of gene expression in immune cells. *Nat Immunol* **9**, 1091-1094 (2008). https://doi.org/10.1038/ni1008-1091

85 Monaco, G. *et al.* RNA-seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. *Cell Rep* **26**, 1627-1640 e1627 (2019). https://doi.org/10.1016/j.celrep.2019.01.041

86 Zhang, Z. *et al.* SCINA: a semi-supervised subtyping algorithm of single cells and bulk samples. *Genes (Basel)* **10** (2019). https://doi.org/10.3390/genes10070531

87 Diaz-Mejia, J. J. *et al.* Evaluation of methods to assign cell type labels to cell clusters from single-cell RNA-sequencing data. *F1000Res* **8** (2019). https://doi.org/10.12688/f1000research.18490.3

88 Karlsson, M. *et al.* A single-cell type transcriptomics map of human tissues. *Sci Adv* **7** (2021). https://doi.org/10.1126/sciadv.abh2169

89 Oelen, R. *et al.* Single-cell RNA-sequencing of peripheral blood mononuclear cells reveals widespread, context-specific gene expression regulation upon pathogenic exposure. *Nat Commun* **13**, 3267 (2022). https://doi.org/10.1038/s41467-022-30893-5

90 Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat Commun* **10**, 380 (2019). https://doi.org/10.1038/s41467-018-08023-x

91 Jew, B. *et al.* Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nat Commun* **11**, 1971 (2020). https://doi.org/10.1038/s41467-020-15816-6

92 Dietrich, A. *et al.* SimBu: bias-aware simulation of bulk RNA-seq data with variable cell-type composition. *Bioinformatics* **38**, ii141-ii147 (2022). https://doi.org/10.1093/bioinformatics/btac499

93 Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. 2nd ed. edn, (Springer-Verlag, 2016).

94 Phipson, B. *et al.* propeller: testing for differences in cell type proportions in single cell data. *Bioinformatics* **38**, 4720-4726 (2022). https://doi.org/10.1093/bioinformatics/btac582

95 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015). https://doi.org/10.1093/nar/gkv007

96 Davis, J. R. *et al.* An efficient multiple-testing adjustment for eQTL studies that accounts for linkage disequilibrium between variants. *Am J Hum Genet* **98**, 216-224 (2016). https://doi.org/10.1016/j.ajhg.2015.11.021

97 Delaneau, O. *et al.* A complete tool set for molecular QTL discovery and analysis. *Nat Commun* **8**, 15452 (2017). https://doi.org/10.1038/ncomms15452

98 Berry, M. P. *et al.* An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature* **466**, 973-977 (2010). https://doi.org/10.1038/nature09247

99 Zak, D. E. *et al.* A blood RNA signature for tuberculosis disease risk: a prospective cohort study. *Lancet* **387**, 2312-2322 (2016). https://doi.org/10.1016/S0140-6736(15)01316-1

100 Singhania, A. *et al.* A modular transcriptional signature identifies phenotypic heterogeneity of human tuberculosis infection. *Nat Commun* **9**, 2308 (2018). https://doi.org/10.1038/s41467-018-04579-w
101 López Herráez, D. *et al.* Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. *PLoS ONE* **4**, e7888 (2009). https://doi.org/10.1371/journal.pone.0007888
102 Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu Rev Genomics Hum Genet* **10**, 387-406 (2009). https://doi.org/10.1146/annurev.genom.9.081307.164242
103 Puyskens, A. *et al.* Aryl hydrocarbon receptor modulation by tuberculosis drugs impairs host defense and treatment outcomes. *Cell Host Microbe* **27**, 238-248 e237 (2020). https://doi.org/10.1016/j.chom.2019.12.005
104 Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-1195 (2012). https://doi.org/10.1126/science.1222794
105 Castaño, D., García, L. F. & Rojas, M. Increased frequency and cell death of CD16+ monocytes with *Mycobacterium tuberculosis* infection. *Tuberculosis (Edinb)* **91**, 348-360 (2011). https://doi.org/10.1016/j.tube.2011.04.002
106 Estévez, O. *et al.* Multi-parameter flow cytometry immunophenotyping distinguishes different stages of tuberculosis infection. *J Infect* **81**, 57-71 (2020). https://doi.org/10.1016/j.jinf.2020.03.064
107 Adane, T. *et al.* Accuracy of monocyte to lymphocyte ratio for tuberculosis diagnosis and its role in monitoring anti-tuberculosis treatment: Systematic review and meta-analysis. *Medicine (Baltimore)* **101**, e31539 (2022). https://doi.org/10.1097/MD.0000000000031539
108 Hillman, H. *et al.* Single-cell profiling reveals distinct subsets of CD14+ monocytes drive blood immune signatures of active tuberculosis. *Front. Immunol.* **13**, 1087010 (2022). https://doi.org/10.3389/fimmu.2022.1087010
109 Roy Chowdhury, R. *et al.* A multi-cohort study of the immune factors associated with *M. tuberculosis* infection outcomes. *Nature* **560**, 644-648 (2018). https://doi.org/10.1038/s41586-018-0439-x
110 Hand, T. W. & Kaech, S. M. Intrinsic and extrinsic control of effector T cell survival and memory T cell development. *Immunol Res* **45**, 46-61 (2009). https://doi.org/10.1007/s12026-008-8027-z
111 Scriba, T. J., Coussens, A. K. & Fletcher, H. A. Human immunology of tuberculosis. *Microbiol Spectr* **5** (2017). https://doi.org/10.1128/microbiolspec.TBTB2-0016-2016
112 Joosten, S. A. *et al.* Patients with tuberculosis have a dysfunctional circulating B-cell compartment, which normalizes following successful treatment. *PLoS Pathog* **12**, e1005687 (2016). https://doi.org/10.1371/journal.ppat.1005687
113 Strober, W. & Watanabe, T. NOD2, an intracellular innate immune sensor involved in host defense and Crohn's disease. *Mucosal Immunol.* **4**, 484-495 (2011). https://doi.org/10.1038/mi.2011.29
114 Ferwerda, G. *et al.* NOD2 and toll-like receptors are nonredundant recognition systems of *Mycobacterium tuberculosis*. *PLoS Pathog* **1**, 279-285 (2005). https://doi.org/10.1371/journal.ppat.0010034
115 Cubillos-Angulo, J. M. *et al.* The influence of single nucleotide polymorphisms of *NOD2* or *CD14* on the risk of *Mycobacterium tuberculosis* diseases: a systematic review. *Syst Rev* **10**, 174 (2021). https://doi.org/10.1186/s13643-021-01729-y
116 Divangahi, M. *et al.* NOD2-deficient mice have impaired resistance to *Mycobacterium tuberculosis* infection through defective innate and adaptive immunity. *J. Immunol.* **181**, 7157-7165 (2008). https://doi.org/10.4049/jimmunol.181.10.7157
117 Kauffman, K. J. *et al.* Delipidation of mammalian Atg8-family proteins by each of the four ATG4 proteases. *Autophagy* **14**, 992-1010 (2018). https://doi.org/10.1080/15548627.2018.1437341

118 Nguyen, T. N. *et al.* ATG4 family proteins drive phagophore growth independently of the LC3/GABARAP lipidation system. *Mol. Cell* **81**, 2013-2030 e2019 (2021). https://doi.org/10.1016/j.molcel.2021.03.001
119 Yamamoto, H., Zhang, S. & Mizushima, N. Autophagy genes in biology and disease. *Nat Rev Genet* **24**, 382-400 (2023). https://doi.org/10.1038/s41576-022-00562-w
120 Gutierrez, M. G. *et al.* Autophagy is a defense mechanism inhibiting BCG and *Mycobacterium tuberculosis* survival in infected macrophages. *Cell* **119**, 753-766 (2004). https://doi.org/10.1016/j.cell.2004.11.038
121 Deretic, V. & Wang, F. Autophagy is part of the answer to tuberculosis. *Nat Microbiol* **8**, 762-763 (2023). https://doi.org/10.1038/s41564-023-01373-3
122 Golovkine, G. R. *et al.* Autophagy restricts *Mycobacterium tuberculosis* during acute infection in mice. *Nat Microbiol* **8**, 819-832 (2023). https://doi.org/10.1038/s41564-023-01354-6
123 Qu, Y. *et al.* MiR-129-3p favors intracellular BCG survival in RAW264.7 cells by inhibiting autophagy via Atg4b. *Cell. Immunol.* **337**, 22-32 (2019). https://doi.org/10.1016/j.cellimm.2019.01.004
124 Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10**, e1004383 (2014). https://doi.org/10.1371/journal.pgen.1004383
125 Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* **47**, 1091-1098 (2015). https://doi.org/10.1038/ng.3367
126 Price, A. L. *et al.* Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genet* **7**, e1001317 (2011). https://doi.org/10.1371/journal.pgen.1001317
127 Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177-1186 (2017). https://doi.org/10.1016/j.cell.2017.05.038

Data availability

The n = 240 RNA-seq paired-end fastq files derived from the n = 48 individuals are available from the Sequence Read Archive (SRA) (GEO accession number: GSE157657)[41]. The reference genome used for the alignment of the RNA-seq reads can be found at the following link: GRCh38_full_analysis_set_plus_decoy_hla. The DeepVariant variant calling model[39,40] is available at the following link: https://github.com/google/deepvariant. The GIAB data sets can be accessed at the following address: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release. The PBMC scRNA-seq data from n = 3 active TB patients and n = 2 LTBI patients used in this study are available on the SRA (accession number SRP247583; SRR IDs SRR11038989–SRR11038993)[75]. The filtered and imputed variants for all n = 48 samples in addition to the raw *cis*-eQTL, reQTL and ieQTL results will be made publicly available upon publication.

Code availability

The computer code and scripts used in this study are available at the following GitHub link: https://github.com/jfogrady1/Human_TB_reQTL.

Acknowledgements

J.F.O'G was supported by Research Ireland through the Research Ireland Centre for Research Training in Genomics Data Science (grant no. 18/CRT/6214). This study was also supported by Science Foundation Ireland (SFI) Investigator Programme Awards to D.E.M. and S.V.G. (grant nos. SFI/08/IN.1/B2038 and SFI/15/IA/3154), and University College Dublin (UCD) College of Health and Agricultural Sciences seed funding awarded to J.F.O'G. Note: Since the 1st of August 2024, Science Foundation Ireland (SFI) has been part of Taighde Éireann Ù Research Ireland (www.researchireland.ie).

The authors would like to thank Joe Keane, John Browne, Thomas Hall, Adnan Khan, and Kseniia Maximova for useful discussions.

Author contributions

J.F.O'G., A.S.L., H.P., I.C.G, and D.E.M. conceived and designed the study. J.F.O'G., S.V.G., I.C.G. and D.E.M. acquired funding for the study. J.F.O'G. and A.S.L. performed bioinformatics analyses in collaboration with L.F., H.L., and H.P. J.F.O'G. wrote the first draft of the manuscript, and prepared all figures, tables, and supplementary information with input from D.E.M. All authors read and approved the final manuscript.

Declaration of interests

The authors declare no competing interests.

Consent for publication

Not applicable

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 02OGradyetal2025SuppInfo.docx
- 03OGradyetal2025SuppTableS1S12.xlsx
- 04OGradyetal2025SuppTableS13S17.xlsx
- 05OGradyetal2025SuppTableS18S32.xlsx