Identifying gene regulatory signature in individuals susceptible to tuberculosis

John D. Blischak1,2, Ludovic Tailleux3, Marsha Myrthil1, Luis B. Barreiro4,5,\*, Yoav Gilad1,6,\*

1Department of Human Genetics, University of Chicago, Chicago, Illinois, USA

2Committee on Genetics, Genomics, and Systems Biology, University of Chicago, Chicago, Illinois, USA

3Mycobacterial Genetics Unit, Institut Pasteur, Paris, France

4Department of Genetics, CHU Sainte-Justine Research Center, Montreal, Québec, Canada

5Department of Pediatrics, University of Montreal, Montreal, Québec, Canada

6Department of Medicine, University of Chicago, Chicago, Illinois, USA

\*Correspondence should be addressed to YG (gilad@uchicago.edu) and LBB (luis.barreiro@umontreal.ca).

**Tuberculosis is a deadly infectious disease, which kills millions of people every year. The causative pathogen, *Mycobacterium tuberculosis* (MTB), is estimated to have infected up to a third of the world’s population; however, only approximately 10% of healthy individuals progress to active TB disease. Despite evidence for heritability, it is not currently possible to predict whether a healthy person is susceptible to TB. To explore approaches to classify susceptibility to TB, we infected dendritic cells (DCs) from individuals known to be susceptible or resistant to TB with MTB, and measured genome-wide gene expression levels in infected and uninfected cells. We found that gene expression differences between susceptible and resistant individuals were primarily present in the non-infected cells. Intersecting our results with genome wide association analyses (GWAS), we demonstrated an enrichment of low GWAS p-values for the differentially expressed genes, and also identified two interesting candidate genes: *CCL1* and *UNC13A*. Lastly, we trained a classifier based on the gene expression levels in the non-infected state and demonstrated decent performance on our data and an independent data set. Overall, our promising results from this small study suggest that training a classifier on a larger cohort could enable us to accurately predict TB susceptibility, which would inform the prevention and treatment of TB infections.**

# Introduction

Tuberculosis (TB) is a major public health issue. Worldwide, over a million people die of TB annually, and millions more currently live with the disease \cite{WHO2015a, WHO2015b, Glaziou2015}. Successful treatment requires months of antibiotic therapy \cite{Sotgiu2015}, and the difficulty of adhering to the full treatment regimen has lead to the emergence of drug-resistant strains of *Mycobacterium tuberculosis* (MTB) \cite{Seung2015}.

Approximately a third of the world’s population has been infected with MTB, but most are asymptomatic. While these naturally resistant individuals are able to avoid active disease, MTB persists in a dormant state inside their innate immune cells, known as a latent TB infection \cite{Munoz2015}. In contrast, approximately 10% of individuals will develop active TB after infection with MTB \cite{North2004, OGarra2013}. Unfortunately, we are currently unable to predict if an individual is susceptible. While twin and family studies have indicated a heritable component of TB susceptibility \cite{Kallmann1943, Comstock1978, Cobat2010, Moller2010}, genome wide association studies (GWAS) have only identified a few loci with low effect size \cite{Thye2010, Mahasirimongkol2012, Thye2012, Png2012, Chimusa2014, Curtis2015, Sobota2016}. Due to the highly polygenic architecture, it may be informative to examine differences between susceptible and resistant individuals at a higher level of organization, e.g. gene regulatory networks. Indeed, a previous study has successfully identified gene expression differences in innate immune cells isolated from individuals known be susceptible or resistant to an infectious disease \cite{Thuong2008, Bryant2014}.

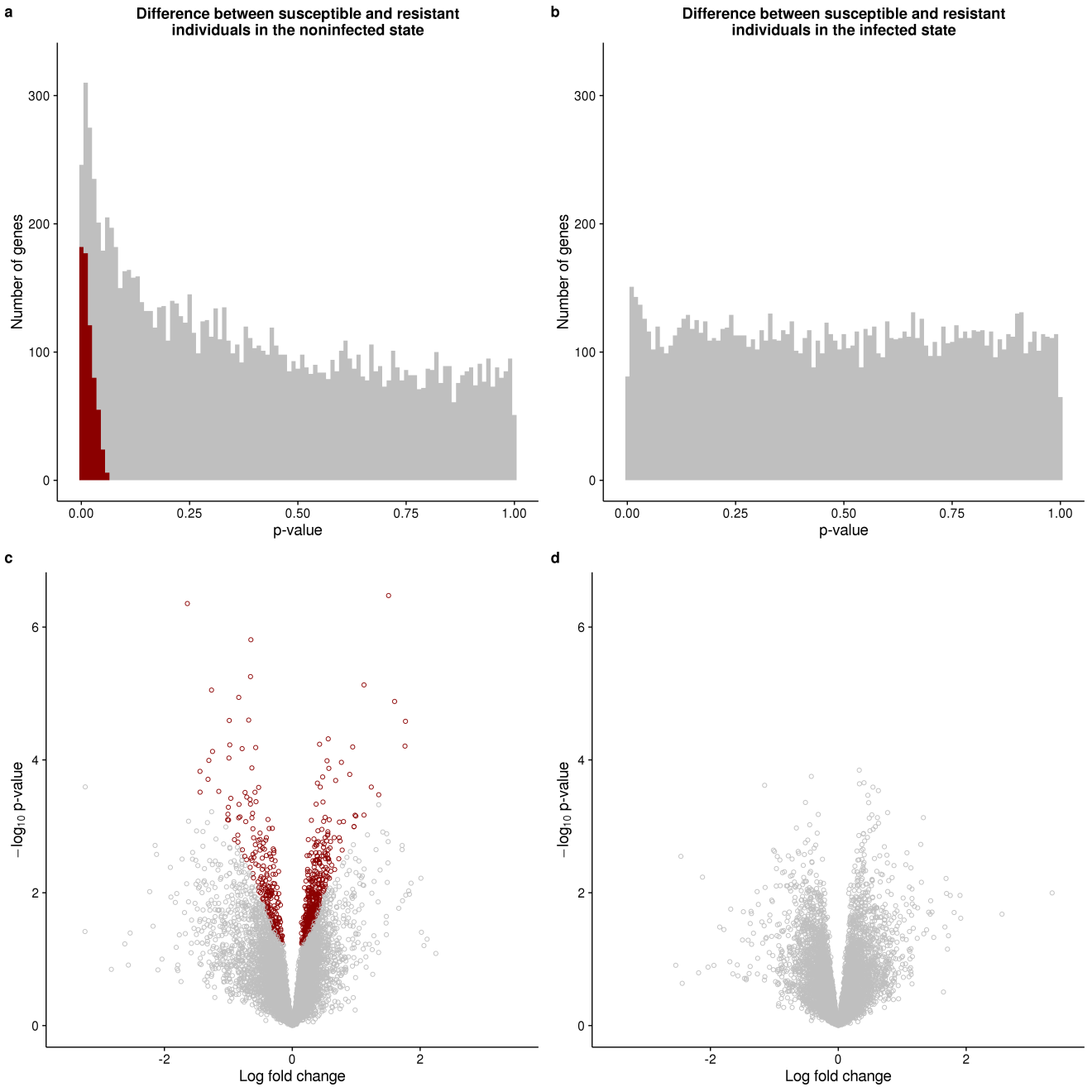
These differences in gene expression in innate immune cells have the potential to classify individuals based on their susceptibility to develop an active TB infection. To our knowledge, there has been no study reporting such an attempt. In this study, we isolated innate immune cells from individuals that are resistant or susceptible to TB and infected them with MTB. We discovered that the gene expression differences between resistant and susceptible innate immune cells were present primarily in the noninfected state, that these differentially expressed genes were enriched for nearby SNPs with low p-values in TB susceptibility GWAS, and furthermore that these gene expression levels could be used to classify individuals based on their susceptibility status.

# Results

## Susceptible individuals have an altered transcriptome in the non-infected state

We obtained whole blood samples from 25 healthy individuals. Six of the donors had recovered from a previous active TB infection, and are thus susceptible. The remaining 19 tested positive for a latent TB infection without ever experiencing symptoms of active TB, and are thus resistant. We isolated dendritic cells (DCs) and treated them with *Mycobacterium tuberculosis* (MTB) or a mock control for 18 hours. To measure genome-wide gene expression levels in infected and non-infected samples, we isolated and sequenced RNA using a processing pipeline designed to minimize the introduction of unwanted technical variation (Supplementary Fig. \ref{fig:process}). We obtained a mean ($\pm$ SEM) of 48 $\pm$ 6 million raw reads per sample. We performed quality control analyses to remove non-expressed genes (Supplementary Fig. \ref{fig:gene}), identify and remove outliers (Supplementary Fig. \ref{fig:heat-all}, \ref{fig:heat-filt}, \ref{fig:outliers}), and check for confounding batch effects (Supplementary Fig. \ref{fig:batch}, \ref{fig:infection}). Ultimately 6 samples failed the quality checks and were removed from all downstream analyses (Supplementary Fig. \ref{fig:outliers}).

We performed a standard differential expression analysis using a linear modeling framework, defined in equation (\ref{eq:limma}). As expected, there was a strong response to infection with MTB in both resistant and susceptible individuals (Supplementary Fig. \ref{fig:limma-supp}). Considering the resistant individuals, we identified 3,486 differentially expressed (DE) genes between the non-infected and infected states at a q-value of 10% and an arbitrary absolute log-fold change greater than 1. Similarly, 3,789 genes were DE between the non-infected and infected states for resistant individuals at a q-value of 10% and an absolute log fold change greater than 1. The DE genes included the important immune response factors *IL12B*, *REL*, and *TNF*. While the treatment effect was obvious in all individuals, of most interest were the patterns of gene expression differences between susceptible and resistant individuals in either the non-infected or infected states (Fig. \ref{fig:limma}). We identified 645 DE genes between resistant and susceptible individuals in the non-infected state at a q-value of 10%, including *ATPV1B2*, *FEZ2*, *PSMA2*, *TNFRSF25*, and *TRIM38*. In contrast, no genes were DE between resistant and susceptible individuals in the non-infected state (q-value of 10%).



\begin{figure}[p]

\centering

\includegraphics[width=\linewidth]{../figure/limma.pdf}

\caption{

Differential expression analysis. The top panel contains the distribution of unadjusted p-values after testing for differential expression between susceptible and resistant individuals in the (a) noninfected or (b) infected state. The bottom panel contains the corresponding volcano plots for the (c) noninfected and (d) infected states. The x-axis is the log fold change in gene expression level between susceptible and resistant individuals and the y-axis is the –log10 p-value. Red indicates genes which are significant differentially expressed with a q-value less than 10%.

}

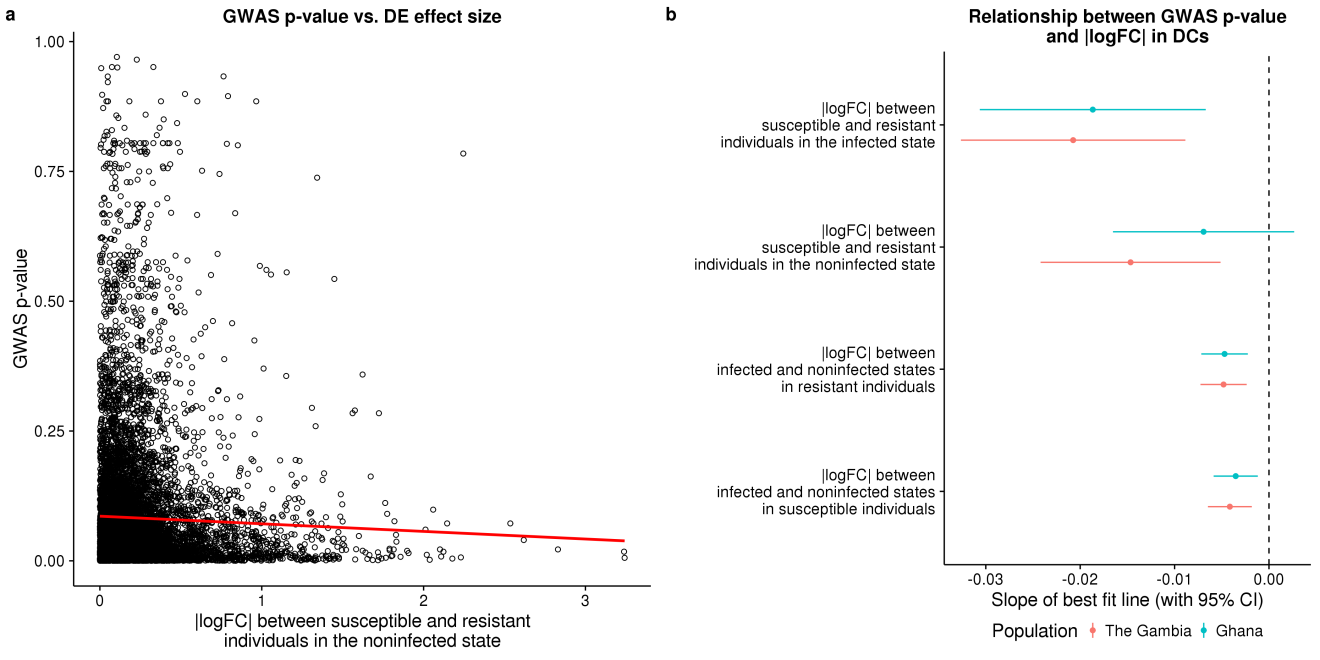
\label{fig:limma}

\end{figure}

## Differentially expressed genes are enriched with TB susceptibility loci

We next asked for evidence that genes classified as DE between individuals sensitive and resistant to TB in our *in vitro* experimental system play a role in determining susceptibility to TB. To do this, we compared our results to those from a TB susceptibility GWAS conducted in The Gambia and Ghana \cite{Thye2010}. Specifically, for each gene we considered the SNP with the lowest GWAS p-value among all SNPs located within 50 kb of its transcription start site (TSS). If DE genes in our experiment were enriched for TB susceptibility loci, we expected a negative correlation between the absolute values of the log fold changes in our experiment and the GWAS p-values. We have indeed observed this pattern (Fig. \ref{fig:gwas}).

Interestingly, we observed the steepest negative slopes for the tests comparing differential expression between susceptible and resistant individuals in the non-infected or infected states (Fig. \ref{fig:gwas}b). However, the slopes of the best fit lines for the tests of the effect of treatment in either resistant or susceptible individuals were also negative. Of particular interest as potential genes involved in TB susceptibility were the genes that had a p-value less than 0.01 in both The Gambia and Ghana GWAS and an absolute log fold change greater than 2 between susceptible and resistant individuals in the noninfected state (Fig. \ref{fig:gwas}a). Only 2 genes met these criteria: *CCL1* and *UNC13A*.



\begin{figure}[ht]

\centering

\includegraphics[width=\linewidth]{../figure/gwas.pdf}

\caption{

Comparison of differential expression and GWAS results. (a) The relationship between p-values from the GWAS in The Gambia \cite{Thye2010} and the absolute values of the log fold changes between susceptible and resistant individuals in the noninfected state. The red line is the least squares regression. (b) The slopes (with 95% confidence interval) of the regression lines for each test. The results from the GWAS in Gambia are in red and those from Ghana in blue. All slopes are significantly different than 0 (t-test *P* < 0.05), except for the slope between the Ghana GWAS p-values and the absolute values of the log fold changes between susceptible and resistant individuals in the noninfected state.

}

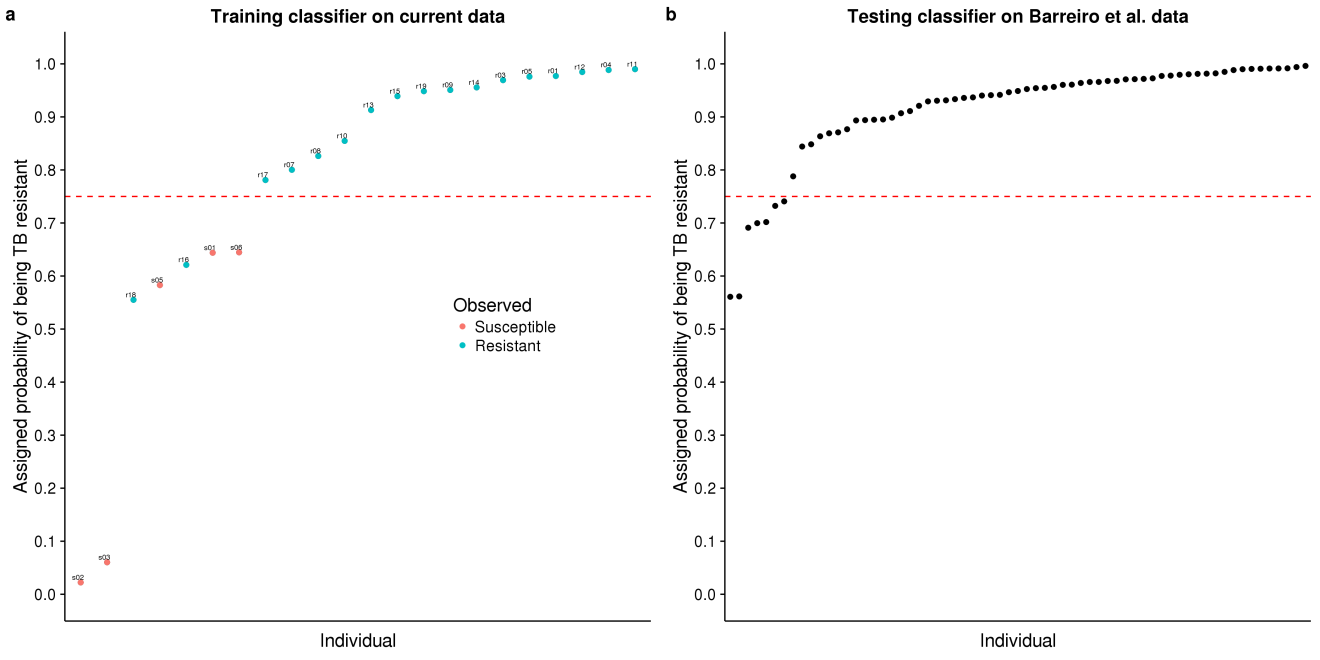
\label{fig:gwas}

\end{figure}

## Can gene expression levels in the non-infected state predict TB susceptibility status?

Next we attempted to build a gene expression-based classifier to predict TB susceptibility status. We focused on the gene expression levels measured in the non-infected state both because this is where we observed the largest gene regulatory differences between susceptible and resistant individuals (Fig. \ref{fig:limma}ac), and also because, from the perspective of translational application, it is more practical to obtain gene expression data from non-infected DCs. We trained a support vector machine using the 99 genes that were differentially expressed between resistant and susceptible individuals in the non-infected state at a q-value less than 5% (see Methods for a full description of how we selected this model). Encouragingly, we observed a clear separation between susceptible and resistant individuals when comparing the predicted probability of being resistant to TB for each sample obtained from leave-one-out-cross-validation (Fig. \ref{fig:classifier}a). Using a cutoff of 0.75 for the predicted probability of being resistant to TB, we obtained a sensitivity of 100% (5 out of 5 susceptible individuals classified as susceptible) and a specificity of ~71% (5 out of 7 individuals classified as susceptible were true positives).

Unfortunately our current data set is too small to properly split into separate training and testing sets (it is challenging to collect samples from previous TB patients, who are healthy and have no medical reason to go back for a GP visit). To our knowledge, there are also no other similar data sets available. Thus, in order to further assess the plausibility of our model, we applied the classifier to an independent study, which collected genome-wide gene expression levels in DCs from 65 healthy individuals \cite{Barreiro2012}, none with a previous history of TB. Using the cutoff of 0.75 for the probability of being resistant to TB (determined to be optimal in the training set), ~11% (7 of 65) of the individuals were classified as susceptible to TB. This result is intriguing similar to the estimate that roughly 10% of the general population is susceptible to TB (Fig. \ref{fig:classifier}b).



\begin{figure}[ht]

\centering

\includegraphics[width=\linewidth]{../figure/classifier-svm.pdf}

\caption{

Classifying TB susceptible individuals using a support vector machine model. (a) The estimates of predicted probability of TB resistance from the leave-one-out-cross-validation for individuals in the current study. The red circles represent individuals known to be susceptible to TB, and blue those resistant to TB. The horizontal dashed red line at a probability of 0.75 separates susceptible and resistant individuals. (b) The estimates of predicted probability of TB resistance from applying the classifier trained on the data from the current study to a test set of independently collected healthy individuals \cite{Barreiro2012}.

}

\label{fig:classifier}

\end{figure}

# Discussion

We obtained dendritic cells (DCs) from individuals that were known to be susceptible or resistant to developing active tuberculosis (TB) and measured genome-wide gene expression levels in non-infected DCs and DCs infected with *Mycobacterium tuberculosis* (MTB) for 18 hours. As expected, there were large changes in gene expression due to MTB infection in both resistant and susceptible individuals (Supplementary Fig. \ref{fig:limma-supp}). We identified 645 genes, which were differentially expressed (DE) between susceptible and resistant individuals in the non-infected state; whereas, we did not observe any DE genes between susceptible and resistant individuals in the infected state (Fig. \ref{fig:limma}). This suggests that the differences in the transcriptomes between DCs of resistant and susceptible individuals are present pre-infection, and affect the initial response to MTB. Yet, 18 hours after infection gene expression profiles in both susceptible and resistant individuals have converged to the same gene regulatory network to fight the active infection. We chose to measure gene expression 18 hours post-infection because this time point was previously associated with a large change in genome-wide gene expression levels \cite{Tailleux2008}. Given our observations, however, future studies investigating the difference in the innate immune response between individuals resistant and susceptible to TB may want to focus on earlier time points post infection.

Among the 645 DE genes between resistant and susceptible individuals in the non-infected state, there were many interesting genes involved in important innate immune activities critical for fighting MTB and other pathogens such as autophagy \cite{Deretic2014, Castrejon-Jimenez2015}, phagolysosomal acidification, and antigen processing. In particular, *FEZ2*, a suppressor of autophagosome formation \cite{Spang2014}, was down-regulated when DCs were infected with MTB; however, in the non-infected DCs, this gene has elevated expression level in susceptible compared with resistant individuals. In turn, *ATP6V1B2*, a gene coding for a subunit of the proton transporter responsible for acidifying phagolysosomes \cite{Sturgill-Koszycki1994, Hornef2002, Hestvik2005}, has increased expression in susceptible individuals compared to resistant in the non-infected state. Lastly, genes coding for nine subunits of the proteasome, which is critical for processing of MTB antigens to be presented via major histocompatibility complex (MHC) class I molecules \cite{Flynn1992, Grotzke2009, Grotzke2010, LindestamArlehamn2014}, have increased expression in susceptible individuals compared to resistant in the non-infected state. These genes are candidates for future functional studies investigating the mechanisms of TB susceptibility.

To our knowledge, our study was only the second to collect data from *in vitro* MTB infected innate immune cells isolated from individuals known to be susceptible to MTB (Thuong et al., 2008). However, there were substantial differences between our study and that of Thuong et al., 2008 \cite{Thuong2008}. First, they isolated and infected macrophages, the primary target host cell in which MTB resides; whereas, we infected DCs, which play a larger role in stimulating the adaptive immune response to MTB. Second, the susceptible individuals in Thuong et al., 2008 had an active TB infection at the time the cells were isolated; whereas, our individuals had recovered from a past TB infection. Third, we collected samples from a larger number of resistant individuals (19 versus 4), increasing our power to distinguish between the gene expression profiles of susceptible and resistant individuals.

We observed that DE genes in our *in vitro* experimental system were enriched for lower GWAS p-values (Fig. \ref{fig:gwas}). This suggests that such *in vitro* approaches are informative for interrogating the genetic basis of disease susceptibility. That being said, we recognized multiple caveats with this analysis. First, assigning SNPs to their nearest gene on the linear chromosome is problematic because regulatory variants can have longer range effects. Second, the negative slopes of the least squares regression line were quite small because there were many SNPs with low p-values nearby genes with low effect sizes in our experiment. It is possible that these variants contribute to TB susceptibility by affecting gene expression in other cell types or environmental conditions. Third, the individuals in our study were Europeans; whereas, the GWAS were conducted in Africans. Nevertheless…

Not only did this analysis identify a global enrichment of TB susceptibility loci, but by intersecting our DE results and the GWAS we were able to identify two genes (*CCL1* and *UNC13A*) which were marginally significant in both. Interestingly, both of these genes have prior evidence of playing important roles in MTB infection. *CCL1* is a chemokine that stimulates migration of monocytes \cite{Miller1992}. In our study, it was upregulated in susceptible individuals compared to resistant in both the noninfected and infected states (but did not reach statistical significance in either) and was statistically significantly upregulated with MTB treatment. The previous differential expression study of TB susceptibility mentioned above found that *CCL1* was upregulated to a greater extent 4 hours post MTB-infection in macrophages isolated from individuals with an active TB infection (i.e. susceptible) compared to individuals with a latent TB infection (i.e. resistant) \cite{Thuong2008}. Additionally they performed a candidate gene association study and found that SNPs nearby *CCL1* were associated with TB susceptibility. In a previous study from our lab, we discovered that *CCL1* was one of only 288 genes that were differentially expressed in macrophages 48 hours post-infection with MTB and related mycobacterial species but not unrelated virulent bacteria \cite{Blischak2015}. *UNC13A* is involved in vesicle formation \cite{Sudhof2004}. In our study, it was downregulated in susceptible individuals compared to resistant in both the noninfected and infected states (but did not reach statistical significance in either) and was statistically significantly upregulated with MTB treatment. In our past study mapping expression quantitative trait loci (eQTLs) in DCs 18 hours post-infection with MTB, *UNC13A* was one of only 98 genes which was associated with an eQTL post-infection but not pre-infection, which we called an MTB-specific eQTL \cite{Barreiro2012}. Thus our new results increased the evidence that *CCL1* and *UNC13A* play important roles in TB susceptibility.

**Subtitle?**

Previous attempts to use gene expression based classifiers in the context of TB have focused on predicting the status of an infection rather than the susceptibility status of an individual \cite{Berry2010, OGarra2013, Blankley2014}. In other words, the goal of most previous study was to detect individuals in an early stage of an active TB infection when antibiotic intervention would be most effective or to monitor the effectiveness of a treatment regimen \cite{Maertzdorf2015}. In contrast, our goal was not to distinguish between an active or latent infection, but instead to be able to determine susceptibility status before individuals have an active TB infection. Even with our small sample size, we were able to successfully train a classier with high sensitivity and decent specificity. Because such a classification of susceptibility status could affect the decision of whether or not to take antibiotics to treat a latent MTB infection \cite{Munoz2015}, false negatives (susceptible individuals mistakenly classified as resistant) would be much more harmful than false positives (resistant individuals mistakenly classified as susceptible), which is why we emphasized sensitivity over specificity.

At this time, we are not aware of any other data set from healthy individuals known to be sensitive to TB, with which we can further test our classifier. When we applied our classifier to an independent set of non-infected DCs isolated from healthy individuals of unknown susceptibility status, our model predicted that ~11% of the individuals were susceptible TB, which reassuringly is similar to the average in the general population (10%). Despite this success, our results must be interpreted cautiously as a proof-of-principle due to our very small sample size of only 5 susceptible individuals. That said, our promising results in this small study suggest that collecting blood samples from a larger cohort of susceptible individuals would enable building a gene expression based classifier able to confidently assess risk of TB susceptibility. By reducing the number of resistant individuals receiving treatment for a latent TB infection, we can eliminate the adverse health effects of a 6 month regimen of antibiotics for these individuals and also reduce the selective pressures on MTB to develop drug resistance.

# Methods

## Ethics Statement

We recruited 25 subjects to donate a blood sample for use in our study. All methods were carried out in accordance with relevant guidelines and regulations. The experimental protocols were approved by the Institutional Review Boards of the University of Chicago (10-504-B) and the Institut Pasteur (IRB00006966). All study participants provided written informed consent.

## Sample collection

We collected whole blood samples from healthy Caucasian male individuals living in France. The putatively resistant individuals tested positive for a latent TB infection in an interferon-$\gamma$ release assay, but had never developed active TB. The putatively sensitive individuals had developed active TB in the past, but were currently healthy.

## Isolation and infection of dendritic cells

We performed these experiments as previously described \cite{Barreiro2012}. Briefly, we isolated mononuclear cells from the whole blood samples using Ficoll-Paque centrifugation, extracted monocytes via CD14 positive selection, and differentiated the monocytes into dendritic cells (DCs) by culturing them for 5 days 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). Next we infected the DCs with *Mycobacterium tuberculosis* (MTB) H37Rv at a multiplicity of infection of 1-to-1 for 18 hours.

## RNA extraction and sequencing

We extracted RNA using the Qiagen miRNeasy Kit and prepared sequencing libraries using the Illumina TruSeq Kit. We sent the master mixes to the University of Chicago Functional Genomics Facility to be sequenced on an Illumina HiSeq 4000. We designed the batches for RNA extraction, library preparation, and sequencing to balance the experimental factors of interest and thus avoid potential technical confounders (Supplementary Fig. \ref{fig:process}).

## Read mapping

We mapped reads to human genome hg38 (GRCh38) using Subread \cite{Liao2013} and discarded non-uniquely mapping reads. We downloaded the exon coordinates of 19,800 Ensembl \cite{Yates2016} protein-coding genes (Ensembl 83, Dec 2015, GRCh38.p5) using the R/Bioconductor \cite{Huber2015} package biomaRt \cite{Durinck2005, Durinck2009} and assigned mapped reads to these genes using featureCounts \cite{Liao2014}.

## Quality control

First we filtered genes by their expression level by removing all genes with a median log2 counts per million (cpm) less than zero. This resulted in a final set of 11,336 genes for downstream analysis (Supplementary Fig. \ref{fig:gene}, Supplementary Data S1). Next we used principal components analysis (PCA) and hierarchical clustering to identify and remove 6 outliers (Supplementary Fig. \ref{fig:heat-all}, \ref{fig:heat-filt}, \ref{fig:outliers}). We did this systematically by removing any sample that did not fall within two standard deviations of the mean for the first six PCs. Furthermore, for the first PC which separated the samples by treatment, we calculated a separate mean for the noninfected and infected samples.

After filtering lowly expressed genes and removing outliers, we recalculated the PCA to check for any potential confounding technical batch effects (Supplementary Fig. \ref{fig:batch}). Reassuringly, the major sources of variation in the data were from the biological factors of interest. PC1 was strongly correlated with the effect of treatment, and PCs 2-6 were correlated with inter-individual variation. The only concerning technical factor was the infection experiments, which were done in 12 separate batches (Supplementary Fig. \ref{fig:process}). Infection batch correlated with PCs 3 and 5; however, we verified that this variation was not confounded with our primary outcome of interest, TB susceptibility (Supplementary Fig. \ref{fig:infection}).

## Differential expression analysis

We used limma+voom \cite{Smyth2004, Law2014, Ritchie2015} to implement the following linear model to test for differential expression:

\begin{equation} \label{eq:limma}

Y\ \sim \beta\_{0} + X\_{treat}\beta\_{treat} + X\_{status}\beta\_{status} + X\_{treat,status}\beta\_{treat,status} + I + \epsilon

\end{equation}

where $\beta\_{0}$ is the mean expression level in noninfected cells of resistant individuals, $\beta\_{treat}$ is the fixed effect of treatment in resistant individuals, $\beta\_{status}$ is the fixed effect of susceptibility status in noninfected cells, $\beta\_{treat,status}$ is the fixed interaction effect of treatment in susceptible individuals, and $I$ is the random effect of individual. The random individual effect was implemented using the limma function duplicateCorrelation \cite{Smyth2005}. To jointly model the data with voom and duplicateCorrelation, we followed the recommended best practice of running both voom and duplicateCorrelation twice in succession \cite{Liu2015}.

We used the model to test 4 separate hypotheses (Supplementary Data S2). We identified genes which were differentially expressed (DE) between infected and noninfected DCs of resistant individuals by testing $\beta\_{treat} = 0$, genes which were DE between infected and noninfected DCs of susceptible individuals by testing $\beta\_{treat} + \beta\_{treat,status} = 0$, genes which were DE between susceptible and resistant individuals in the noninfected state by testing $\beta\_{status} = 0$, and genes which were DE between susceptible and resistant individuals in the infected state by testing $\beta\_{status} + \beta\_{treat,status} = 0$. We corrected for multiple testing using q-values estimated via adaptive shrinkage \cite{Stephens2016} and considered differentially expressed genes as those with a q-value less than 10%.

## Comparison to GWAS results

The GWAS p-values were from a study of TB susceptibility conducted in The Gambia and Ghana \cite{Thye2010}. To compare our differential expression results to these genetic associations, we assigned each gene the p-value of the SNP with the minimum p-value out of all the SNPs located within 50 kb up or downstream of its transcription start site. Specifically, we obtained the genomic coordinates of the SNPs with the R/Bioconductor \cite{Huber2015} package SNPlocs.Hsapiens.dbSNP144.GRCh38 and matched SNPs to nearby genes using GenomicRanges \cite{Lawrence2013}. 10,260 of the 11,336 were assigned an association p-value. For each of the 4 hypotheses we tested, we performed least squares regression of the differential expression effect sizes (the log fold changes) and the assigned GWAS p-values. We assessed the statistical significance of these regressions using the standard t-test and reported the slope of each regression line (Fig. \ref{fig:gwas}, Supplementary Data S3). Lastly, we also observed a negative correlation between the GWAS p-value assigned to a gene and the number of SNPs tested nearby that gene (Supplementary Fig. \ref{fig:gwas-n-snps}). However, we could not think of an explanation for why genes with a larger log fold change in our *in vitro* experimental system would have more nearby genetic variation, and thus we do not believe this relationship biased our observation of a negative correlation between GWAS p-value and log fold change.

## Classifier

The training set included the 44 high-quality noninfected samples from this study with known susceptibility status. The test set included the 65 noninfected samples from one of our previous studies in which the susceptibility status is unknown \cite{Barreiro2012}, and thus assumed to be similar to that in the general population (~10%). Because the two studies are substantially different, we took multiple steps to make them comparable. First, we subset to include only those 9,450 genes which were assayed in both. Second, because the dynamic range obtained from RNA-seq (current study) and microarrays (previous study \cite{Barreiro2012}) were very different, we normalized the gene expression levels to a standard normal with $\mu = 0$ and $\sigma = 1$ (Supplementary Fig. \ref{fig:combined-dist}). Third, we corrected for the large, expected batch effect between the two studies by regressing out the first PC of the combined expression data using the limma function removeBatchEffect \cite{Ritchie2015} (Supplementary Fig. \ref{fig:combined-pca}).

To identify genes to use in the classifier, we performed a differential expression analysis on the normalized, batch-corrected data from the current study using the same approach described above (with the exception that we no longer used voom \cite{Law2014} since the data were no longer counts). Specifically, we tested for differential expression between susceptible and resistant individuals in the noninfected state and identified sets of genes to use in the classifier by varying the q-value cutoff. Cutoffs of 5%, 10%, 15%, 20%, and 25% corresponded to gene set sizes of 99, 385, 947, 1,934, and 3,697, respectively. We used the R package caret \cite{Kuhn2008} to train 3 different machine learning models: elastic net \cite{Friedman2010}, support vector machine \cite{Karatzoglou2004}, and random forest \cite{Liaw2002} (the parameters for each individual model were selected using the Kappa statistic). To assess the results of the model on the training data, we performed leave-one-out-cross-validation (LOOCV). In order to choose the model with the best performance, we calculated the difference between the mean of the LOOCV-estimated probabilities of being TB resistant for the samples known to be TB resistant and the corresponding mean for the samples known to be TB susceptible. This metric emphasized the ability to separate the susceptible and resistant individuals into two separate groups. Using this metric, the best performing model was the support vector machine with the 99 genes that are significantly differentially expressed at a q-value of 5% (Supplementary Fig. \ref{fig:class-compare}, Supplementary Data S4); however, both the elastic net (Supplementary Fig. \ref{fig:class-en}) and random forest (Supplementary Fig. \ref{fig:class-rf}) had similar performance. Lastly, we tested the classifier by predicting the probability of being TB resistant in the 65 healthy samples (Fig. \ref{fig:classifier}b). For evaluating the predictions on the test set of individuals with unknown susceptibility status, we used a relaxed cutoff of the probability of being TB resistant of 0.75, which was based on the ability of the model at this cutoff to classify all TB susceptible individuals in the training set as susceptible with only 2 false positives. As expected, the 99 genes used in the classifier had similar normalized, batch-corrected median expression levels in the noninfected state across both studies (Supplementary Fig. \ref{fig:class-exp}).

## Software implementation

We automated our analysis using Python (https://www.python.org/) and Snakemake \cite{Koster2012}. Our processing pipeline used the general bioinformatics software FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), MultiQC \cite{Ewels2016}, samtools \cite{Li2009}, and bioawk (https://github.com/lh3/bioawk). We used R \cite{R2015} for all statistics and data visualization. We obtained gene annotation information from the Ensembl \cite{Yates2016} and Lynx \cite{Sulakhe2016} databases. The computational resources were provided by the University of Chicago Research Computing Center. All code is available for viewing and reuse at https://github.com/jdblischak/tb-suscept.

## Data availability

The raw fastq files have been deposited in NCBI's Gene Expression Omnibus \cite{Edgar2002} and are accessible through GEO Series accession number GSEXXXXX (http://www.ncbi.nlm.nih.gov/geo/ query/acc.cgi?acc=GSEXXXXX). The RNA-seq gene counts and other summary data sets are included as Supplementary Data and are also available for download at https://github.com/jdblischak/tb-suscept/data.

# Acknowledgements

We thank T. Thye for sharing the GWAS data with us. This study was funded by National Institutes of Health (NIH) Grant AI087658 to YG and LT. JDB was supported by NIH T32GM007197. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

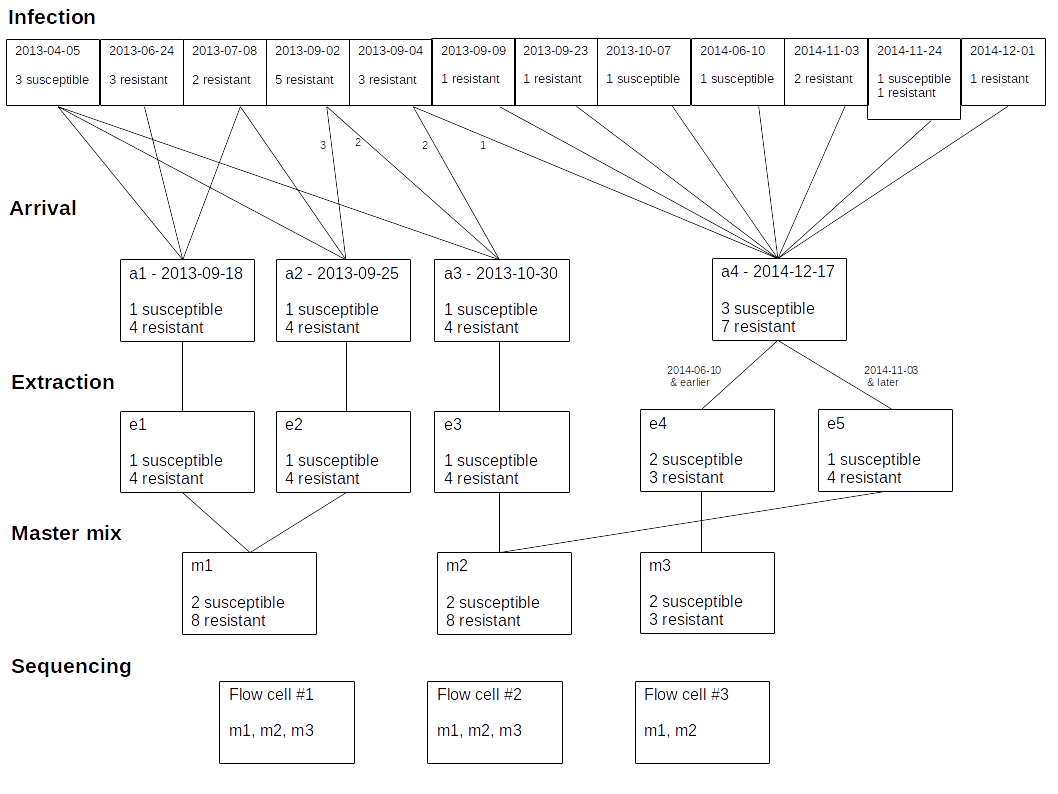
# Author Contributions

YG, LT, and LBB conceived of the study and designed the experiments. LT coordinated sample collection and performed the infection experiments. MM extracted the RNA and prepared the sequencing libraries. JDB analyzed the results. LBB and YG supervised the project. JDB and YG wrote the original draft. All authors reviewed the manuscript.

\bibliography{references}

# Supplementary Information

## Supplementary Figures



\begin{figure}[ht]

\centering

\includegraphics[width=\linewidth]{../figure/processing.pdf}

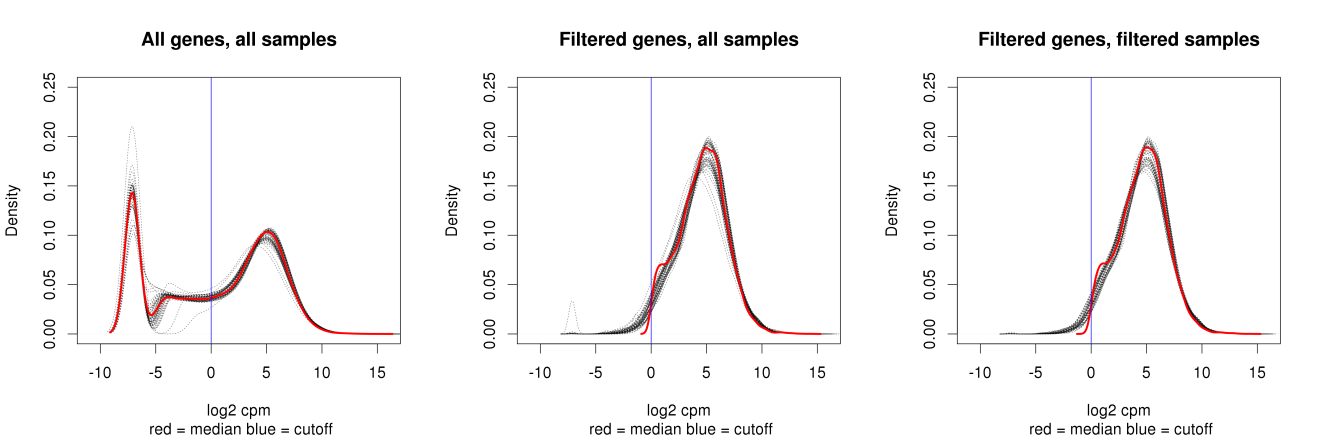
\caption{

Batch processing. We designed the processing of the samples to minimize the introduction of technical batch effects. Specifically, we attempted to balance the processing of samples obtained from susceptible and resistant individuals. In the diagram, each box represents a batch. “Infection” labels the batches of the infection experiments, “Arrival” labels the batch shipments of cell lysates arrived in Chicago, USA from Paris, France, “Extraction” labels the batches of RNA extraction, “Master Mix” labels the batches of library preparation, and “Sequencing” labels the batches of flow cells. Each master mix listed in a flow cell batch was sequenced on only one lane of that flow cell.

}

\label{fig:process}

\end{figure}



\begin{figure}[ht]

\centering

\includegraphics[width=\linewidth]{../figure/gene-exp-distribution.pdf}

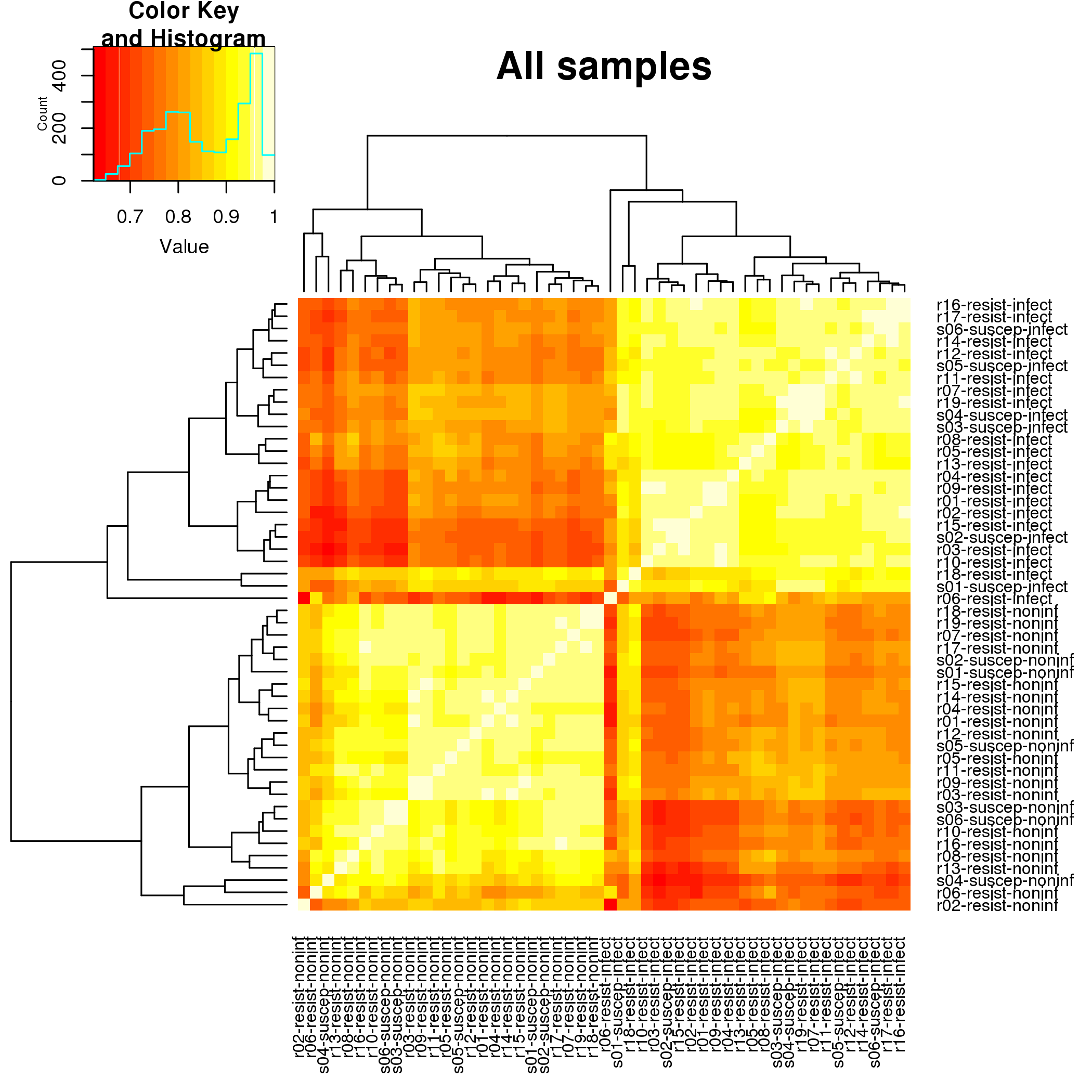
\caption{

Gene expression distributions before and after filtering genes and samples. The log2 counts per million (cpm) of each sample is plotted as a dashed gray line. The solid red line represents the median value across all the samples. The vertical solid blue line at $x = 0$ represents the cutoff used to filter lowly expressed genes based on their median log2 cpm. The left panel is the data from all 19,800 genes and 50 samples, the middle panel is the data from the 11,336 genes remaining after removing lowly expressed genes, and the right panel is the data from 11,336 genes and the 44 samples remaining after removing outliers.

}

\label{fig:gene}

\end{figure}



\begin{figure}[ht]

\centering

\includegraphics[width=\linewidth]{../figure/heatmap-all-samples.pdf}

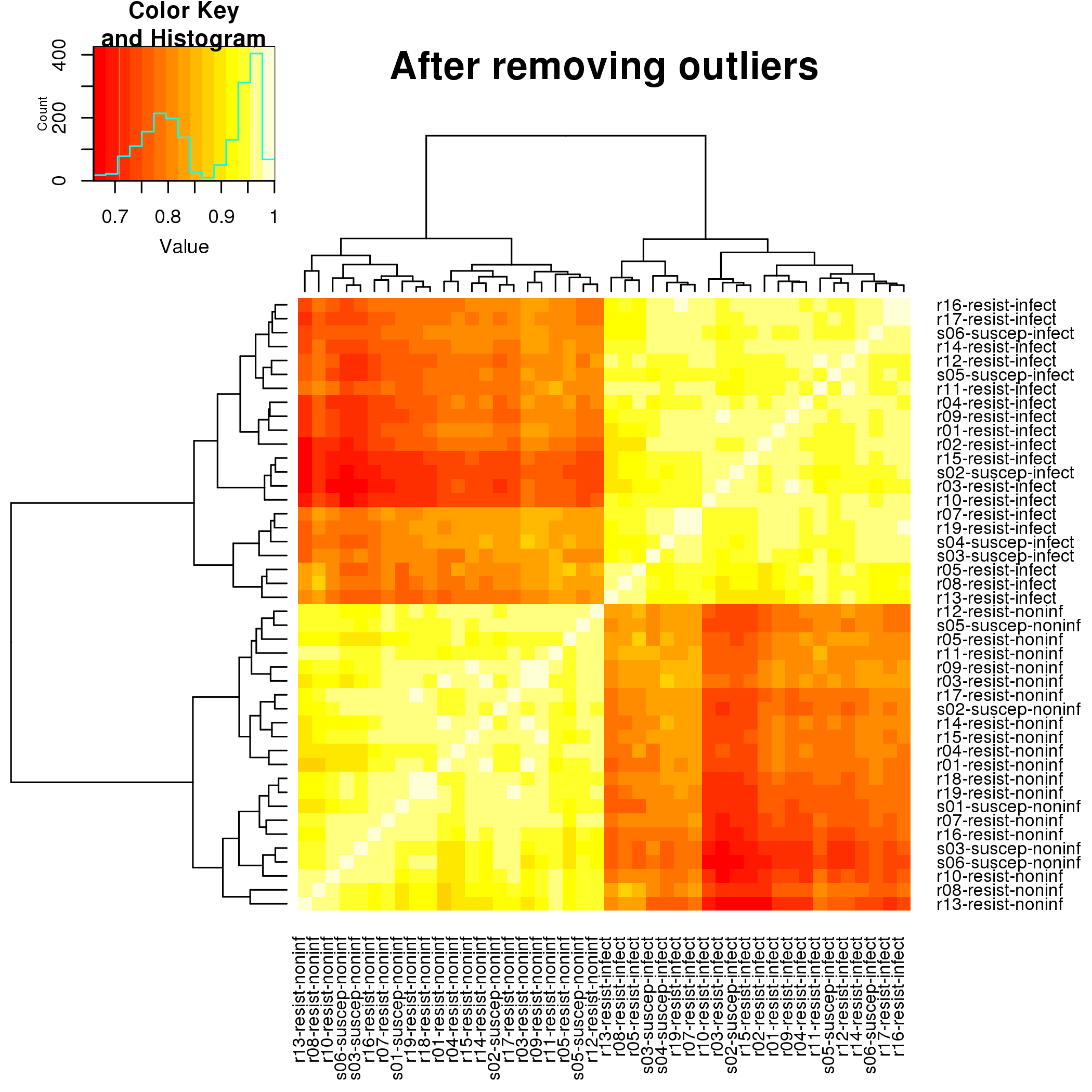
\caption{

Heatmap of correlation matrix of samples. Each square represents the Pearson correlation between the log2 cpm expression values of two samples. Red indicates a low correlation of zero and white represents a high correlation of 1. The dendrogram displays the results of hierarchical clustering with the complete linkage method. The outliers of the noninfected samples are s04-suscept-noninf, r02-resist-noninf, and r06-resist-noninf. The outliers of the infected samples are s01-suscep-infect, r06-resist-infect, and r18-resist-infect.

}

\label{fig:heat-all}

\end{figure}



\begin{figure}[ht]

\centering

\includegraphics[width=\linewidth]{../figure/heatmap-no-outliers.pdf}

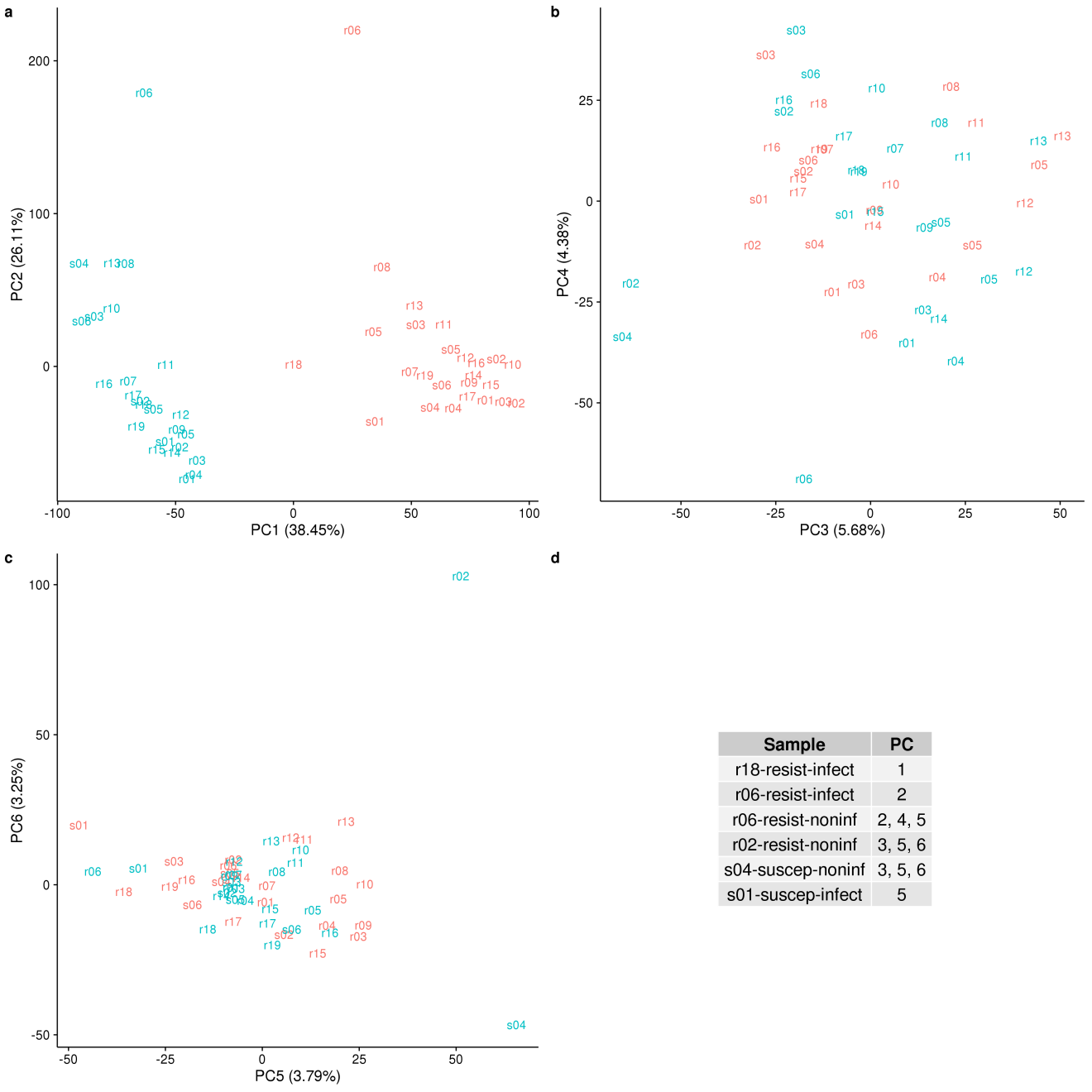
\caption{

Heatmap of correlation matrix after removing outliers. Each square represents the Pearson correlation between the log2 cpm expression values of two samples. Red indicates a low correlation of zero and white represents a high correlation of 1. The dendrogram displays the results of hierarchical clustering with the complete linkage method.

}

\label{fig:heat-filt}

\end{figure}



\begin{figure}[ht]

\centering

\includegraphics[width=\linewidth]{../figure/outliers.pdf}

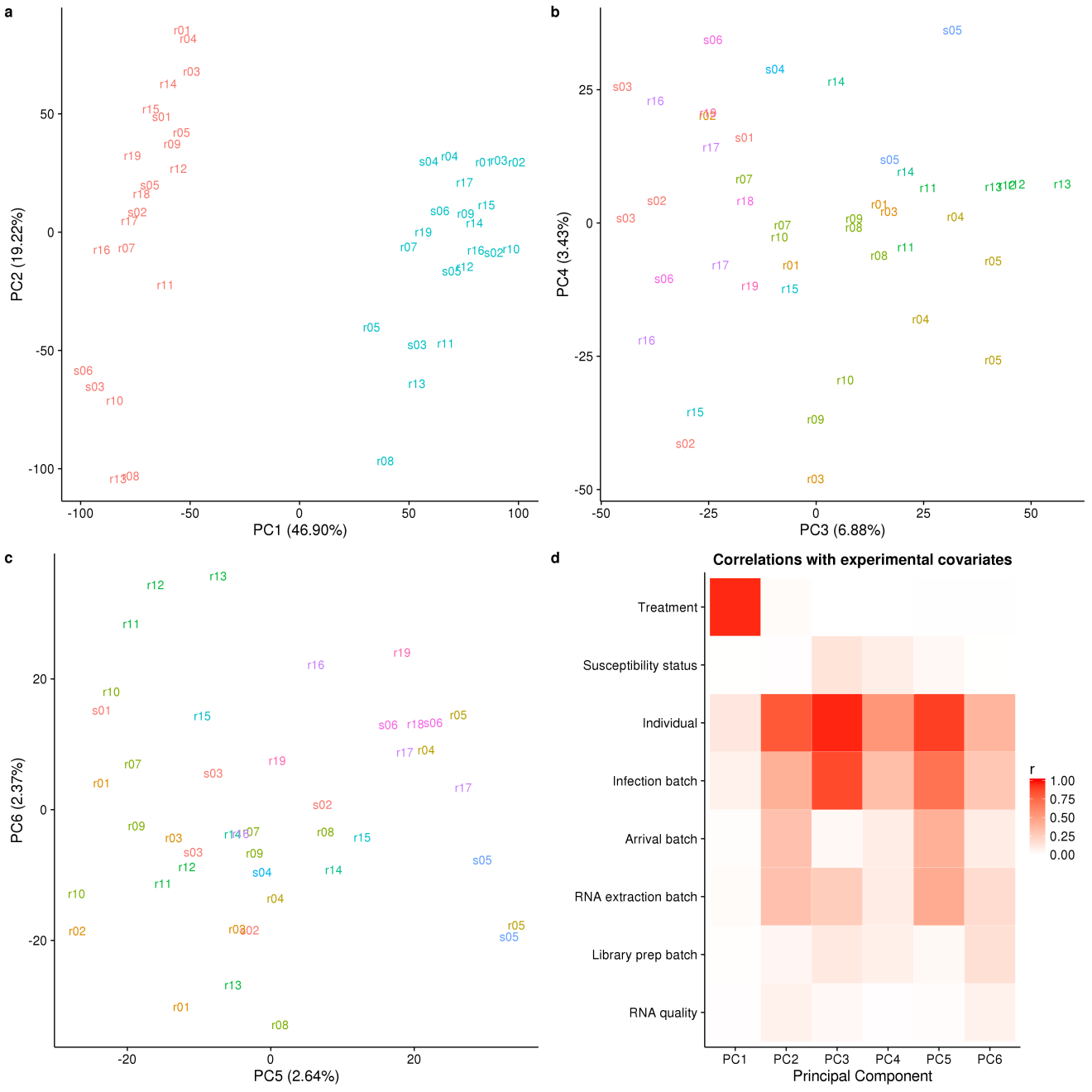
\caption{

Principal components analysis (PCA) to identify outliers. PC1 versus PC2 (a), PC3 versus PC4 (b), and PC5 versus PC6 (c). Each sample is represented by its 3-letter ID. “s” stands for susceptible and “r” for resistant, and the text is colored on the basis of treatment status (blue is noninfected; red is infected). The value is parentheses in each axis is the percentage of total variation accounted for by that PC. The outliers are listed in (d). These samples do not fall within 2 standard deviations of the mean value of the PCs listed in the right column. Note that a separate mean was calculated for the noninfected and infected samples for PC1 only.

}

\label{fig:outliers}

\end{figure}



\begin{figure}[ht]

\centering

\includegraphics[width=\linewidth]{../figure/batch-pca.pdf}

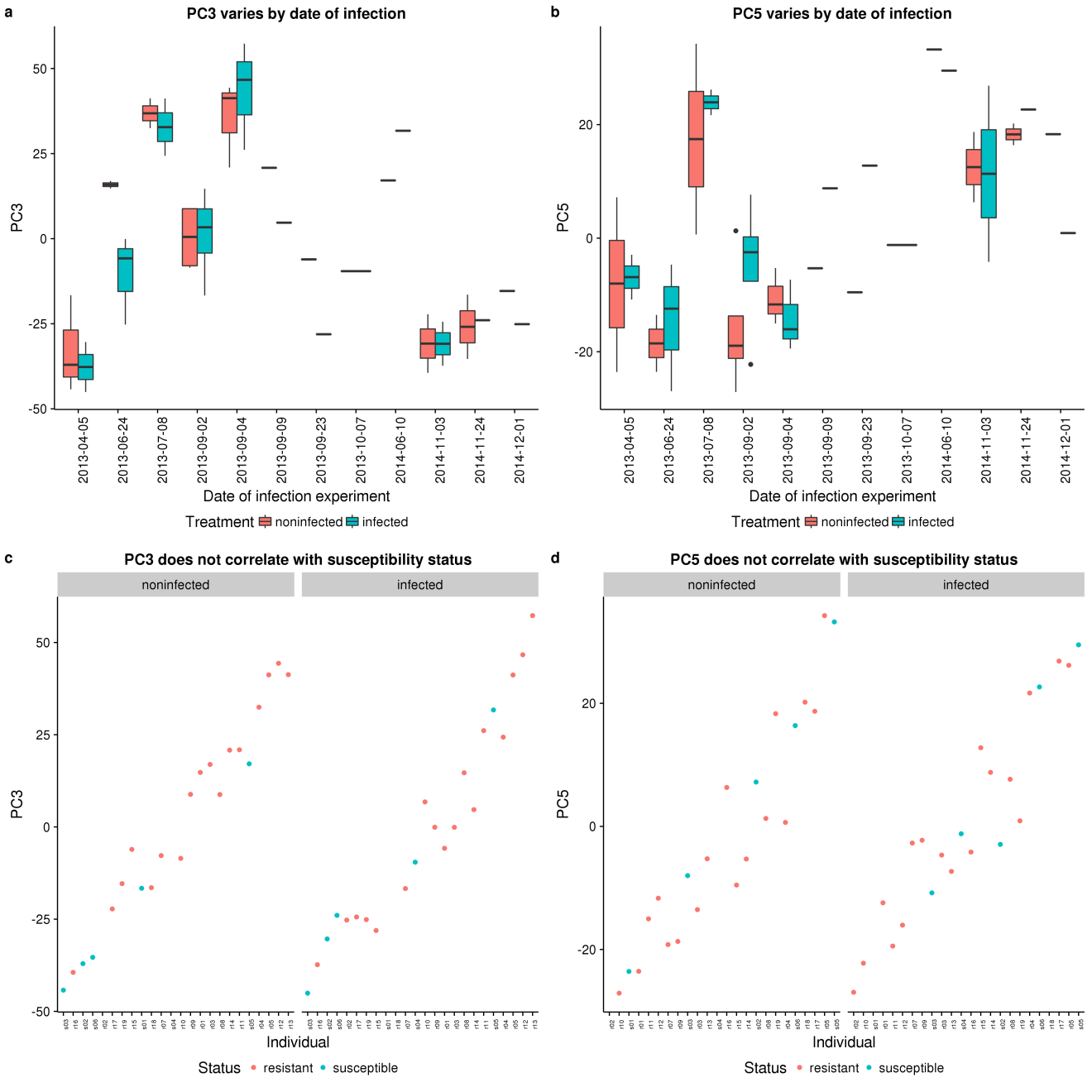
\caption{

Check for technical batch effects using principal components analysis (PCA). (a) PC1 versus PC2. The text labels are the individual identifiers. Red indicates noninfected samples and blue indicates infected. (b) PC3 versus PC4. The colors indicate the different infection batches. (c) PC5 versus PC6. The colors indicate the different infection batches. (d) The Pearson correlation of PCs 1-6 with each of the recorded biological and technical covariates. The correlations vary from 0 (white) to 1 (red).

}

\label{fig:batch}

\end{figure}



\begin{figure}[ht]

\centering

\includegraphics[width=\linewidth]{../figure/batch-infection.pdf}

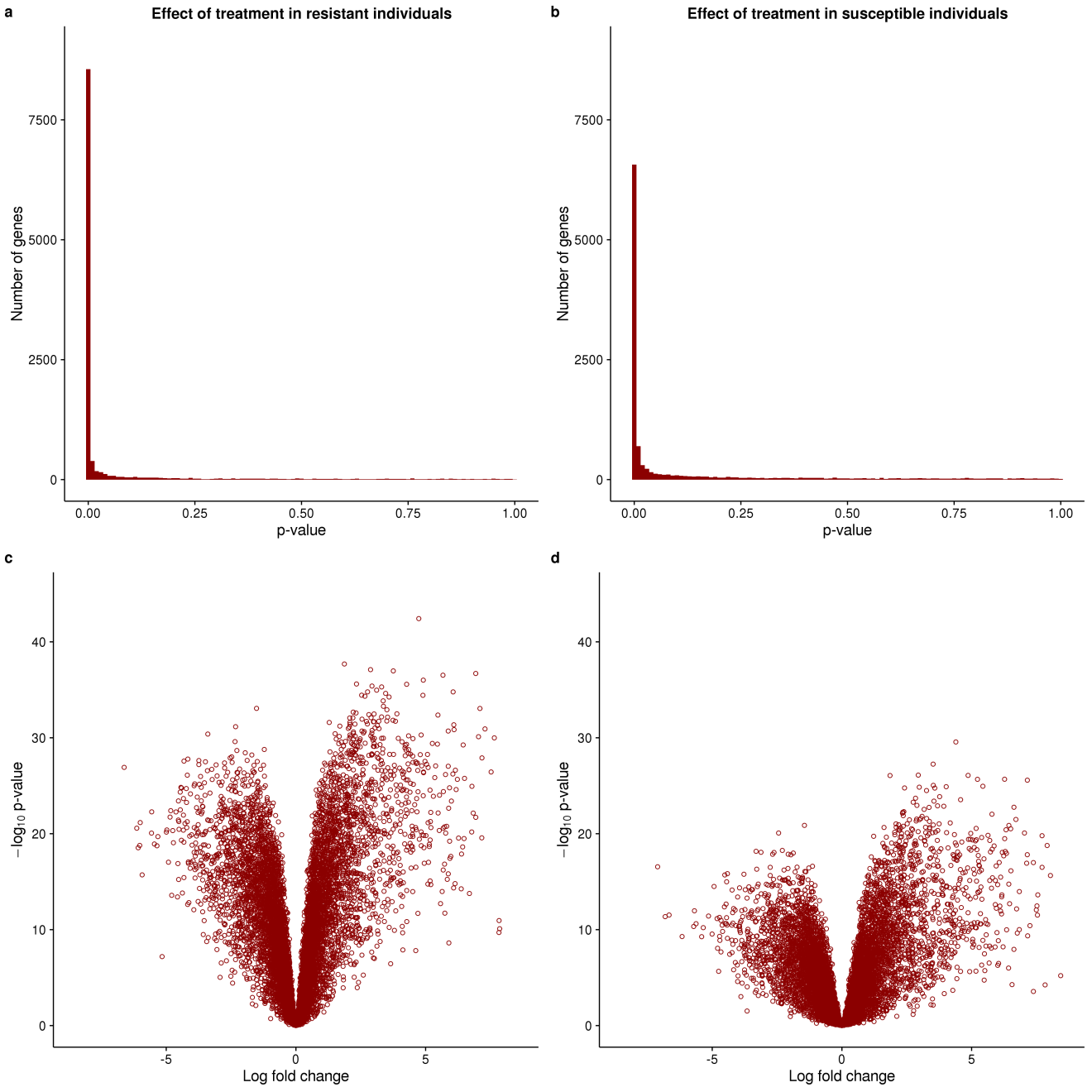
\caption{

Check for confounding effect of infection batch. PC3 (a) and PC5 (b) varied by the date of infection. Noninfected samples are in red and infected samples in blue. Importantly, however, this technical variation arising from infection batch did not correlate with the susceptibility status of the individuals (c and d). Resistant individuals are in red and susceptible individuals in blue.

}

\label{fig:infection}

\end{figure}



\begin{figure}[ht]

\centering

\includegraphics[width=\linewidth]{../figure/limma-supp.pdf}

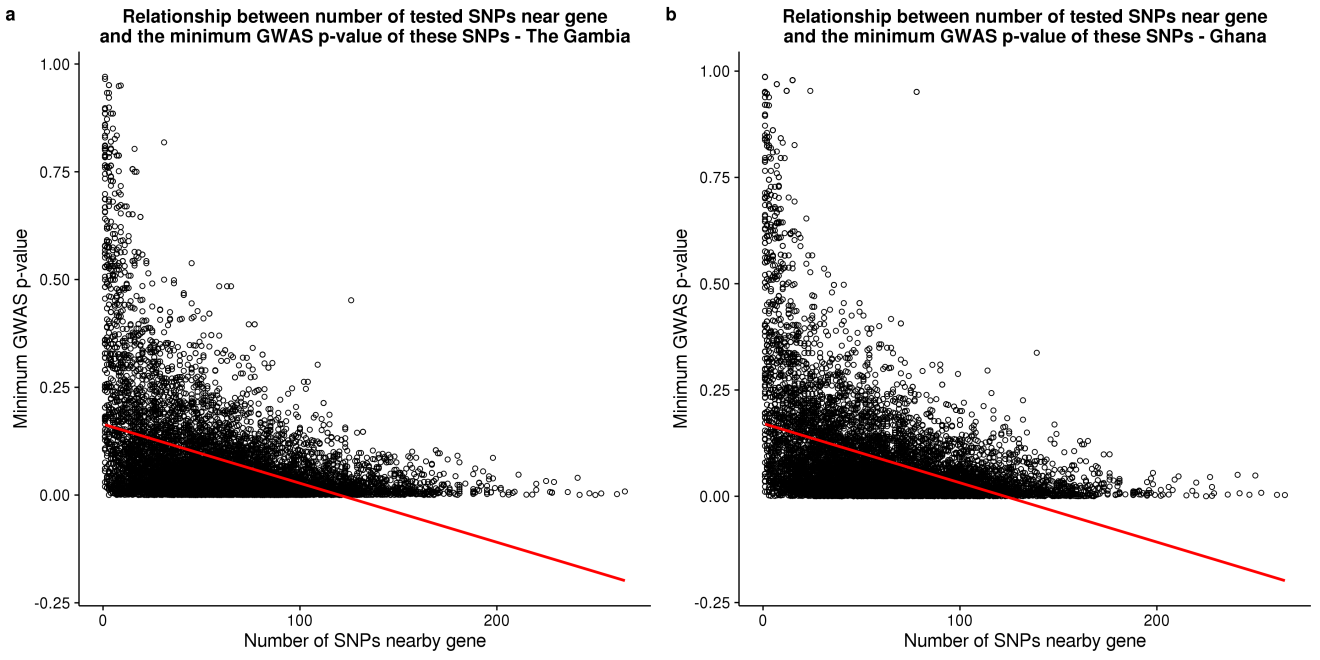
\caption{

Effect of treatment with MTB. The top panel contains the distribution of unadjusted p-values after testing for differential expression between the noninfected and infected states in (a) resistant and (b) susceptible individuals. The bottom panel contains the corresponding volcano plots for the (c) resistant and (d) susceptible individuals. The x-axis is the log fold change in gene expression level between susceptible and resistant individuals and the y-axis is the –log10 p-value. Red indicates genes which are significant differentially expressed with a q-value less than 10%. Because of the extremely skewed p-value distribution, all genes are significantly differentially expressed at this false discovery rate.

}

\label{fig:limma-supp}

\end{figure}



\begin{figure}[ht]

\centering

\includegraphics[width=\linewidth]{../figure/gwas-n-snps.pdf}

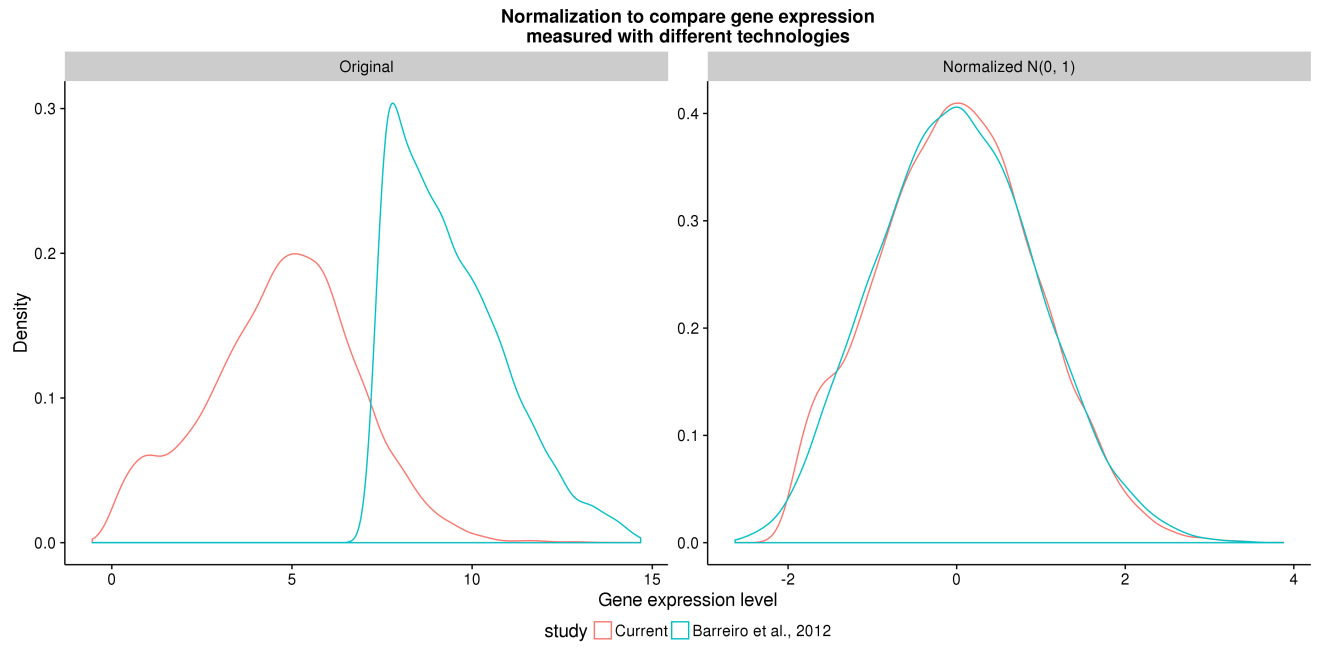
\caption{

Relationship between the minimum GWAS p-value assigned to a gene and the number of SNPs nearby that gene. The x-axis is the number of SNPs tested in the GWAS that are located 50 kb up or downstream of the TSS of each gene. The y-axis is the minimum GWAS p-value of the SNPs nearby each gene. The red line is the least squares regression line. The slope is significantly different than zero (t-test *P* < 0) for both the (a) The Gambia and (b) Ghana p-values.

}

\label{fig:gwas-n-snps}

\end{figure}



\begin{figure}[ht]

\centering

\includegraphics[width=\linewidth]{../figure/combined-distributions.pdf}

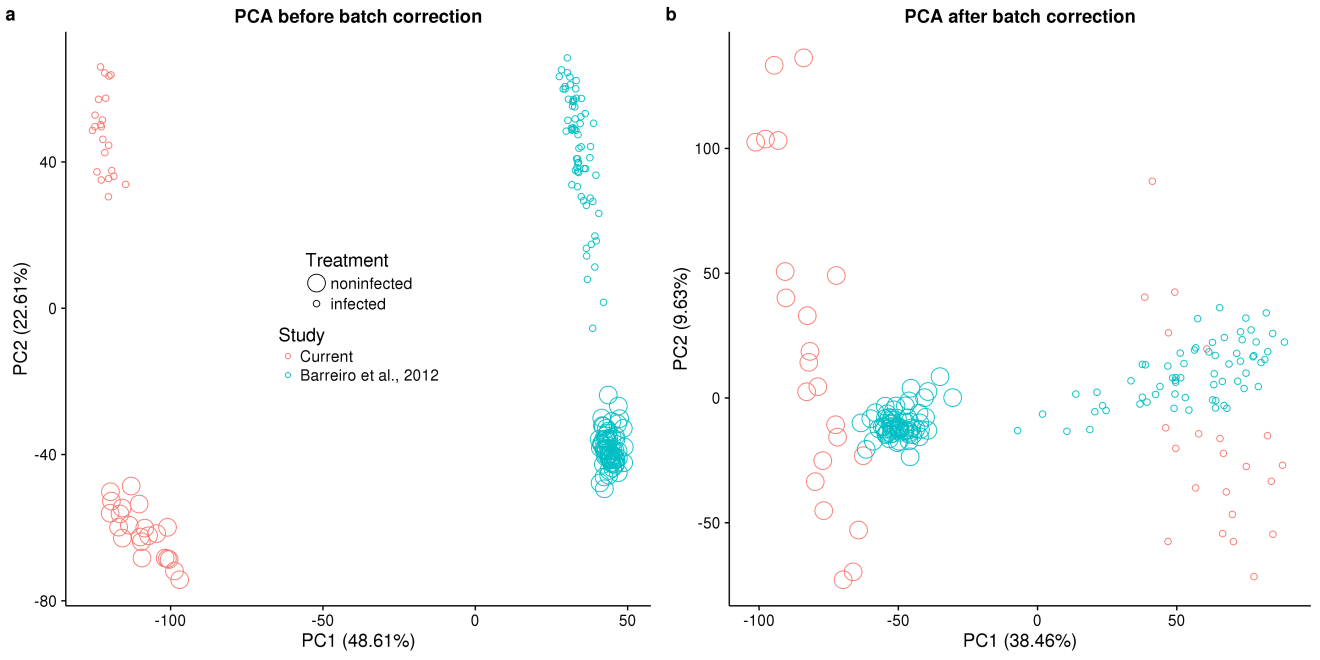
\caption{

Normalizing gene expression distributions. (left) The distribution of the median log2 cpm of the RNA-seq data from the current study in red compared to the distribution of the median gene expression levels of the microarray data from Barreiro et al., 2012 \cite{Barreiro2012} in blue. (right) The distributions of the same data sets after normalizing each sample to a standard normal distribution.

}

\label{fig:combined-dist}

\end{figure}



\begin{figure}[ht]

\centering

\includegraphics[width=\linewidth]{../figure/combined-pca.pdf}

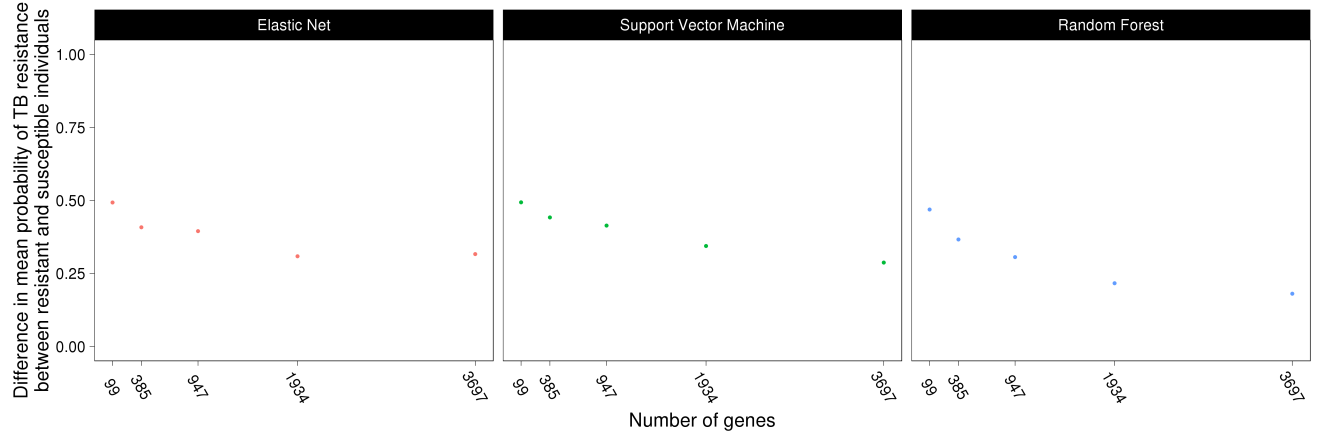
\caption{

Principal components analysis (PCA) of combined data sets. (a) PC1 versus PC2 of the combined data set of the RNA-seq data from the current study (red) and the microarray data from Barreiro et al., 2012 \cite{Barreiro2012} (blue). The large circles are noninfected samples, and the small circles are infected samples. The value in parentheses is the percentage of the total variation accounted for by that PC. (b) The same data after regressing the original PC1 in (a).

}

\label{fig:combined-pca}

\end{figure}



\begin{figure}[ht]

\centering

\includegraphics[width=\linewidth]{../figure/classifier-compare.pdf}

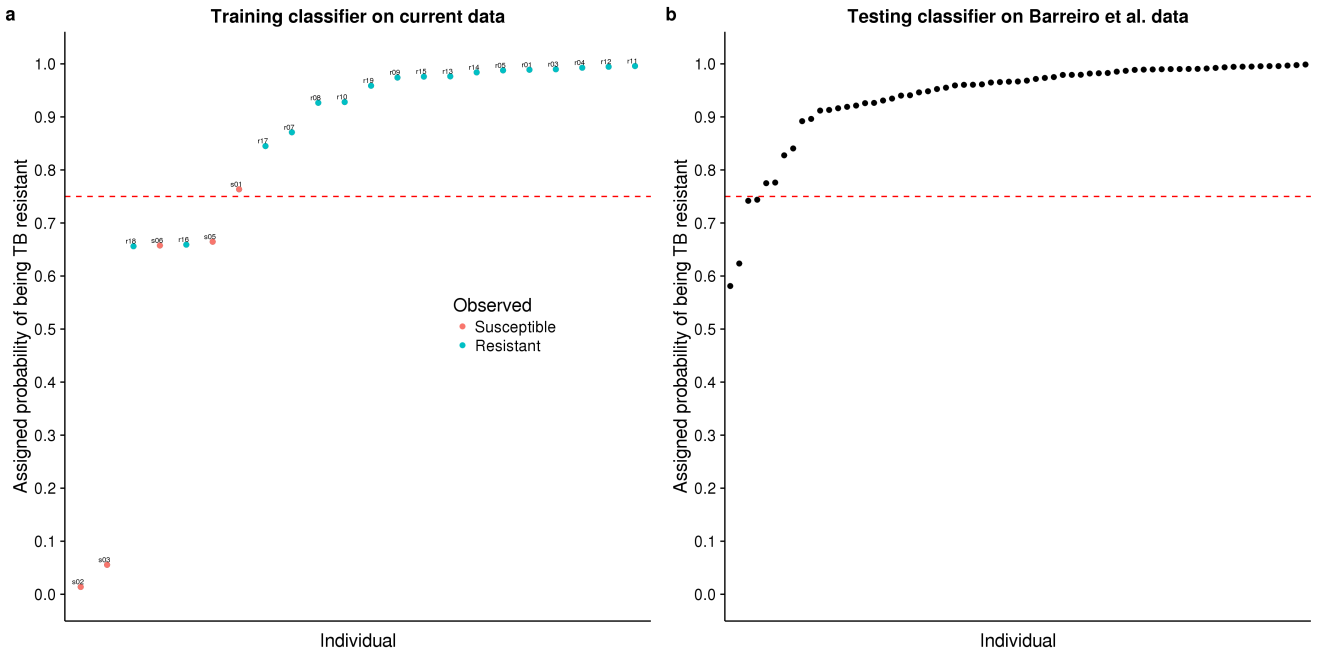
\caption{

Comparing the classification results of different methods and number of input genes. We compared 3 different machine learning methods (elastic net, support vector machine, random forest) and used 5 different sets of input genes. The input genes (x-axis) were obtained by varying the q-value cutoff for differential expression between susceptible and resistant individuals in the noninfected state from 5% to 25%. The evaluation metric (y-axis) was the difference of the mean assigned probability of being TB resistant between the known resistant and susceptible individuals in the current study.

}

\label{fig:class-compare}

\end{figure}



\begin{figure}[ht]

\centering

\includegraphics[width=\linewidth]{../figure/classifier-en.pdf}

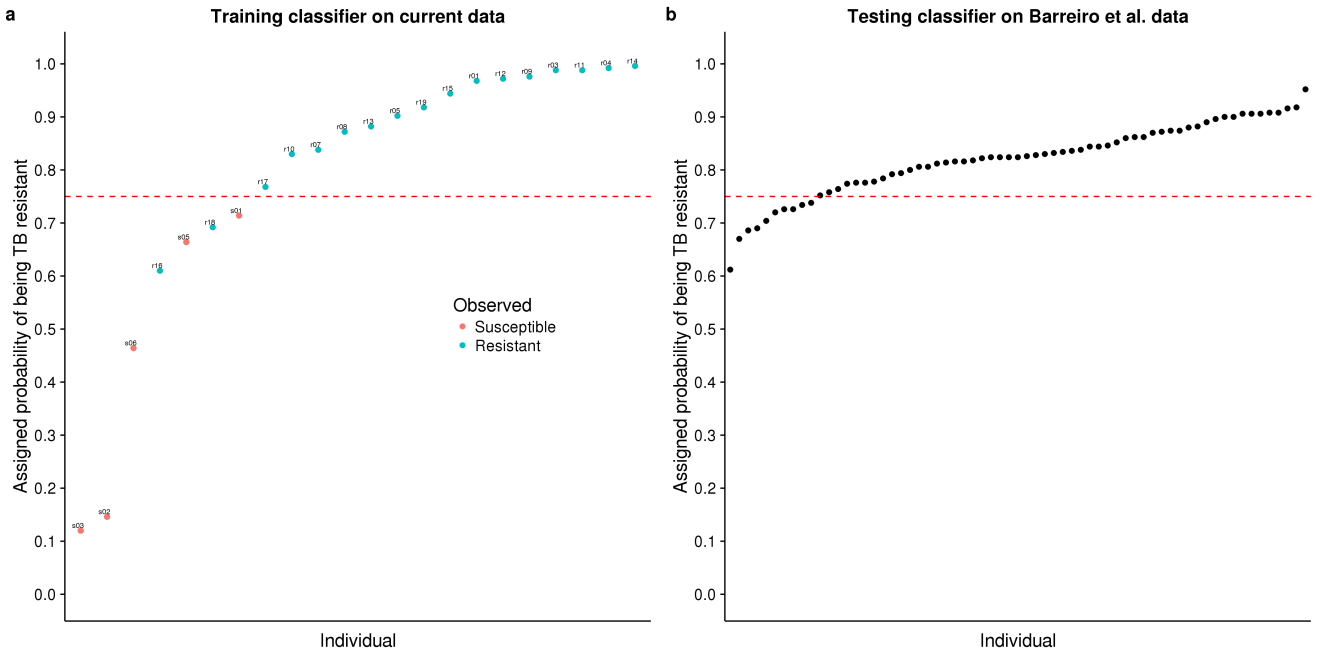
\caption{

Classifying TB susceptible individuals using an elastic net model. (a) The estimates of predicted probability of TB resistance from the leave-one-out-cross-validation for individuals in the current study. The red circles represent individuals known to be susceptible to TB, and blue those resistant to TB. The horizontal blue line at a probability of 0.75 almost separates susceptible and resistant individuals. (b) The estimates of predicted probability of TB resistance from applying the classifier trained on the data from the current study to a test set of independently collected healthy individuals \cite{Barreiro2012}.

}

\label{fig:class-en}

\end{figure}



\begin{figure}[ht]

\centering

\includegraphics[width=\linewidth]{../figure/classifier-rf.pdf}

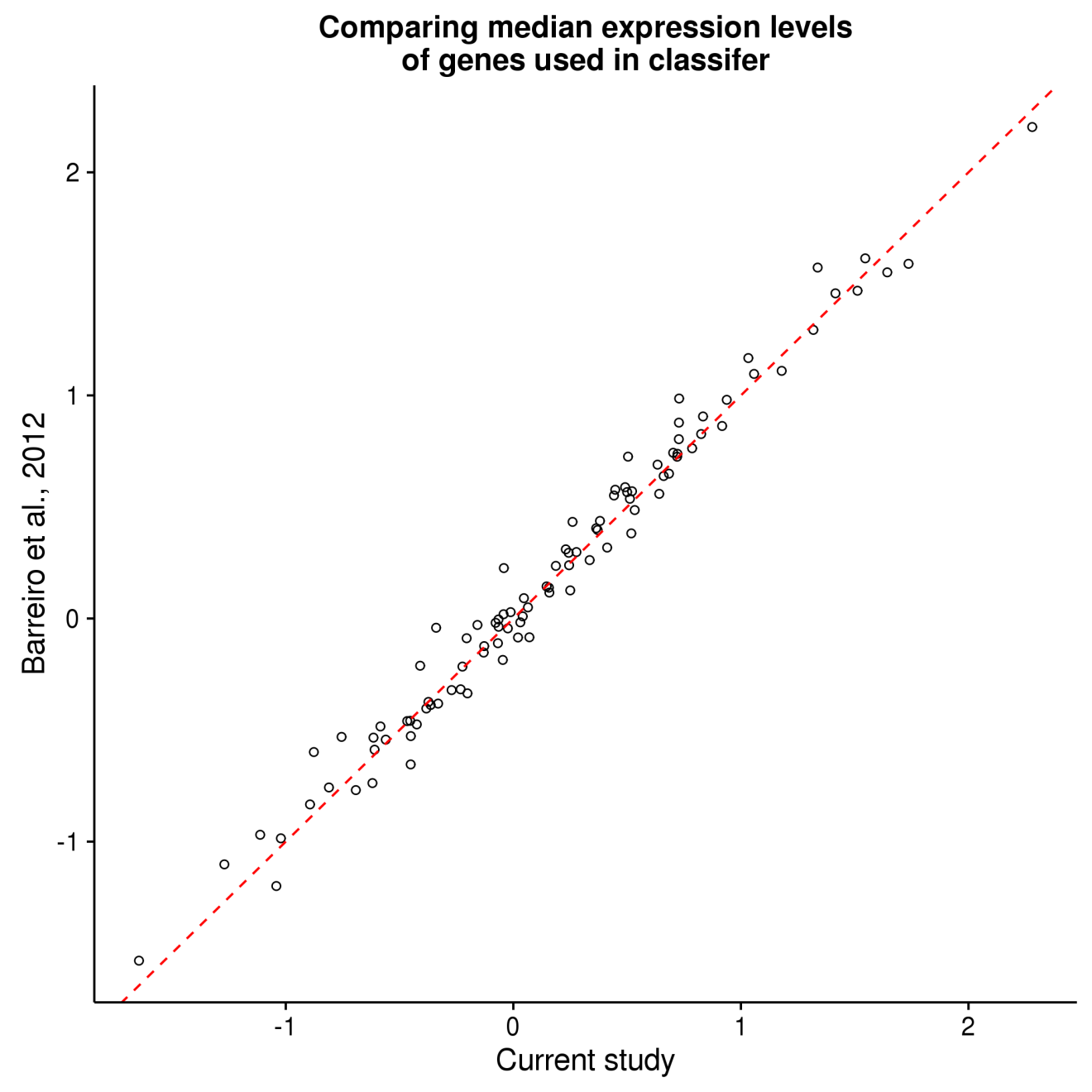
\caption{

Classifying TB susceptible individuals using a random forest model. (a) The estimates of predicted probability of TB resistance from the leave-one-out-cross-validation for individuals in the current study. The red circles represent individuals known to be susceptible to TB, and blue those resistant to TB. The horizontal blue line at a probability of 0.75 separates susceptible and resistant individuals. (b) The estimates of predicted probability of TB resistance from applying the classifier trained on the data from the current study to a test set of independently collected healthy individuals \cite{Barreiro2012}.

}

\label{fig:class-rf}

\end{figure}



\begin{figure}[ht]

\centering

\includegraphics[width=\linewidth]{../figure/classifier-exp.pdf}

\caption{

Comparing gene expression between the two studies. After normalization and batch-correction, the median expression levels of the 99 genes used in the classifier were similar between the samples in the current study and those in Barreiro et al., 2012 \cite{Barreiro2012}. The dashed red line is the 1:1 line.

}

\label{fig:class-exp}

\end{figure}

## Supplementary Data

### Supplementary Data S1

Supplementary Data S1 contains the gene expression counts for the 11,336 genes after filtering lowly expressed genes for all 50 samples. Each row is a gene labeled with its Ensembl gene ID. Each column is a sample. Each sample is labeled according to the pattern “x##-status-treatment”, where x is “r” for resistant or “s” for susceptible, ## is the ID number, status is “resist” for resistant or “suscep” for susceptible, and treatment is “noninf” for noninfected or “infect” for infected. (tds)

### Supplementary Data S2

Supplementary Data S2 contains the results of the differential expression analysis with limma (Fig. \ref{fig:limma}). The workbook contains 4 sheets corresponding to the 4 tests performed. “status\_ni” is the test between resistant and susceptible individuals in the noninfected state, “status\_ii” is the test between resistant and susceptible individuals in the infected state, “treat\_resist” is the test between the noninfected and infected states for resistant individuals, and “treat\_suscep” is the test between the noninfected and infected states for susceptible individuals. Each sheet has the same columns. “id” is the Ensembl gene ID, “gene” is the gene name, “logFC” is the log fold change from limma, “AveExpr” is the average log expression from limma, “t” is the t-statistic from limma, “P.Value” is the p-value from limma, “adj.P.Val” is the adjusted p-value from limma, “qvalue” is the q-value calculated with adaptive shrinkage, “chr” is the chromosome where the gene is located, “description” is the description of the gene from Ensembl, “phenotype” is the associated phenotype(s) assigned my Ensembl, “go\_id” is the associated GO term(s) assigned by Ensembl, and “go\_description” is the corresponding name(s) of the GO term(s). (xlsx)

### Supplementary Data S3

Supplementary Data S3 contains the results of the GWAS comparison analysis (Fig. \ref{fig:gwas}). The first sheet “input-data” contains the data for the 10,260 genes which were assigned a SNP in the studies from The Gambia and Ghana. “gwas\_p\_ghana” is the minimum p-value from the GWAS in Ghana, “gwas\_p\_gambia” is the minimum p-value from the GWAS in The Gambia, and “n\_snps” is the number of GWAS SNPs within 50 kb of the transcription start site. The columns status\_ni, status\_ii, treat\_resist, and treat\_suscep refer to the tests described for Supplementary Data S2 and contain the absolute log fold changes for each comparison. All the other gene annotation columns are the same as described for Supplementary Data S2. The second sheet “regression-results” contains the results of the linear regression between the GWAS p-values and the log fold changes (or the number of SNPs). “t” is the t-statistic, “p” is the p-value, “slope” is the slope of the regression line, and “slope\_se” is the standard error of the slope. (xlsx)

### Supplementary Data S4

Supplementary Data S4 contains the results of the classifier analysis. Specifically it contains the results from the support vector machine using the genes with a qvalue less than 0.05 (Fig. \ref{fig:classifier}). The sheet “gene-list” contains information about the genes used for the classifier (the columns are described in the section for Supplementary Data S2). The sheet “training-input” contains the input gene expression data for training the model. The sheet “training-results” contains the results of the leave-one-out-cross-validation when training the model on the samples from the current study. The sheet “testing-input” contains the input gene expression data for testing the model. The sheet “testing-results” contains the results from testing the model on the samples from Barreiro et al., 2012 \cite{Barreiro2012}. The column “prob\_tb\_resist” is the probability of being resistant to TB assigned by the model. (xlsx)