Jeannie She

20.260

Professors Lauffenburger and White

15 May 2024

20.260 Final Project

Introduction to the Biological Problem

Tuberculosis (TB) results in over 1.5 million deaths each year, but despite it being the world's top infectious cause of death (WHO, n.d.), no preventative treatments are available worldwide. Much is yet to be understood about the pathogenesis of the disease, and the fact that 80% of its cases plague developing countries is a significant obstacle to maintaining urgency for TB research (WHO, n.d.). In their paper, Moreira-Teixeira et al. (2020) analyzed mouse models of TB infection to draw parallels with human patients who have active TB infection. It was revealed that the blood transcriptional signature of infected humans is most similarly captured by HN878-infected (the clinical TB isolate) C3HeB/FeJ (TB-susceptible) mice (hereon referred to as the "most representative mouse"), which naturally encourages further analysis of this representative mouse model for TB pathogenesis research.

Insight Gained from the Original Study

Moreira-Teixeira et al. (2020) acquired three datasets from their experimental mice: lung and blood RNA-seq data, as well as blood array data. Performing principal component analysis (PCA) allowed them to conclude the specific TB strain and mouse line that reflected the human blood modular transcriptional TB signature, which was informed by data from another study (Singhania et al., 2018) with patients from London and South Africa who were uninfected, had

an active TB infection, or had latent TB infection (LTBI), meaning that they are infected without symptoms or risk of transmission. Despite the fact that the immunology of mice and humans differ in a few notable ways (Mestas and Hughes, 2004), the authors discovered that a few sets of granulocyte-associated genes were highly expressed in blood of the most representative mouse as well as humans. They further performed histological analysis on the mouse lungs to visualize the development of neutrophils, T cells, and B cells in the disease model. Then, they showed that mouse blood signatures were largely informative of the mouse lung pathology, and pursued this finding by qualifying the same relationship in human data.

A Systems Biology Approach and its Biological Significance

Phase 1: Predicting the tuberculosis strain that caused infection

One of the tricky features of tuberculosis which makes it difficult to eradicate is its rapid evolution of resistance to antibiotics. While current TB treatment involves intense rounds of all-encompassing antibiotics, in the future, knowing which TB strain infected an individual could mean that they are prescribed a strain-specific treatment, which could prove to be simultaneously more effective on eradicating the disease *and* be less taxing on the body. With this idea in mind, I wanted to uncover which set of mouse transcriptomic data, blood or lung, would best predict the strain of TB the mouse was infected with: HN878, the clinical isolate, or H37Rv, the laboratory strain. To begin, I preprocessed all of my data by z-scoring it.

For the blood transcriptomic data, I performed PCA to visualize how the data would cluster. Two principal components (PCs) appeared to explain 50% of the variance (Fig. 1a), and plotting the data on axes representing the first two PCs showed that the first PC split the data into the desired clusters (Fig. 1b). Then, I performed unsupervised k-means clustering on the same

data, which resulted in 3 clusters that disagreed with the PCA clustering, but was still interesting to visualize (Fig. 1c). The clusters “0” and “1” align with HN878 infected/Uninfected and H37Rv infected mice respectively, which may suggest possible similarities between the HN878 infected and uninfected groups. The separation between these clusters is distinct, as seen in the silhouette plot (Fig. 1d). Finally, I created a PLS-DA model with a test dataset consisting of 30% of the entire data. This model had an accuracy score of 0.814 and an f1-score of 0.94, and its prediction was visualized using a confusion matrix (Fig. 1e). Then, I performed 10-fold cross validation, which resulted in an average accuracy score of the model of 0.721.

For the lung transcriptomic data, I did all of the same analyses as for the blood transcriptomic data. On this data, the first PC already explains 60% of the variance (Fig. 2a), so seeing that the clustering was distinct in Figure 2b felt promising. The unsupervised k-means clustering and its respective silhouette plot (Figs. 2c and 2d) supported the PC clustering, which was a confirmation that the PC clustering was well-defined. The PLS-DA model, which ran on 30% of the data as the test dataset, performed with an 0.741 accuracy score, a 0.90 f1-score, and a confusion matrix as seen in Figure 2e. Finally, after 10-fold cross validation, the average accuracy score of the model was 0.688.

Both confusion matrices (Fig 1e and 2e) appear to distinguish the HN878-infected samples easily from the H37Rv-infected samples. Moreira-Teixeira et al. (2020) infected the experimental mice via an inhalation exposure system, and TB usually manifests in the lungs for humans. Therefore, I initially expected the model trained on lung data to be more accurate than the model trained on blood data, but the systems biology analysis proved otherwise. This contradiction could be explained either by the uniqueness of this specific dataset, or possibly by

a hypothesis that TB infection from any strain impacts the lungs in a similar fashion, but strain-specific effects might manifest more clearly by biomarkers in the bloodstream.

Phase 2: Predicting the tuberculosis disease state

Another aspect of TB that remains to be known is the characteristics that make one individual experience *latent* TB infection instead of *active* TB infection. For the second phase of my project, I explored the legitimacy of developing a predictive model on the severity of a human's disease state from their blood transcriptomic data. The data sourced from Singhania, et al. (2018) included data from a sample set in London and another in South Africa.

First, I looked only at the London dataset to explore clustering and predictive accuracies. I ran PCA and noted that the first two PCs only describe 25% of the variance in the data (Fig. 3a). Not surprisingly then, the respective clustering on the PC1 and PC2 axes was not very clear (Fig. 3b). The unsupervised k-means clustering (Fig. 3c) provided a rather unsatisfying split into three clusters, so I further performed t-SNE (Fig. 3d) and UMAP (Fig. 3e) clustering to see if other nonlinear clustering methods may generate clearer separations between my groups. The conclusion seems the same across all methods of clustering: the conglomeration of LTBI and Control samples, generally apart from the Active TB samples, may suggest that the transcriptomic profiles of those groups are more similar than those of the Active TB state. This hypothesis is noteworthy for next steps. Then, I wanted to examine the predictive power of this model trained on 70% of the data and testing on 30% of the data. The accuracy of this PLS-DA model was 0.706, with a confusion matrix shown in Figure 3f, and after 10-fold cross validation, the average accuracy was 0.62.

I wished to examine how a model trained on the London data would perform to predict the outcome of the South Africa data. I was curious to see if the differences in geography and human diversity would preclude the ability to create a useful predictive model for TB disease state. The South Africa data only had Active TB and LTBI samples, so I trained the model on London data pruned to only those two groups. Upon using the South Africa data as my test dataset, the model was shown to have 0.745 accuracy and a confusion matrix shown in Figure 3g. It was surprising to see such a highly accurate model, which may suggest that the TB disease state in humans manifests similarly across continents and is not dependent on individual physiology.

Finally, I wondered what genes had the most contribution to the latent variables in this predictive PLS-DA model. I plotted the top 50 genes with the highest loadings in latent variables 1 and 2 (Figs. 4 and 5) to examine the biological significance of what separated the latently infected individuals from the actively infected ones. After running the list of these genes in gProfiler, the two pathways with the smallest adjusted p-values of significance were "nucleobase-containing small molecule metabolic process," a rather general pathway, and "CD95 death-inducing signaling complex/ripoptosome," a cell-death-inducing pathway responsible for responding to genotoxic stress (G:Profiler, n.d.). Although little literature exists relating these pathways to the pathogenesis of tuberculosis, this may be an indication of having uncovered possible pathways that have yet to have a relationship to the disease established.

Possible Experimental Next Steps

I would be remiss not to mention the potential of performing cross-species translation using the data from Moreira-Teixeira et al. (2020). It is true that the mouse and human genome

differ widely, but using a method like TransCompR to map mouse transcriptomic data onto human transcriptomic data could help us develop even more accurate pre-clinical models of various tuberculosis disease states. We could examine the corresponding human transcriptomic "translation" of mice profiles infected with a known TB strain, and use it to understand how that specific strain impacts humans differently. Currently, mouse models for tuberculosis are largely only representative of an active tuberculosis infection. In the future, experiments could be done where genes such as the ones with the highest loadings in Figures 4 and 5 are knocked out in mice to try and induce a latent or active infection. Perhaps the disease state is related to the strain or environmental conditions as well, so these mice with certain sets of gene knockouts could be further tested by infection with various strains and variable environmental conditions.

Conclusions

In the future, I envision that these regression models have the potential to become robust and even assist physicians who are busy treating patients. These models could help inform what kinds of prescriptive treatment are ideal for any given patient and help give a more accurate prognosis to the patient and their family. Tuberculosis continues to be a disease that ravages humankind, but with the right tools and creativity, we can confidently be on our way to addressing and eradicating it entirely.

Figures

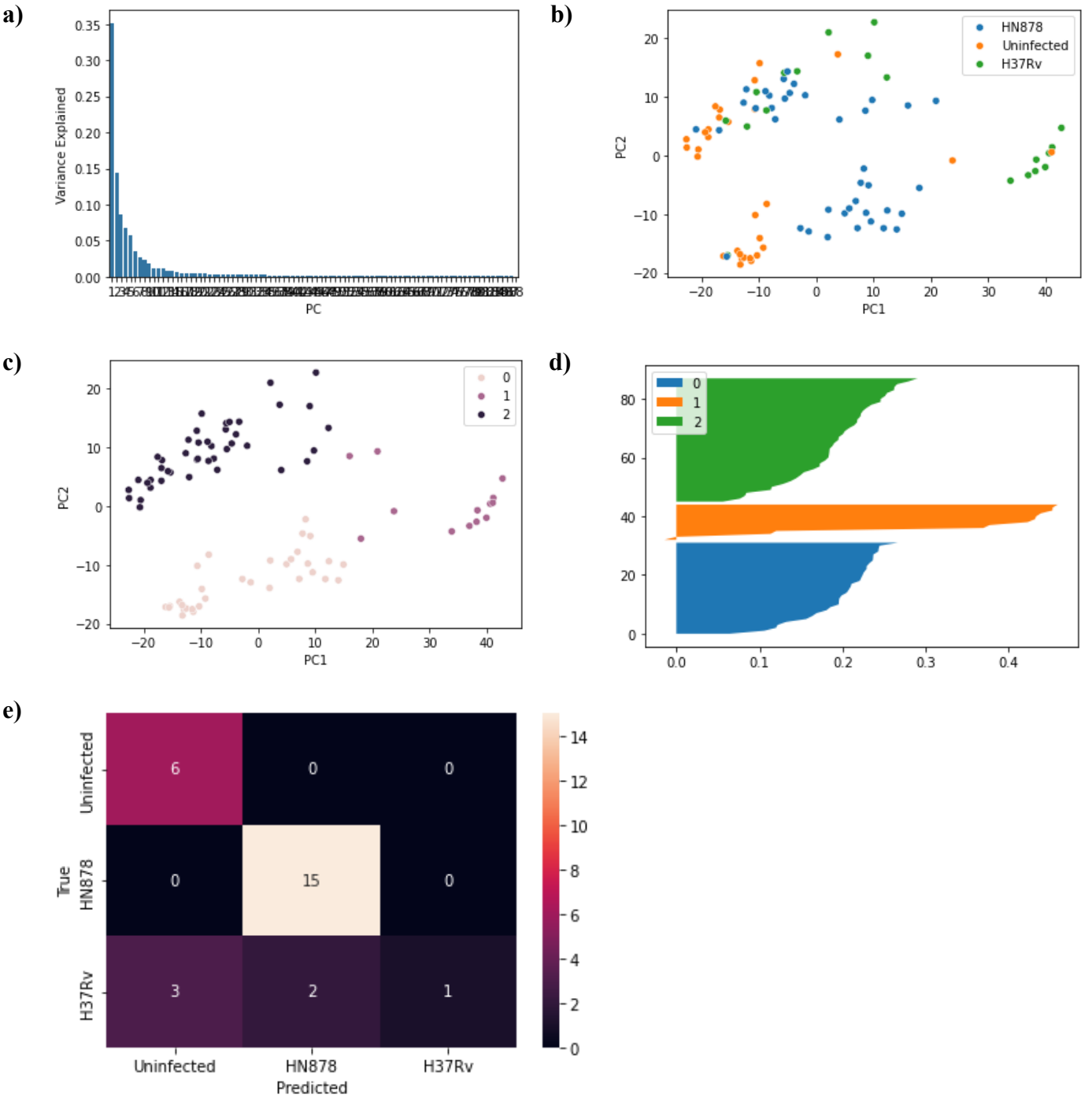


Figure 1 | Analysis of mouse blood transcriptomic data from Moreira-Teixeira et al., 2020.

PCA performed to visualize how much variance each PC explains in the data (**a**). Data plotted against the first two PCs and colored according to TB strain that infected the sample (**b**). Data undergoing unsupervised k-means clustering with 3 clusters (**c**). Corresponding silhouette plot for k-means clustering (**d**). Confusion matrix resulting from the PLS-DA regression with a 70/30 split of training and testing data (**e**).

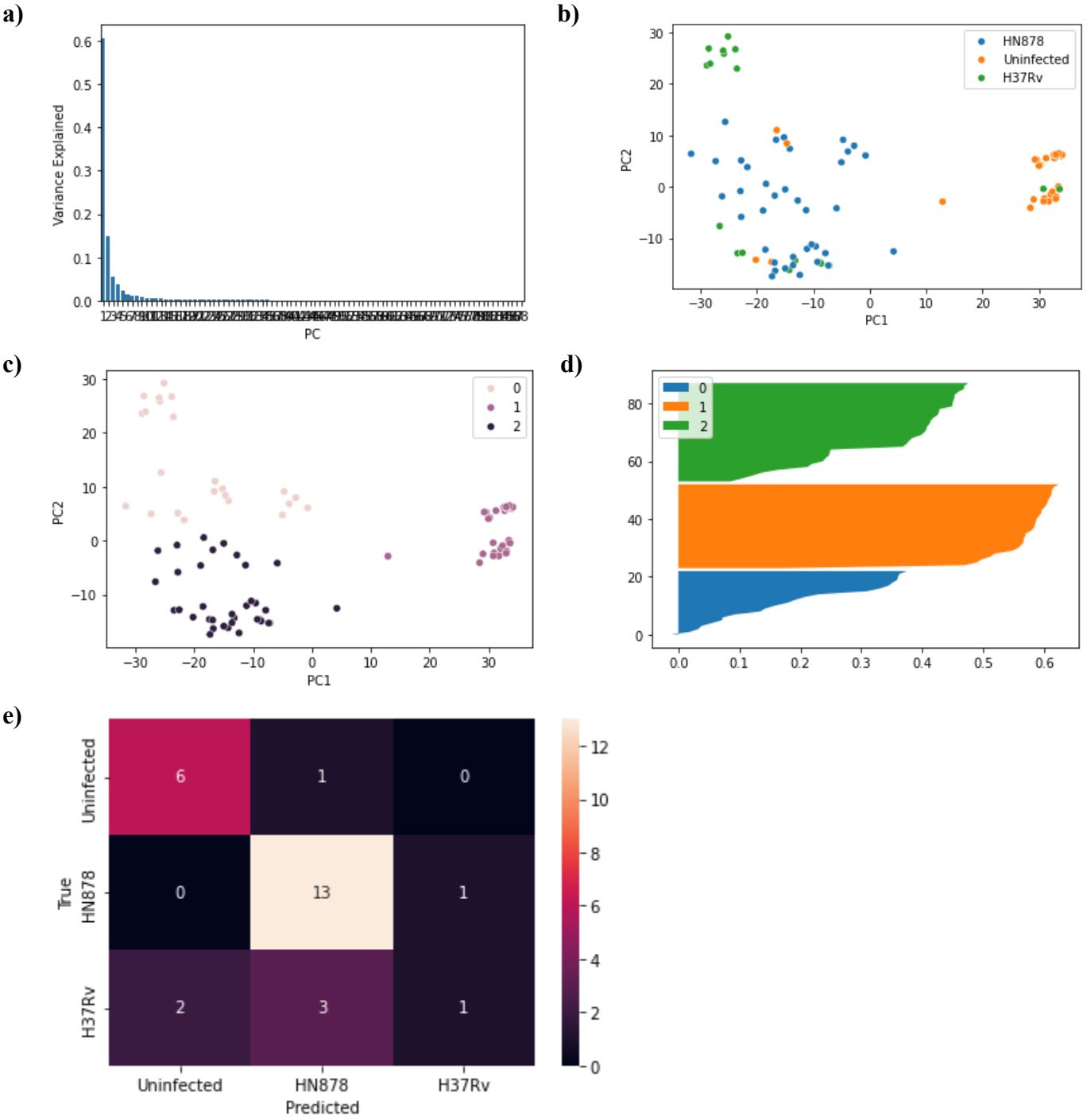
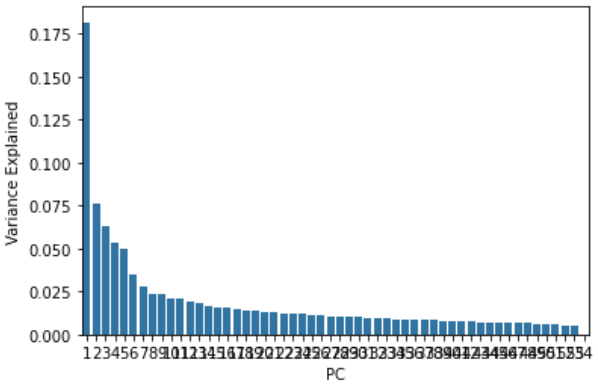


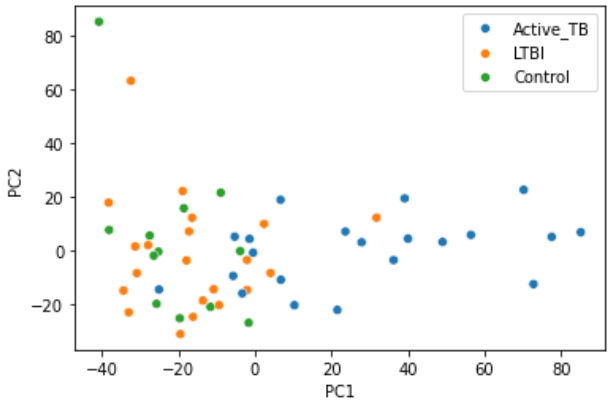
Figure 2 | Analysis of mouse lung transcriptomic data from Moreira-Teixeira et al., 2020.

PCA performed to visualize how much variance each PC explains in the data (**a**). Data plotted against the first two PCs and colored according to TB strain that infected the sample (**b**). Data undergoing unsupervised k-means clustering with 3 clusters (**c**). Corresponding silhouette plot for k-means clustering (**d**). Confusion matrix resulting from the PLS-DA regression with a 70/30 split of training and testing data (**e**).

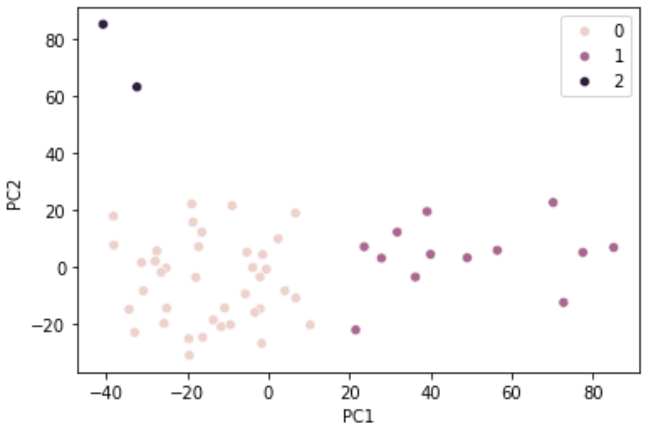
a)


Variance Explained
PC

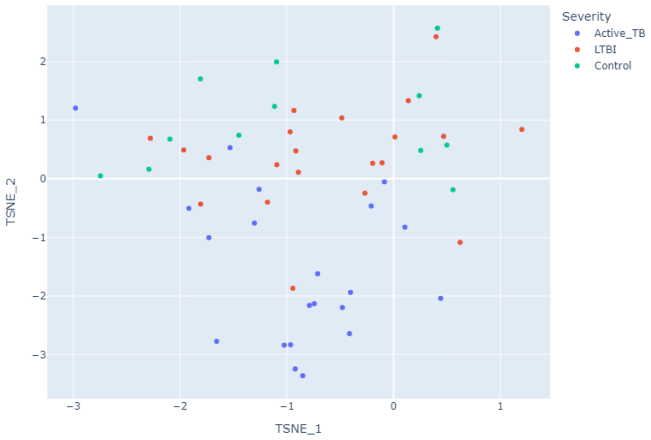
b)


Active_TB
LTBI
Control
PC2
PC1

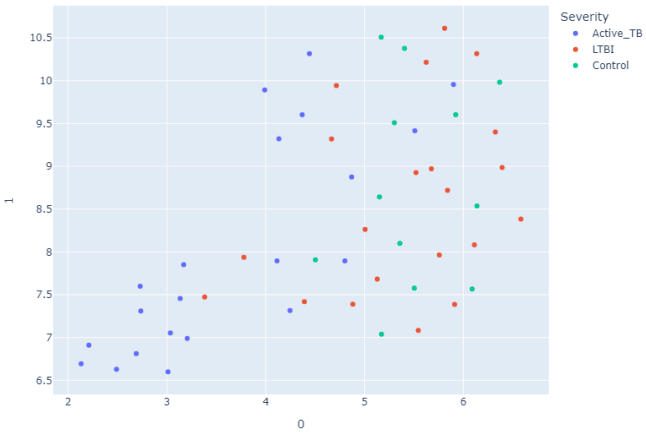
c)


0
1
2
PC2
PC1

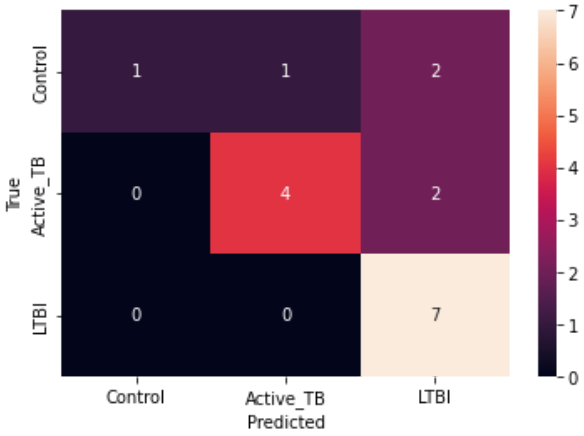
d)


Severity
Active_TB
LTBI
Control
TSNE_2
TSNE_1

e)


Severity
Active_TB
LTBI
Control

f)


Control
Active_TB
LTBI
True
Predicted

g)

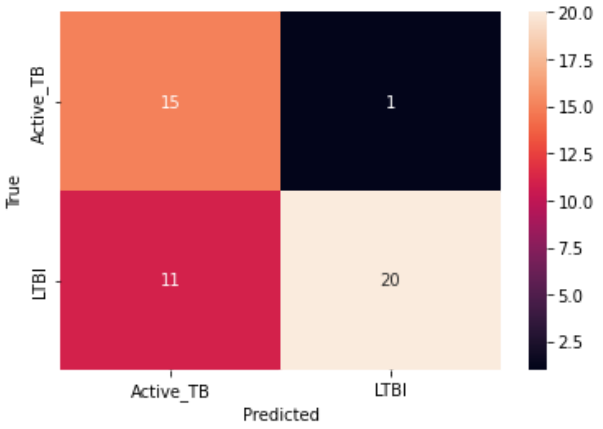

Active_TB
LTBI
True
Predicted

Figure 3 | Analysis of human tuberculosis transcriptomic data in London and South Africa from Singhania et al., 2018.

PCA performed to visualize how much variance each PC explains in the London data (**a**). London data plotted against the first two PCs and colored according to TB disease state, active, latent (LTBI), or uninfected (**b**). London data undergoing unsupervised k-means clustering with 3 clusters (**c**). London data undergoing t-SNE clustering plotted on two axes (**d**). London data undergoing UMAP clustering plotted on two axes (**e**). Confusion matrix resulting from the PLS-DA regression with a 70/30 split of training and testing data, using solely the London data (**f**). Confusion matrix resulting from the PLS-DA regression using London data as the training data and South Africa data as the testing data (**g**). The model was trained on and predicted only two groups, active and LTBI, since the South Africa data did not have any uninfected samples.

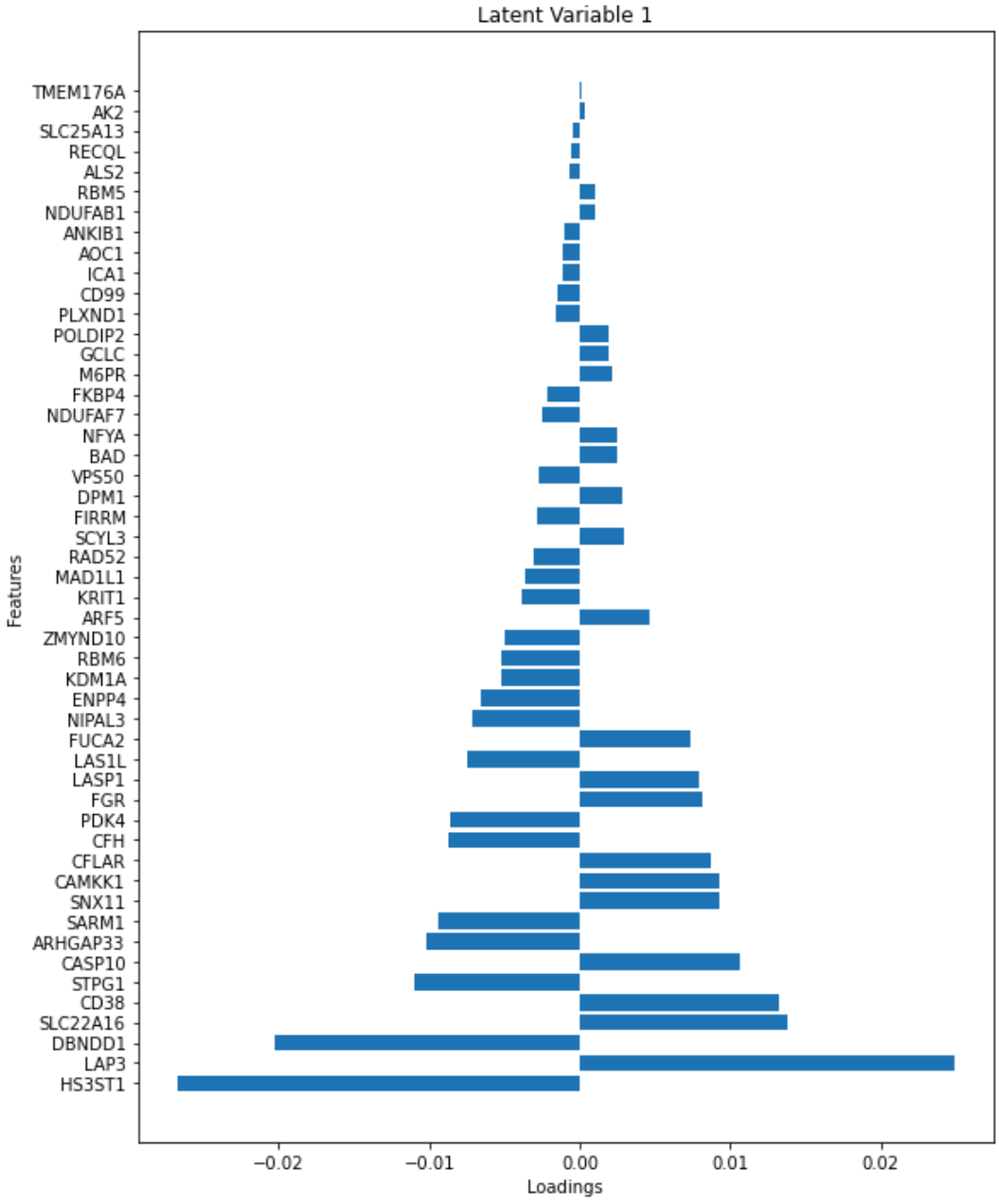


Figure 4 | Genes having the top 50 loadings in Latent Variable 1 of the PLS-DA model trained on London data filtered to contain only active TB and LTBI individuals.

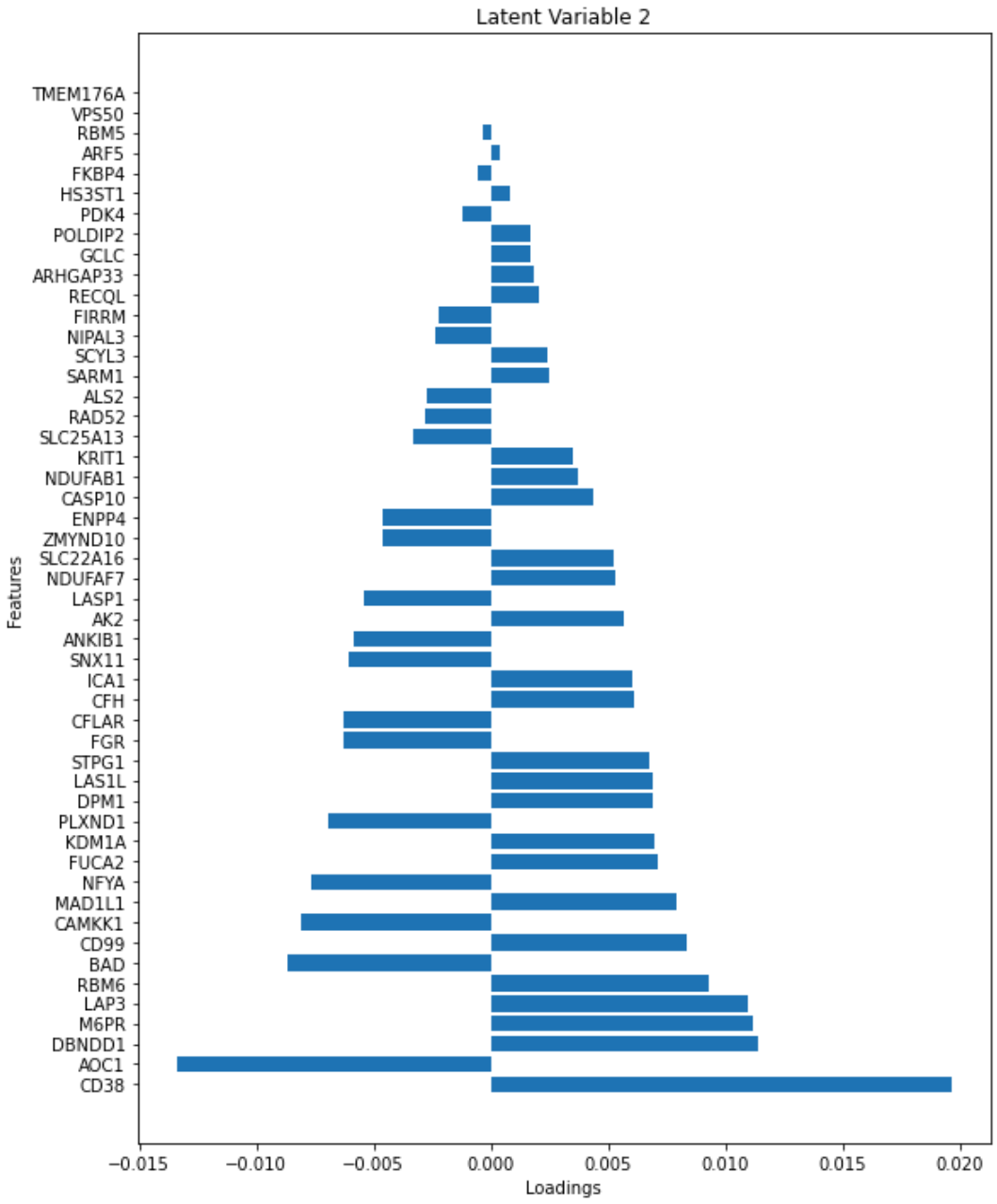


Figure 5 | Genes having the top 50 loadings in Latent Variable 2 of the PLS-DA model trained on London data filtered to contain only active TB and LTBI individuals.

Works Cited

G:Profiler – a web server for functional enrichment analysis and conversions of gene lists. (n.d.). https://biit.cs.ut.ee/gprofiler/gost

Javier Mestas, Christopher C. W. Hughes; Of Mice and Not Men: Differences between Mouse and Human Immunology. *J Immunol* 1 March 2004; 172 (5): 2731–2738.

Moreira-Teixeira, L., Tabone, O., Graham, C.M. et al. Mouse transcriptome reveals potential signatures of protection and pathogenesis in human tuberculosis. *Nat Immunol* 21, 464–476 (2020). https://doi.org/10.1038/s41590-020-0610-z

Singhania, A., Verma, R., Graham, C.M. et al. A modular transcriptional signature identifies phenotypic heterogeneity of human tuberculosis infection. *Nat Commun* 9, 2308 (2018). https://doi.org/10.1038/s41467-018-04579-w

WHO. (n.d.). *Tuberculosis*. World Health Organization. https://www.who.int/health-topics/tuberculosis#tab=tab_1