“Smoking-Induced Genetic Drivers of Latent Tuberculosis Reactivation Unveiled by Single-cell RNA Sequencing Analysis”

A DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

in

BIOTECHNOLOGY

Submitted by

ISHITA SEHGAL
2k22/MSCBIO/58

Under the Supervision of
Dr. ASMITA DAS

DEPARTMENT OF BIOTECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi - 110042
June, 2024

ACKNOWLEDGMENT

I would like to express my deepest gratitude to my supervisor, Dr. Asmita Das, for her constant support, encouragement, and invaluable guidance throughout this research. I am sincerely thankful for the direction and insights she provided, which significantly shaped the course of this research.

I also extend my heartfelt thanks to the Department of Biotechnology at Delhi Technological University (DTU) for providing all the essential facilities and resources. Their support was crucial for the computational work and overall progress of this study.

Ishita Sehgal

DEPARTMENT OF BIOTECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, **Ishita Sehgal, Roll No. 2k22/MSCBIO/58,** hereby certify that the work which is being presented in the thesis entitled **"Smoking-Induced Genetic Drivers of Latent Tuberculosis Reactivation Unveiled by Single-cell RNA Sequencing Analysis"** is in partial fulfillment of the requirement for the award of the Degree of Master of Science, submitted by me to the Department of Biotechnology, Delhi Technological University, Delhi is an authentic record of my own work carried out during the period from January 2024 to May 2024 under the supervision of Dr. Asmita Das.
The matter presented in the thesis has not been submitted by me for the award of any degree from this or any other institute.

My conference paper has cleared its first screening phase in Scopus indexed journal with the following details:
Title of the paper: "Comparative docking studies of drugs and phytocompounds with DosR protein as a therapeutic approach for Latent TB."
Name of Authors: Ishita Sehgal, Vanshika Choudhary, Asmita Das
Journal name: International Conference on Control, Computing, Communication and Materials 2024 (IEEE-2024)
Journal Indexing: Scopus

Place: Delhi **Ishita Sehgal**
Date: **2k22/MSCBIO/58**

DEPARTMENT OF BIOTECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

Certified that **Ishita Sehgal (2k22/MSCBIO/58)** has carried out their search work presented in this thesis entitled **“Smoking-Induced Genetic Drivers of Latent Tuberculosis Reactivation Unveiled by Single-cell RNA Sequencing Analysis”** for the award of the degree of Master of Science and submitted to the Department of Biotechnology, Delhi Technological University, Delhi, under my supervision. This thesis embodies the results of original work and studies carried out by the student herself, and the contents of the thesis do not form the basis for the award of any other degree to the candidate or anybody else from this or any other Institution.

Prof. Yasha Hasija
Head of Department
Department of Biotechnology
Delhi Technological University

Dr. Asmita Das
Supervisor
Department of Biotechnology
Delhi Technological University

Date:

“Smoking-Induced Genetic Drivers of Latent Tuberculosis Reactivation Unveiled by Single-cell RNA Sequencing Analysis”

Ishita Sehgal
Delhi Technological University, Delhi, India
Email: ishi.sehgal2000@gmail.com

ABSTRACT

Tuberculosis (TB), a formidable infectious disease caused by *Mycobacterium tuberculosis*, remains a significant global health challenge. While many individuals harbor a latent form of the infection, certain risk factors, such as cigarette smoking, can result in the reactivation of latent tuberculosis into an active and transmissible state. This phenomenon underscores the critical need to elucidate the molecular mechanisms by which smoking modulates the host immune response, rendering individuals more susceptible to tuberculosis reactivation.

Single-cell RNA sequencing (scRNA-seq) was employed to comprehensively profile the transcriptomic landscape of immune cells from individuals with active tuberculosis, including a subset of smokers. By leveraging cutting-edge computational approaches, we identified distinct gene expression signatures and cellular subpopulations that were differentially expressed in smokers compared to non-smokers.

Through rigorous bioinformatic analyses, key genes and pathways were uncovered associated with smoking-induced dysregulation of inflammatory responses and immune cell function, providing mechanistic insights into their heightened susceptibility to tuberculosis reactivation.

Our findings pinpoint potential therapeutic targets and biomarkers for early detection and personalized management of tuberculosis, particularly in the context of smoking-related risk.

LIST OF PUBLICATIONS

1. The conference paper “Comparative docking studies of drugs and phytocompounds with DosR protein as a therapeutic approach for Latent TB” has cleared its first screening phase at the International Conference on Control, Computing, Communication and Materials 2024 (IEEE-2024), which will be held in June 2024.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

Symbol/ Abbreviations	**Meaning/Full form**
TB	Tuberculosis
M. tb.	*Mycobacterium tuberculosis*
LTBI	Latent TB Infection
scRNA-seq	Single-cell RNA Sequencing
HBA1	Hemoglobin Subunit Alpha 1
TUBB1	Tubulin Beta 1
HBA2	Hemoglobin Subunit Alpha 2
HBB	Hemoglobin Subunit Beta
PBMC	Peripheral blood mononuclear cells
CD14	Cluster of Differentiation 14
CD93	Cluster of Differentiation 93
CCR2	CC Chemokine Receptor 2

CHAPTER 1

INTRODUCTION

Tuberculosis (TB), a persistent global health threat caused by *Mycobacterium tuberculosis* (*M. tb.*), continues to afflict millions of people globally, disproportionately impacting vulnerable populations. While significant advancements have been made in understanding the disease pathogenesis, a major challenge remains in elucidating the cellular and molecular mechanisms underlying the reactivation of latent TB infection (LTBI), particularly in individuals who are exposed to the risk of smoking.

Smoking is a potent risk factor that increases the likelihood of LTBI reactivation and progression to active TB disease. However, the intricate interplay between smoking, host immune response, and the transcriptional dysregulation of immune cells, primarily monocytes, remains poorly understood. This critical gap hinders the development of targeted interventions and diagnostic approaches tailored for smokers with LTBI, a population that is at elevated risk of disease reactivation.

Single-cell RNA sequencing (scRNA-seq) was performed to address this particular challenge. This cutting-edge technology enables the transcriptional profiling of individual cells at an unprecedented resolution. The primary objective was to delineate the distinct transcriptional signatures and regulatory pathways in monocytes that are affected by chronic exposure to smoking, consequently influencing their phenotypes, functional plasticity, and susceptibility to subversion by *M. tuberculosis*.

By employing a comprehensive scRNA-seq workflow, the transcriptomes of Peripheral blood mononuclear cells (PBMCs) isolated from a cohort of TB patients stratified by smoking status were analyzed. This approach allowed for the deconvolution of discrete monocyte activation states, differentiation trajectories, and functional phenotypes influenced by smoking and TB infection. Integrating computational approaches for dimensionality reduction, clustering, and differential expression analysis, we identified a panel of smoking-modulated genes and pathways that may contribute to the heightened risk of LTBI reactivation observed in smokers.

Our findings revealed a constellation of transcriptional alterations in essential monocyte markers, including CD93, CD14, and CCR2, which are implicated in immune regulation, inflammatory responses, and bacterial clearance mechanisms. Notably, the altered expression of these genes was significantly associated with smoking exposure and suggested their potential role as biomarkers or therapeutic targets for overcoming the detrimental effects of smoking on TB control.

CHAPTER 2

LITERATURE REVIEW

2.1 The Enduring Challenge of Tuberculosis

Mycobacterium tuberculosis (M. tb.) is the bacillus responsible for causing Tuberculosis (TB), which continues to be a persistent health threat in many developing nations. TB remains a significant public health concern, affecting vulnerable populations disproportionately in settings with limited resources despite substantial advancements in diagnosis and treatment (Pai et al., 2016). A major obstacle is Latent TB Infection (LTBI), where individuals harbor the bacteria without active symptoms. The WHO estimates that around a quarter of the global population has LTBI, serving as a reservoir for potential active cases, which makes the diagnosis and treatment of LTBI a crucial step in the course of action to stimulate an overall decline in worldwide TB infection and achieve its elimination to prevent future disease (Houben & Dodd, 2016).

2.2 Latent TB Infection

M. tb. infection occurs when the tubercle bacilli, originating from an individual with active disease, enter the lungs of an uninfected individual (Delogu et al., 2013). The bacteria replicate within immune cells called macrophages, leading to their spread and eventual granuloma formation. The granuloma, referred to as the Ghon complex, serves as *M. tb.'*s protective "sanctuary" during latent tuberculosis infection, where the bacteria remain dormant for years or even a lifetime, establishing a dynamic balance with the host's immune system (Alsayed & Gunosewoyo, 2023; Kiazyk & Ball, 2017; Tufariello et al., 2003). Given that immunosuppression, smoking, and other predisposing factors can cause the dormant bacteria to reactivate and cause active TB disease, LTBI represents a persistent reservoir for potential reactivation (Alavi-Naini et al., 2012),.

2.3 Smoking as a potential risk factor for Tuberculosis Reactivation

Smoking is a major recognized risk factor for lung cancer, chronic obstructive pulmonary disease (COPD), and other respiratory infections like tuberculosis (TB) (Alavi-Naini et al., 2012).

The World Health Organization (WHO) states that exposure to ambient air pollutants from multiple sources, such as particulate matter, burning biomass fuels, and tobacco smoking, has been associated with a number of harmful health effects in humans. Studies conducted across the globe have shown that exposure to ambient air pollutants, such as smoke from wildfires and tobacco smoke, is linked to a higher risk of tuberculosis. Several theories have been proposed to explain the potential mechanisms underlying this association, suggesting that ambient particulate matter may impair the human immune system, thereby increasing the risk of latent tuberculosis infection (Yanbaeva et al., 2007; Stevenson et al., 2007). Specifically, smoking has been postulated to compromise pulmonary immunity against the bacteria (Lin et al., 2007; Maurya et al., 2002; Slama et al., 2007; *Toxics | Free Full-Text | Smoking Exposure and the Risk of Latent Tuberculosis Infection*, n.d.).

2.3.1 Cells Affected

One important protein, CD14, is mostly expressed by monocyte/macrophage lineage cells. It is essential for the innate immune response against intracellular pathogens. (Ayaslioglu et al., 2013; Layre et al., 2014).

Studies have found that high tobacco consumption or cigarette smoking has been associated with elevated levels of leukocytes, monocytes, neutrophils, and lymphocytes, where monocytes show the highest dysregulation (Pedersen et al., 2019). However, current research in this area is limited by the diverse nature of monocytes and the inability to accurately identify specific transcriptional programs within different subpopulations that may drive smoking-induced susceptibility to TB reactivation (Scriba et al., 2022).

2.4 TB Diagnosis challenges

Latent TB is recognized by the existence of *Mycobacterium tuberculosis* in the host without an active illness, and it poses a diagnostic challenge due to the bacteria's dormancy. Tuberculin skin tests (TST) and interferon-γ release assays (IGRA) are widely being used as diagnostic techniques for detecting tuberculosis (Mwaba et al., 2020). Unfortunately, both of these methods have several limitations when it comes to identifying latent TB infection. This underscores the necessity for a comprehensive diagnostic approach that integrates multiple tools and a meticulous clinical evaluation to accurately identify latent TB cases, especially in the context of smokers (Dey & Bishai, 2014; Qiu et al., 2021).

2.5 RNA Sequencing

The entire collection of RNA transcripts that are present in a cell at a particular stage of development or physiological state is referred to as the transcriptome (Conesa et al., 2016; Stark et al., 2019). Deciphering the transcriptome is essential in understanding the genome's functional components, clarifying the molecular makeup of cells and tissues, and learning more about the mechanisms underlying disease and developmental processes (Love et al., 2014; Trapnell et al., 2013).

Numerous technologies, including microarrays, have been developed to infer and quantify the transcriptome. The reliance on pre-existing genomic sequence data, high background noise from cross-hybridization, and a constrained dynamic range of detection because of background signals and signal saturation are some of the limitations of these techniques. Transcriptome analysis has been completely transformed by the introduction of new high-throughput sequencing techniques, such as RNA Sequencing.

These techniques offer a potent method for mapping and quantifying transcriptomes, which has clear advantages over conventional techniques (Z. Wang et al., 2009).

2.5.1 Overcoming the limitations of bulk RNA sequencing with Single-Cell RNA Sequencing

Bulk RNA sequencing (Bulk RNA-seq) is a crucial tool for comprehending gene expression in intricate biological samples. Nonetheless, it can overshadow the disparities between different cell types and states by averaging the gene expression of numerous cell types. This drawback sparked the advent of single-cell RNA sequencing (scRNA-seq). This innovation facilitated the separation of individual cells, the amplification of their RNA, and the following sequencing and examination of their transcriptomes (Hwang et al., 2018).

In the context of this study to determine the impact of smoking on Tuberculosis, scRNA-seq holds immense potential to determine the molecular mechanisms underlying the detrimental effects of cigarette smoke on the host’s immune response to *M. tb.* infection (Burusie et al., 2020). By profiling the transcriptomes of individual immune cell types from individuals with LTBI, with and without smoking exposure, a comprehensive understanding of the smoking-induced cellular and molecular changes that may contribute to TB reactivation and disease progression can be achieved.

2.5.2 Advantages of single-cell RNA sequencing over Bulk RNA sequencing

Single-cell RNA sequencing is a powerful tool that allows us to discover and define new cell types and subtypes that were previously undetectable in large-scale analyses. By studying individual cells, this technique reveals the transitional states that cells go through during their processes of differentiation and disease progression (Jovic et al., 2022; Yu et al., 2021).

Studying the active pathways and gene interactions in individual cells can uncover new potential treatment targets. Creating computational tools to analyze single-cell information has been essential, with methods like clustering and dimensionality reduction being the key. Integrating single-cell data from techniques like scRNA-seq with other forms, such as single-cell proteomics, offers a complete view of cellular biology. By studying individual cells, we can find clues about disease markers and possible treatment targets. This knowledge can be used to create personalized diagnostics and treatment methods that are linked to each patient's specific cellular characteristics.(Li & Wang, 2021; Yu et al., 2021).

2.5.3 Single-cell RNA sequencing Data Analysis using R

With its robust ecosystem of specialized packages and libraries, R is an open-source programming language and software environment for the graphical and statistical analysis of data. It has gained widespread adoption as a tool for scRNA-seq data analysis (Huber et al., 2015; Lun et al., 2016).

2.6 Study Objectives and Significance

This study is designed to understand the underlying mechanisms of what is happening in people who smoke and have TB by using a new set of data (molecular/genomic approaches), analyzing the manner by which tobacco smoke is signaling the reactivation of tuberculosis, and the subsequent immune responses that follow. This is to achieve the following goals:

1. Analyze the effect of smoking exposure on the transcriptional profile of immune cells from latent tuberculosis-infected individuals.

2. Determine the functional states and crosstalk of immune cell subsets, including macrophages, T cells, and many other immune-related factors that act together to combat *M. tb.* infection under the influence of smoking.

3. Uncover potential molecular markers or signatures that may be associated with the risk of LTBI reactivation or disease progression in smokers as new therapeutic biomarkers.

CHAPTER 3

METHODOLOGY

3.1 Data Collection

The sequencing data for this analysis was obtained from the Single Cell Portal database, a publicly accessible repository for high-throughput genomic data (Tarhan et al., 2023). Before proceeding with the analysis, several preprocessing steps were performed to determine the sequencing data's quality and reliability. These included trimming the low-quality bases, removing adaptor sequences, and filtering out the reads that did not meet specific quality thresholds.

3.2 Data Analysis

The dataset was analyzed using the R package “Seurat”, a widely used package designed for QC analysis, and exploration of scRNA-seq data. Some of the analysis was done using the Python package Scanpy, a scalable toolkit for analyzing single-cell gene expression data (Amezquita et al., 2020).

3.3 Functionality of scRNA Workflow

To determine the heterogeneity from single-cell transcriptomic measurements and to integrate diverse types of single-cell data, Seurat uses advanced statistical models:

- **Seurat Object:** It acts as a container that stores the single-cell data, including the Count Matrix (containing the number of RNA molecules (reads) for each gene in each cell), Metadata (information about the cells), and Analysis Results (Outputs from various computations performed on the data like dimensionality reduction—PCA or clustering results) (Butler et al., 2018).

- **Quality Control and Filtering:** Seurat allows you to easily explore QC metrics and filter cells based on any user-defined criteria (Subramanian et al., 2022).

  A few QC metrics commonly used include:

  - Number of unique genes detected in each cell
  - Low-quality cells
  - Cell doublets or multiplets
  - Percentage of reads mapping to mitochondrial genome
  - Low-quality/dying cell

- **Normalization:** Normalization is an essential step in analyzing the single-cell data in order to address variations in sequencing depth and technical noise, which can impact the accuracy of subsequent analysis. Log Normalize, a function in the Seurat R package, normalizes single-cell RNA sequencing information. It eliminates technical discrepancies among the cells. To achieve this, the counts for each gene in a cell are divided by the total counts for that specific cell, accounting for the differences in sequencing depth. A scale factor then adjusts the resulting values to calibrate the overall scale of normalized data. Finally, the log1p function is applied to these normalized values (Hafemeister & Satija, 2019).

- **Finding Variable Features:** The vst method in Seurat's FindVariableFeatures function is used to identify highly variable genes in the data (Peng et al., 2024).

- **Scaling the data:** The data was scaled prior to performing dimensionality reduction, such as PCA. The ScaleData() function adjusts the expression of each gene to have a mean of 0 and variance of 1 across the cells (Hafemeister & Satija, 2019). This is done in order to ensure that highly expressed genes do not have more influence in the subsequent analyses.

- **Dimensionality Reduction:** It is a technique used to reduce the number of variables or features in a dataset. It is useful in dealing with high-dimensional data, as it can help overcome the problem of dimensionality, reduce computational complexity, and improve model performance. The Seurat package provides several dimensionality reduction methods, such as the Principal Component Analysis (PCA), which identifies a set of orthogonal linear combinations (principal components) that explain the majority of the variance in the data (Sun et al., 2019). These components capture the most significant sources of variation in the dataset. The first component is the primary direction that accounts for the greatest amount of variation in the data, followed by the rest of the components with less variations.

- **Find Neighbours and Clustering:** Higher resolution results in an increased number of clusters, with a default value of 0.8. This process involves identifying distinct subgroups within the dataset, where cells within the same cluster are highly similar while those in different clusters are markedly different (Zhuang et al., 2022). The K-nearest neighbor of each cell can be computed using the FindNeighbors() function. Then, we conducted graph clustering with the FindClusters() function. This will provide each cell with a specific number based on its cluster (Zhang et al., 2023).

- **Non-Linear Dimensionality Reduction:** Seurat uses multiple types of non-linear dimensionality reduction methods, including the UMAP method:

    - **Uniform Manifold Approximation and Projection (UMAP)** is a method for reducing the dimensionality of data while maintaining its underlying structure. This technique measures similarities between data points based on their topology and creates a graph to approximate the data (Yousuff et al., 2024). Therefore, UMAP creates a lower-dimensional representation of the original data.

- **Differential Expression Analysis:** Differential expression using the Seurat package's statistical methods was performed to identify marker genes that distinguish specific cell populations or subpopulations. This analysis enabled us to characterize the unique transcriptional signatures associated with distinct cell states or conditions, such as smoking status.

3.4 Visualization

Different visualization techniques were utilized to interpret the scRNA-seq data. I generated Violin plots, Feature plots, Dim plots, Variable Feature plots, Elbow plots, Heat maps, UMI plots, and more to visualize the analysis's results.

CHAPTER 4

RESULT AND DISCUSSION

4.1 Visualization of the number of RNA, number of features, percentage of mitochondria, and percentage of ribosomal proteins via Violin Plot (Vln Plot) and Feature Scatter plot

The number of counts of RNA (represented by nCount_RNA), number of features (represented by nFeature _RNA), percentage of mitochondria (represented by percent.mt), and percentage of ribosomal proteins (represented by percent.rb) per cell across the sample is represented via Violin and Feature Scatter plots.

A number of quality control metrics are first evaluated to determine the integrity and reliability of single-cell RNA sequencing data, such as counts of RNA (nCount_RNA), the number of features (nFeature_RNA), the percentage of mitochondrial gene expression (percent.mt), and the percentage of ribosomal protein genes (percent.rb) per cell across the sample. Violin plots and feature scatter plots are used to visualize and examine the distribution of these metrics. Violin plots provide a comprehensive view of the probability density of the data and allow for the identification of potential outliers or skewed distributions. Feature scatter plots give a two-dimensional representation of the relationship between any two variables and enable the assessment of potential correlations or patterns within the dataset (Mothe & Martha, 2023).

Moreover, these visual representations help researchers gain insights into the quality of scRNA-seq data. For instance, cells with a high percentage of mitochondrial gene expression of ribosomal protein genes may indicate issues like cellular stress or degradation. Similarly, cells with extremely low or high counts of RNA or features may be indicative of technical faults or biological outliers.

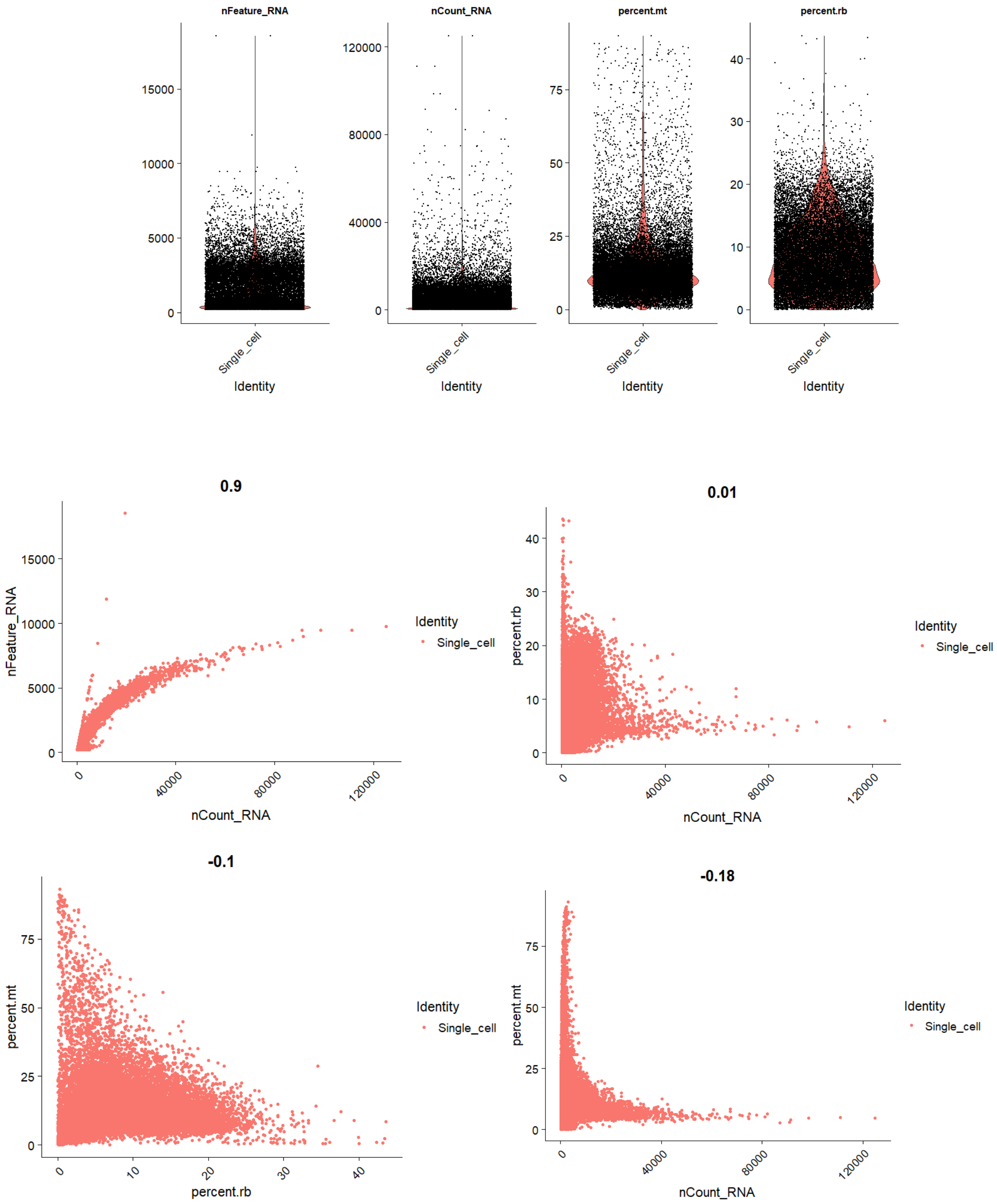


Fig. 4.1: a) Violin Plot representing nCount_RNA, nFeature_RNA, percent.mt, percent.rb across the samples; b) Feature Scatter Plot for nCount_RNA vs nFeature_RNA; c) Feature Scatter Plot for percent.rb vs nCount_RNA; d) Feature Scatter Plot for percent.mt vs percent.rb; e) Feature Scatter Plot for percent.mt vs nCount_RNA

4.2 Visualization of the Top Highly Variable Genes via the Variable Feature Plot

The Variable feature plot, also known as the mean-variance plot or the dispersion plot, is a visualization tool used to analyze single-cell RNA sequencing data. It helps to identify the highly variable genes or features across the dataset (Y. Wang et al., 2023).

The relationship between the geometric mean (x-axis) and the residual variance (y-axis) of expression of each gene or feature is shown in this plot (Fig. 2). The residual variance quantifies the departure from the expected mean-variance relationship, indicating the degree of variability for each gene, whereas the geometric mean of expression represents the average expression level across all cells. As seen in Fig. 2, the most variable genes or factors are those with moderate to high expression levels. These genes or features are often highlighted or colored differently than others. The plot highlights several important genes or cell markers, including CLC, HBA1/TUBB1, HBA2, and HBB.

This remarkable diversity of genes is particularly interesting because it is possible that they will be differentially expressed in different cell types or conditions. This provides important information about the cellular mechanisms by which smoking affects and reactivates tuberculosis.

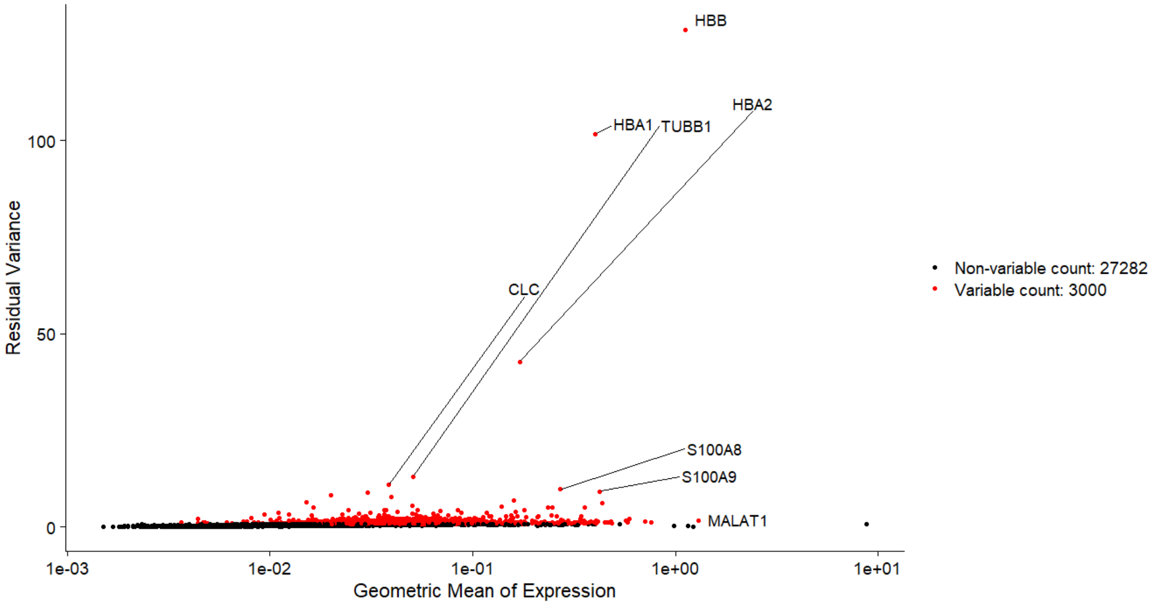


Fig. 4.2: Variable Feature Plot expressing the Top Highly Variable Genes across the sample data

4.3 Visualization of Linear Dimensional Reduction (PCA) via VizDimLoadings Plot, Dim Plot, and Dim Heatmap Plot

A popular method for linear dimensionality reduction used in the analysis of single-cell RNA sequencing (scRNA-seq) data is Principal Component Analysis (PCA) (Tsuyuzaki et al., 2020). This method allows one to visualize and investigate the underlying cellular heterogeneity and potential subpopulations by identifying the most important sources of variation in the high-dimensional gene expression data.

In this study, the PCA results are visualized with the help of three plots: the VizDimLoadings plot (represented by Fig. 3a), the Dim plot (represented by Fig. 3b), and the Dim Heatmap Plot (represented by Fig. 3c). These plots provide insights into the gene loadings, cell distribution, and individual gene contributions to the principal component (PCs) loadings, respectively.

A. The VizDimLoadings plot represents the gene loadings for each principal component. It can identify the genes that contribute the most to each PC's evolution. For example, the top genes in PC_1 exhibit high loadings, suggesting that they may be relevant in incorporating the observed variation in the data set.

B. Dim plot is a scatter plot of the distribution of cells along the first two principal components (PC_1 and PC_2). This plot shows the possible clusters or subpopulations because cells with similar transcriptome profiles tend to group together.

C. Lastly, the Dim Heatmap plot depicts the cell embeddings across multiple principal components. Each panel in the plot represents a two-dimensional projection of the cells onto the respective PC pairs. The intensity of the colors in the plot can be compared to the density of cells in a region to identify potential substructures or trajectories in the data.

With the help of these complimentary PCA visualizations, insights can be gained into the gene signatures responsible for cellular heterogeneity, identify possible subpopulations/groups, and determine relationships between cells based on their individual transcriptome profiles (Townes et al., 2019).

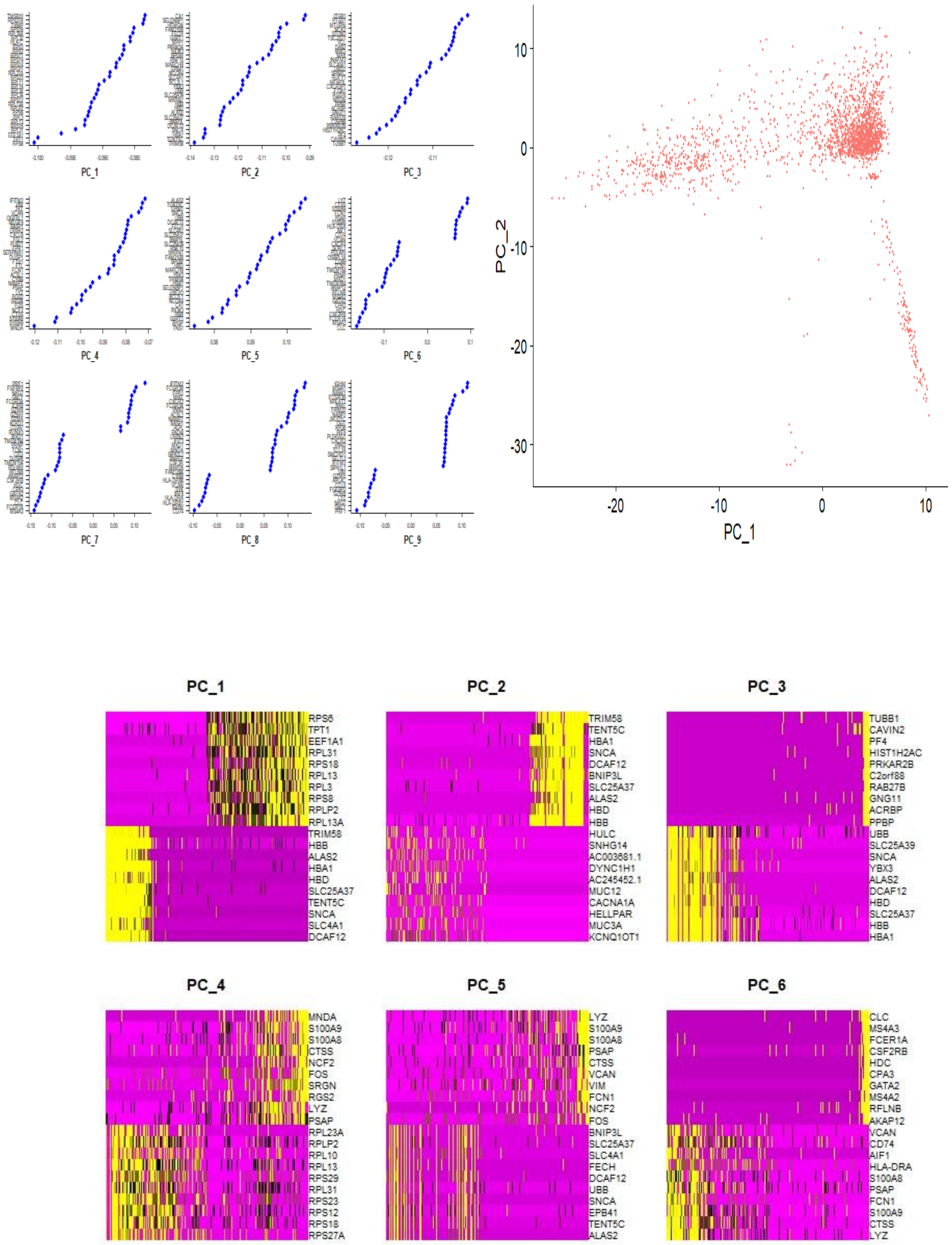


Fig. 4.3: a) VizDimPlot Representing grouped genes in PCs; b) DimPlot PCA; c) DimHeatmap for PCs (1:6)

4.4 Visualization of Dimensionality of the Dataset via Elbow Plot

The dataset's Dimensionality is determined via an Elbow Plot. It shows the number of clusters forming an elbow-like shape and the possible number of groups. It helps determine how many PCs we need to capture the majority of the variations in the data, which in this case is 10.

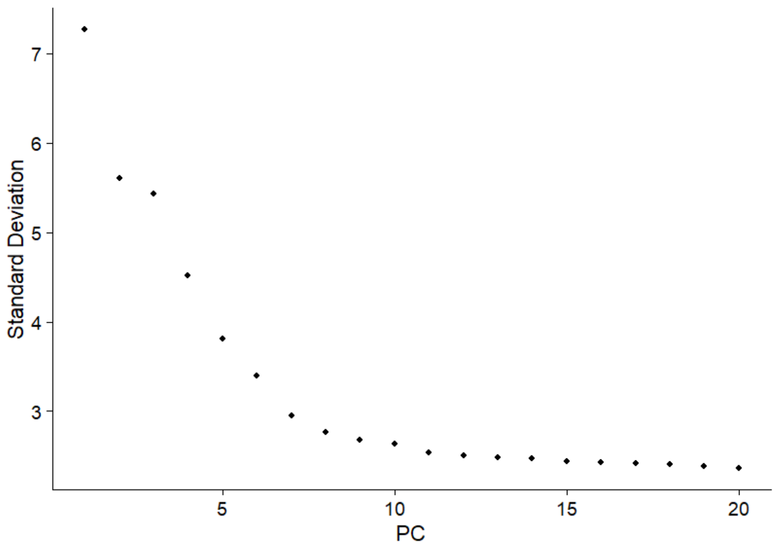


Fig. 4.4: Visualization of Elbow Plot for determining the Dimensionality of the dataset

4.5 Visualization of Non-Linear Dimensionality Reduction and Comparison with Linear Dimensional Reduction via Dim Plot

The Non-Linear Dimensionality Reduction is expressed via Dim Plot using the UMAP algorithm. The high-dimensional dataset is reduced to a low-dimensional plot that retains much of the original information (Yousuff et al., 2024). Single-cell RNA sequencing datasets contain thousands of gene expression counts for each cell. This information is then condensed down to two or three dimensions, allowing each to be represented on a two- or three-dimensional plot. The UMAP and t-SNE algorithms are very similar as they are manifold learning algorithms.

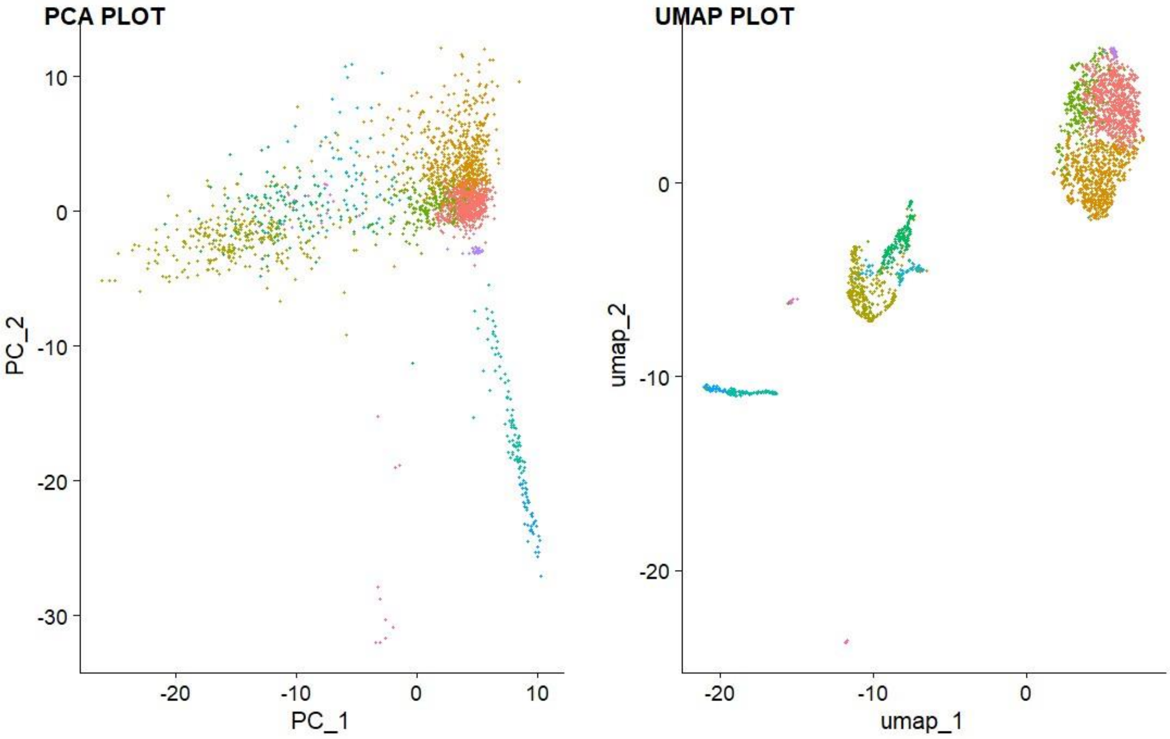


Fig. 4.5: Visualization of PCA Plot and UMAP Plot via Dim Plot

4.6 Visualization of Clusters

Using FindAllMarkers reveals all the markers present, and grouping them by clusters reveals all the types of biomarkers and the number of clusters present, which is expressed via Dim Plot. According to our study, specific biomarkers and genes of interest are visually expressed using a Violin Plot and Feature Plot.

Cluster Analysis reveals ten different clusters present across the sample dataset. Tissue-specific markers are highly expressed in the cluster for at least one tissue and not highly expressed in some other tissue. Cluster marker analysis is done with expression to associate gene expression with cluster identity.

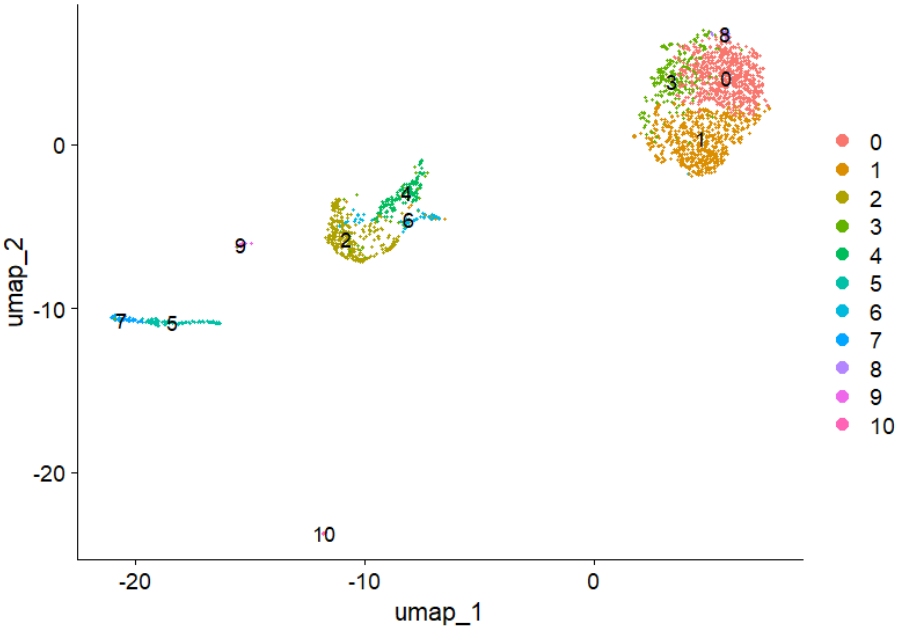


Fig. 4.6: Visualization of Clusters using Dim Plot

4.7 Visualization of Top Markers Expression

Fig. 4.7 displays violin plots, which serve as a reliable visual tool for quantifying the expression levels of important cell markers (namely, lung monocyte markers) across various experimental settings or cells in the smoker's TB dataset (Tanious & Manolov, 2022). The violin plots display the expression levels of many important markers in certain cell types, including CD93, CD14, FPR1, CCR2, LTF, and HLA-DRA.

The violin plot of **CD93**, a marker linked to immune cell activation and differentiation, demonstrates stable expression levels across the dataset, with minimal fluctuations.

Likewise, the expression of the model recognition receptor **CD14**, which is implicated in the innate immune response, has a pretty consistent pattern, while certain identities display slightly elevated levels of expression.

Chemotaxis and inflammatory response-associated formyl peptide receptor 1 (**FPR1**) exhibit highly variable expression patterns between identities, suggesting that changes in activation or functional status can occur in specific cell subsets.

Furthermore, due to a few potential outliers, the expression of **CCR2**, a chemokine receptor involved in inflammation and monocyte recruitment, appears to be relatively constant in most of the identities.

The antimicrobial and immunomodulatory lactoferrin (**LTF**) marker shows a wider range of expression levels, indicating possible differences in the functional states of specific cell subsets.

Lastly, there is a distinct expression pattern for the major histocompatibility complex class II molecule **HLA-DRA**, which is essential for antigen presentation and adaptive immune responses. One identity, representing a particular cell type (CD14+ monocyte in this case), shows significantly higher levels than the others.

Notably, out of the six markers, CD93, CD14, and CCR2 show the maximum consistency in terms of expression levels across the entire dataset.

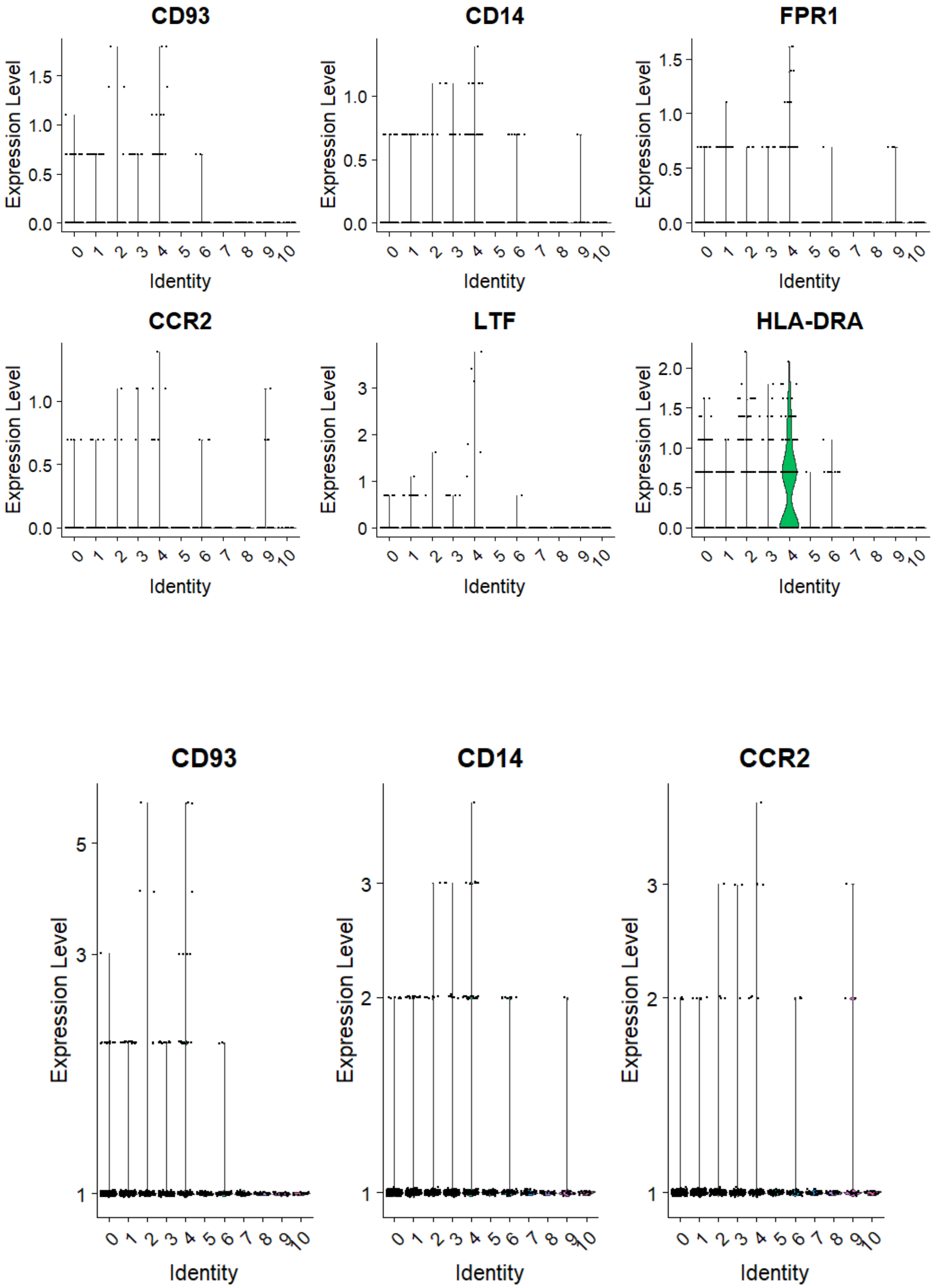


Fig. 4.7: Highly Expressed Lung Monocytes Markers visualized using Violin Plot

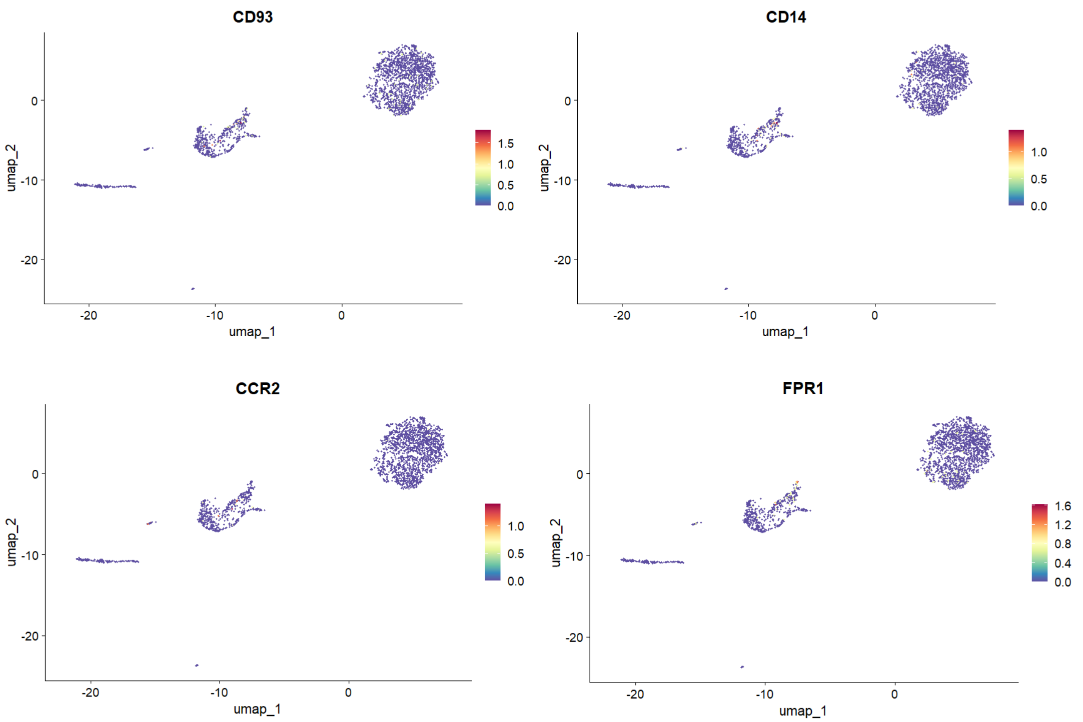


Fig. 4.8: Visualization of Top Markers Expression via Feature Plots

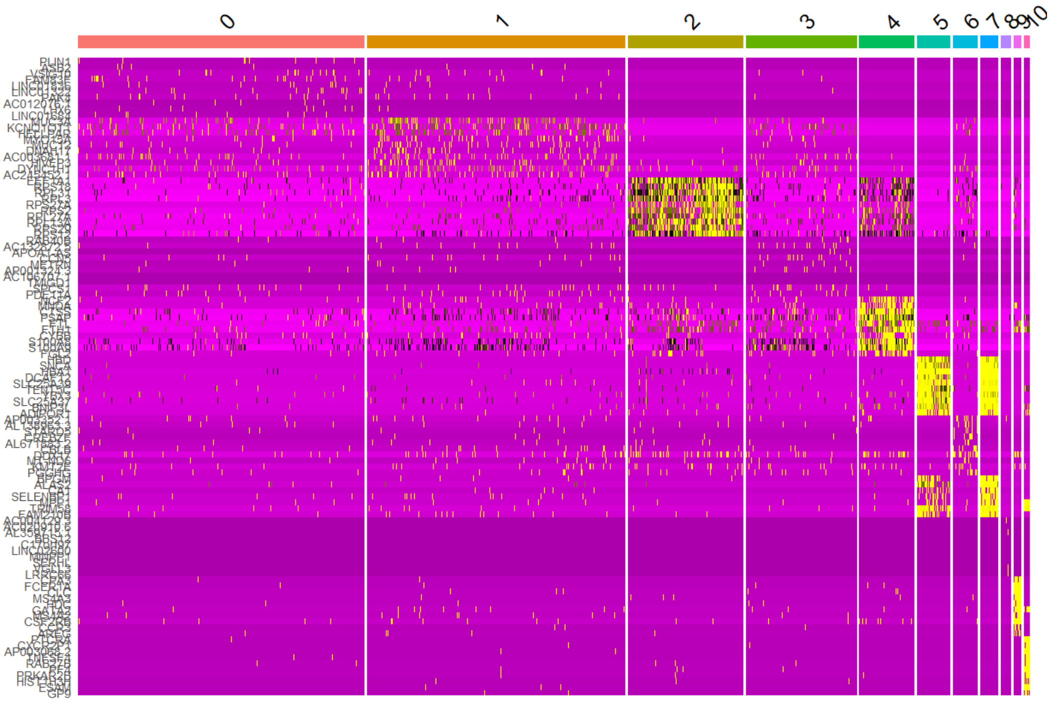


Fig. 4.9: Heat Map representing the sample expressions of genes across all 10 clusters

4.8 Visualization of Cell Proportion and Unique Molecular Identifier (UMI) Plot

The cell proportions plot shows the relative abundance of various cell types within the sample. This plot shows that CD14+ monocytes are the most common cell population in the smokers' tuberculosis dataset, indicating their possible significance in the immune system's reaction to smoking-related illnesses and tuberculosis.

The Unique Molecular Identifier (UMI) plot shows the distribution of UMI counts across cells, allowing one to evaluate the complexity and quality of the data (Chen et al., 2018; Sena et al., 2018). The UMI plot in this dataset supports the number of CD14+ monocytes, emphasizing their importance in smoking and tuberculosis.

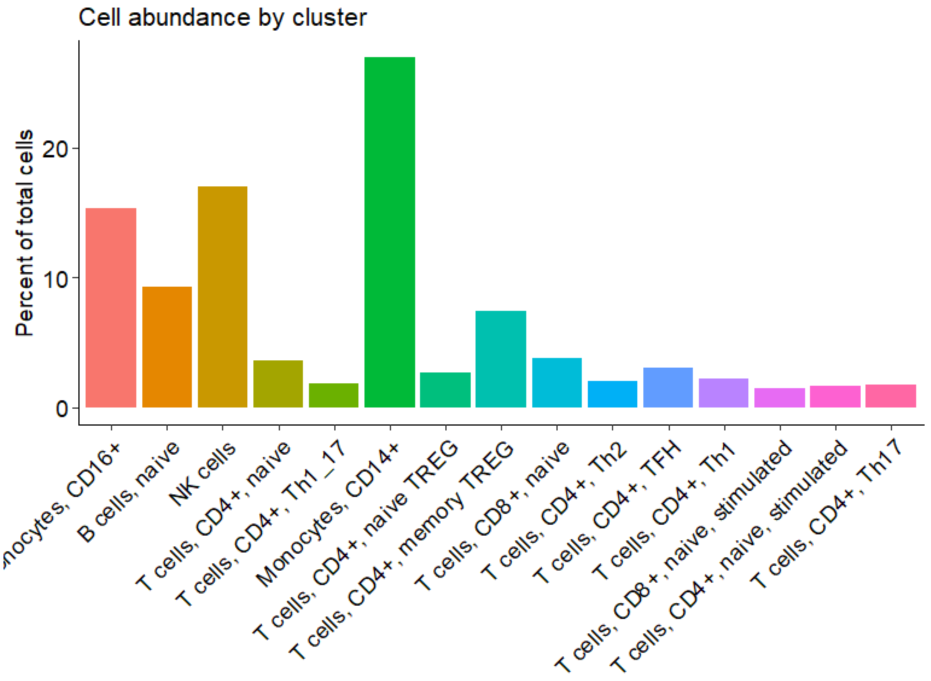


Fig. 4.10: UMI Plot showing CD14+ Monocytes as the most abundant cells present in the sample

CHAPTER 5

CONCLUSION AND FUTURE PROSPECTS

An important discovery from the Single-cell RNA Sequence analysis of the TB dataset from smokers is the upregulation of **CCR2, CD14,** and **CD93** expression. These can act as potential lung biomarkers associated with smoking-induced latent TB reactivation.

To trigger the natural immune response against diseases such as Mycobacterium TB, it is crucial to have monocytes and macrophages that carry the pattern recognition receptor **CD14**. Individuals who smoke experience elevated levels of CD14+ monocyte expression in their lungs. These results are consistent with previous studies showing that smoking can recruit immature inflammatory monocytes into the airspaces, creating an environment that is favorable for *M. tb.* proliferation (Ayaslioglu et al., 2013; Corleis et al., 2022).

The migration and trafficking of monocytes depend on the chemokine receptor **CCR2**, which is generated by monocytes and other immune cells. Research indicates that smokers have higher-than-normal levels of CCR2 in their lungs, which may help attract monocytes to the site of inflammation and increase the inflammatory response (Samstein et al., 2013). This procedure upsets the delicate equilibrium that controls the sickness and may lead to a recurrence of tuberculosis.

The transmembrane protein **CD93**, which is connected to immunology and endothelial cell integrity, may be able to recognize a portion of recently recruited pulmonary monocytes linked to smoking, according to research by Jagatia et al (Jagatia & Tsolaki, 2021). Given the high expression of this gene in the data set, chemokines that bind to CCR2 most likely have an impact on monocyte recruitment. CD93 may influence the development of the immune system throughout the pathophysiology of TB, given its association with both inflammation and the immunological response (Mälarstig et al., 2011).

Examining the ways in which smoking, CD93, and TB reactivation interact and contribute to the disruption of the host-pathogen balance is necessary in order to fully understand the interaction between these three variables. Elevated levels of **CD93** might potentially affect the stability of endothelial cells and the vascular response, both of which are critical for immune cell migration to infection sites. However, long-term inflammation caused by smoking and dysregulated CD93 may break the delicate balance needed to manage latent tuberculosis, which might hinder immune cell migration and worsen endothelial dysfunction. Consequently, the compromised function of the endothelium and altered immune environment promote the reactivation of latent TB, highlighting the dual role of CD93 in both avoiding infection and potentially aggravating the effects of the disease.

The results of this work on the connection between smoking, immunological dysregulation, and latent tuberculosis reactivation using single-cell RNA sequencing analysis provide interesting directions for further investigation and possible treatment approaches (Pan et al., 2023). The identification of important molecular players, including CD14, CCR2, and CD93, has opened the door to more study of the underlying mechanisms and the potential use of these targets as therapies.

The findings of the study provide opportunities for developing better biomarker panels and diagnostic instruments in addition to therapeutic choices. Researchers could create more precise and comprehensive risk assessment models that would identify people who are more likely to experience latent tuberculosis reactivation due to smoking or other environmental exposures by combining the identified biomarkers with other pertinent factors, such as clinical and demographic data.

Furthermore, combining scRNA-seq data with other omics techniques, such as proteomics and metabolomics, may help us better comprehend the cellular and molecular environment of smoking-induced tuberculosis reactivation. By using a multi-omics approach, we can find new pathways, biomarkers, and therapeutic targets that could help us better understand this intricate disease process.

REFERENCES

1. Alavi-Naini, R., Sharifi-Mood, B., & Metanat, M. (2012). Association Between Tuberculosis and Smoking. *International Journal of High Risk Behaviors & Addiction*, *1*(2), 71–74. https://doi.org/10.5812/ijhrba.5215
2. Alsayed, S. S. R., & Gunosewoyo, H. (2023). Tuberculosis: Pathogenesis, Current Treatment Regimens and New Drug Targets. *International Journal of Molecular Sciences*, *24*(6), 5202. https://doi.org/10.3390/ijms24065202
3. Amezquita, R. A., Lun, A. T. L., Becht, E., Carey, V. J., Carpp, L. N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C., Waldron, L., Pagès, H., Smith, M. L., Huber, W., Morgan, M., Gottardo, R., & Hicks, S. C. (2020). Orchestrating single-cell analysis with Bioconductor | Nature Methods. *Nature Methods*, *17*(2), 137–145. https://doi.org/10.1038/s41592-019-0654-x
4. Ayaslioglu, E., Kalpaklioglu, F., Kavut, A. B., Erturk, A., Capan, N., & Birben, E. (2013). The role of CD14 gene promoter polymorphism in tuberculosis susceptibility. *Journal of Microbiology, Immunology and Infection*, *46*(3), 158–163. https://doi.org/10.1016/j.jmii.2012.05.008
5. Burusie, A., Enquesilassie, F., Addissie, A., Dessalegn, B., & Lamaro, T. (2020). Effect of smoking on tuberculosis treatment outcomes: A systematic review and meta-analysis. *PLoS ONE*, *15*(9), e0239333. https://doi.org/10.1371/journal.pone.0239333
6. Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, *36*(5), 411–420. https://doi.org/10.1038/nbt.4096
7. Chen, W., Li, Y., Easton, J., Finkelstein, D., Wu, G., & Chen, X. (2018). UMI-count modeling and differential expression analysis for single-cell RNA sequencing. *Genome Biology*, *19*(1), 70. https://doi.org/10.1186/s13059-018-1438-9

8. Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*(1), 13. https://doi.org/10.1186/s13059-016-0881-8
9. Corleis, B., Tzouanas, C., Wadsworth, M., Cho, J., Linder, A., Schiff, A., Dickey, A., Medoff, B., Shalek, A., & Kwon, D. (2022). *Tobacco smoke exposure results in recruitment of inflammatory airspace monocytes and accelerated growth of Mycobacterium tuberculosis*. https://doi.org/10.1101/2022.12.21.521304
10. Delogu, G., Sali, M., & Fadda, G. (2013). THE BIOLOGY OF MYCOBACTERIUM TUBERCULOSIS INFECTION. *Mediterranean Journal of Hematology and Infectious Diseases*, *5*(1), e2013070. https://doi.org/10.4084/mjhid.2013.070
11. Dey, B., & Bishai, W. R. (2014). Crosstalk between Mycobacterium tuberculosis and the host cell. *Seminars in Immunology*, *26*(6), 486–496. https://doi.org/10.1016/j.smim.2014.09.002
12. Hafemeister, C., & Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, *20*(1), 296. https://doi.org/10.1186/s13059-019-1874-1
13. Houben, R. M. G. J., & Dodd, P. J. (2016). The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling. *PLOS Medicine*, *13*(10), e1002152. https://doi.org/10.1371/journal.pmed.1002152
14. Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleś, A. K., … Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, *12*(2), 115–121. https://doi.org/10.1038/nmeth.3252
15. Hwang, B., Lee, J. H., & Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine*, *50*(8), 1–14. https://doi.org/10.1038/s12276-018-0071-8

16. Jagatia, H., & Tsolaki, A. G. (2021). The Role of Complement System and the Immune Response to Tuberculosis Infection. *Medicina*, *57*(2), 84. https://doi.org/10.3390/medicina57020084

17. Jovic, D., Liang, X., Zeng, H., Lin, L., Xu, F., & Luo, Y. (2022). Single-cell RNA sequencing technologies and applications: A brief overview. *Clinical and Translational Medicine*, *12*(3), e694. https://doi.org/10.1002/ctm2.694

18. Kiazyk, S., & Ball, T. (2017). Latent tuberculosis infection: An overview. *Canada Communicable Disease Report*, *43*(3/4), 62–66. https://doi.org/10.14745/ccdr.v43i34a01

19. Layre, E., Lee, H. J., Young, D. C., Martinot, A. J., Buter, J., Minnaard, A. J., Annand, J. W., Fortune, S. M., Snider, B. B., Matsunaga, I., Rubin, E. J., Alber, T., & Moody, D. B. (2014). Molecular profiling of Mycobacterium tuberculosis identifies tuberculosinyl nucleoside products of the virulence-associated enzyme Rv3378c. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(8), 2978–2983. https://doi.org/10.1073/pnas.1315883111

20. Li, X., & Wang, C.-Y. (2021). From bulk, single-cell to spatial RNA sequencing | International Journal of Oral Science. *International Journal of Oral Science*, *13*(1), 36. https://doi.org/10.1038/s41368-021-00146-0

21. Lin, H.-H., Ezzati, M., & Murray, M. (2007). Tobacco smoke, indoor air pollution and tuberculosis: A systematic review and meta-analysis. *PLoS Medicine*, *4*(1), e20. https://doi.org/10.1371/journal.pmed.0040020

22. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

23. Lun, A. T. L., McCarthy, D. J., & Marioni, J. C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research*, *5*, 2122. https://doi.org/10.12688/f1000research.9501.2

24. Mälarstig, A., Silveira, A., Wågsäter, D., Öhrvik, J., Bäcklund, A., Samnegård, A., Khademi, M., Hellenius, M.-L., Leander, K., Olsson, T., Uhlén, M., de Faire, U., Eriksson, P., & Hamsten, A. (2011). Plasma CD93 concentration is a

potential novel biomarker for coronary artery disease. *Journal of Internal Medicine*, *270*(3), 229–236. https://doi.org/10.1111/j.1365-2796.2011.02364.x

25. Maurya, V., Vijayan, V. K., & Shah, A. (2002). Smoking and tuberculosis: An association overlooked. *The International Journal of Tuberculosis and Lung Disease: The Official Journal of the International Union Against Tuberculosis and Lung Disease*, *6*(11), 942–951.
26. Mothe, R., & Martha, S. (2023). Analysis of transcriptome of single-cell RNA sequencing data using machine learning. *Soft Computing*, *27*, 1–11. https://doi.org/10.1007/s00500-023-08432-1
27. Mwaba, P., Chakaya, J. M., Petersen, E., Wejse, C., Zumla, A., & Kapata, N. (2020). Advancing new diagnostic tests for latent tuberculosis infection due to multidrug-resistant strains of Mycobacterium tuberculosis—End of the road? *International Journal of Infectious Diseases*, *92*, S69–S71. https://doi.org/10.1016/j.ijid.2020.02.011
28. Pai, M., Behr, M. A., Dowdy, D., Dheda, K., Divangahi, M., Boehme, C. C., Ginsberg, A., Swaminathan, S., Spigelman, M., Getahun, H., Menzies, D., & Raviglione, M. (2016). Tuberculosis. *Nature Reviews Disease Primers*, *2*(1), Article 1. https://doi.org/10.1038/nrdp.2016.76
29. Pan, J., Chang, Z., Zhang, X., Dong, Q., Zhao, H., Shi, J., & Wang, G. (2023). Research progress of single-cell sequencing in tuberculosis. *Frontiers in Immunology*, *14*, 1276194. https://doi.org/10.3389/fimmu.2023.1276194
30. Pedersen, K. M., Çolak, Y., Ellervik, C., Hasselbalch, H. C., Bojesen, S. E., & Nordestgaard, B. G. (2019). Smoking and Increased White and Red Blood Cells. *Arteriosclerosis, Thrombosis, and Vascular Biology*, *39*(5), 965–977. https://doi.org/10.1161/ATVBAHA.118.312338
31. Peng, M., Lin, B., Zhang, J., Zhou, Y., & Lin, B. (2024). scFSNN: A feature selection method based on neural network for single-cell RNA-seq data. *BMC Genomics*, *25*(1), 264. https://doi.org/10.1186/s12864-024-10160-1
32. Qiu, B., Tao, B., Liu, Q., Li, Z., Song, H., Tian, D., Wu, J., Wu, Z., Zhan, M., Lu, W., & Wang, J. (2021). A Prospective Cohort Study on the Prevalent and

Recurrent Tuberculosis Isolates Using the MIRU-VNTR Typing. *Frontiers in Medicine*, *8*, 685368. https://doi.org/10.3389/fmed.2021.685368

33. Samstein, M., Schreiber, H. A., Leiner, I. M., Sušac, B., Glickman, M. S., & Pamer, E. G. (2013). Essential yet limited role for CCR2+ inflammatory monocytes during Mycobacterium tuberculosis-specific T cell priming. *eLife*, *2*, e01086. https://doi.org/10.7554/eLife.01086
34. Scriba, T. J., Dinkele, R., Warner, D. F., & Mizrahi, V. (2022). Challenges in TB research. *The Journal of Experimental Medicine*, *219*(12), e20221334. https://doi.org/10.1084/jem.20221334
35. Sena, J. A., Galotto, G., Devitt, N. P., Connick, M. C., Jacobi, J. L., Umale, P. E., Vidali, L., & Bell, C. J. (2018). Unique Molecular Identifiers reveal a novel sequencing artefact with implications for RNA-Seq based gene expression analysis | Scientific Reports. *Scientific Reports*, *8*(1), 13121. https://doi.org/10.1038/s41598-018-31064-7
36. Slama, K., Chiang, C.-Y., Enarson, D. A., Hassmiller, K., Fanning, A., Gupta, P., & Ray, C. (2007). Tobacco and tuberculosis: A qualitative systematic review and meta-analysis. *The International Journal of Tuberculosis and Lung Disease: The Official Journal of the International Union Against Tuberculosis and Lung Disease*, *11*(10), 1049–1061.
37. Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: The teenage years. *Nature Reviews. Genetics*, *20*(11), 631–656. https://doi.org/10.1038/s41576-019-0150-2
38. Subramanian, A., Alperovich, M., Yang, Y., & Li, B. (2022). Biology-inspired data-driven quality control for scientific discovery in single-cell transcriptomics. *Genome Biology*, *23*(1), 267. https://doi.org/10.1186/s13059-022-02820-w
39. Sun, S., Zhu, J., Ma, Y., & Zhou, X. (2019). Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biology*, *20*(1), 269. https://doi.org/10.1186/s13059-019-1898-6
40. Tanious, R., & Manolov, R. (2022). Violin Plots as Visual Tools in the Meta-Analysis of Single-Case Experimental Designs. *Methodology European Journal*

of Research Methods for the Behavioral and Social Sciences, *18*, 221–238. https://doi.org/10.5964/meth.9209

41. Tarhan, L., Bistline, J., Chang, J., Galloway, B., Hanna, E., & Weitz, E. (2023). Single Cell Portal: An interactive home for single-cell genomics data. *bioRxiv*, 2023.07.13.548886. https://doi.org/10.1101/2023.07.13.548886
42. Townes, F. W., Hicks, S. C., Aryee, M. J., & Irizarry, R. A. (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology*, *20*(1), 295. https://doi.org/10.1186/s13059-019-1861-6
43. *Toxics | Free Full-Text | Smoking Exposure and the Risk of Latent Tuberculosis Infection: Results from NHANES 2011–2012*. (n.d.). Retrieved June 3, 2024, from https://www.mdpi.com/2305-6304/12/1/94
44. Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., & Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, *31*(1), 46–53. https://doi.org/10.1038/nbt.2450
45. Tsuyuzaki, K., Sato, H., Sato, K., & Nikaido, I. (2020). Benchmarking principal component analysis for large-scale single-cell RNA-sequencing. *Genome Biology*, *21*(1), 9. https://doi.org/10.1186/s13059-019-1900-3
46. Tufariello, J. M., Chan, J., & Flynn, J. L. (2003). Latent tuberculosis: Mechanisms of host and bacillus that contribute to persistent infection. *The Lancet Infectious Diseases*, *3*(9), 578–590. https://doi.org/10.1016/S1473-3099(03)00741-2
47. Wang, Y., Sarfraz, I., Pervaiz, N., Hong, R., Koga, Y., Akavoor, V., Cao, X., Alabdullatif, S., Zaib, S. A., Wang, Z., Jansen, F., Yajima, M., Johnson, W. E., & Campbell, J. D. (2023). Interactive analysis of single-cell data using flexible workflows with SCTK2. *Patterns*, *4*(8), 100814. https://doi.org/10.1016/j.patter.2023.100814
48. Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

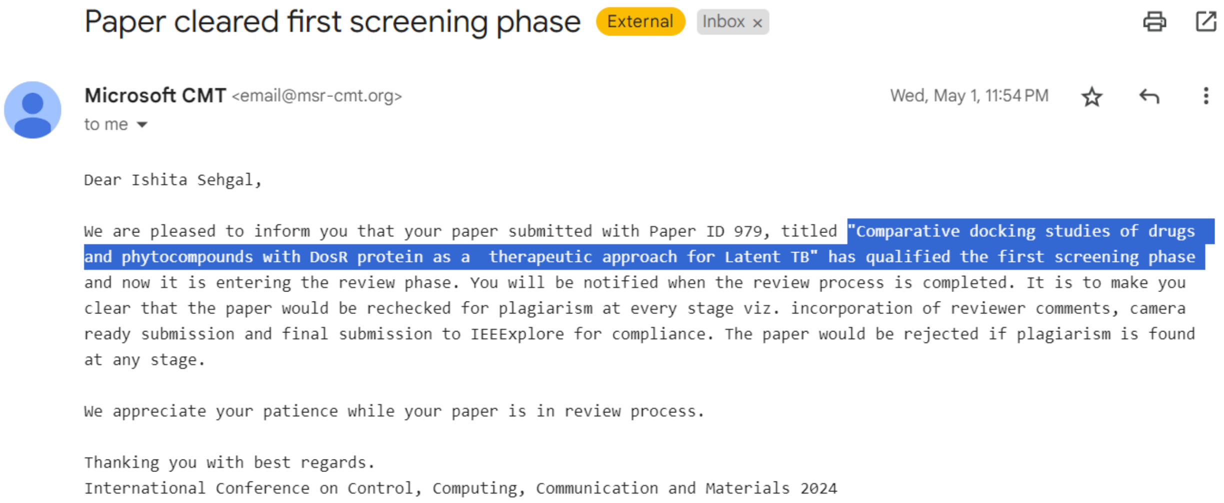
49. Yousuff, M., Babu, R., & Rathinam, A. (2024). Nonlinear dimensionality reduction based visualization of single-cell RNA sequencing data. *Journal of Analytical Science and Technology*, *15*(1), 1. https://doi.org/10.1186/s40543-023-00414-0

50. Yu, X., Abbas-Aghababazadeh, F., Chen, Y. A., & Fridley, B. L. (2021). Statistical and Bioinformatics Analysis of Data from Bulk and Single-Cell RNA Sequencing Experiments. *Methods in Molecular Biology (Clifton, N.J.)*, *2194*, 143–175. https://doi.org/10.1007/978-1-0716-0849-4_9

51. Zhang, S., Li, X., Lin, J., Lin, Q., & Wong, K.-C. (2023). Review of single-cell RNA-seq data clustering for cell-type identification and characterization. *RNA*, *29*(5), 517–530. https://doi.org/10.1261/rna.078965.121

52. Zhuang, H., Wang, H., & Ji, Z. (2022). findPC: An R package to automatically select the number of principal components in single-cell analysis. *Bioinformatics*, *38*(10), 2949–2951. https://doi.org/10.1093/bioinformatics/btac235

LIST OF PUBLICATIONS



Paper cleared first screening phase

Microsoft CMT <email@msr-cmt.org>

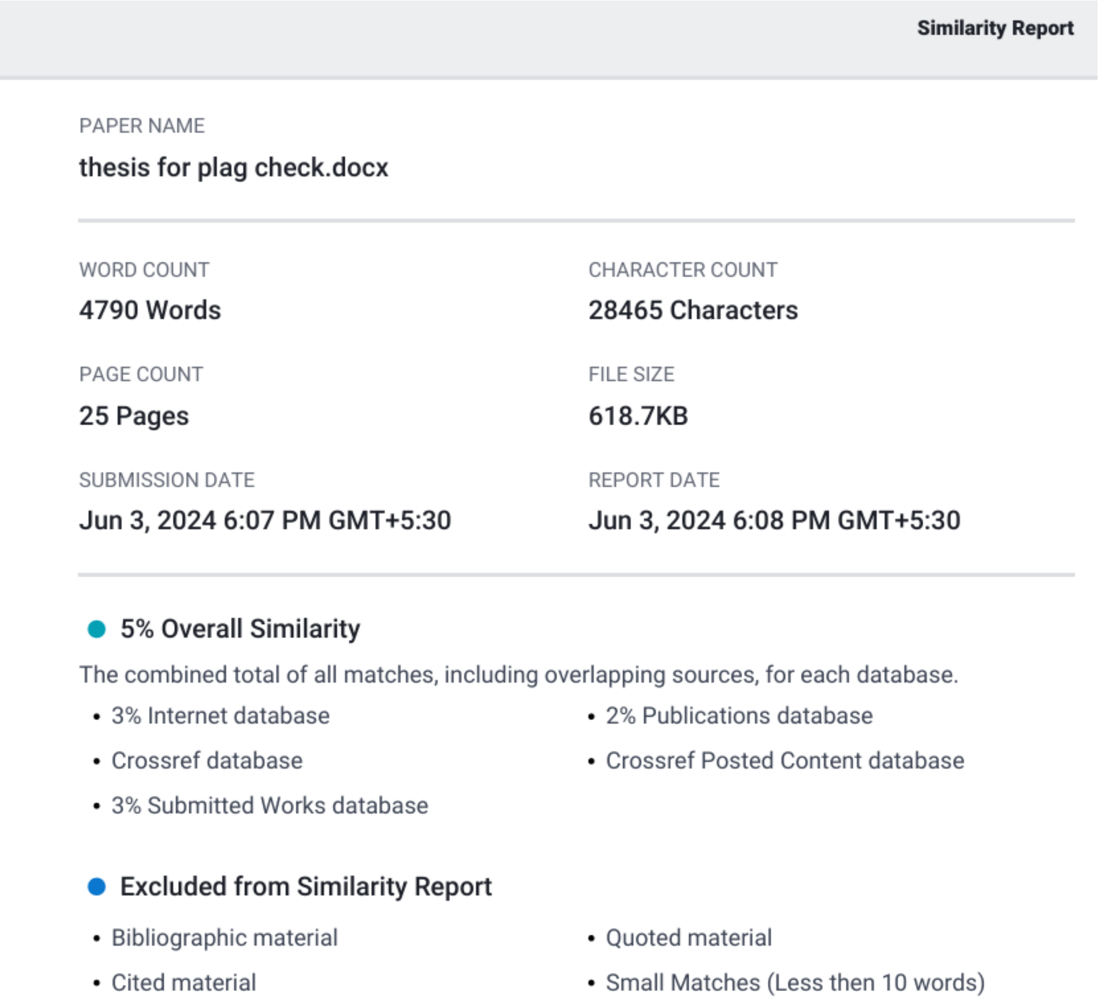
Wed, May 1, 11:54 PM

Dear Ishita Sehgal,

We are pleased to inform you that your paper submitted with Paper ID 979, titled "Comparative docking studies of drugs and phytocompounds with DosR protein as a therapeutic approach for Latent TB" has qualified the first screening phase and now it is entering the review phase. You will be notified when the review process is completed. It is to make you clear that the paper would be rechecked for plagiarism at every stage viz. incorporation of reviewer comments, camera ready submission and final submission to IEEExplore for compliance. The paper would be rejected if plagiarism is found at any stage.

We appreciate your patience while your paper is in review process.

Thanking you with best regards.
International Conference on Control, Computing, Communication and Materials 2024



Similarity Report

PAPER NAME

thesis for plag check.docx

WORD COUNT 4790 Words	CHARACTER COUNT 28465 Characters
PAGE COUNT 25 Pages	FILE SIZE 618.7KB
SUBMISSION DATE Jun 3, 2024 6:07 PM GMT+5:30	REPORT DATE Jun 3, 2024 6:08 PM GMT+5:30

● **5% Overall Similarity**

The combined total of all matches, including overlapping sources, for each database.

- 3% Internet database
- Crossref database
- 3% Submitted Works database
- 2% Publications database
- Crossref Posted Content database

● **Excluded from Similarity Report**

- Bibliographic material
- Cited material
- Quoted material
- Small Matches (Less then 10 words)

ISHITA SEHGAL

+91 8588804931 ishi.sehgal2000@gmail.com
2k22/MSCBIO/58

EDUCATION

M.Sc. Biotechnology	2022-2024	Delhi Technological University	CGPA- 9.22 (Ist year)
B.Sc. (Hons) Biochemistry	2019-2022	Sri Venkateswara College, Delhi University	CGPA- 9.06
CBSE (Class XII)	2019	DAV Public School, Sreshtha Vihar, Delhi	94.0 %
CBSE (Class X)	2017	DAV Public School, Sreshtha Vihar, Delhi	CGPA- 9.40 (89.30%)

INTERNSHIPS

- **Bioinformatics Co-Pilot Internship** [01/02/2024 – Current]
  Under the supervision of **Dr. Pritam Kumar Panda** (German Cancer Research Center, Heidelberg, Germany)
  Computational analysis, Autodock Vina, R sequencing (Bulk and Single-cell RNA-seq), Nextflow, Snakemake, Transcriptomics, Docker, Proteomics LFQ pipeline in R, Chimera X visualization, GROMACS
- **Rapture Biotech** [01/12/2022 – 15/12/2022]
  Completed wet-lab training under the "Advanced Technical Module" on the topics: I) Genomics and Recombinant DNA Technology, II) Medical and Clinical Microbiology.
- **Research Trainee at Sri Venkateswara College** [01/06/2021– 30/12/2021]
  Project I: Comparative docking studies of drugs and phytocompounds for emerging variants of SARS-CoV-2.
  Project II: In-Silico approach to identify and characterize naturally occurring ligands as potential drugs against Dengue Virus.
- **Metvy Research Program Intern** [01/12/2021 – 31/12/2021]
  Understanding research fundamentals, scientific writing, and plagiarism.

COURSES AND CERTIFICATES

- BS Data Science and Programming from IIT Madras [12/2023 – Current]
- Google Data Analytics
  Data Analysis, SQL, Spreadsheet software, Tableau, Data Visualization, Data Management
- Add on course in Bioinformatics and Computational Biology [09/2020 – 01/2021]

PUBLICATIONS

- Sehgal, I.; Chugh, A.; Khurana, N.; Verma, K.; Rolta, R.; Vats, P.; Phartyal, R.; Salaria, D.; Kaushik, N.; Choi, E.H.; et al. Changing Dynamics of SARS-CoV-2: A Global Challenge. Appl. Sci. 2022, 12, 5546. https://doi.org/10.3390/app12115546 [Review article]
- Sehgal, I.; Chugh, A., Khurana, N. et al. Comparative docking studies of drugs and phytocompounds for emerging variants of SARS-CoV-2. 3 Biotech 13, 36 (2023). https://doi.org/10.1007/s13205-022-03450-6 [Research article]

SCIENCE EVENTS

- Poster presentation at INSCR [11/2022] - Presented a poster titled "**Identification of phytocompounds against SARS-CoV-2 variants through comparative docking studies**".
- Poster presentation at the 5th Undergraduate Research Conference, titled "**In Silico approach to identify and characterize naturally occurring ligands as potential drugs against Dengue Virus**".
- Abstract presentation at INSCR Presented an abstract titled "**Combating SARS-CoV-2 using drugs and phytocompounds**".
- MINDSPAR 2.0 Participated and presented a novel approach on the project titled: "**Organic Pesticidal biscuits from Neem oil"** at **MindSpar 2.0** "Meeting New Challenges" Ideathon, organized by the Department of Biochemistry, Sri Venkateswara College in association with Regional Centre for Biotechnology (Haryana).

LABORATORY SKILLS

- **Wet lab**- UV-visible spectrophotometry, Enzymatic assays, Protein purification, Chromatography, Plasmid extraction from bacteria, Isolation of nucleic acids from plant and animal tissues, Sub-Cellular Fractionation, Temporary hematological slide preparation, staining, and cell counting.
- **Bioinformatics and Computational Biology-** Docking using Autodock Vina; Protein structure visualization and analysis using Jmol, PyMol, and Chimera; Homology Modeling using online software; Sequence Alignment and Phylogenetic analysis: BLAST, Clustal-Omega, Bioedit.
- **Scientific writing-** Literature review, Review and Research Paper writing.
- **Other skills**- SQL, Canva, MS Word, MS Excel, and MS PowerPoint.

POSITIONS OF RESPONSIBILITY

- **Placement Coordinator** at Delhi Technological University.
- Conduct of Amalgam’22
  Member of the organizing team in "Amalgam 2022" held in March 2022 organized by the Department of Biochemistry, Sri Venkateswara College, University of Delhi.

EXTRA-CURRICULAR ACTIVITIES AND ACHIEVEMENTS

- Active member of Effulgence-The Photography Society of Sri Venkateswara College [01/2021 – 08/2022].
- Fundraising Internship at the MUSKURAHAT FOUNDATION.

OTHER INFORMATION

LinkedIn profile - www.linkedin.com/in/ishita-sehgal-82b254218

DECLARATION

I hereby declare that the details furnished above are true and correct to the best of my knowledge and belief.