Final Project

Transcriptomic Profile of Memory CD4 T Cells in Latent TB: A Principal Components Analysis

**Kathleen Ashbaker**

**University of Washington, Seattle, WA**

**Introduction:**

This study, conducted by Burel et al. at the La Jolla Institute for Allergy and Immunology in 2018, aims to identify specific gene signatures in memory CD4 T cells that can differentiate between individuals with latent TB infection and healthy controls. The dataset, sourced from the NCBI Gene Expression Omnibus Series GSE99373, includes data from 20 subjects with latent TB and 19 healthy controls. These cells were derived from cryopreserved peripheral blood mononuclear cells (PBMCs).1

The RNA sequencing was performed on these cells without any specific treatment involved. Cells were directly isolated and lysed after thawing the cryopreserved samples

The extracted RNA was amplified and sequenced using high-throughput sequencing technologies. The process included staining with antibodies, sorting using a cell sorter directly into Trizol, and processing with the Smart-seq2 protocol for cDNA amplification and sequencing.

**Methods:**

Principal Component Analysis (PCA) was employed as an initial step in this study to quantify and preserve the variance in transcript expression in both healthy controls and latent-TB infected samples. The rationale behind using PCA is to reduce the dimensionality of the dataset while retaining the most significant sources of variation. By transforming the original variables into a new set of orthogonal components (principal components), PCA captures the maximum variance in the data with the fewest number of components, making it easier to visualize and interpret the underlying patterns.

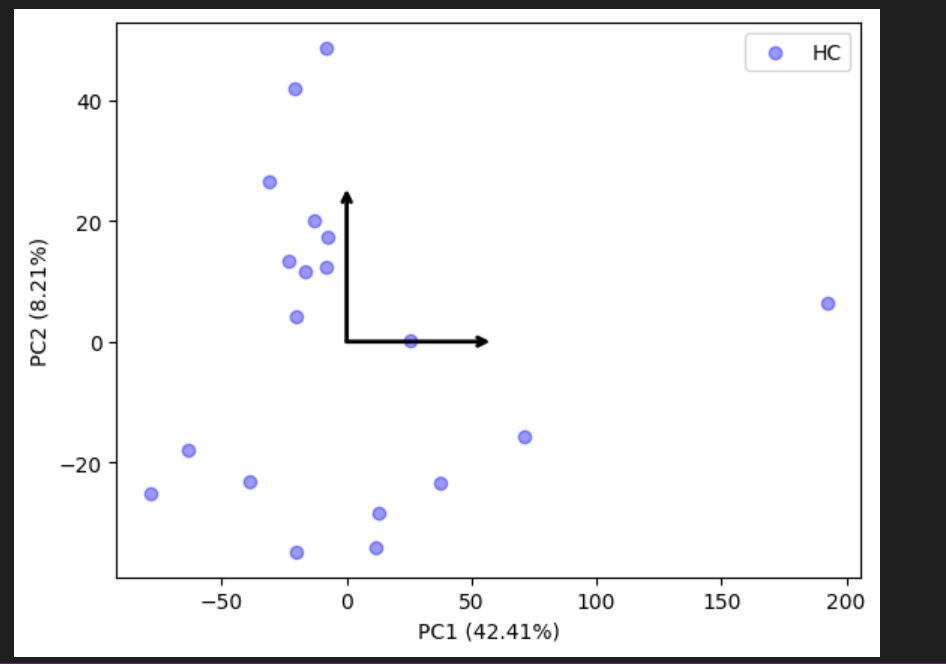
In this study, PCA was conducted using Python version 3.11.8, leveraging libraries such as NumPy and Matplotlib. The process involved the following steps:

*Data Preparation*: The transcript expression data for memory CD4 T cells from both healthy controls and latent-TB infected individuals were normalized and standardized.

*PCA Computation*: The standardized data were subjected to PCA, and the principal components were computed, capturing the majority of variance in the dataset.

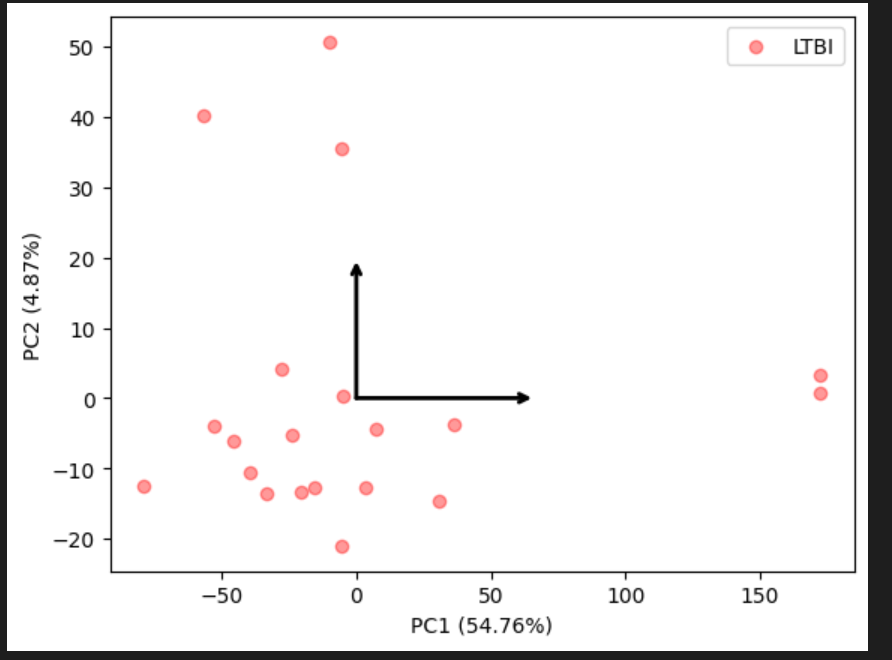
*Visualization:* The first two principal components were plotted to create a scatter plot, allowing for visual differentiation between the two groups. Genes strongly associated with each principal component were identified and annotated on the plot.

**Results:**

****

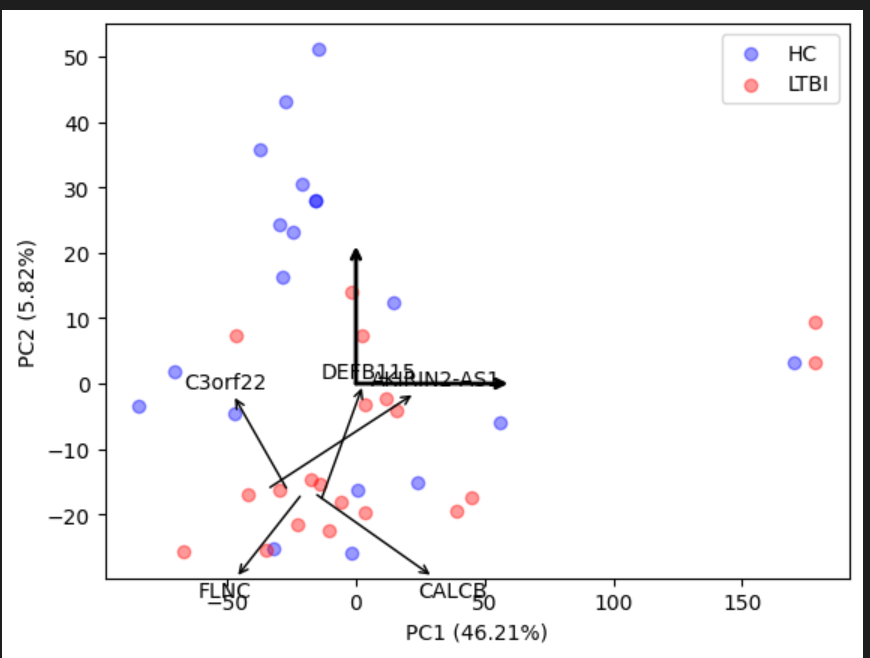
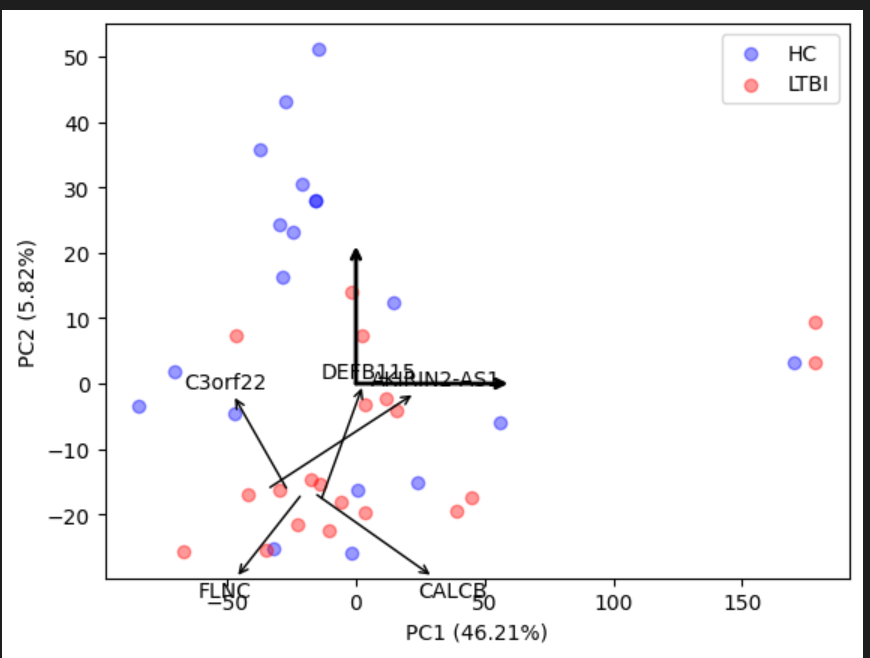
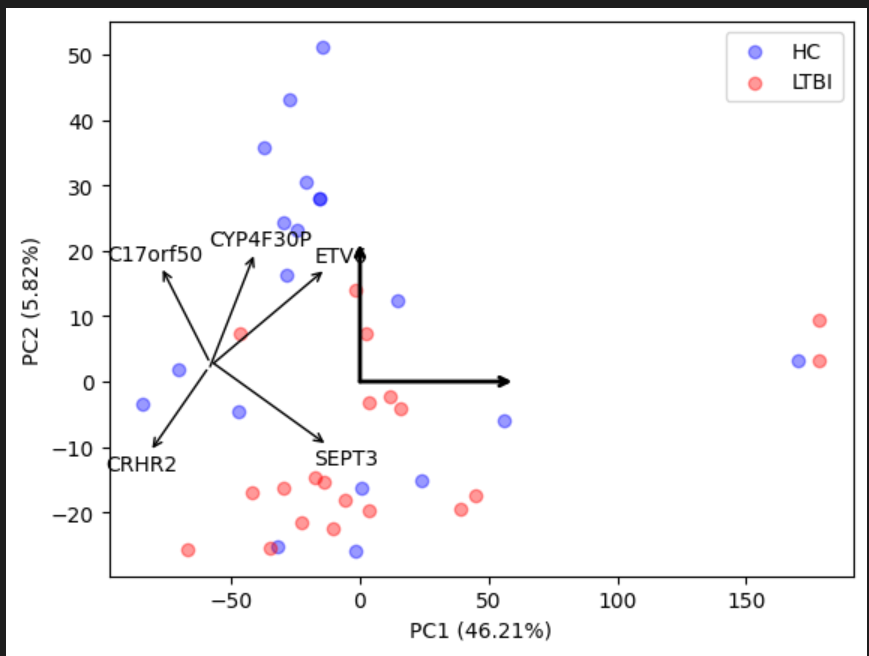
**Figure 1: Principal Components Analysis of Healthy Control Samples**

Healthy Controls (HC): The blue points are scattered, showing some variance along both PC1 and PC2, with PC1 explaining 42.41% of the variance and PC2 explaining 8.21%.

****

**Figure 2: Principal Components Analysis of Latent-TB Infected Samples**

Latent TB Infection (LTBI): The red points show a different distribution pattern with PC1 explaining 54.76% of the variance and PC2 explaining 4.87%.

****

**Figure 3: Comparison of Principal Component Analysis (PCA) with Genes of Interest: PC1 vs. PC2**

The PCA plots reveal distinct clustering patterns between HC and LTBI samples, indicating different gene expression profiles in these two groups.

**Discussion:**

The PCA visuals provided in this study offer a comprehensive comparison of the gene expression profiles in healthy controls (HC) and latent tuberculosis infection (LTBI) samples. The first set of PCA plots illustrates the separation of HC and LTBI samples along the principal components, with each component capturing significant variance in the data. The subsequent PCA plots highlight genes most strongly associated with PC1 and PC2, providing insights into the key drivers of variance between the two groups.

*PCA with Genes Associated with PC1:*

The genes ETV6, C17orf50, SEPT3, CRHR2, and CYP4F30P were identified as the top contributors to PC1. These genes are annotated in the PCA plot, showing their significant role in differentiating HCC from LTBI along PC1.

*PCA with Genes Associated with PC2*:

The genes AKIRIN2-AS1, C3orf22, CALCB, FLNC, and DEFB115 are identified as the top contributors to PC2. The PCA plot with these gene annotations highlights their influence on the variance captured by PC2.

*Key Findings:*

The PCA findings reveal both commonalities and differences between the healthy controls (HC) and latent TB infection (LTBI) groups. For HC, PC1 explains 42.41% of the variance and PC2 explains 8.21%, while for LTBI, PC1 explains 54.76% and PC2 explains 4.87%, indicating that in both groups, a significant portion of the variance is captured by the first principal component. Both plots show samples spread across the axes, indicating variability within each group, and include arrows representing the principal component vectors. The LTBI group has a higher percentage of variance captured by PC1, suggesting that PC1 is more dominant in explaining variability in LTBI. The distribution of samples shows HC samples more spread out, indicating more variability, while LTBI samples are more concentrated along PC1. The genes most strongly associated with PC1 in HC include C17orf50, CYP4F30P, ETV, CRHR2, and SEPT3, while for LTBI, genes associated with PC2 include AKIRIN2-AS1, C3orf22, CALCB, FLNC, and DEFB115. The HC plot shows a more even spread of samples, indicating a heterogeneous gene expression profile, whereas the LTBI plot shows some separation along PC1, suggesting distinct patterns in LTBI samples.

*Caveats with PCA:*

*Linear Assumption, Variance* *and Data Scaling*: PCA assumes linear relationships between variables, which may not always capture complex, non-linear interactions in gene expression data. In addition, PCA focuses on maximizing variance, which might lead to the exclusion of biologically relevant features that do not contribute significantly to variance. The results of PCA are sensitive to the scaling of data. Different scaling methods can lead to different principal components.

*Interpretation:* While PCA provides a useful dimensionality reduction technique, the biological interpretation of principal components requires careful consideration and further validation.In conclusion, the PCA analysis in this study effectively highlights the key differences in gene expression profiles between healthy controls and latent TB infection samples. The identification of top genes associated with PC1 and PC2 offers valuable insights into the underlying biological processes. However, the limitations of PCA should be acknowledged, and complementary methods should be employed to validate and extend these findings.

A broader implication of this research is to further elucidate the genetic signature of latent TB, which can help alleviate the global and local burden of TB disease by enhancing diagnostics for latent TB infection and facilitating the development of targeted therapeutics and treatments for TB.

**References**

1. Burel JG, Lindestam Arlehamn CS, Khan N, Seumois G et al. Transcriptomic Analysis of CD4(+) T Cells Reveals Novel Immune Signatures of Latent Tuberculosis. *J Immunol* 2018 May 1;200(9):3283-3290. PMID: [29602771](https://www.ncbi.nlm.nih.gov/pubmed/29602771)