Dendritic cells isolated from individuals susceptible to tuberculosis have an altered gene expression profile

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

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

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

Write abstract here…

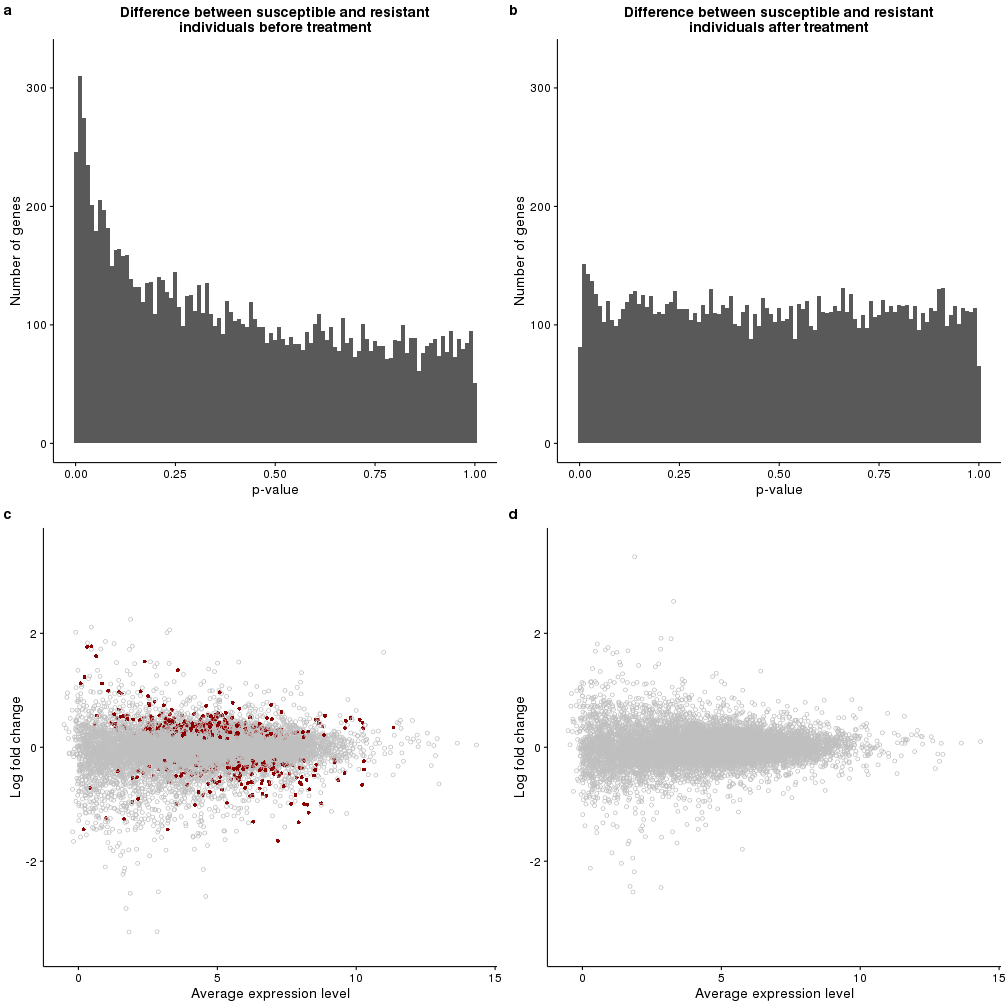
# Introduction

# Results

## Susceptible individuals have an altered transcriptome in the noninfected 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. To measure genome-wide gene expression levels, we sequenced the RNA at 18 hours post-infection, using a processing pipeline designed to minimize the introduction of unwanted technical variation, and obtained a mean of X $\pm$ X million raw reads per sample. We performed quality control analyses to remove non-expressed genes (Supplementary Fig. \ref{fig:genes}), identify and remove outliers (Supplementary Fig. \ref{fig:outliers}), and check for confounding batch effects (Supplementary Fig. \ref{fig:batch}, \ref{fig:infection}). Ultimately 6 samples were removed from all downstream analyses (Supplementary Table \ref{tab:outliers}).

Next we performed a standard differential expression analysis using a linear modeling framework, defined in equation (\ref{eq:limma}). Of most interest are genes which were differentially expressed (DE) between susceptible and resistant individuals in the noninfected and infected states (Fig. \ref{fig:limma}). After correcting for multiple testing, we identified X DE genes in the noninfected state, and none in the infected state. <Insert GO results and a few examples>



\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 MA plots for the (c) noninfected and (d) infected states. The x-axis is the average gene expression level and the y-axis is the log fold change in expression level between susceptible and resistant individuals. Red solid dots indicate genes which are significant differentially expressed after applying a multiple testing correction.

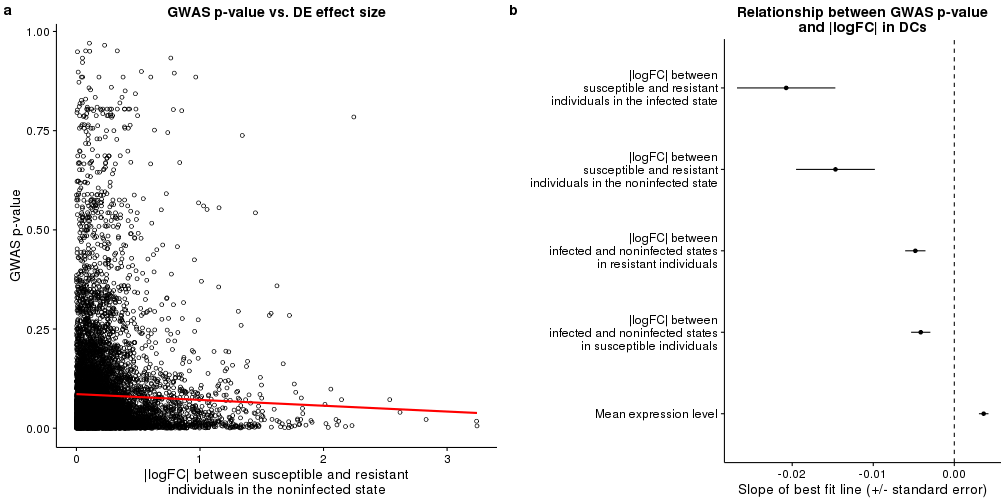
}

\label{fig:limma}

\end{figure}

## Differentially expressed genes are enriched for TB susceptibility loci

We next sought to investigate whether the differentially expressed genes we had identified in our *in vitro* experimental system were important for the genetic basis of TB susceptibility. To do this, we compared our results to a TB susceptibility GWAS conducted in Gambia and Ghana \cite{Thye2010}. Specifically, for each gene we assigned the SNP with the lowest p-value among all SNPs located within 50 kb of its transcription start site (TSS). If the differentially expressed genes are enriched for TB susceptibility loci, we expect a negative correlation between the absolute values of the log fold changes in our experiment and the GWAS p-values. Indeed, this is what we observed (results from Gambia GWAS in Fig. \ref{fig:gwas}, results from Ghana GWAS in Supplementary Fig. \ref{fig:ghana}). We fit a best fit linear line using least squares regression for each of our differential expression tests. Interestingly, we observed the steepest negative slopes for the tests comparing differential expression between susceptible and resistant individuals in the noninfected 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 was were also negative. Reassuringly, we did not observe a negative relationship between the GWAS p-values and the mean expression levels (in fact it this had a positive slope). Therefore the negative slopes we observed were not an artifact due to the power to call differential expression in our experimental system.



\begin{figure}[ht]

\centering

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

\caption{

Comparison of differential expression and GWAS results. (a) The relationship between GWAS p-values \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 ($\pm$ standard error) of the regression lines for each test. Additionally it includes the comparison of the mean expression level to the GWAS p-values as a control. All slopes are significantly different than 0 (F-test *P* $<$ 0.05).

}

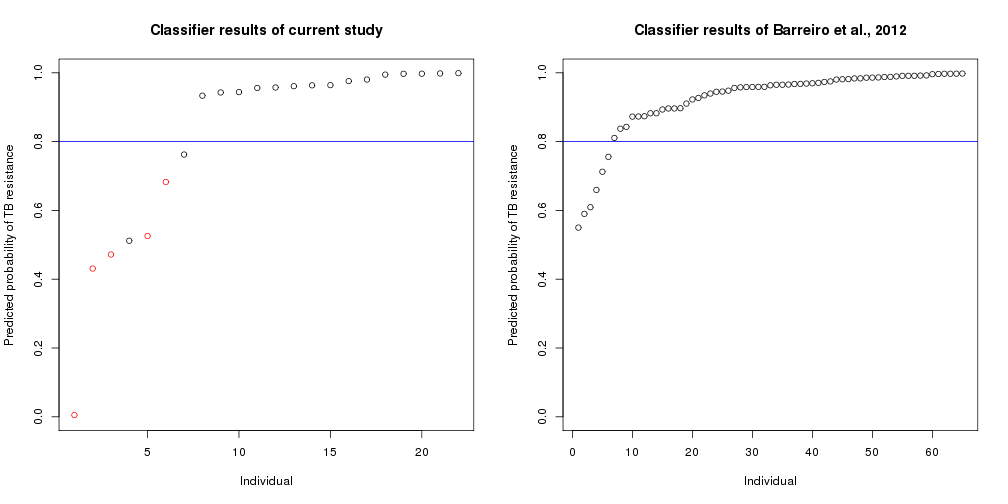
\label{fig:gwas}

\end{figure}

## Gene expression levels in the noninfected state can predict susceptibility status

Next we attempted to build a gene expression based classifier to predict susceptibility status. We focused on the gene expression levels both because this is where we observed the largest differences between susceptible and resistant individuals (Fig. \ref{fig:limma}ac) and also since it is much more practical to obtain gene expression data from noninfected DCs compared to MTB-infected DCs. We trained an elastic net classifier using the 1,057 genes that were differentially expressed at an ASH s-value \cite{Stephens2016} less than 5\%, using leave-one-out-cross-validation to select the optimal model parameters. 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 the cross validation (Fig. \ref{fig:classifier}). Using a cutoff of 0.8 for the predicted probability of being resistant to TB infection, we obtained a sensitivity of 100\% (5 out of 5 susceptible individuals labeled as susceptible) and a specificity of ~71\% (5 out of 7 individuals classified as susceptible were true positives).

Unfortunately our current data set was too small to properly split into separate training and testing sets. In order to 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}. Using the same cutoff of 0.8 for the probability of being resistant to TB that was determined to be optimal in the training set, ~9\% (6 out of 65) of the individuals were classified as being susceptible to TB, similar to the general estimate that 10\% of the population is susceptible to TB (Fig. \ref{fig:classifier}).



\begin{figure}[ht]

\centering

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

\caption{

Classifying TB susceptible individuals. (left) The estimates of predicted probability of TB resistance from the leave-one-out-cross-validation for individuals in the current study. The red open circles represent individuals known to be susceptible to TB. The horizontal blue line at a probability of 0.8 separates susceptible and resistant individuals. (right) 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

To do…

Previous work in TB \cite{Thuong2008}

Previous work using gene expression to understand susceptibility \cite{Bryant2014}

# Methods

## 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. Our protocol was approved by the Institutional Review Boards of the University of Chicago (10-504-B) and the Institut Pasteur (IRB00006966).

## Isolation and infection of dendritic cells

## 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.

## Read mapping

We mapped reads to human genome hg38 (GRCh38) using Subread and discarded non-uniquely mapping reads. We assigned mapped reads to 19,800 Ensembl protein-coding genes (Ensembl 83, Dec 2015, GRCh38.p5) using featureCounts.

## 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. Sx). Next we used principal components analysis (PCA) and hierarchical clustering to identify and remove 6 outliers (Supplementary Fig. Sx, Sx, Sx; Supplementary Table \ref{tab:outliers}).

## 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}

## Comparison to GWAS results

## Classifier

# 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 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 Tables

|  |  |  |
| --- | --- | --- |
| ID | Description | PCs |
| 18-contact-infected | Resistant individual after infection | 1 |
| 06-contact-none | Resistant individual before infection | 2, 4, 5 |
| 06-contact-infected | Resistant individual after infection | 2 |
| 04-tb-none | Susceptible individual before infection | 3, 5, 6 |
| 02-contact-none | Resistant individual before infection | 3, 5, 6 |
| 01-tb-infected | Susceptible individual after infection | 5 |

\begin{table}[ht]

\centering

\caption{

The following outliers were removed from all analyses based on their discordant clustering observed with PCA and hierarchical clustering.

}

\begin{tabular}{|l|l|l|}

\hline

ID & Description & PCs \\

\hline

18-contact-infected & Resistant individual after infection & 1 \\

\hline

06-contact-none & Resistant individual before infection & 2, 4, 5 \\

\hline

06-contact-infected & Resistant individual after infection & 2 \\

\hline

04-tb-none & Susceptible individual before infection & 3, 5, 6 \\

\hline

02-contact-none & Resistant individual before infection & 3, 5, 6 \\

\hline

01-tb-infected & Susceptible individual after infection & 5 \\

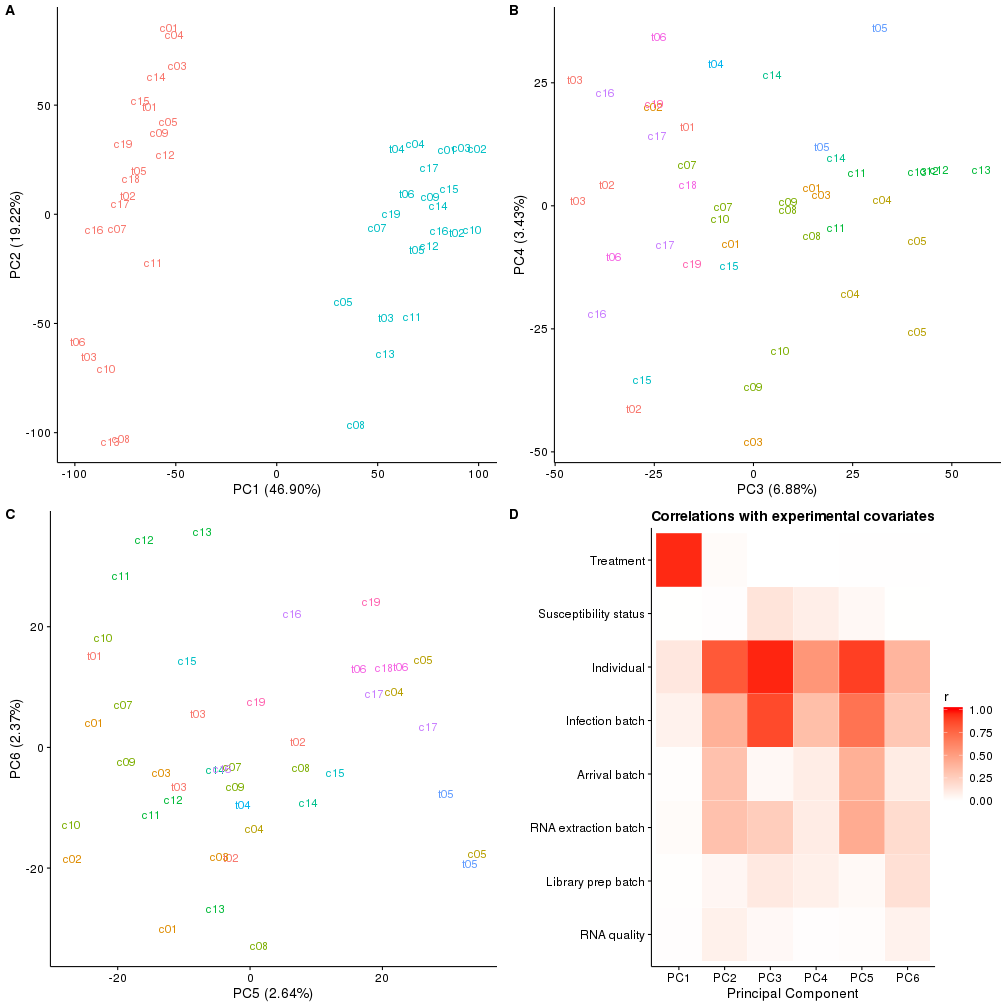
\hline

\end{tabular}

\label{tab:outliers}

\end{table}

## Supplementary Figures



\begin{figure}[ht]

\centering

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

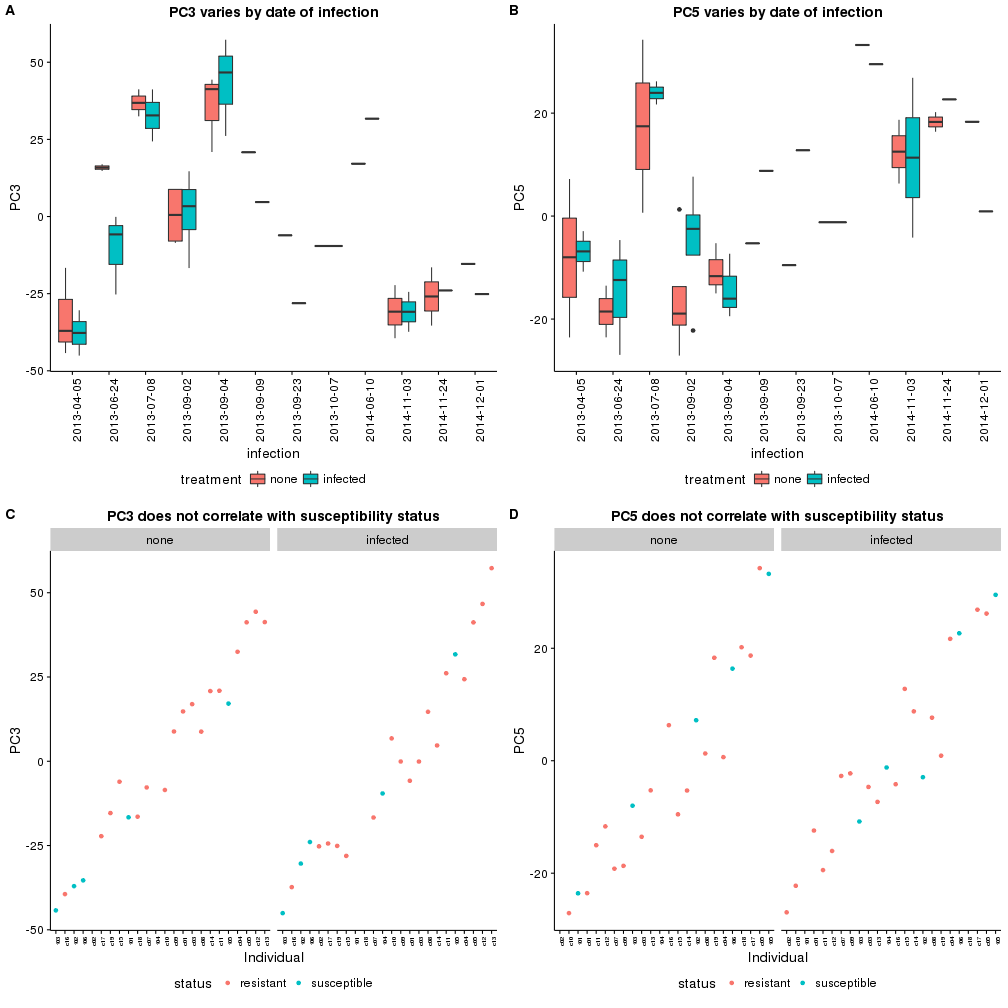
\caption{

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.

}

\label{fig:batch}

\end{figure}



\begin{figure}[ht]

\centering

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

\caption{

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

}

\label{fig:infection}

\end{figure}