Supporting Information

Barreiro et al. 10.1073/pnas.1115761109

SI Materials and Methods

Sample Collection. Blood samples from 68 healthy donors were obtained from Research Blood Components. A signed written consent was obtained from all of the participants. All individuals recruited in this study were healthy Caucasian males between the ages of 21 and 55 y old. We decided to focus on only one sex to avoid the potentially confounding effects of sex-specific differences in gene expression level on response phenotypes (1, 2). We chose males because gene expression levels are known to differ more among females, due to estrus cycling (e.g., refs. 3, 4), an effect that would reduce the power to identify eQTL. Only individuals self-reported as currently healthy, not under medication, and with no history of diseases such as malaria, tuberculosis, cancer, or hepatitis were included in the study. In addition, each donor's blood was tested for standard blood-borne pathogens, and only samples negative for all of the pathogens tested were used.

***Mycobacterium tuberculosis* Preparation.** We infected dendritic cells (DCs) with a *Mycobacterium tuberculosis* (MTB) strain expressing green-fluorescent protein (H37Rv). This recombinant strain carries a pEGFP plasmid, which encodes a gene that confers resistance to hygromycin and harbors the *GFP* gene under the control of the mycobacterial Phsp60 constitutive promoter. Importantly, our work (5), as well as that of others (6) has shown that the presence of GFP in MTB does not alter growth or virulence of the bacilli under axenic conditions, relative to wild-type MTB. *M. tuberculosis* H37Rv was grown from a frozen stock to midlog phase in 7H9 medium (BD) supplemented with albumin-dextrose-catalase (ADC; Difco). We tested the virulence of the bacteria in the frozen stock by infecting C57BL/6 mice intranasally with 10^3 bacilli. After 21 and 42 d, we estimated a load of 10^7 bacteria in the mice lungs, indicating that the bacteria did not lose its natural virulence (7).

Isolation and Infection of DCs. Blood mononuclear cells from healthy volunteers were isolated by Ficoll-Paque centrifugation. Blood monocytes were purified from peripheral blood mononuclear cells by positive selection with magnetic CD14 MicroBeads (Miltenyi Biotech). Monocytes were then cultured for 5 d in RPMI 1640 (Invitrogen) supplemented with 10% heat-inactivated FCS (Dutscher), L-glutamine (Invitrogen), GM-CSF (20 ng/mL; Immunotools), and IL-4 (20 ng/mL; Immunotools). Cell cultures were fed every 2 d with complete medium supplemented with the cytokines previously mentioned. Before infection, we systematically checked the differentiation/activation status of the monocyte-derived DCs by flow cytometry, using antibodies against CD1a, CD14, CD83, and HLA-DR. All antibodies were purchased from Becton Dickinson. Only samples presenting the expected phenotype for nonactivated DCs—$CD1a^+$, $CD14^-$, $CD83^-$, and $HLA\text{-}DR^{low}$—were used in downstream experiments. The resulting monocyte-derived DCs were then infected with MTB for 18 h at a multiplicity of infection of 1-to-1. The choice of 18 h is based on previous work, which revealed that the largest number of transcriptional changes following MTB infection could be captured at 18 h postinfection (8).

DNA Extraction and Genome-Wide Genotyping. DNA from each of the blood donors was extracted from the depleted white cell populations (i.e., T cells, B cells, NK cells, etc.), using the PureGene DNA extraction kit (Gentra Systems). Genotyping of 68 individuals was then performed using Illumina's Omni1-Quad BeadChip array, which interrogates 970,287 SNPs. Genotype calls were extracted from the raw data using BeadStudio. All samples had genotype call rates (CR) above 98%, with the exception of individual TB91 (CR = 80%), who was excluded from further analysis. After applying standard quality control criteria (SNPs with no missing data and nominal P value for testing deviation from Hardy–Weinberg equilibrium $>10^{-4}$), 873,973 SNPs remained for analysis. Because samples were collected anonymously, we tested for relatedness in our sample. To do so, we used PLINK (9) to estimate the pair-wise genome-wide identity by state (IBS) between all possible pairs of individuals.

We found two pairs of individuals that appeared to be genetically identical (i.e., they shared >99.9% of their genotypes), suggesting that two individuals donated blood twice during our recruitment process. We randomly excluded the data of one individual from each of these pairs. All other samples were unrelated as defined by an estimated proportion of IBS <0.2 (second degree relatives). All samples were confirmed to be males on the basis of the genotype data from the X chromosome. Finally, although all our blood donors were self-identified as European Americans, we used principal component analyses (PCA) to confirm their ethnic origin on the basis of the genotype data alone. To do so, we used smart PCA (10) after integrating our samples with the ethnically well-defined HapMap population samples. All our samples clustered tightly together with the European population from HapMap with the exception of four individuals that presented some evidence of admixture with non-European groups (Fig. S5). In the analyses presented in the main text, we kept the data from these "admixed" individuals, but we confirmed that our conclusions remain unaltered by excluding these samples.

In summary, we excluded data from one individual with a low genotype call rate, and data from two pairs of individuals were practically identical (we retained one from each pair). These steps resulted in a final dataset of 65 individuals that were used in the eQTL analysis.

Gene Expression Measurements and Preprocessing of Expression Data. Total RNA was extracted from the noninfected DCs and the MTB-infected DCs using the miRNeasy kit (Qiagen). RNA quantity was evaluated spectrophotometrically, and the quality was assessed with the Agilent 2100 Bioanalyzer (Agilent Technologies). Only samples with no evidence for RNA degradation (RNA integrity number >8) were kept for further experiments. Genome-wide gene expression profiling of untreated and infected DCs was obtained by hybridizing the RNA to the Illumina HumanHT-12 v4 Expression BeadChips arrays. The cDNA synthesis, labeling, and subsequent hybridization to the microarrays were performed by the Southern California Genotyping Consortium at the University of California at Los Angeles. Two technical replicates were performed for each sample yielding data from 260 expression arrays (65 individuals × 2 conditions × 2 technical replicates). We found that the gene expression estimates obtained from technical replicates were highly correlated (median Pearson's $r = 0.98$) indicating excellent reproducibility. On the basis of an analysis of the pairwise correlation of technical replicate data, we found two clear outlier arrays, which were excluded from subsequent analyses. For six individuals, we also performed the infection experiments in triplicate (three treated and three untreated cultures of DCs from each of the six individuals) to evaluate the degree of variation associated with our experimental setup. We found very high correlations (Pearson's $r > 0.96$) between biological replicates (i.e., in-

dependently untreated or treated replicate samples of DCs from the same individuals), demonstrating that our cell culture and infection procedures are highly replicable and consistent.

Low-level microarray analyses were performed in R, using the Bioconductor software package lumi (11). We first applied a variance stabilizing transformation to all arrays (12) and then quantile normalized the data. After normalization, we removed probes with intensities indistinguishable from background noise (as measured by the negative controls present on each array). We next annotated the significantly expressed probes by mapping them to RefSeq gene sequences using BLAT. Only probes that mapped to unique gene IDs were kept for downstream analyses. In addition, to avoid spurious associations between specific genotypes and gene expression measurements, we excluded all probes that contained one or more HapMap SNPs. Finally, we removed probes that mapped to putative and/or nonwell characterized genes (i.e., genes without an Ensembl gene ID). After these preprocessing steps, data from 17,017 probes corresponding to 12,958 well-annotated Ensembl genes were available for association analysis.

Identifying Genes Differentially Expressed After MTB Infection. To identify genes whose expression levels were altered following MTB infection of DCs, we used a linear modeling-based approach. Specifically, we used the Bioconductor limma package (13) to fit, for each gene, a linear model with individual treatment (i.e., MTB infection) and batch as fixed effects. We included a batch effect because the RNA samples were hybridized in two separate batches (first batch, 180 arrays; second batch, 80 arrays, each with a balanced number of infected and noninfected samples). For each gene, we subsequently used the empirical Bayes approach of Smyth (13) to calculate a moderated t statistic and P value. We corrected for multiple testing using the false discovery rate (FDR) approach of Benjamini and Hochberg (14).

Gene Ontology (GO) and Pathway Enrichment Analysis. We used GeneTrail (15) to test for enrichment of functional annotations among differentially expressed genes after MTB infection, using all expressed genes (i.e., 12,958 genes) as a background set. The tests were performed using all GO categories and Kyoto Encyclopedia of Genes and Genomes pathways. P values were calculated by comparing the observed data with the quantiles of a hypergeometric distribution, and we used the approach of Benjamini and Hochberg (14) to control the false discovery rate.

Quantification of Cytokine and Chemokine Levels in Supernatants. We used the Bio-Plex Pro Human Cytokine 27-plex (Bio-Rad) to measure the levels of 27 different cytokines/chemokines in the supernatants of untreated and infected DCs. We chose this assay because it includes the most important cytokines currently known to be involved in protective immunity against tuberculosis (e.g., IFN-γ, IL-12, IL-17, or TNF-α). The assay was performed at the Flow Cytometry Facility at the University of Chicago, according to the manufacturer's recommendations. Each sample was assayed in two technical replicates. For each protein, the average quantity across technical replicates was calculated and used for all subsequent analyses. To reduce the effects of outliers on the protein QTL mapping, the secretion values of each protein were quantile normalized so that they followed a $N(0,1)$ distribution across individuals using the "qqnorm" function in R (both for infected and noninfected samples). Ties due to estimated secretion levels of zero were broken randomly.

Because samples were assayed in four different 96-well plates (with a balanced number of infected and noninfected samples in each plate) we used linear regression to remove the potential "plate-effect" confounder from the measurement of each protein and the corrected data were used in all subsequent analyses. Of the 27 proteins tested, 4 showed nondetectable (i.e., extremely small) secretion levels (IL-5, IL-7, Eotaxin, and FGF) and two presented secretion levels above our maximum detection limits (MIP-1a and M1P-1b). These 6 proteins were excluded from our analyses. We also excluded from all analyses GM-CSF and IL-4 because the measured secretion levels could be biased as we added those two cytokines to the culture media to derive DCs.

Genotype–Phenotype Association Analysis. We limited the eQTL analysis to data from 11,996 genes, which are a subset (93%) of the 12,958 genes that we classified as expressed in DCs. We excluded 962 genes (of the set of 12,958 genes expressed in DCs) either because: (*i*) they were located on a sex chromosome (457 on the X chromosome and 11 on the Y chromosome), which limits the power to detect eQTL given that all our samples are males (i.e., for these genes we have half the number of genotyped chromosomes) or (*ii*) the genes were poorly annotated and we could not identify reliable transcription start site (TSS) positions (and therefore we could not define a putative "*cis*"-eQTL region).

We examined associations between SNP genotypes and either transcript or protein by using a linear regression model in which phenotype was regressed against genotype. In all cases, we assumed that alleles affecting either transcript or protein expression levels did so in an additive manner. We mapped infected and noninfected DCs separately. All regressions were performed using a Python script, whereas downstream analyses were carried out using the R statistical framework. We only tested associations with SNPs with a minor allele frequency greater than 10% because, given our limited sample size, we have low power to detect eQTL or pQTL for rare variants. When looking for variants putatively associated with gene expression levels or protein secretion *in cis*, we tested for an association between expression levels and genotypes at SNPs located within a 200-kb window centered on the gene's TSS. We recorded the minimum P value (i.e., the strongest association) observed for each gene, which we used as statistical evidence for the presence of at least one eQTL for that gene.

To estimate an FDR, we permuted the phenotypes (expression levels) three times, reperformed the linear regressions, and recorded the minimum P values for the gene for each permutation. These sets of minimum P values were used as our empirical null distribution. We then compared the observed distribution of the minimum P values to the null distribution to estimate the FDR, as previously described (16). Briefly, we found the P value i such that $\Pr(P_{\text{permuted}} < i)/\Pr(P_{\text{real}} < i) = 0.01$, where 0.01 corresponds to the FDR of 1% used in our study, $\Pr(P_{\text{permuted}} < i)$ is the proportion of minimum P values from the permutations that fall below the P value threshold, and $\Pr(P_{\text{real}} < i)$ is the proportion of minimum P values from the real data that fall below the P value threshold. In our data, an FDR of 1% corresponded to a value of i equal to 1.4×10^{-5}.

Consistent with previous reports (16, 17) we found that we could increase the power to detect *cis*-eQTL by accounting for unmeasured—surrogate—confounders (e.g., related to technical effects or sample quality biases). To do this, we first determined the principal components of the correlation matrix for the noninfected and infected gene expression data. Subsequently, for each gene we regressed out the first five principal components (PCs) or eight PCs from the noninfected and MTB-infected data, respectively, before performing the association analysis. The numbers of PCs to regress out were chosen because they empirically led to the identification of the largest number of eQTL in each of the conditions. Importantly, whereas the PC corrections clearly increase power to detect eQTL, they do not affect the underlying structure of the expression data. Indeed, >87% of the eQTL observed before any PC correction are also observed after PC correction at the same FDR cutoff.

Identifying Response eQTL. In principle, after independently classifying eQTL in the untreated and infected DCs we could look for response eQTL by simply comparing the lists of eQTL in each class of DCs. However, a naive comparison of the lists of eQTL ignores the fact that evidence for eQTL in one class of DC provides information about the likelihood of an eQTL in the other class. Thus, using a single arbitrary statistical cutoff in independent analyses of the untreated and infected DCs is likely to result in a high rate of falsely identified response eQTL. Instead, we classified response eQTL by using a two-step FDR cutoff.

Specifically, we first used a stringent FDR cutoff of 1% to classify eQTL in either the untreated or the infected DCs. Subsequently, the threshold for classifying corresponding eQTL in the other class of DCs was relaxed to an FDR of 50%. This approach results in a conservative classification of response eQTL. We note that the choice of statistical cutoffs was arbitrary (as is typically the case, regardless of the use of one or two cutoffs).

Importantly, our observations are robust to the method used to identify response eQTL. Indeed, an alternative approach used to identify response eQTL is to treat the changes in gene expression levels after a treatment, in our case MTB infection, as the quantitative trait to be mapped (18, 19). This approach makes the assumption that interaction effects result in additive changes in gene regulation and for that reason we chose not to present it as the main analysis (our approach allows for threshold effects, which are known to be common in gene regulatory networks). In addition, the approach based on mapping the gene expression response has low power to detect a significant interaction when the genotype effect on expression levels in untreated and infected DCs, independently, is of different magnitude but has the same direction. On the other hand, the approach of mapping the regulatory change has the advantage of not relying on the choice of two arbitrary cutoffs.

Reassuringly, the lists of response eQTL identified using either approach were highly similar (Fig. S4 and Dataset S3). Moreover, response eQTL identified using either approach were also significantly enriched for genome-wide association study (GWAS) P values <0.05 (1.8-fold enrichment, $P = 0.01$; Fig. S4).

1. Ranz JM, Castillo-Davis CI, Meiklejohn CD, Hartl DL (2003) Sex-dependent gene expression and evolution of the Drosophila transcriptome. *Science* 300:1742–1745.
2. Rinn JL, Snyder M (2005) Sexual dimorphism in mammalian gene expression. *Trends Genet* 21:298–305.
3. Kang J, et al. (2005) Expression of human prostaglandin transporter in the human endometrium across the menstrual cycle. *J Clin Endocrinol Metab* 90:2308–2313.
4. Sarkar MA, Vadlamuri V, Ghosh S, Glover DD (2003) Expression and cyclic variability of CYP3A4 and CYP3A7 isoforms in human endometrium and cervix during the menstrual cycle. *Drug Metab Dispos* 31:1–6.
5. Tailleux L, et al. (2003) Constrained intracellular survival of Mycobacterium tuberculosis in human dendritic cells. *J Immunol* 170:1939–1948.
6. Kremer L, Baulard A, Estaquier J, Poulain-Godefroy O, Locht C (1995) Green fluorescent protein as a new expression marker in mycobacteria. *Mol Microbiol* 17: 913–922.
7. Tanne A, et al. (2009) A murine DC-SIGN homologue contributes to early host defense against Mycobacterium tuberculosis. *J Exp Med* 206:2205–2220.
8. Tailleux L, et al. (2008) Probing host pathogen cross-talk by transcriptional profiling of both Mycobacterium tuberculosis and infected human dendritic cells and macrophages. *PLoS ONE* 3:e1403.
9. Purcell S, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575.
10. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2:e190.
11. Du P, Kibbe WA, Lin SM (2008) lumi: A pipeline for processing Illumina microarray. *Bioinformatics* 24:1547–1548.
12. Lin SM, Du P, Huber W, Kibbe WA (2008) Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Res* 36:e11.
13. Smyth GK (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3:Article3.
14. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Ser A Stat Soc* 57:289–300.
15. Backes C, et al. (2007) GeneTrail—advanced gene set enrichment analysis. *Nucleic Acids Res* 35:W186–192.
16. Pickrell JK, et al. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464:768–772.
17. Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* 3:1724–1735.
18. Maranville JC, et al. (2011) Interactions between glucocorticoid treatment and cis-regulatory polymorphisms contribute to cellular response phenotypes. *PLoS Genet* 7: e1002162.
19. Smirnov DA, Morley M, Shin E, Spielman RS, Cheung VG (2009) Genetic analysis of radiation-induced changes in human gene expression. *Nature* 459:587–591.

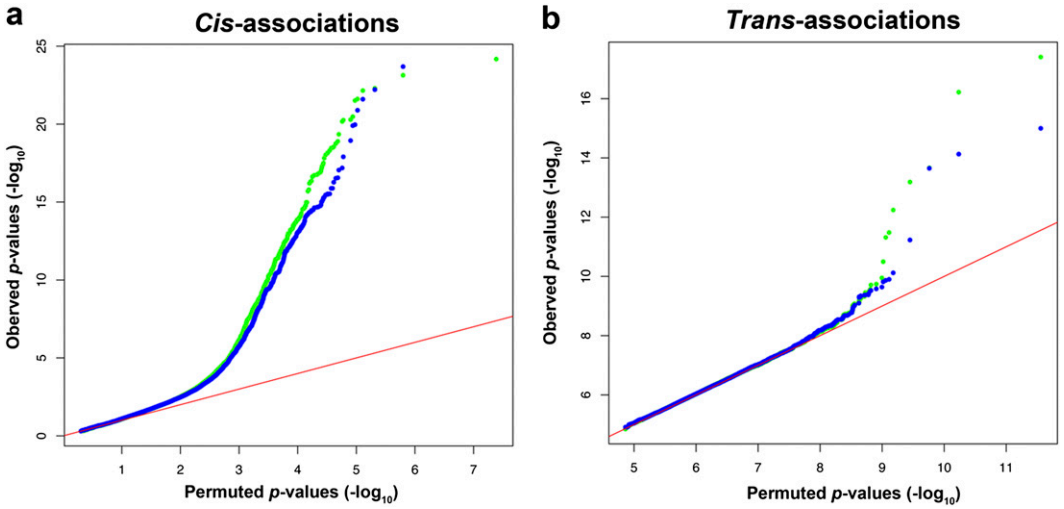

Fig. S1. Most SNPs associated with gene expression levels act *in cis*. (*A*) Quantile–quantile plot of *P* values obtained when testing for an association between gene expression estimates and all SNPs located in a 200-kb window centered on a gene's transcription starting site (TSS) (*y* axis) compared with *P* values obtained by permuting the gene expression measurement (*x* axis). (*B*) Quantile–quantile plot of *P* values obtained when testing for an association between gene expression estimates and all SNPs located more than 500 kb away from the TSS of the gene being tested (*y* axis) compared with *P* values obtained by permuting the gene expression measurement (*x* axis).

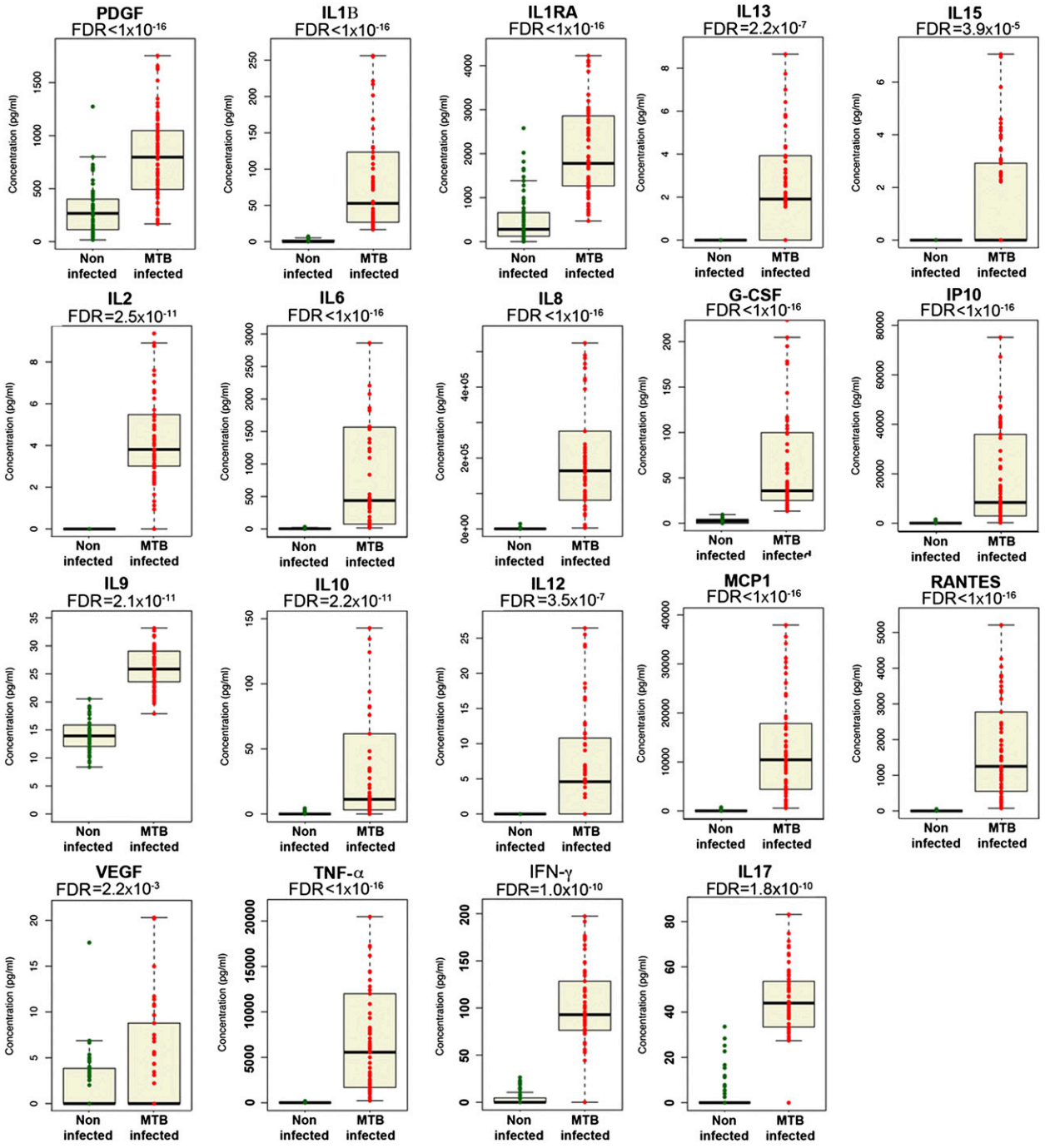

Fig. S2. Protein level measurements from untreated (green) and infected (red) DCs for the 19 cytokines/chemokines tested.

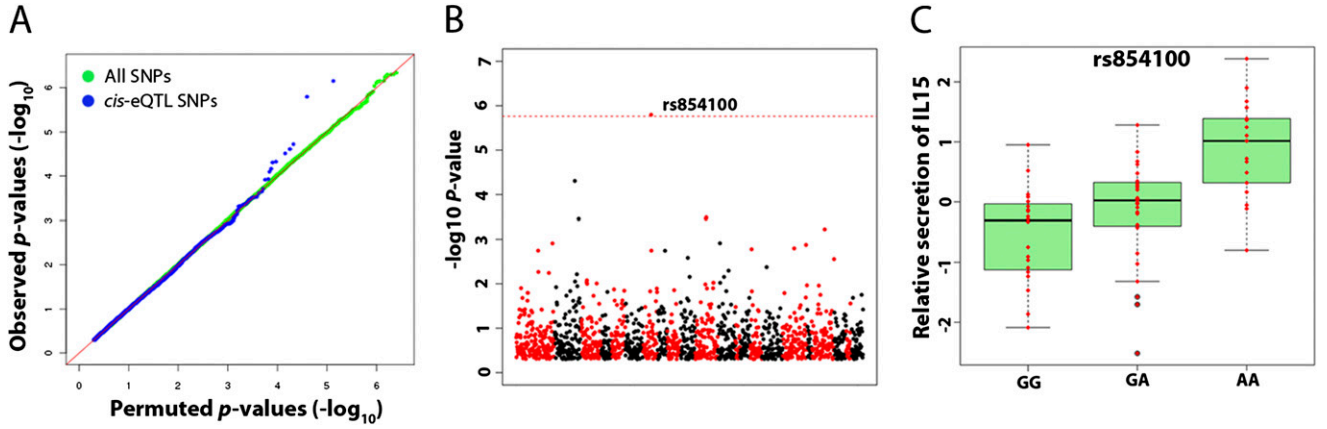

Fig. S3. Mapping protein QTL. (*A*) Quantile–quantile plot of *P* values obtained when testing for an association between protein measurement against all SNPs (green) and against SNPs identified as being *cis*-QTL (blue) (*y* axis) compared with *P* values obtained by permuting 10 times the protein measurement (*x* axis). (*B*) Manhattan plot (*Left*) showing the negative $\log_{10}$ transformed *P* values (*y* axis) for the association between all SNPs classified as *cis*-eQTL and the secretion levels of IL15 measured in the supernatant of infected DCs. (*C*) Correlation between genotypes at rs854100 and the relative secretion levels of IL15.

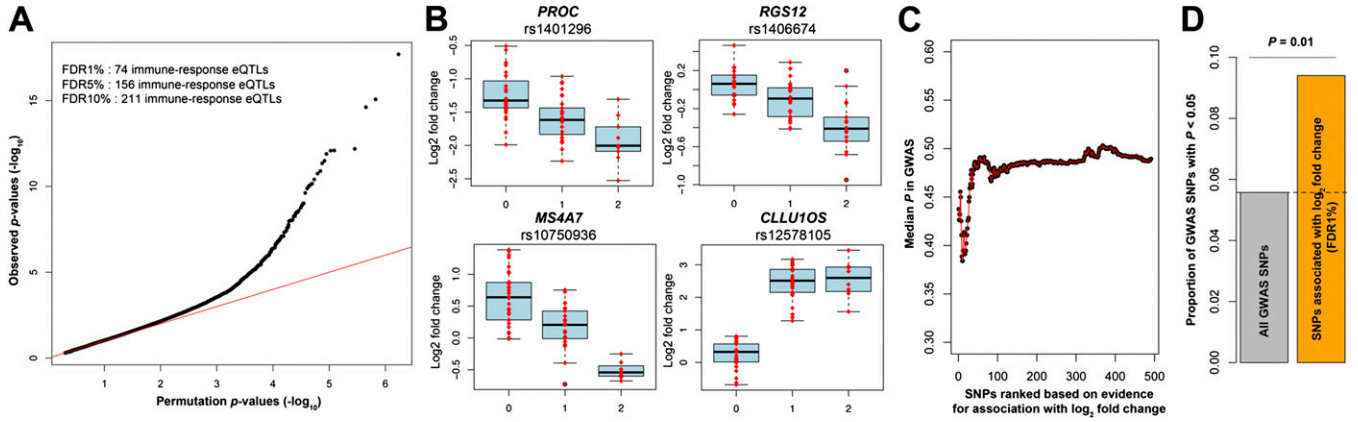

Fig. S4. Response eQTL identified when treating the changes in gene expression levels after MTB infection as the quantitative trait to be mapped. (*A*) Quantile–quantile plot summarizing the results from tests for genetic variation associated with changes in gene expression levels after MTB infection. We plot the observed *P* values (*y* axis) against the *P* values obtained by permuting three times the phenotypes (i.e., fold-change in expression levels; *x* axis). (*B*) Examples of response eQTL identified when mapping the changes in expression levels after MTB infection. (*C*) MTB-response eQTL identified by mapping changes in gene expression levels after MTB infection are enriched for susceptibility alleles to TB. The median GWAS *P* value for an expanding window of genes is plotted. We used the GWAS *P* values obtained when combining the Ghana and Gambia cohorts. Genes are ordered by the strength of evidence supporting an MTB-response eQTL based on the lowest *P* value obtained when testing an association between *cis*-SNPs (i.e., SNPs located in 200-kb window centered on the TSS of proximal genes) and changes in gene expression levels after MTB infection. To avoid positional biases we restricted our analyses to the set of *cis*-SNPs that was tested in our study. (*D*) Histogram of the proportion of GWAS SNPs with nominal *P* values <0.05 among all GWAS SNPs (gray), and among response eQTL (orange).

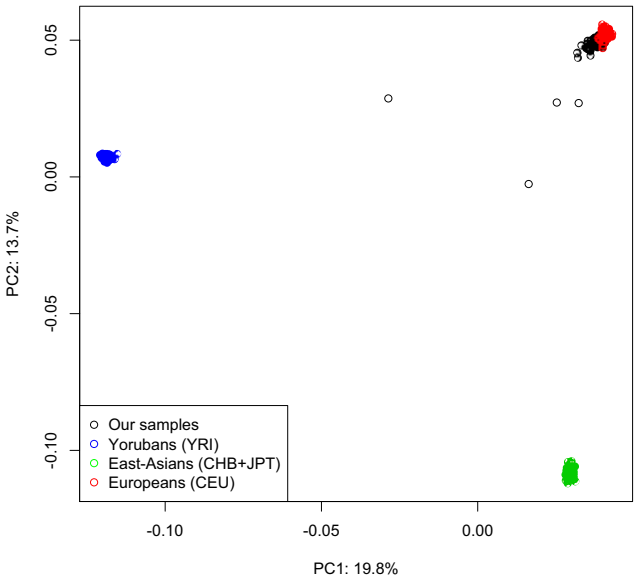

Fig. S5. Principal component analysis integrating data on our samples with data from the ethnically well-defined HapMap populations. The analysis was done using a set of 634,191 SNPs that were genotyped in our samples as well as in the HapMap samples.

Table S1. List of genes with response eQTL that are also associated with SNPs with a nominal *P* value <0.05 in the combined genome-wide association study (GWAS) data

Ensembl gene ID	Hugo gene ID	SNP, rs no.	*P* eQTL NS	*P* eQTL MTB	*P* GWAS
ENSG00000161326	DUSP14	rs712039	9.61E-06	0.046899177	3.30E-06
ENSG00000158022	TRIM63	rs868551	3.14E-10	0.394820211	0.001585
ENSG00000160593	AMICA1	rs4537793	0.006919186	4.47E-06	0.002643
ENSG00000168268	NT5DC2	rs3755806	8.21E-08	0.179097087	0.005162
ENSG00000159788	RGS12	rs1406674	0.076363479	5.67E-09	0.006678
ENSG00000138604	GLCE	rs10162608	0.355105618	7.01E-07	0.01147
ENSG00000205057	CLLU1OS	rs432771	0.014540672	3.52E-06	0.01285
ENSG00000095209	TMEM38B	rs10739209	4.84E-08	0.133219781	0.01615
ENSG00000181458	TMEM45A	rs7616839	2.35E-08	0.012969461	0.01636
ENSG00000164465	DCBLD1	rs7746536	3.75E-06	0.011329278	0.01661
ENSG00000109099	PMP22	rs1380179	0.076313816	1.77E-10	0.01767
ENSG00000104312	RIPK2	rs40457	0.088217122	1.86E-07	0.02165
ENSG00000171792	C12orf32	rs1860434	0.016617327	4.15E-06	0.02413
ENSG00000117151	CTBS	rs12143652	1.15E-08	0.005546015	0.02438
ENSG00000187164	KIAA1598	rs10886017	3.26E-06	0.435779466	0.02507
ENSG00000149499	EML3	rs3809079	0.009468462	9.84E-06	0.02563
ENSG00000168234	TTC39C	rs1843839	3.80E-06	0.203439918	0.02649
ENSG00000144182	LIPT1	rs11688004	7.28E-07	0.007389465	0.02656
ENSG00000188056	TREML4	rs9349180	0.030824819	3.58E-10	0.0285
ENSG00000082497	SERTAD4	rs2485903	9.34E-07	0.02513761	0.02868
ENSG00000156475	PPP2R2B	rs9325026	2.48E-06	0.092747289	0.02886
ENSG00000091157	WDR7	rs559998	0.077327867	5.22E-06	0.03239
ENSG00000189046	ALKBH2	rs7135947	6.86E-08	0.223122089	0.03252
ENSG00000116791	CRYZ	rs12120636	0.016141668	1.81E-06	0.03619
ENSG00000115718	PROC	rs2069933	1.44E-10	0.113634564	0.03698
ENSG00000123415	SMUG1	rs6580976	0.284568281	4.84E-06	0.03729
ENSG00000137312	FLOT1	rs9468830	2.79E-06	0.063806996	0.04063
ENSG00000185344	ATP6V0A2	rs10744162	0.347680074	7.46E-07	0.04222
ENSG00000211456	SACM1L	rs2673028	0.073164192	4.99E-07	0.044
ENSG00000135124	P2RX4	rs1169727	0.381229529	1.73E-06	0.04481
ENSG00000160712	IL6R	rs4379670	1.37E-06	0.202474043	0.04541
ENSG00000166927	MS4A7	rs2233253	0.006554554	2.31E-12	0.0457
ENSG00000176734	TRIL	rs505532	0.254563981	4.38E-06	0.04617
ENSG00000131797	MGC3020	rs3751742	0.227688314	3.01E-09	0.0471
ENSG00000119865	CNRIP1	rs2120334	1.60E-06	0.017934224	0.04806
ENSG00000103064	SLC7A6	rs8056893	1.24E-10	0.03452262	0.04879

Other Supporting Information Files

Dataset S1 (XLS)
Dataset S2 (XLS)
Dataset S3 (XLS)