FINDING GENDER-SPECIFIC DIFFERENCES IN GENE EXPRESSION OF ACTIVE TUBERCULOSIS PATIENTS

Project Report Submitted in partial fulfilment of

Bachelor of Technology

in

Biotechnology

By

Janki Insan (181824)

Under the Supervision of

Dr. Rahul Shrivastava

To

Department of Biotechnology and Bioinformatics,

Jaypee University of Information Technology, Solan

Himachal Pradesh (173234)

CERTIFICATE

This is to certify that the work titled "**Finding Gender-Specific Differences in Gene Expression of Active Tuberculosis Patients**", submitted by **Janki Insan (181824)** in partial fulfillment for the award of degree of B. Tech in Biotechnology at Jaypee University of Information Technology, Solan has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Dr. Rahul Shrivastava
Associate Professor
Department of Biotechnology and Bioinformatics
Jaypee University of Information Technology (JUIT)
Waknaghat, Solan, India - 173234

Date:

CANDIDATE'S DECLARATION

I hereby declare that the work presented in this report entitled "**Finding Gender-Specific Differences in Gene Expression of Active Tuberculosis Patients**" in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in Biotechnology submitted in the Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from August 2021 to May 2022 under the supervision of **Dr. Rahul Shrivastava** (Associate Professor), Department of Biotechnology and Bioinformatics, JUIT.

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Janki Insan, 181824

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Dr. Rahul Shrivastava
Associate Professor
Department of Biotechnology and Bioinformatics
Jaypee University of Information Technology (JUIT)
Waknaghat, Solan, India - 173234

Date:

ACKNOWLEDGEMENT

I am very grateful to all those, who have indirectly or directly, assisted and guided me towards the completion of the project.

I would like to thank my supervisor, **Dr. Rahul Shrivastava,** for his unwavering support and guidance throughout the course of the project. Without his expertise, constant mentoring and constructive scrutiny, I would not have been able to complete the project credibly and on time. The fruitful discussions I have had with him along with the independence he gave me in terms of designing the study and project, allowed me to grow immensely and learn so much.

I would also like to express my deep gratitude to my Guru, my mother and my family members, who have always been a huge pillar of support. They have been my constant source of motivation and enabled me to accomplish this project.

Janki Insan, 181824

Date:

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ATB	Active Tuberculosis
LTB	Latent Tuberculosis
PTB	Pulmonary Tuberculosis
TB	Tuberculosis
WHO	World Health Organization
Mtb	*Mycobacterium tuberculosis*
PCA	Principal Component Analysis
FDR	False Discovery Rate
GEO	Gene Expression Omnibus
NCBI	National Center for Biotechnology Information
IFN	Interferon
ssGSEA	single-sample Gene Set Enrichment Analysis
GO	Gene Ontology
AST	Antimicrobial Susceptibility Testing

ABSTRACT

Tuberculosis is a leading Infectious Disease killer, claiming millions of lives every year. In most countries, more men develop active TB─ there exists almost a 2:1 ratio of male-to-female patients. This difference could be attributed to either the contrasting physiological and genetic landscapes in men and women, non-uniform access to healthcare, or other life-style factors, such as nutrition, alcohol intake or smoking. Using transcriptomic datasets from different backgrounds, we investigated if there is a biological basis for these differences by determining whether or not sex influences gene expression in Active Tuberculosis patients using statistical approaches. We filtered samples from patients with active pulmonary tuberculosis (<2 months from the start of anti-tuberculosis therapy) and then looked at the differentially expressed genes in men versus women across cohorts. We found differences in several important pathways, including inflammatory responses and interferon signaling. Our results may have implications on the mechanistic understanding of Tuberculosis and the blood transcriptome-based tests currently under development for TB diagnosis.

Keywords: Blood Transcriptomics, Pulmonary Tuberculosis, Sex-specific differences, TB transcriptomics

CHAPTER 1: INTRODUCTION

1.1. Introduction

Tuberculosis affects millions across the world and is one of the biggest Infectious Disease killers globally. It is estimated that one quarter of the global population has a *Mycobacterium tuberculosis* infection, and although most of these are latent infections, one in ten infected individuals will progress to active disease [1]. Despite numerous TB control programmes in progress across the globe, the WHO estimates that approximately 9.9 million people fell ill with TB in 2020 [2]. Moreover, the COVID-19 pandemic has left around 1.4 million people without access to Tuberculosis-related treatment services. Furthermore, as a result of the COVID-19-related disruptions in Tuberculosis treatment and management, an estimated 500,000 additional TB patients could lose their lives in the coming years, which makes the global burden of TB even worse [3].

1.2. Objectives

- To find if there are any important biological differences in men versus women with regards with regards to susceptibility to Tuberculosis.
- Are there significant differences in transcriptomic data from TB between males and females and are the differences primarily lifestyle-derived or is there any biological basis to it?
- To determine if the gene expression landscape for important immune related genes changes in Male vs. Female patients with pulmonary Tuberculosis.

1.3. Rationale

Tuberculosis is one of the biggest Infectious Disease killers in India. Finding any sex-specific differences could impact how the current diagnostic biomarkers may be used and will also give insight in the mechanistic progression of TB. The diagnostic signatures can help in easy diagnosis of an infection and potentially if the latent form of tuberculosis is likely to progress to an active form (while also considering other factors that may impact this transition). Mechanistic signatures, unlike diagnostic signatures, help us understand the sequence of events or in more detail, the immune response to the host infection [4]. This project focuses on finding genes with significant differences in expression and how these genes tie into our current understanding of the mechanistic and diagnostic Transcriptomic signatures of TB.

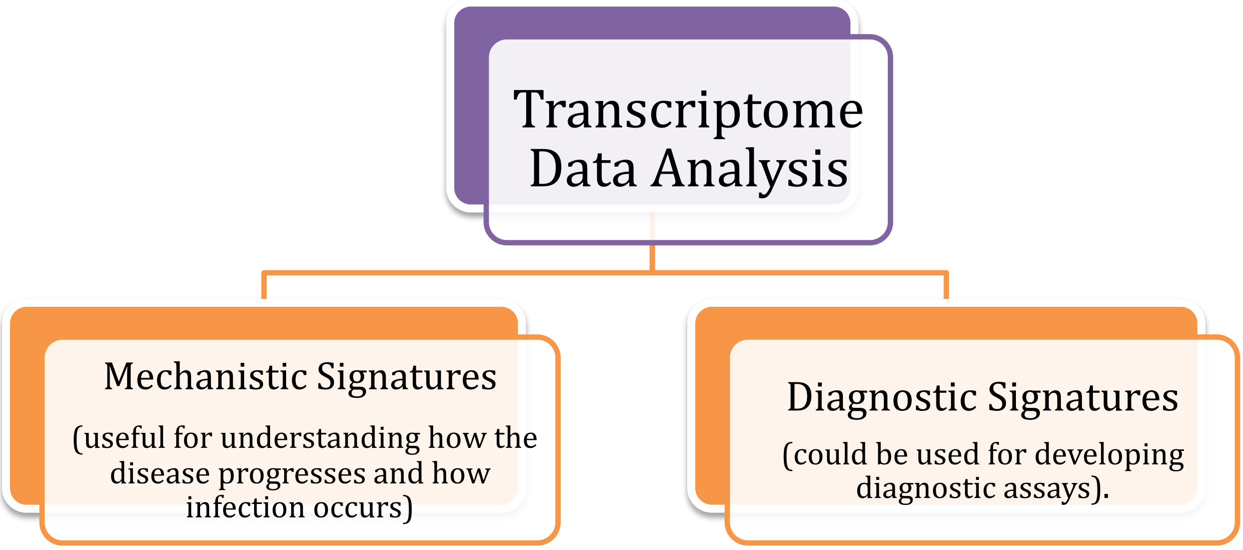

Transcriptome Data Analysis
Mechanistic Signatures
(useful for understanding how the disease progresses and how infection occurs)
Diagnostic Signatures
(could be used for developing diagnostic assays).

Figure 1: Two major uses of signatures derived from transcriptomic data, both of these were analysed in this project.

CHAPTER 2: REVIEW OF LITERARTURE

1.4. Introduction to Tuberculosis

Tuberculosis is an infectious disease, caused by *M. tuberculosis*, a pathogenic bacterium which belongs to the family *Mycobacteriaceae*. The disease, which primarily affects the lungs, causes millions of deaths every year and is one of the biggest infectious disease killers worldwide [2].

Traditionally, any tuberculosis infection is either labelled as 'latent disease' or 'active disease', where in the individuals with Latent Tuberculosis (LTB) do not require either a medical intervention or an antibiotic regimen. Most individuals (~90%) infected with *M. tuberculosis* show no physiological or radiological symptoms [1]. Around 5-15% of infected individuals may progress from LTB to Active Tuberculosis (ATB) and several factors are known to accelerate this progression [2], [5]. These factors include genetic factors, co-morbidities such as Diabetes and HIV, and environmental and lifestyle factors such as alcohol use and smoking. The extent to which these factors impact the progression from LTB to ATB remains an important area of research [6], [7].

Despite most TB infections to this day being labelled as either 'latent' or 'active', recent studies have discovered significant heterogeneity in tuberculosis infections, suggesting that there exists a continuous spectrum of tuberculosis infections [5], [8]. In the last few years, experts have increasingly suggested that dividing tuberculosis infections into only latent and active disease categories does not adequately explain the progression of Tuberculosis from exposure to development of pulmonary disease.

A 2018 review by Paul K. Drain et al. [8] describes three new stages of tuberculosis infections, namely eliminated, incipient and subclinical tuberculosis infections. Herein, an eliminated tuberculosis infection refers to an individual in whom the *M. tuberculosis* infection has been cleared by immune system after prior exposure to the pathogen. Furthermore, the term 'incipient' describes a tuberculosis infection wherein the Mtb pathogen is actively metabolising in the infected person, but the person does not show any microbial, clinical or radiological symptoms. 'Subclinical' disease, on the other hand, describes a state where clinical symptoms are absent but radiological and microbial assays are positive. The authors of the clinical review further emphasize that although describing these states can further develop our understanding of TB infection, this description has its limitations. Discussion of these states and their pathogenesis has been done in the sections that follow.

Overall, the different stages of this disease, the impact of environmental, lifestyle, genetic and health factors, along with increasing incidence of antibiotic resistance in TB patients make tuberculosis a very complex and difficult disease to treat and understand.

2.2. Epidemiology and Pathogenesis of Tuberculosis

2.2.1. Epidemiology

Tuberculosis remains one of the biggest public health burdens for dozens of countries in Africa and Southeast Asia, which is where 70% of the global tuberculosis cases arise. A 2019 report by the Centre for Disease Prevention and Control (CDC), US, also mentions that although the total disease burden is higher for Southeast Asian countries, the proportion of TB-HIV co-infections is higher in Africa (27% for Africa as opposed to 3% for Asia) and the proportion of cases with a form of drug-resistant tuberculosis higher in Europe. Figure 2 shows the share of cases in fifteen countries with the highest TB infection burden as opposed to the rest of the world (the chart was made from the best estimate statistics in the WHO Tuberculosis Report, 2021 [2]). Figure 3 on the following page shows the number of global TB cases area-wise. As can be seen, the number of cases in Africa and Asian are much higher than other regions/continents.

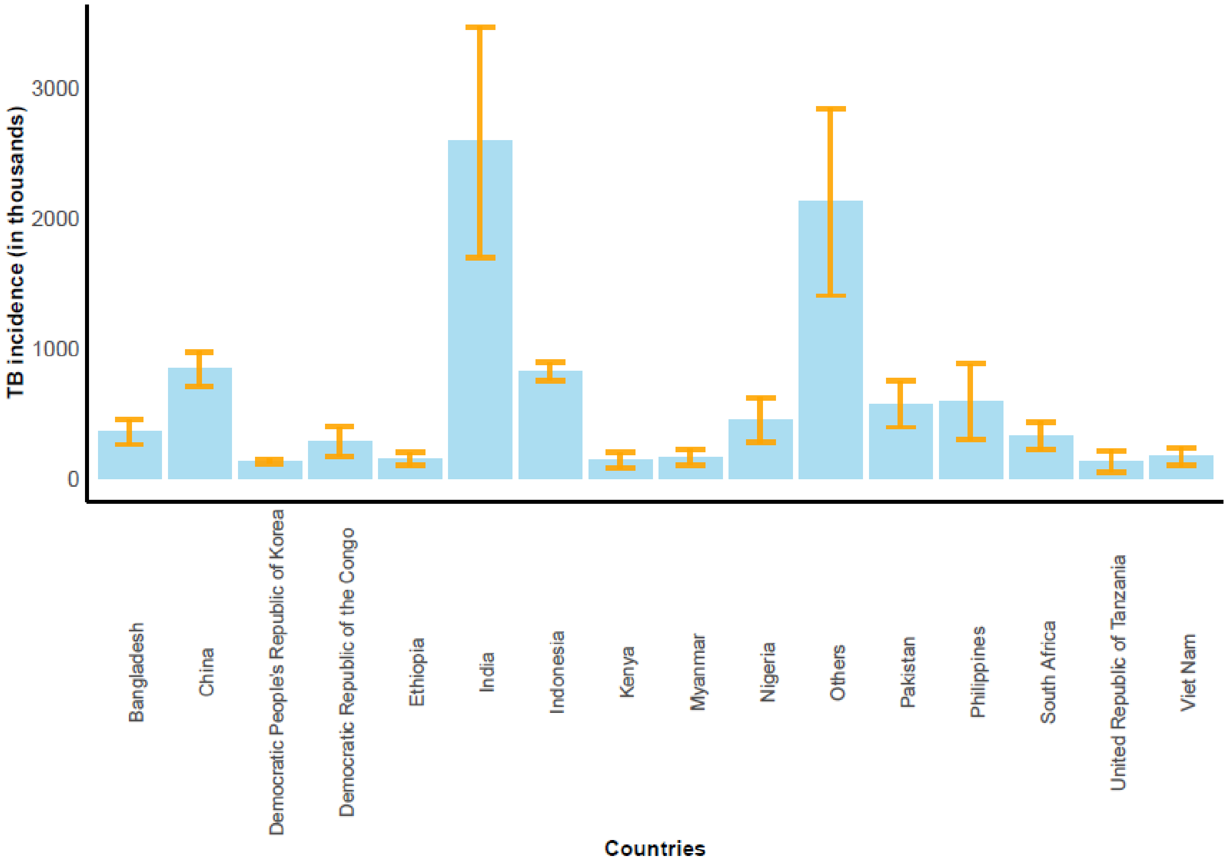


Figure 2: Top 15 countries with the highest TB burden. Data was adapted from the WHO Global Tuberculosis Report, 2021 [2].

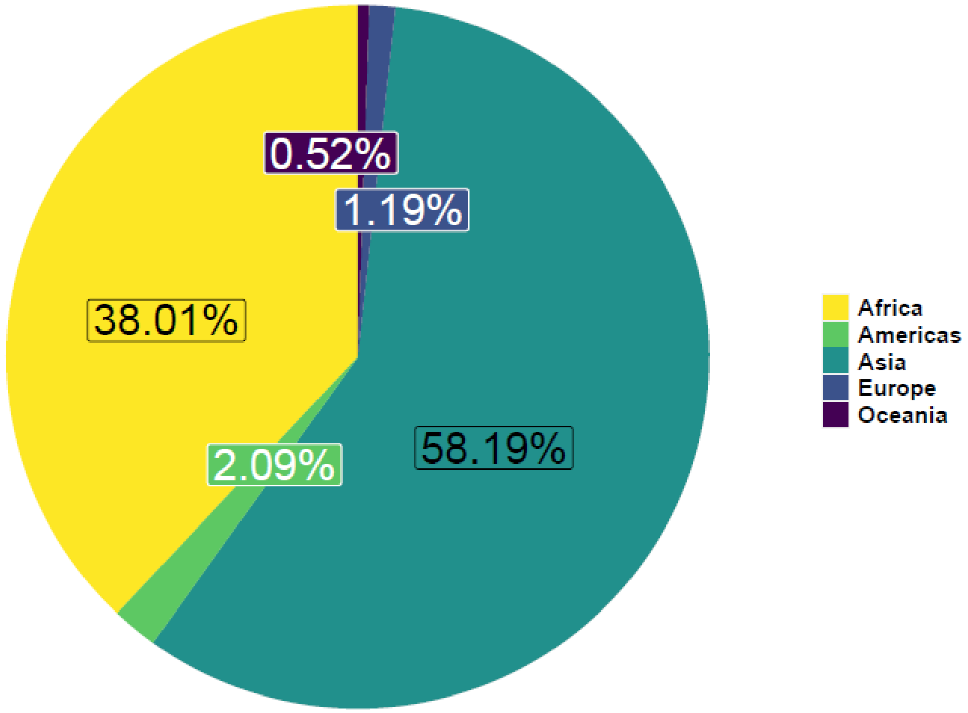


Figure 3: Area-wise distribution of TB cases in 2021. Data adapted from the best estimates of TB infections available in the WHO Global Tuberculosis Report, 2021 [2].

2.2.2. Pathogenesis of Tuberculosis

After an individual is exposed to the Mtb pathogen in the form of aerosolized particles, these pathogen-containing particles travel through the airways, reaching the alveoli inside our lungs. Once in alveoli, the pathogen travels into the tissues in macrophages, where it multiplies and form clusters/aggregates of cells typically referred to as ‘granulomas’. Often, these granulomas protect the pathogen from antibiotics, therefore requiring long-term treatment. The *Mtb* pathogen also can reach a non-replicating state inside the granulomas, becoming resistant to the antibiotics being registered.

An active tuberculosis infection is usually characterised by lesions/granulomas in which a large number of bacteria are found. In immuno-competent individuals, the granulomas or lesions are often highly organized and caveating, whereas, in immuno-compromised individuals are poorly organized and non-caveating. However, the kind of granulomas formed in patients with active disease can differ significantly from individual to individual. In some cases, patients with active disease may also have closed granulomas with hard, central caseum, and fibrotic and calcified lesions with lower bacterial burdens, which are usually associated with latent TB [9].

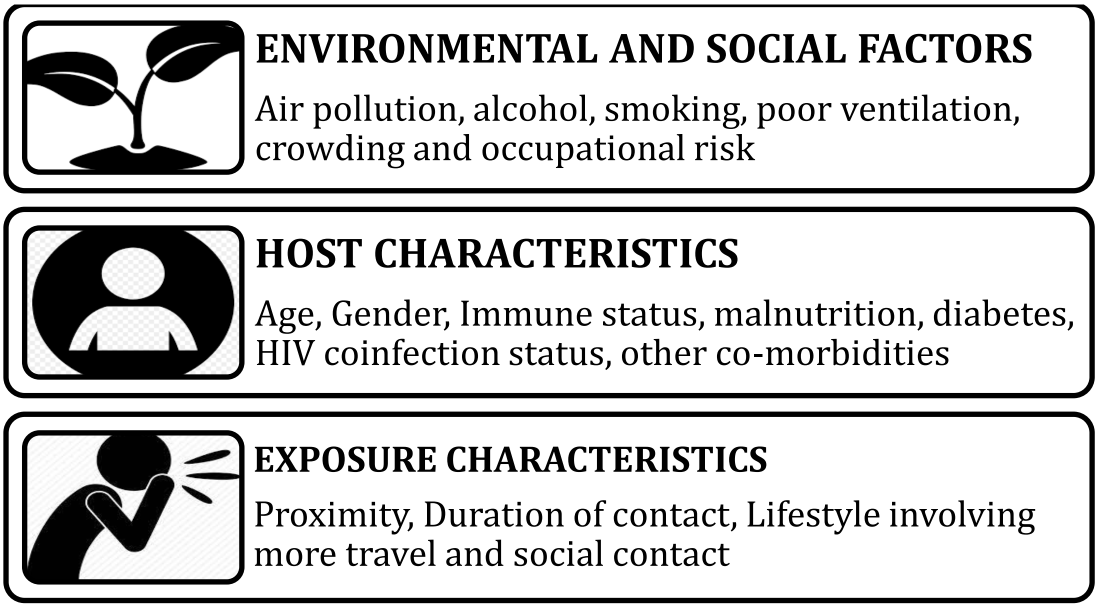
The above-mentioned heterogeneity in pathogenesis of Tuberculosis and the kind of granulomas/lesions formed has been attributed to the differences in the host immune responses and the strain of *M. tuberculosis* the patient has been infected with. Whole genome sequencing analyses of Mtb strains have divided the strains into 7 lineages, each identified from different regions across the globe and causing infections with different characteristics. Infection with some strains from these lineages (Lineage 2, 3) may cause extra-pulmonary tuberculosis, while others (Lineage 1) may show higher inflammatory response or delayed progression from latent to active disease (Lineages 6 and 7). Newer strains of Mtb have also shown to elicit altered immune responses from previous strains, leading to increased bacterial burdens and lung pathology, making them more difficult to treat. Indeed, these differences have led to experts proposing that the Mtb pathogen has evolved alongside humans to have increased rates of transmission and higher speeds of disease progression [8].

Therefore, the progression of TB infection from latent to active disease is not a straightforward as there are a variety of infection sites and a diverse milieu of exptrapulmonary infections that can occur. Most often, infected asymptomatic individuals harbouring subclinical disease show no clinical and pathological symptoms, but have an abnormal number of lung lesions that are visible in PET scans. The mechanisms that Mtb uses to adapt to the surrounding microenvironment, the presence of stresses (for e.g. oxygen and nutrient limitation) also influences the cellular and metabolic processes in the bacilli as well as the outcomes in patients with incipient and subclinical TB disease [8].

Despite this, the presence of lesions exhibiting minimal bactericidal effect after adaptive immunity has kicked in, the evidence of bacterial dissemination and increased inflammation in the lung lesions are characteristic of progression to active disease. Thus, research is now geared towards targeting the bacilli inside the granulomas as it could help prevent progression.

2.3. Risk factors in Tuberculosis

The progression from TB exposure to active disease, which may require long-term therapy, depends on several exogenous and endogenous factors. The exogenous factors determine how high the bacterial load will be in the sputum after exposure to the pathogen through a contact or otherwise. Endogenous factors, on the other hand, are usually host-related and they determine whether the infected individual will ever develop active disease or never show any clinical symptoms. Some of these risk factors have been very well established, while others are being studied [6]. The figure below lists some important risk factors that may influence if exposure leads to infection and consequently disease.



ENVIRONMENTAL AND SOCIAL FACTORS

Air pollution, alcohol, smoking, poor ventilation, crowding and occupational risk

HOST CHARACTERISTICS

Age, Gender, Immune status, malnutrition, diabetes, HIV coinfection status, other co-morbidities

EXPOSURE CHARACTERISTICS

Proximity, Duration of contact, Lifestyle involving more travel and social contact

Figure 4: Important risk factors for Tuberculosis. Figure created from the risk factors described from *Narsimhan et al. (2013)* [6]

2.4. Transcriptomic Biomarkers and their importance in understanding tuberculosis

The immune response to M. tuberculosis is complex and incompletely characterized, hindering development of new diagnostics, therapies and vaccines. Berry et al. (2010) [10] identified a whole blood 393 transcript signature for active TB in intermediate and high burden settings, correlating with radiological extent of disease and reverting to that of healthy controls following treatment. Modular and pathway analysis revealed that the TB signature was dominated by a neutrophil-driven interferon (IFN)-inducible gene profile, consisting of both IFN-γ and Type I IFNαβ signalling.

2.5. TB incidence in Men versus Women: Why the difference?

Gender-specific Tuberculosis prevalence was first discussed in a meta-analysis of 29 prevalence surveys in 14 countries published in the year 2000. The study showed that male/female ratios in almost regions and age-groups were not equal to 1, highlighting that Tuberculosis may impact one gender more than the other. However, this study was severely limited due to the limited data available at the time as well as the reduced rates of TB notification, especially in Asian countries [11].

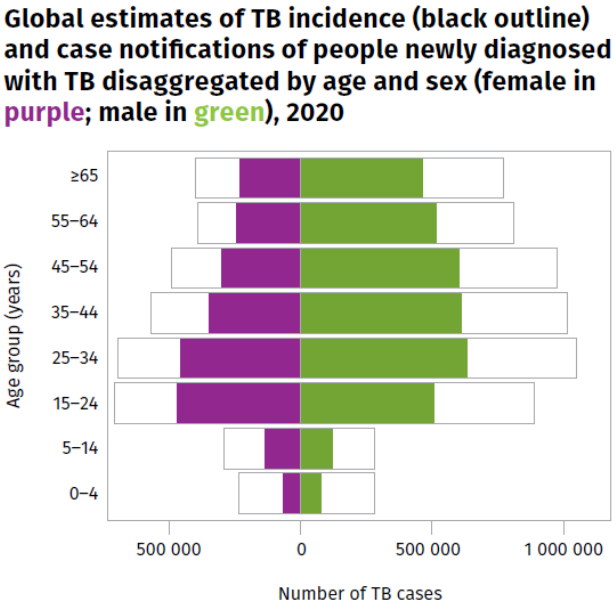


Figure 5: Difference in TB incidence in men vs. women. Figure taken from Global WHO TB dashboard [2].

A similar systematic review and meta-analysis using the data from Low and Middle Income Countries (LMICS), amounting to a total of more than 80 surveys (from 1992 to 106) was done by *Horton et al.* in 2016 [12]. The surveys included in this review looked at TB prevalence using either bacteriological TB testing or smear microscopy. TB prevalence (after accounting for random effects) per 100,000 individuals was found to be 488 (95% CI) among men and 231 (95% CI) among women for bacteriologically positive TB and 314 (95% CI) among men and 129 (95% CI) among women for smear-positive TB. The male-to-female prevalence ratios, after being adjusted for random effects, were also found to be greater than 1 for all surveys except for a few surveys in the Americas. Owing to the results, the authors also called for more focus on TB treatment, prevention and control strategies focused on men.

Several similar studies have also pointed to the fact that the difference in incidence is not just a result of underreported cases in women, or gender-specific disparities in access to healthcare, but the surveys are in fact evidence of how TB affects men more than it affects women. A study by *Neyrolles et al.* in 2009 [13] discussed this sexual inequality in TB incidence, looking at why the notified TB cases in men were almost twice that of men. The study discussed that even though access to healthcare facilities is lower women in certain LMICS (Low and Middle Income Countries), there may be certain biological significance behind the almost constant 2:1 incidence ratio in a majority of countries across the world. However, there are a lot of studies arguing that confounding factors such as access to healthcare and gender stereotypes are the reason that more than 70% of those with active TB are male. But, more evidence leads us to believe that the 2:1 men to women patient ratio may be due to biological differences in between the sexes, including and not limited to sex steroids, antimycobacterial immune response, differences in genetic architecture and nutrition. But, studies with appropriate control groups and defined patient groups need to be undertaken to highlight the difference.

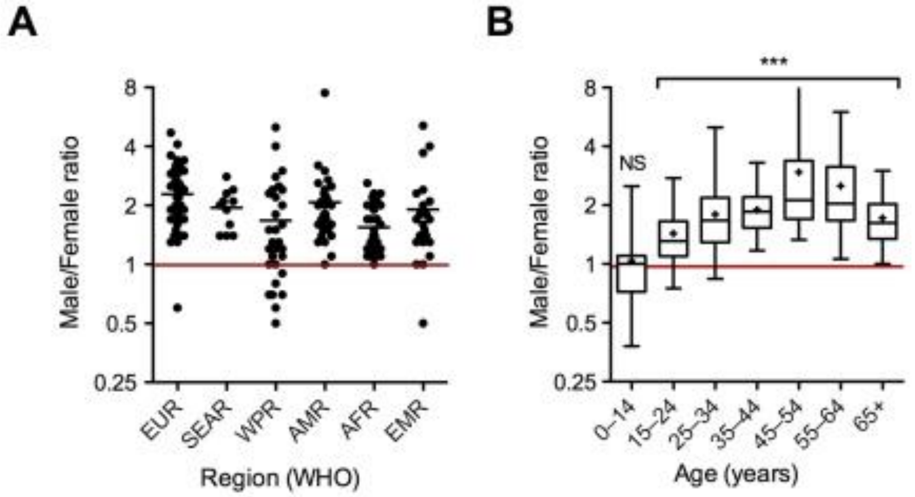


Figure 6: Regional (A) and age-wise (B) distribution of the male/female ratio for new smear-positive TB cases in 2007. Figure from *Neyrolles et al. (2009)* [13]

Several studies have shown that testosterone increases susceptibility to Mycobacterium species in mice compared to female mice [14]. There is a possibility that differences in the genetic architecture as well as anatomic, physiological differences, such as metabolic processes, may play a role in humans too.

Another study by *Lin et al.* [15] analysed Taiwanese National Health Data (data included from national reports, interviews and surveys from 43 424 subjects) and found that there is 2.3-fold higher risk of active tuberculosis in men than women, after the data was adjusted for alcohol abuse, smoking, comorbidities and other confounders.

A study by *Nhamoyebonde et al.* (2014) [16] suggests that these differences in the prevalence of TB in males vs. females are a result of either the behavioural/lifestyle differences which consequently impact the exposure and susceptibility to TB, or physiological and biological differences due to gonads, sex steroids and gender-related heterogeneity in immune responses. Both of these mutually exclusive hypotheses are discussed one-by-one.

- **Behaviour and Lifestyle:** Behavioural factors that increase the likelihood of exposure to TB, such as the amount of travel, social contact and having professions which have a higher risk, are considered to be risk-factors that impact men more than women. Lifestyle factors, such as smoking and alcohol consumptions also influence the susceptibility to the pathogen and progression of tuberculosis infections [16]. A 2006 paper explains that smoking explains only one-third of the difference in incidence.

- **Differences in access to diagnosis, treatment and consequent outcomes due to gender:** Authors in the Taiwanese study cited above suggest that "*future tuberculosis control programmes should particularly target the male population*" [15]. And although TB incidence in men is almost twice than in women, a gender inequality gap still exists in terms of access to treatment and social stigma. Reports indicate that women with TB, especially those in the Low and Middle Income Countries, where the incidence of TB is much higher, face harsh criticisms and social stigma. Indian women with TB may also be ostracized and divorced after their TB diagnosis. Addressing gender inequality in TB diagnosis and treatment remains a challenge to tackle [17].

CHAPTER 3: METHODS

3.1. Summary

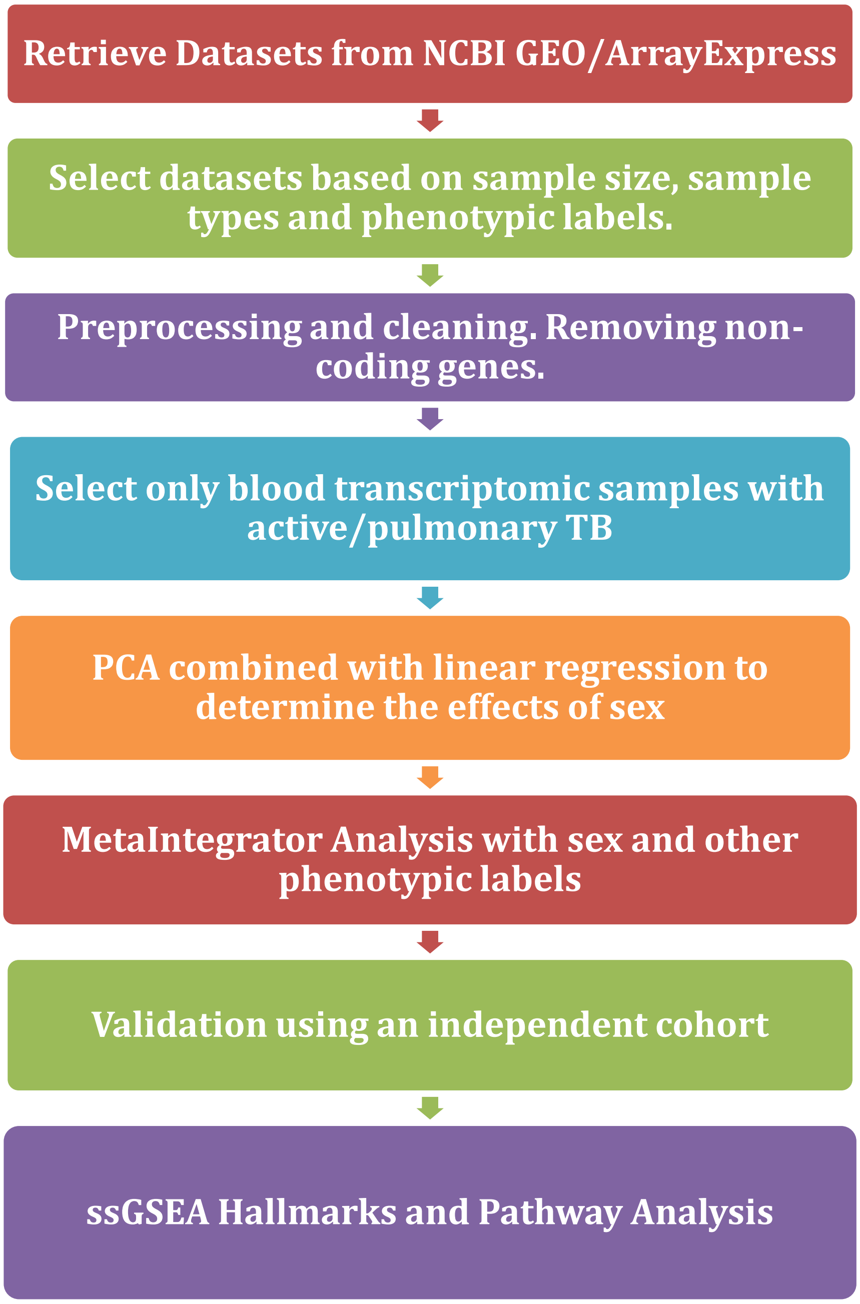

Retrieve Datasets from NCBI GEO/ArrayExpress
Select datasets based on sample size, sample types and phenotypic labels.
Preprocessing and cleaning. Removing non-coding genes.
Select only blood transcriptomic samples with active/pulmonary TB
PCA combined with linear regression to determine the effects of sex
MetaIntegrator Analysis with sex and other phenotypic labels
Validation using an independent cohort
ssGSEA Hallmarks and Pathway Analysis

Figure 7: A brief summary of the methodology used for the analysis.

3.2. WHO data TB incidence in 2021

To determine whether the current incidence of TB in men versus women is similar still follows the trend the studies cited above discuss, TB case data was retrieved from the WHO database, sorted according to regions and the Male: Female ratios of incidence calculated and plotted. The number of cases in countries with the highest incidence of Tuberculosis was also plotted along with a pie-chart distribution of the area-wise case burden of TB. Figure 11 shows the Male/Female TB incidence ratio in different areas of the world as reported in Global Tuberculosis Report, 2021 [2]. From the data available, the best case estimate was utilized for plotting and the plot generated using R studio and the ggplot2, scales, viridis, and ggrepel libraries [18]–[21].

3.3. Dataset evaluation and selection

Transcriptomic/Gene expression data from TB patients was retrieved from the publicly available gene expression data repositories, namely the NCBI GEO (https://www.ncbi.nlm.nih.gov/geo/) and the ArrayExpress from EMBL-EBI (ebi.ac.uk/arrayexpress/). A search of the term, "*Mycobacterium tuberculosis*" on NCBI GEO yielded 11,463 results (on April 25, 2022), upon which the following filters were applied:

- **Study Type:** 'Gene expression profiling by array' and 'Gene expression profiling by high throughput sequencing'.
- **Organism:** *Homo sapiens*
- **Sample Type:** PBMC, Whole blood
- **Sample Size:** >10

The results were refined to get transcriptomic datasets that fit the objectives of the study.

Table 1: A brief description of all datasets evaluated for inclusion in the study. However, only three datasets fulfilled the selection criteria and were used for the analysis.

Accession	Description	Link	Included/Excluded	Reference
GSE19491	Active and Latent tuberculosis patients as well as healthy controls were recruited from several hospitals and centres in the UK and South Africa. RNA extrac-	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=G	**Included**	[10]

	tion and microarray profiling was performed using whole blood samples from these patients and the results analysed to propose gene signatures for TB.	SE19491		
GSE25534	A whole-blood microarray gene expression analysis of a South African cohort of TB patients, including latently as well as uninfected healthy controls, to define biomarkers predictive of susceptibility and resistance.	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25534	Excluded due to the lack of age, gender and other phenotypic labels for each sample	[22]
GSE28623	A Gambian cohort of TB patients (both latent and active) and healthy controls being analysed for pathway and functional association analysis	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28623	**Included.**	[23], [24]
GSE31348	A longitudinal cohort comprising of ex vivo blood samples analysed for 27 first episode pulmonary TB patients, at diagnosis and after 1, 2, 4 and 26 weeks of treatment.	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31348	Excluded for the lack of age, gender and ethnicity labels	[25]
GSE37250	Adults recruited from South Africa and Malawi for genome-wide transcription profiling of Latent and Active Tuberculosis patients	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE37250	Excluded due to missing gender labels.	[26]
GSE79362	A South-African and Gambian cohort of TB contacts, including both progressors and non-	https://www.ncbi.nlm.nih.gov/ge	Excluded since the data available does not have the details	[27]

	progressors.	o/query/acc.cgi?acc=GSE79362	of the stage of TB	
GSE101705	A South Indian cohort of malnourished Tuberculosis patients, having both active and latent disease.	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE101705	Excluded due to lack of gender labels for each sample type.	[28]
GSE157657	A longitudinal RNA-sequencing dataset derived from TB contacts and TB patients recruited at Leicester, UK	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE157657	**Included**	[29]

Several datasets listed above and screened were not included due to the absence of gender labels for the samples. Only three datasets (GSE19491, GSE28623, GSE157657), had sufficient sample sizes, adequate gender labels for all samples and so they were selected for further processing and analysis [10], [23], [24], [29].

3.4. Pre-processing of Datasets

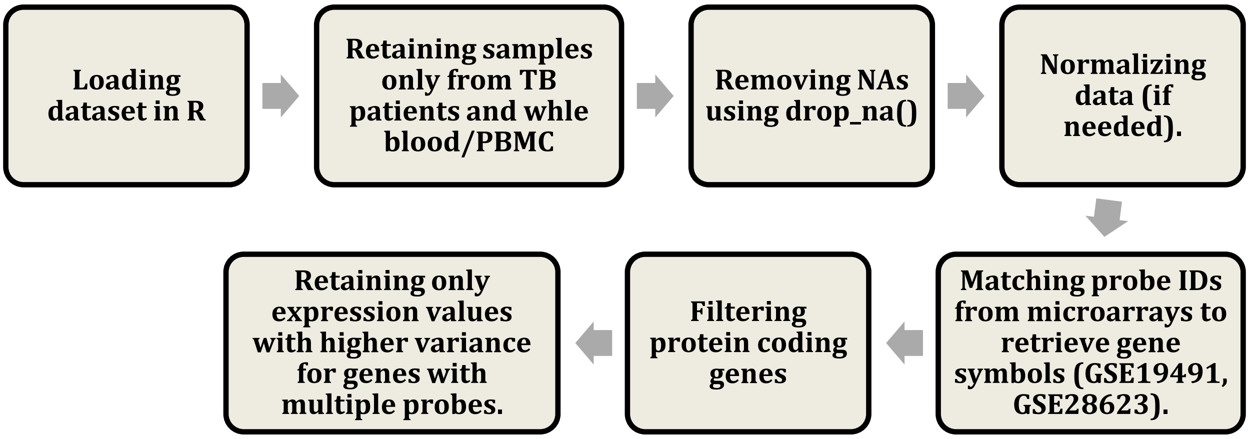

Loading dataset in R
Retaining samples only from TB patients and whle blood/PBMC
Removing NAs using drop_na()
Normalizing data (if needed).
Matching probe IDs from microarrays to retrieve gene symbols (GSE19491, GSE28623).
Filtering protein coding genes
Retaining only expression values with higher variance for genes with multiple probes.

Figure 8: A summary of the pre-processing done on each of the datasets. A detailed description of the methods for each dataset is provided in the sections that follow.

3.4.1. Dataset normalization, cleaning and filtering TB samples

The dataset GSE19491 had intensities corresponding to gene expression of the samples on the Illumina BeadStation 500 and the original study used the Bead Studio v2 by Illumina to obtain normalized values by scaling intensities across chips [10]. However, the normalized values were not available in the GEO datasets repository, therefore all the intensity were scaled to sample-wise means using the scale() function in base R [30]. This was performed after removing any genes containing *NA*'s for expression values with the help of the drop_na() function in *tidyr* [31]. Probe IDs were matched for the GPL platform data file (GPL6947-13512) and the expression dataset available in GEO. For probes with gene symbols, gene symbols replaced the Probe IDs, while the probes which didn't correspond to any gene (no available information) were removed. Since the study was aimed at understanding how mechanisms of tuberculosis response differ in men and women from the point of view of transcriptomic signatures, the samples in GSE19491 containing gene expression values from isolated monocyte/lymphocytes (from TB patients) were removed and only whole blood samples were kept.

For GSE28623, a microarray study, the expression values available in the series matrix were downloaded and used. After the dataset was loaded, drop_na() from *tidyr* was used to remove genes/probes with *NA*s [31]. Expression values were then log2 normalized by using $log2\ (x + a)$ where x deontes the expression values and a is an arbitrary constant higher than the minimum expression value. For this analysis, the value of $a = 10.6$ was taken.

For GSE157657, cDNA libraries were synthesized from blood RNA samples and sequenced using the Illumina HiSeq 4000 platform. The sequences were aligned and mapped to the 'Ensembl GRCh38 (release 95)' build of the human gene to obtain gene-level count data, which was further normalized using the R package DESeq2 by *Tabone et al.* in [29]. The normalized log2 expression values were available as a supplementary file in the NCBI GEO repository and utilised for analysis.

The above-described cleaning and normalization was performed on each of the datasets after selecting only the samples that belong to patients with active TB (in GSE157657, Active TB were labelled as 'PTB' in the series matrix file obtained from GEO). For cohorts mapping longitudinal effects as a result of Tuberculosis patients (GSE157657 and GSE19491), samples from patients who had undergone <=60 days of treatment of tuberculosis. After normalization, the protein-coding genes were filtered (by mapping genes to the list provided in *Piovesan et al. (2019)* [32]) and duplicate genes removed (in case of multiple probes/IDs for a single gene, the expression values having higher variance were retained and the rest removed).

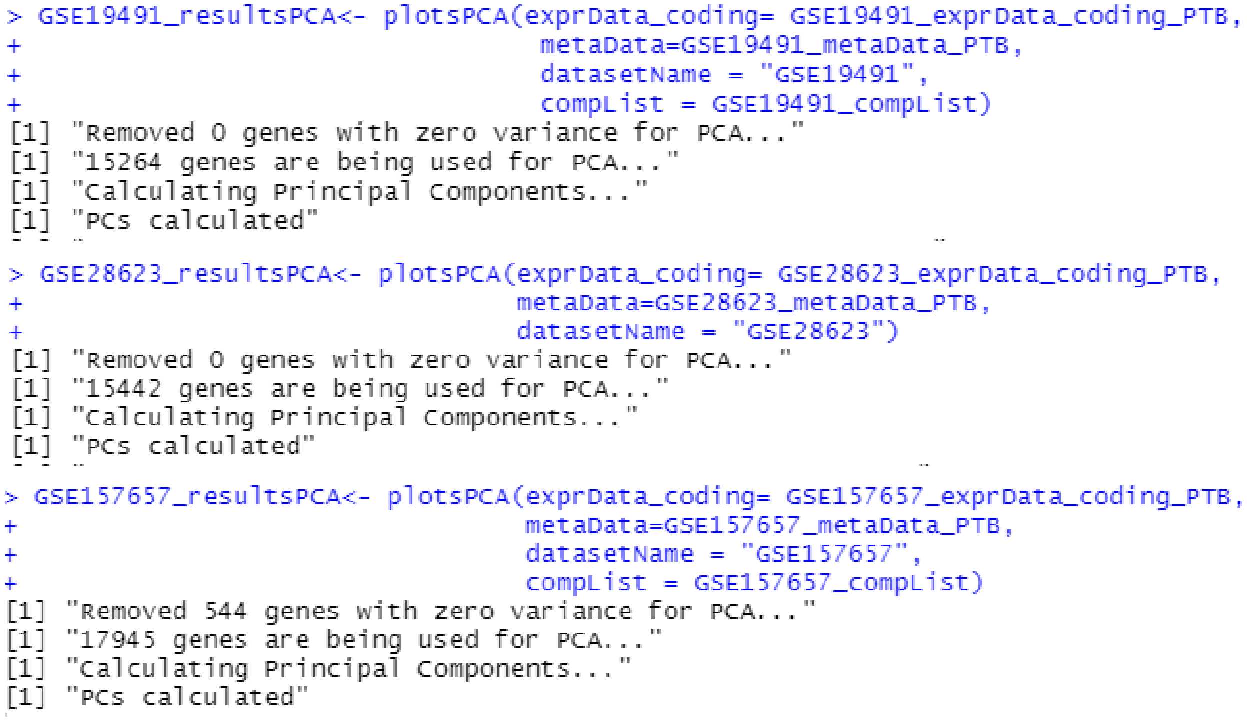
3.4.2. Phenotype labels retained for analysis

Different labels of each sample were retrieved from the series matrix files and input for meta-analysis. Sex and group (indicating the subset of disease) labels were selected for all datasets and for GSE19491, age and ethnicity were included. For GSE157657, in addition to age, ethnicity, sex and group, subgroup, and days_att (indicating days from when the Anti-tuberculosis therapy began) were included.

3.5. Principal Component Analysis and Linear regression

3.5.1. PCA biplots

Next, genes with zero variance were removed to perform Principal Component Analysis (PCA) on the expression values. PCA was then used for dimension reduction of the gene expression datasets and the datasets visualized using biplots, with PC1 on the x-axis and PC2 on the y-axis and the sex/ethnicity labels for grouping samples. To calculate the principal components, the prcomp() function with the ‘centre’ and ‘scale.’ parameters set to TRUE was used and the ggplot2 library utilized to plot the biplots [18]. The proportion of variance explained by each principal component was also plotted for each dataset.



```
> GSE19491_resultsPCA<- plotsPCA(exprData_coding= GSE19491_exprData_coding_PTB,
+                                metaData=GSE19491_metaData_PTB,
+                                datasetName = "GSE19491",
+                                compList = GSE19491_compList)
[1] "Removed 0 genes with zero variance for PCA..."
[1] "15264 genes are being used for PCA..."
[1] "Calculating Principal Components..."
[1] "PCs calculated"
```

```
> GSE28623_resultsPCA<- plotsPCA(exprData_coding= GSE28623_exprData_coding_PTB,
+                               metaData=GSE28623_metaData_PTB,
+                               datasetName = "GSE28623")
[1] "Removed 0 genes with zero variance for PCA..."
[1] "15442 genes are being used for PCA..."
[1] "Calculating Principal Components..."
[1] "PCs calculated"
```

```
> GSE157657_resultsPCA<- plotsPCA(exprData_coding= GSE157657_exprData_coding_PTB,
+                                metaData=GSE157657_metaData_PTB,
+                                datasetName = "GSE157657",
+                                compList = GSE157657_compList)
[1] "Removed 544 genes with zero variance for PCA..."
[1] "17945 genes are being used for PCA..."
[1] "Calculating Principal Components..."
[1] "PCs calculated"
```

Figure 9: Code snippets for the PCA analysis being performed. The plotsPCA function was defined to plot all PCA related graphs and only 544 genes with zero variance were removed from GSE157657 for calculation of Principal Components.

3.5.2. Linear regression: dot plots and box plots

Linear regression modelling was used to model the relationship between the sex and ethnicity features variables and the first ten Principal Components (PCs) derived from the dataset. Since the principal components reduce the dimensions and capture the variation in a dataset, if the p-value of the linear model is <0.05 for any PCs, the correlation effects of the feature (sex or ethnicity) with the dataset were deemed significant. With the help of the ggpot2 library, boxplots and dotplots were also plotted, using the PCs and p-values from the linear regression, to visualize if there were any significant differences [18].

3.5.3. MetaIntegrator Analysis & Pathway enrichment

The MetaIntegrator package available from CRAN offers a pipeline that can be easily employed for the meta-analysis of several gene expression data cohorts. The multi-cohort analysis approach available in MetaIntegrator computes a Hedges g effect for each gene in the dataset and pools the value across datasets. To find the significant differentially expressed genes, the log sum of p-value of up-regulated and down-regulated genes is computed [33].

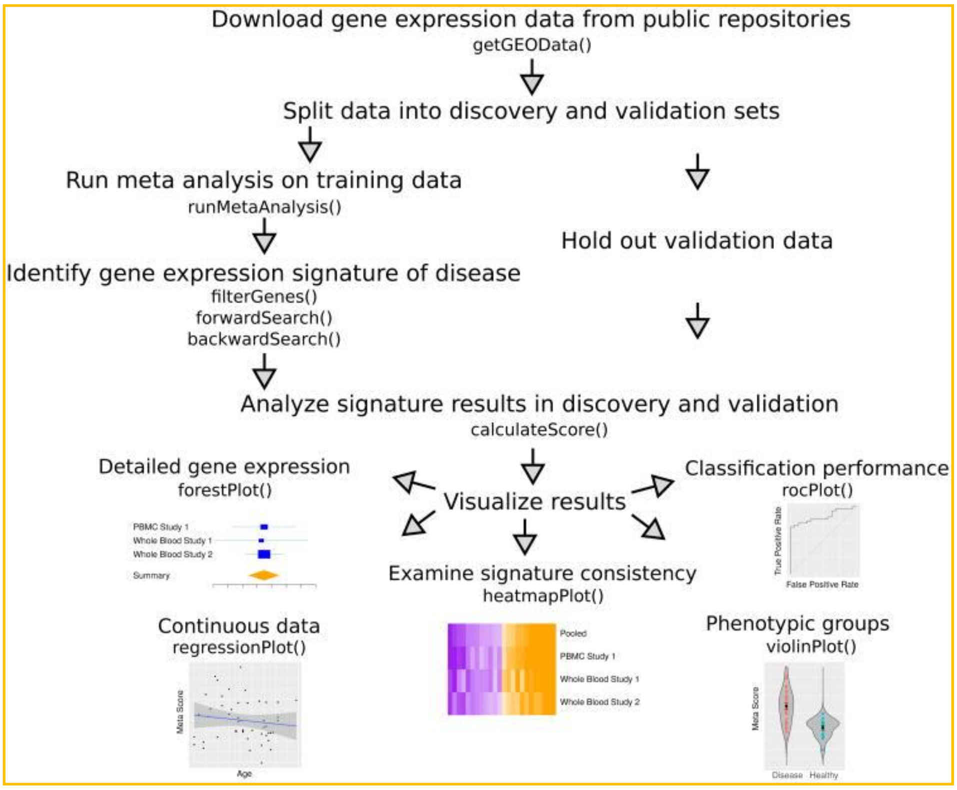


Figure 10: A summary of the pipeline provided by MetaIntegrator provides, from W. A. Haynes et al. (2017) [33].

For this study, we utilised three datasets, out of which GSE19491 and GSE157657 were used as discovery datasets and GSE28623 used for validation. Before running the multi-cohort analysis, classes were assigned to both the genders, with males being assigned '**0**' (or control) and females being assigned **'1'** (or having a disease status). Next, the runMetaAnalysis() function was used to calculate the meta-scores for the discovery datasets.

Table 2: Number of samples and genes used from each dataset for Meta-Analysis

Dataset Type	GEO Accession	Samples			Number of genes
		Males	Females	Total	
Discovery	GSE19491	43	24	67	15,264
	GSE157657	186	103	289	18,489
Validation	GSE28623	25	21	46	15,442

To analyse the up- and down-regulation of different important pathways in the selected datasets, single-sample Gene Set Enrichment Analysis was performed on gene expression datasets using the GSVA library and the hallmark pathway genes available in the Molecular Signature Database, MSigDB [34], [35]. A heatmap from the resulting values, clustered hierarchically (using Euclidean distance, average linkage method) according to the sex of the patient was created using Morpheus (https://software.broadinstitute.org/morpheus/), which is added to the results section.

The significantly up-regulated and down-regulated genes, resulting from meta-analysis, were used for functional analysis using the enrichment analysis tool, Enrichr (https://maayanlab.cloud/Enrichr/).

CHAPTER 4: RESULTS

4.1. WHO Global Incidence for TB, 2021

The **Male: Female Ratio** of TB incidence remains greater than 1 for a majority of countries even in 2020, indicating that the difference is not merely a result of under-reporting in women. The figure below shows the ratios according to different geographical areas.

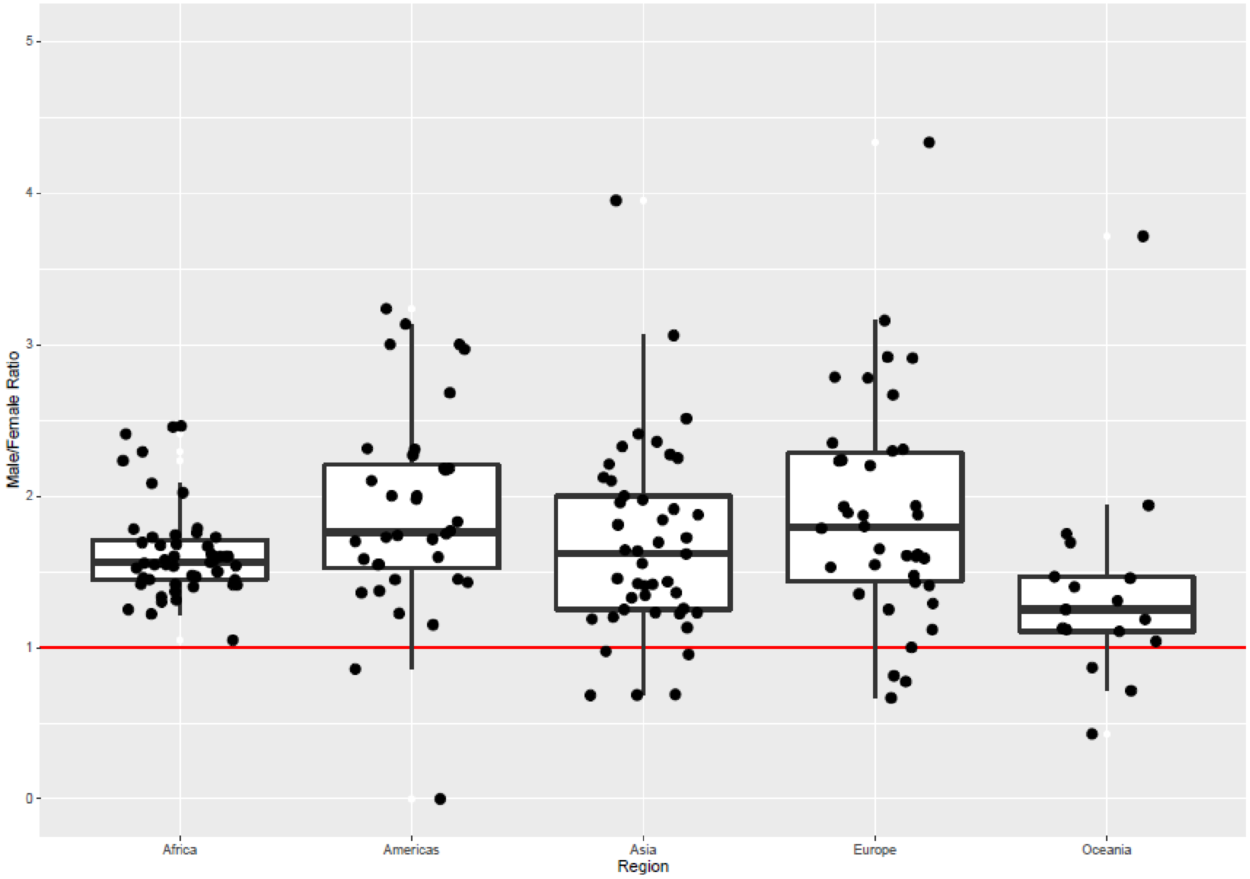
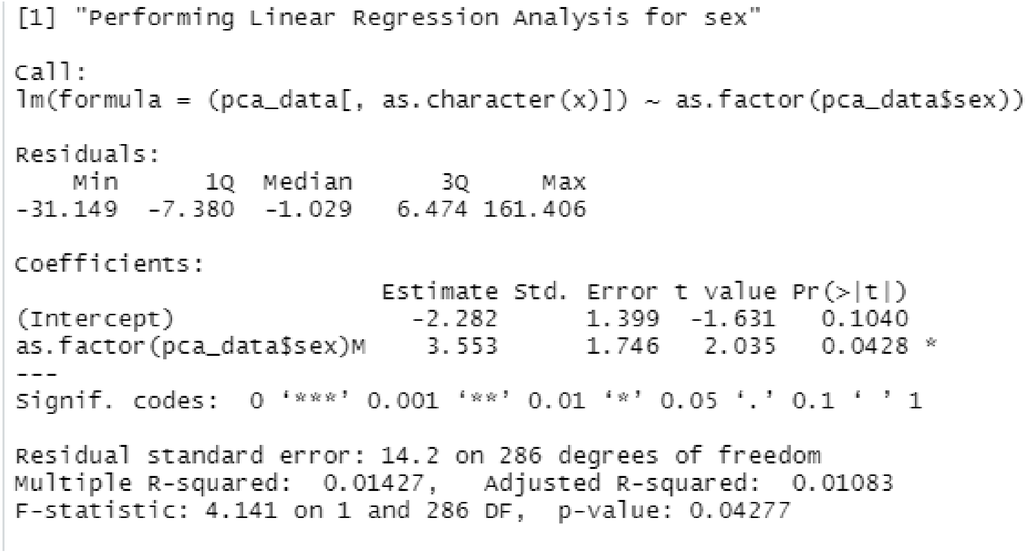


Figure 11: Male: Female Ratios of TB incidence, data from WHO Global TB Report (2021) [2].

4.2. Principal Component Analysis

The PCA biplots visually indicate that the cohorts may be separated on the basis of gender; however, they do not give enough information about the significance of gender in relation to expression. Therefore, linear regression was done and dotplots plotted from the p-values. These plots indicated that there are significant gender-specific differences, especially in GSE157657, but for GSE19491 and GSE28623, there are only one and two Principal Component had a significant value, with p-values <0.05, therefore there is some indication of significance. The figure below shows a function call for linear regression using the lm() function in R.



```
[1] "Performing Linear Regression Analysis for sex"

Call:
lm(formula = (pca_data[, as.character(x)]) ~ as.factor(pca_data$sex))

Residuals:
    Min      1Q  Median      3Q     Max 
-31.149  -7.380  -1.029   6.474 161.406 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)  
(Intercept)               -2.282      1.399  -1.631   0.1040  
as.factor(pca_data$sex)M   3.553      1.746   2.035   0.0428 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.2 on 286 degrees of freedom
Multiple R-squared:  0.01427,	Adjusted R-squared:  0.01083 
F-statistic: 4.141 on 1 and 286 DF,  p-value: 0.04277
```

Figure 12: Output from a linear modelling function call: here PC10 from GSE157657 was modelled on to the gender labels to check the significance of this relation.

The biplots and dotplots generated from the PCA analysis with linear regression are given below.

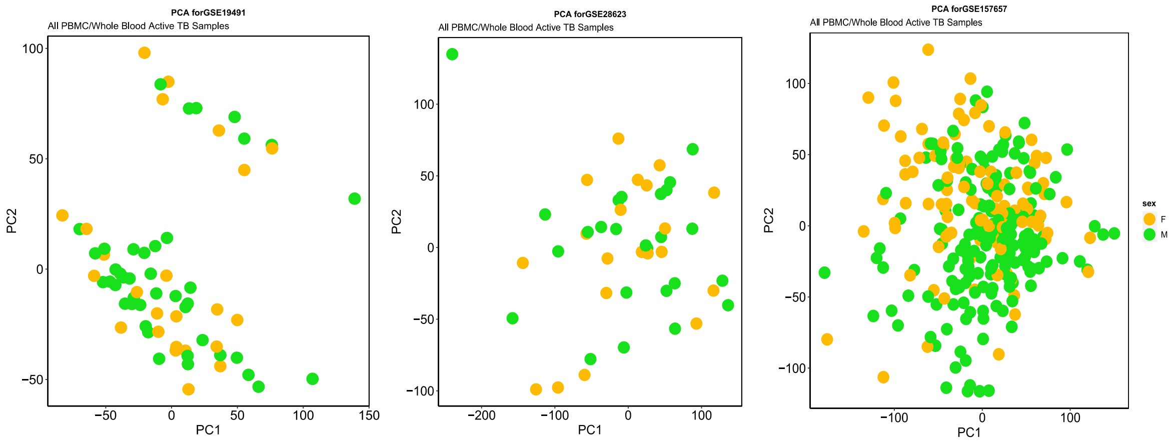


Figure 13: PCA biplots of datasets, labelled by gender: (from left to right), GSE19491, GSE28623, GSE157657.

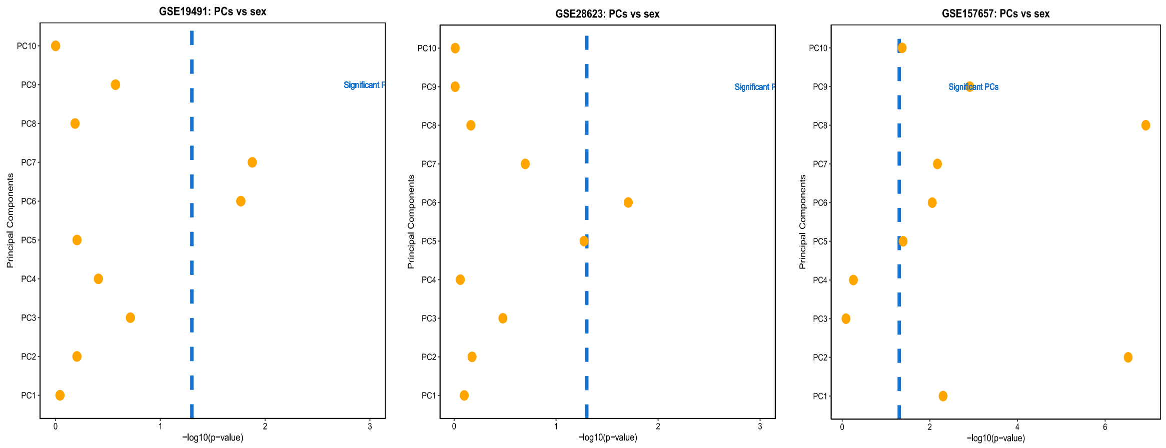


Figure 14: Dotplots with - log10(p-values) on the x-axis and PCs on the y-axis, the dashed horizontal line represents -log10(0.05).

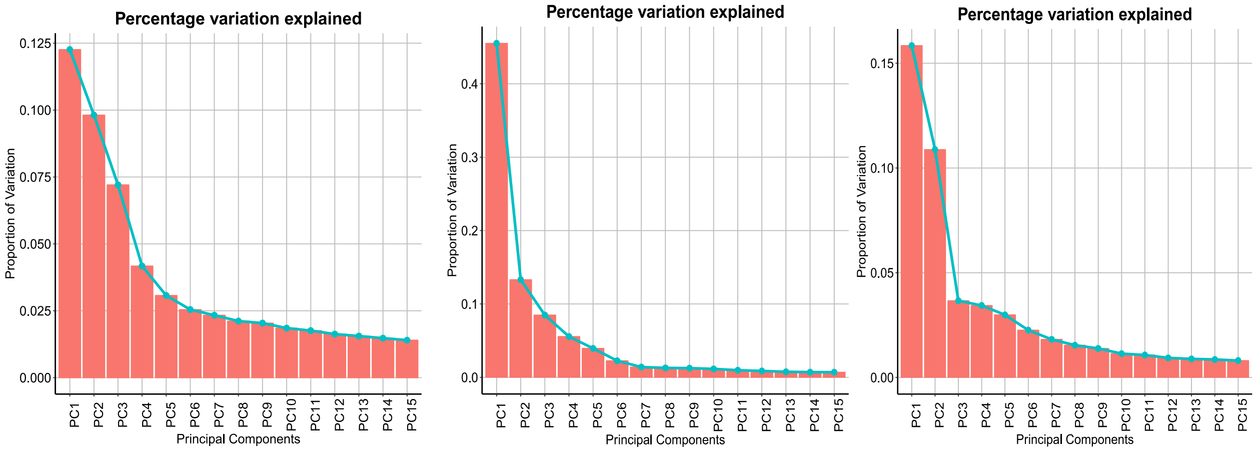


Figure 15: Proportion of variation that each PC explains for the datasets, (from left to right) GSE19491, GSE28623 and GSE157657.

Principal components were also plotted as boxplots to see if gender differences affect the variation in principal components. Below are the first ten principal components derived from GSE19491, GSE28623 and GSE157657; the red boxplots are values from samples of male patients and the blue boxplots have been plotted from female patients’ samples.

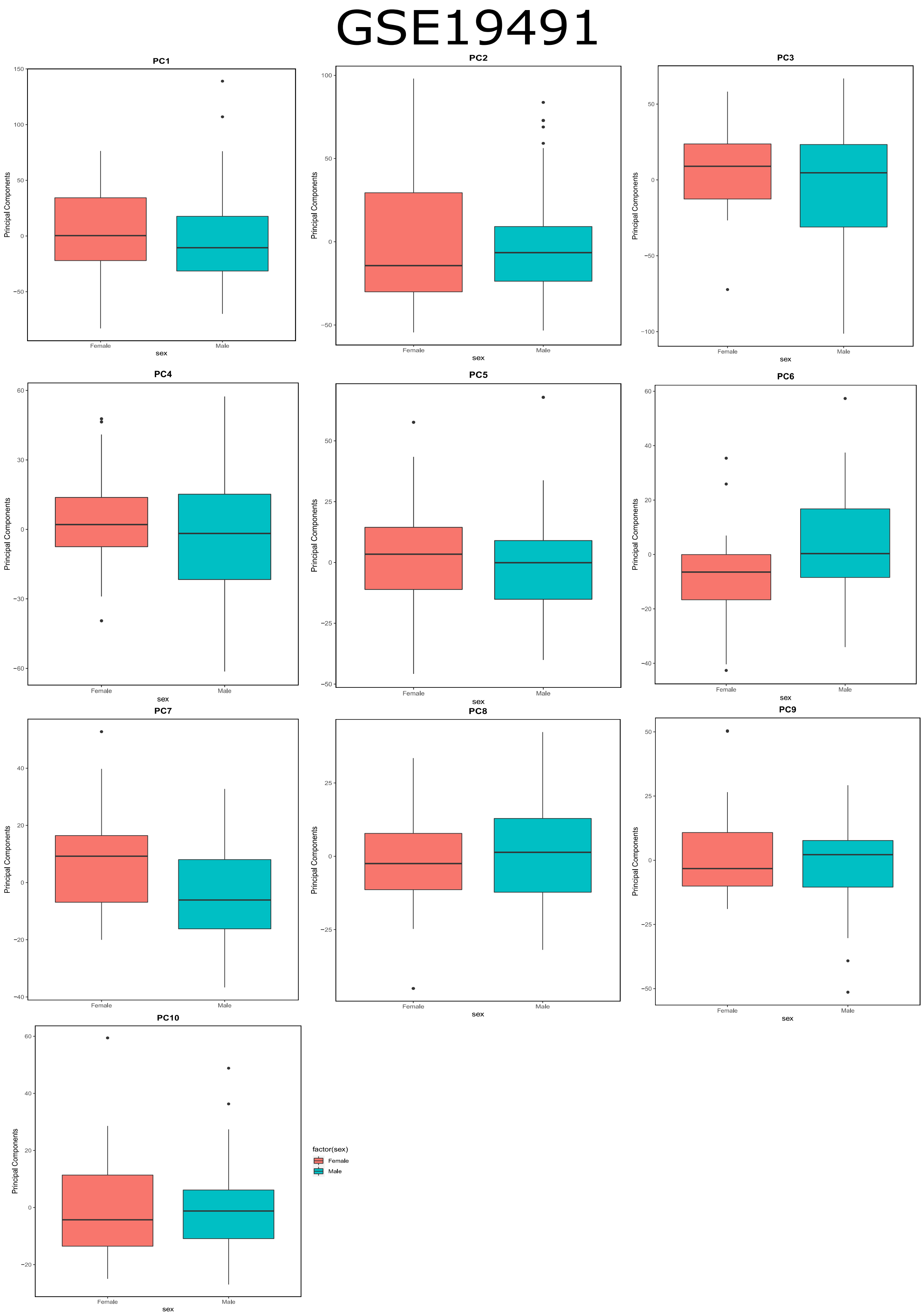


Figure 16: Individual PCs in GSE19491 and how their values differ with gender. The red boxplots indicate females and the blue boxplots indicate males.

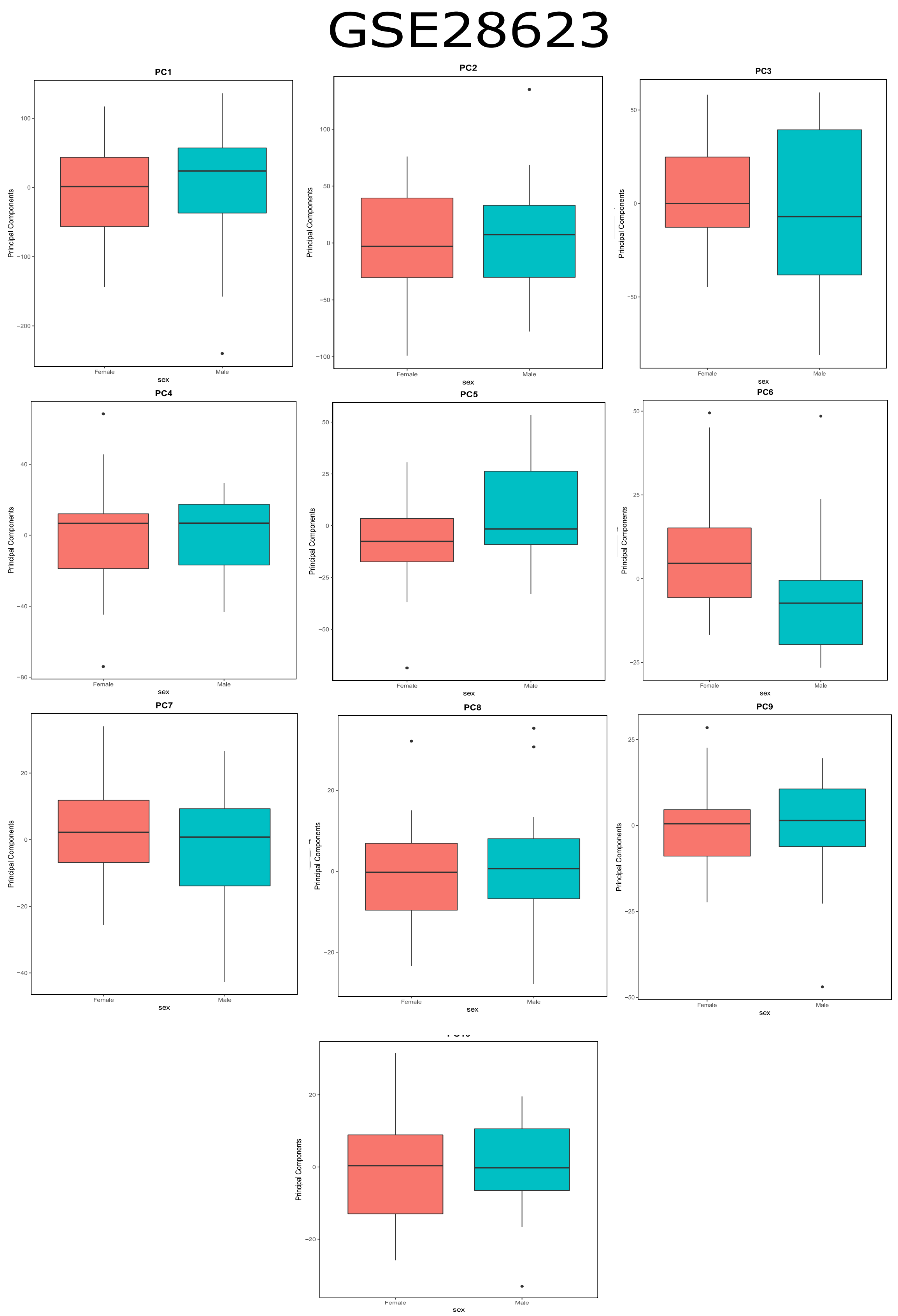


Figure 17: Individual PCs for GSE19491 and how their values differ with gender. The red boxplots indicate females and the blue boxplots indicate males.

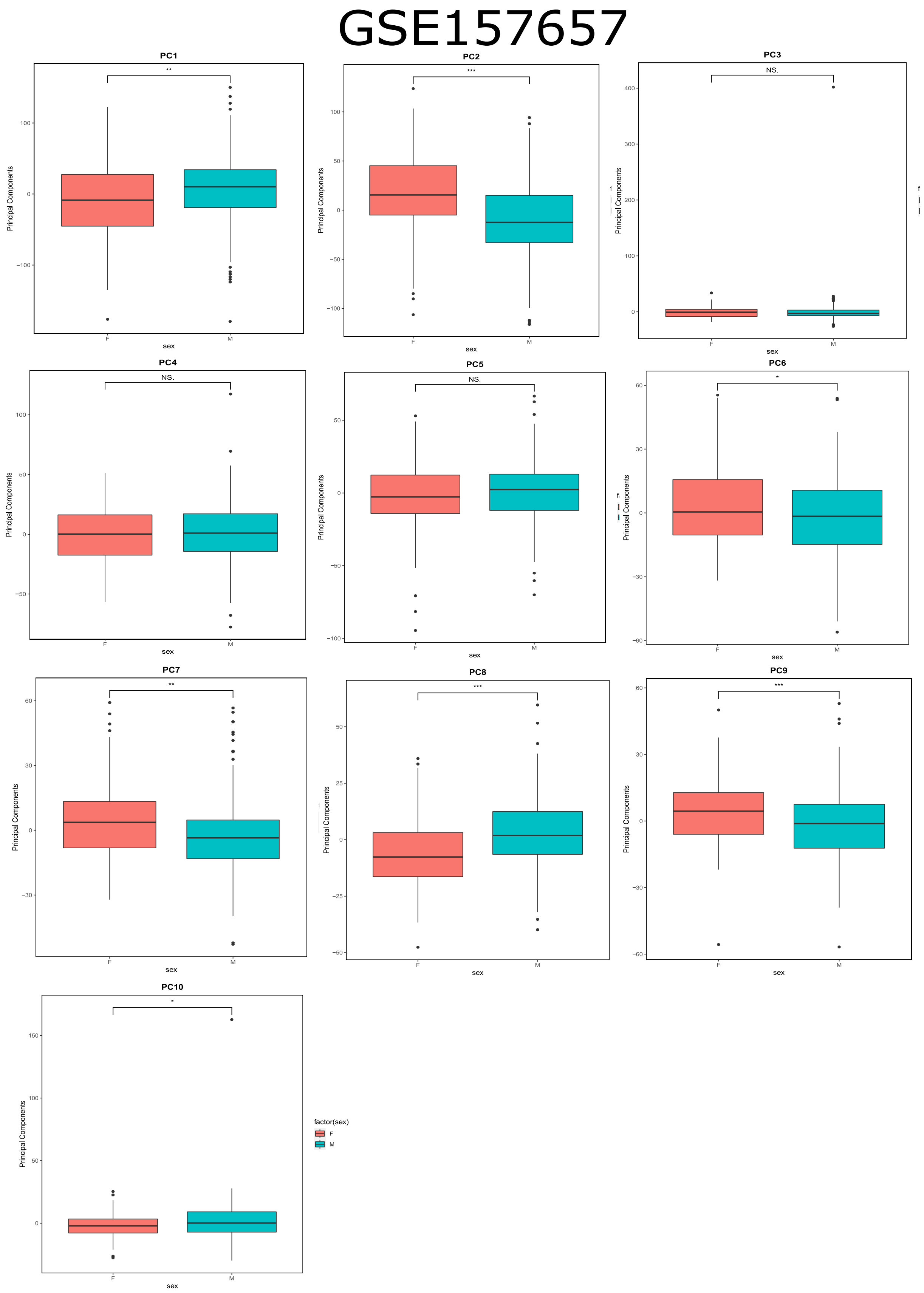


Figure 18: Individual PCs for GSE157657 and how their values differ with gender. The red boxplots indicate females and the blue boxplots indicate males.

Similar PCA analyses were done to find if there exist ethnicity-based differences in datasets which had ethnicity labels for their samples. The plots for GSE19491 and GSE157657 have been added below.

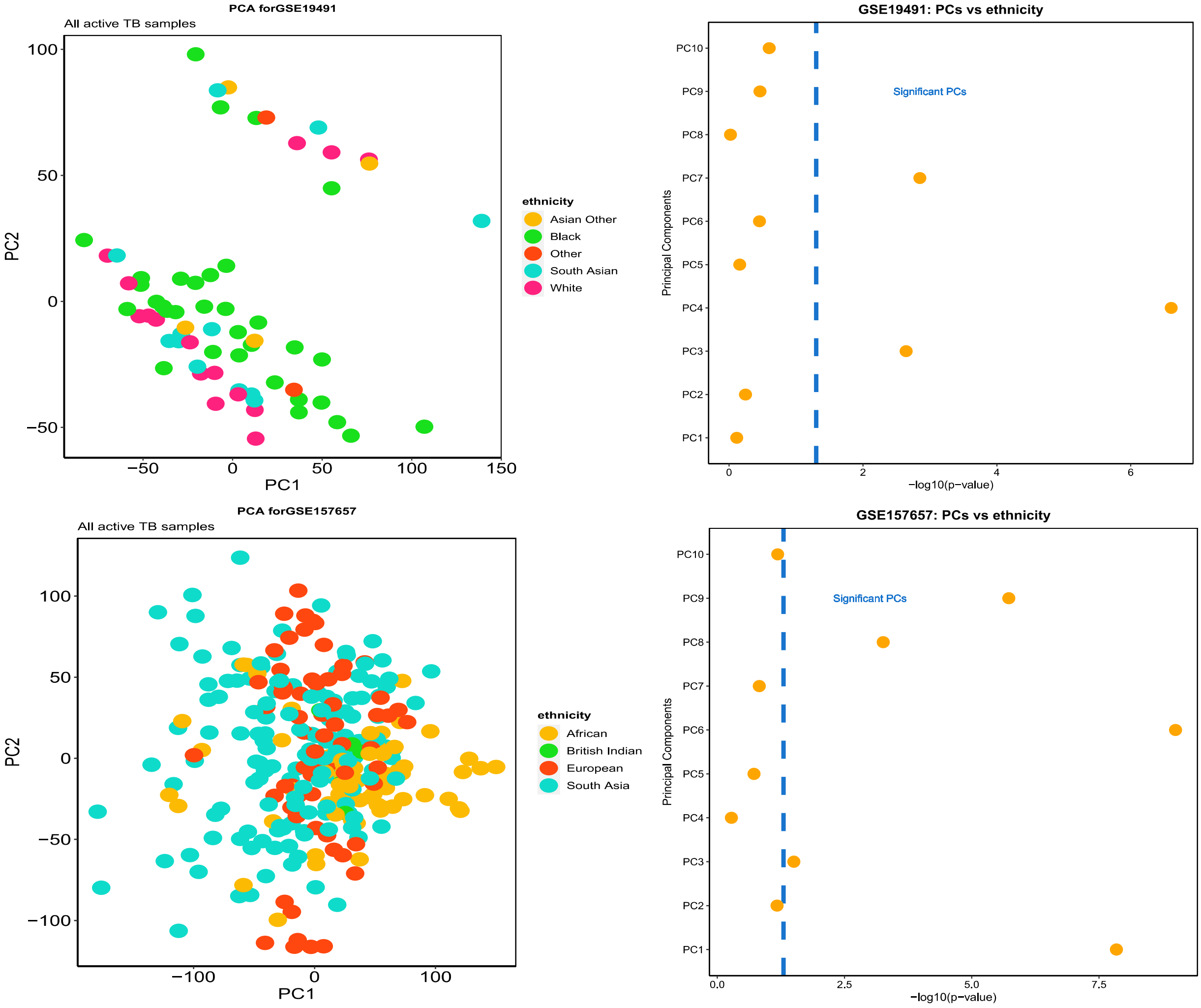


Figure 19: Ethnicity-based biplots and linear-regression dotplots for the samples in GSE19491 (top), GSE157657 (Bottom).

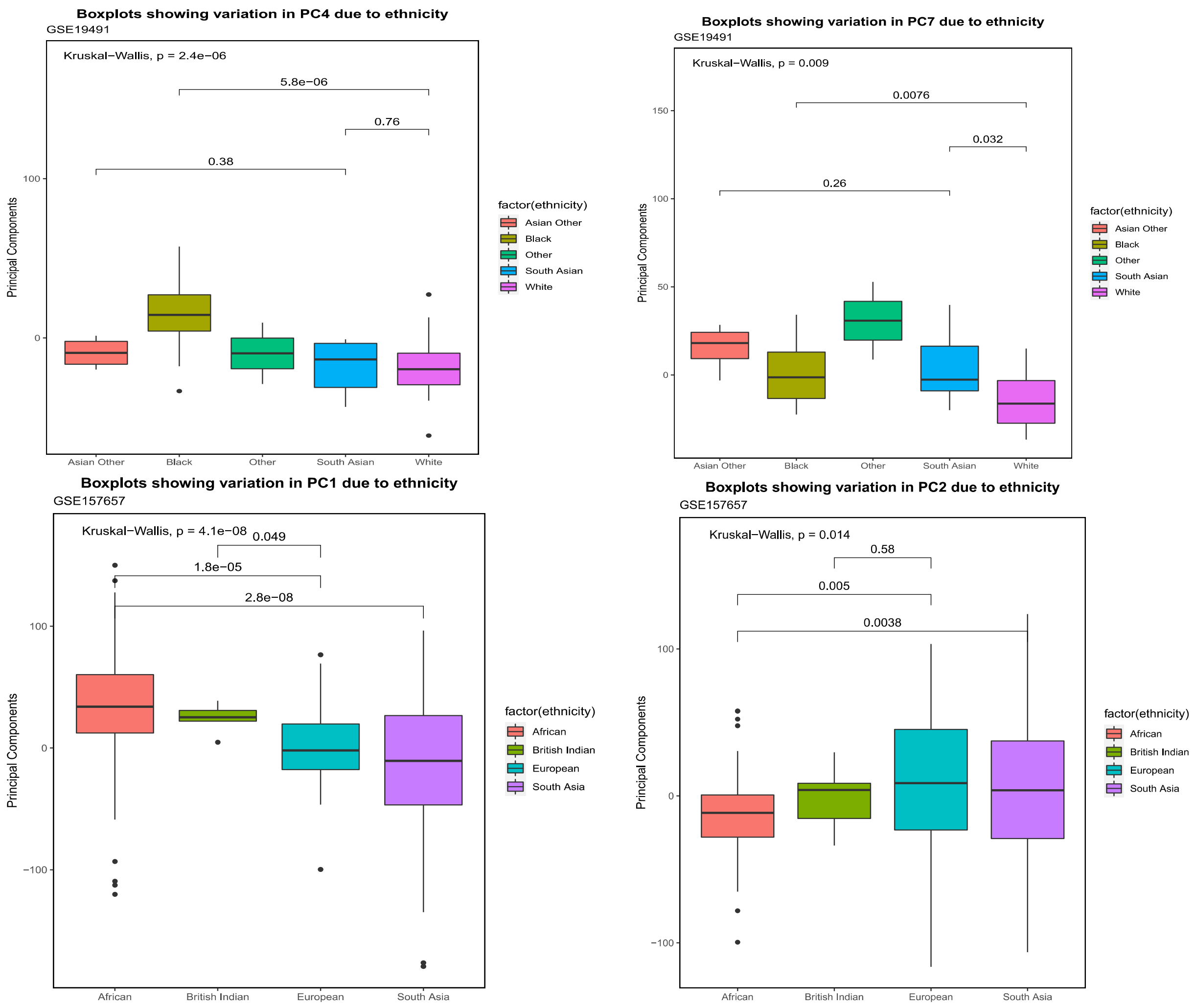
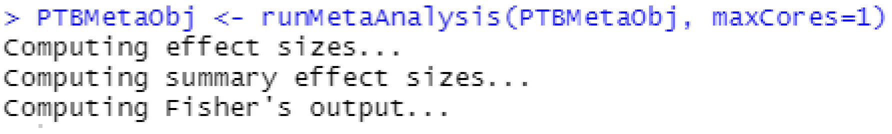


Figure 20: PCs from GSE19491 (top 2) and GSE157657 (bottoms that show significant differences between ethnicities.

4.3. MetaIntegrator Analysis

The MetaIntegrator run Analysis() function was used to find significant genes across cohorts and the significant genes were filtered using FDR values.



```
> PTBMetaObj <- runMetaAnalysis(PTBMetaObj, maxCores=1)
Computing effect sizes...
Computing summary effect sizes...
Computing Fisher's output...
```

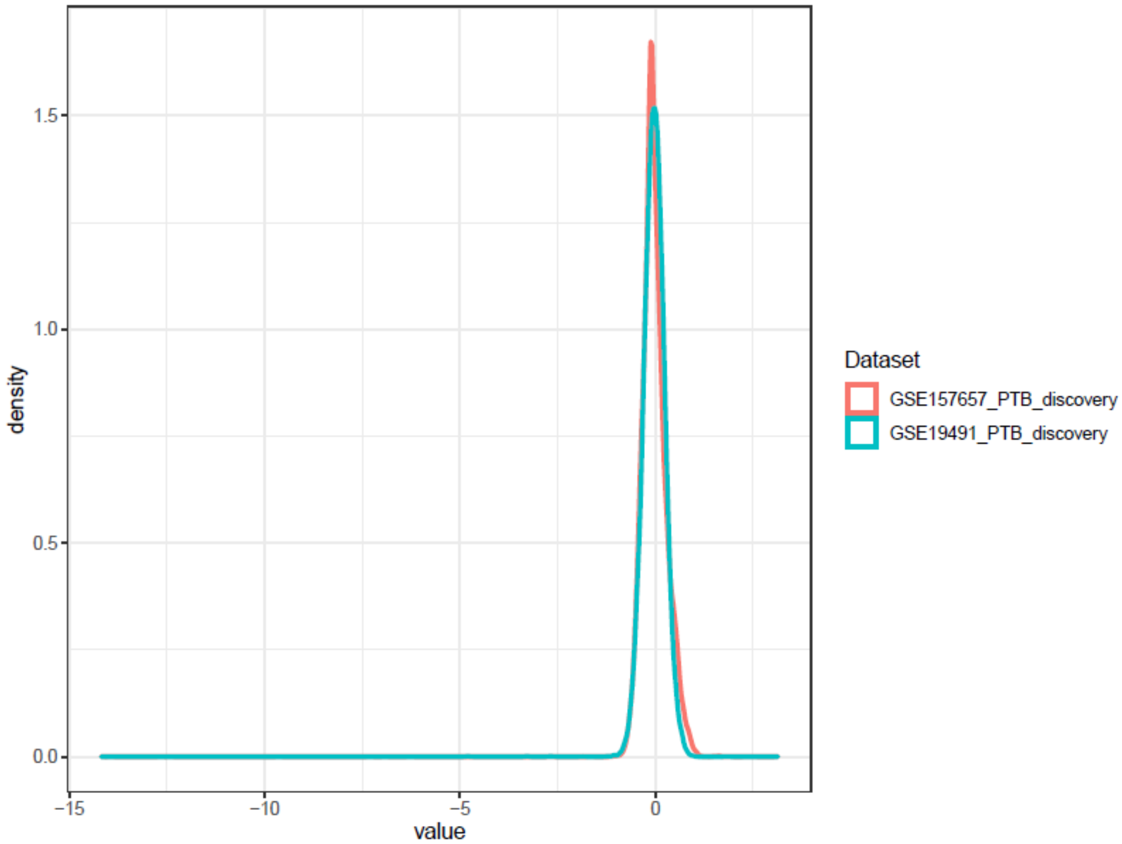


Figure 21: Output from the runmetaAnalysis() function, showing the density of the expression values from the discovery datasets.

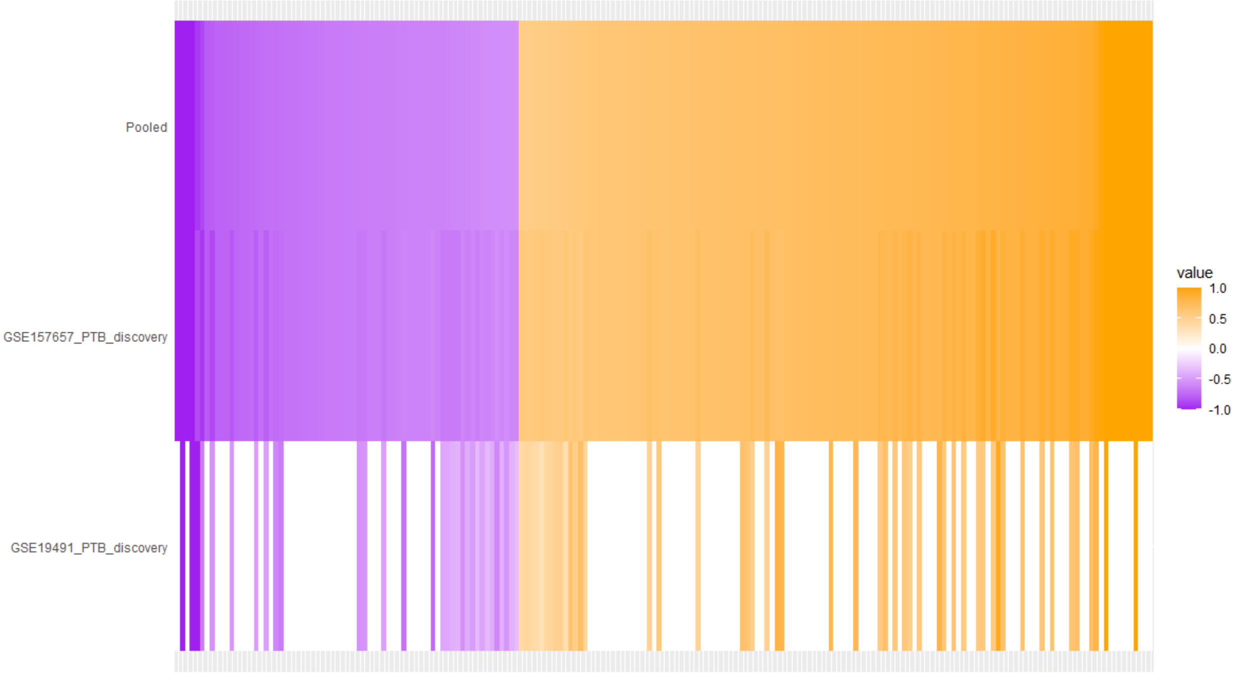


Figure 22: Heat map output from the Meta-analysis, showing the contribution of each dataset to the significant genes.

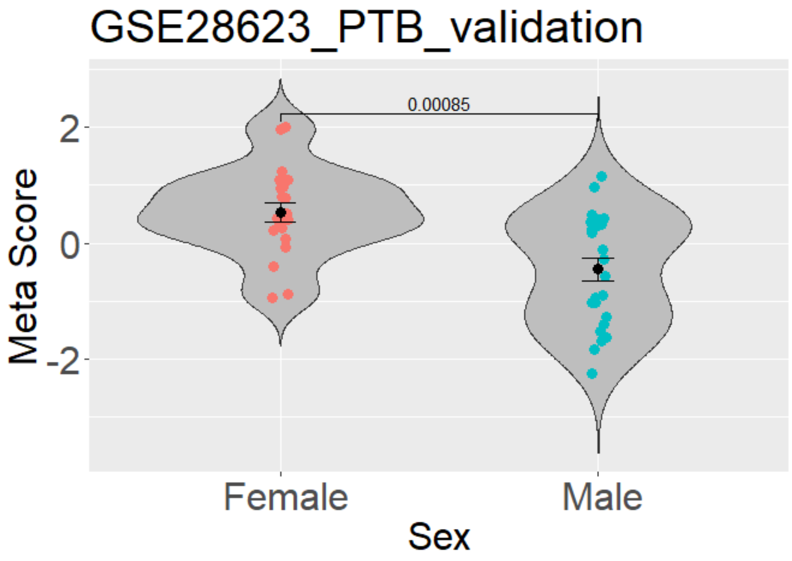


Figure 23: Metascores from the validation dataset, for FDR=0.0001

Genes with significant differential expression values were filtered from the meta-analysis results and the number of genes is given in the table below. These genes were used to plot ROC curves for the validation dataset, GSE28623.

Table 3 shows the number of genes with differential expression.

FDR	Number of positive significant genes (expressed higher in females as compared to males)	Number of negative significant genes (expressed higher in males as compared to females)
0.0001	129	70
0.001	231	181
0.05	738	1045

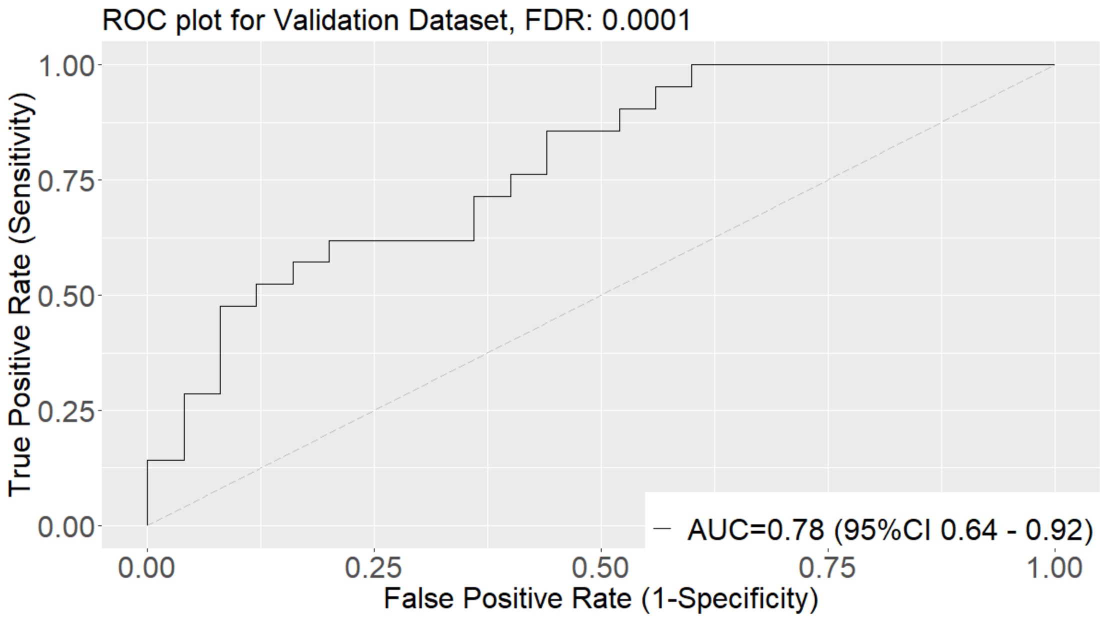


Figure 24: ROC plot for the Validation dataset, showing the accuracy of using significant gene signatures

The functional enrichment analysis shows that interferon alpha and gamma response was higher in females, while oxidative phosphorylation is higher in males. The GO enrichment showed that there may be significant differences in immune responses to a TB infection as several significant genes were related to the Type-1 IFN response, which is a characteristic of tuberculosis.

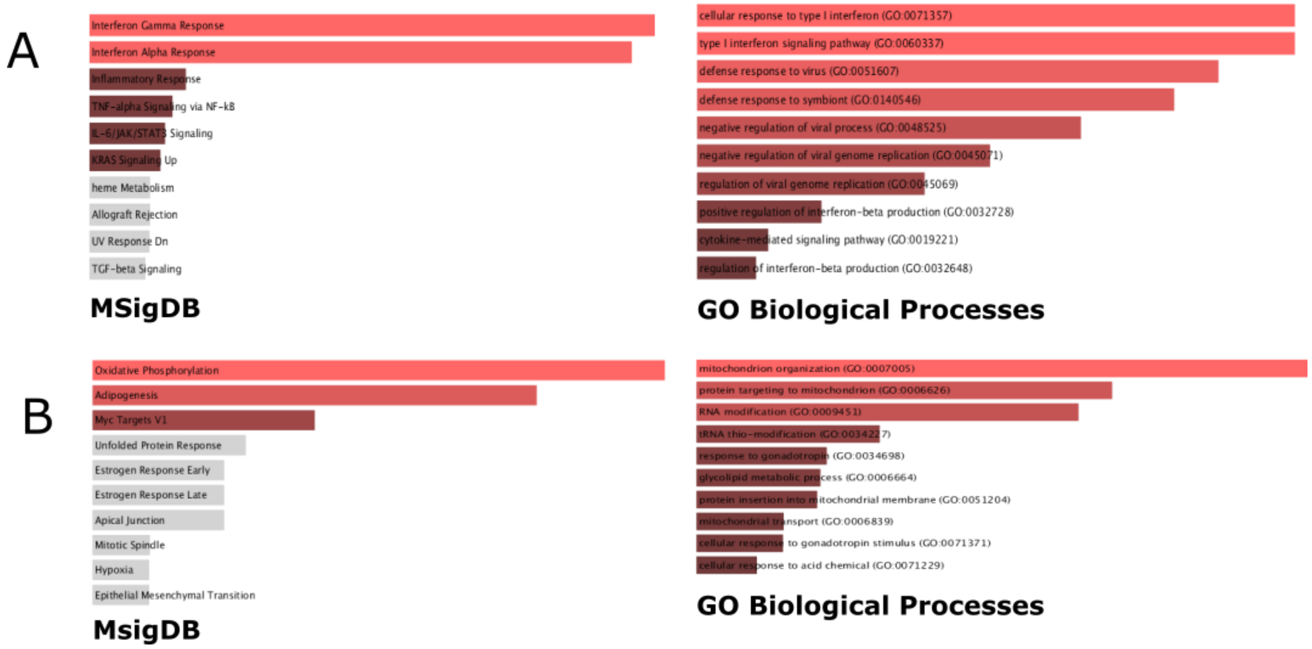


Figure 25: Enrichment analysis for the significant genes using Enrichr

The heatmaps from the ssGSEA analysis of the datasets show no particular up or down-regulation due to gender.

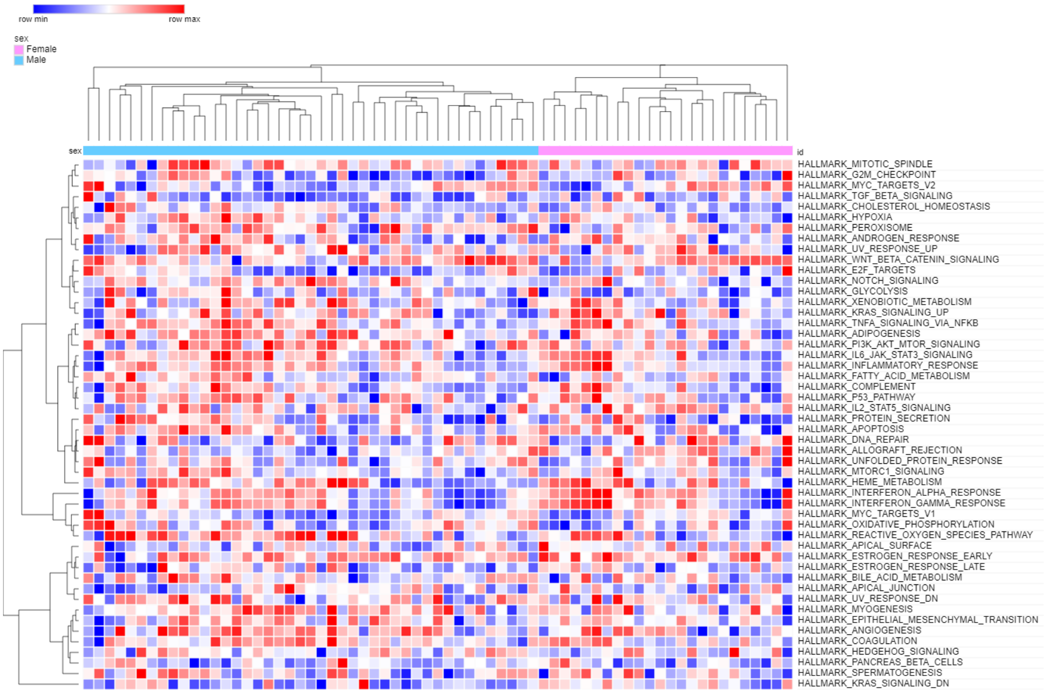


Figure 26 Single-sample gene set enrichment analysis of hallmark pathways for GSE19491.

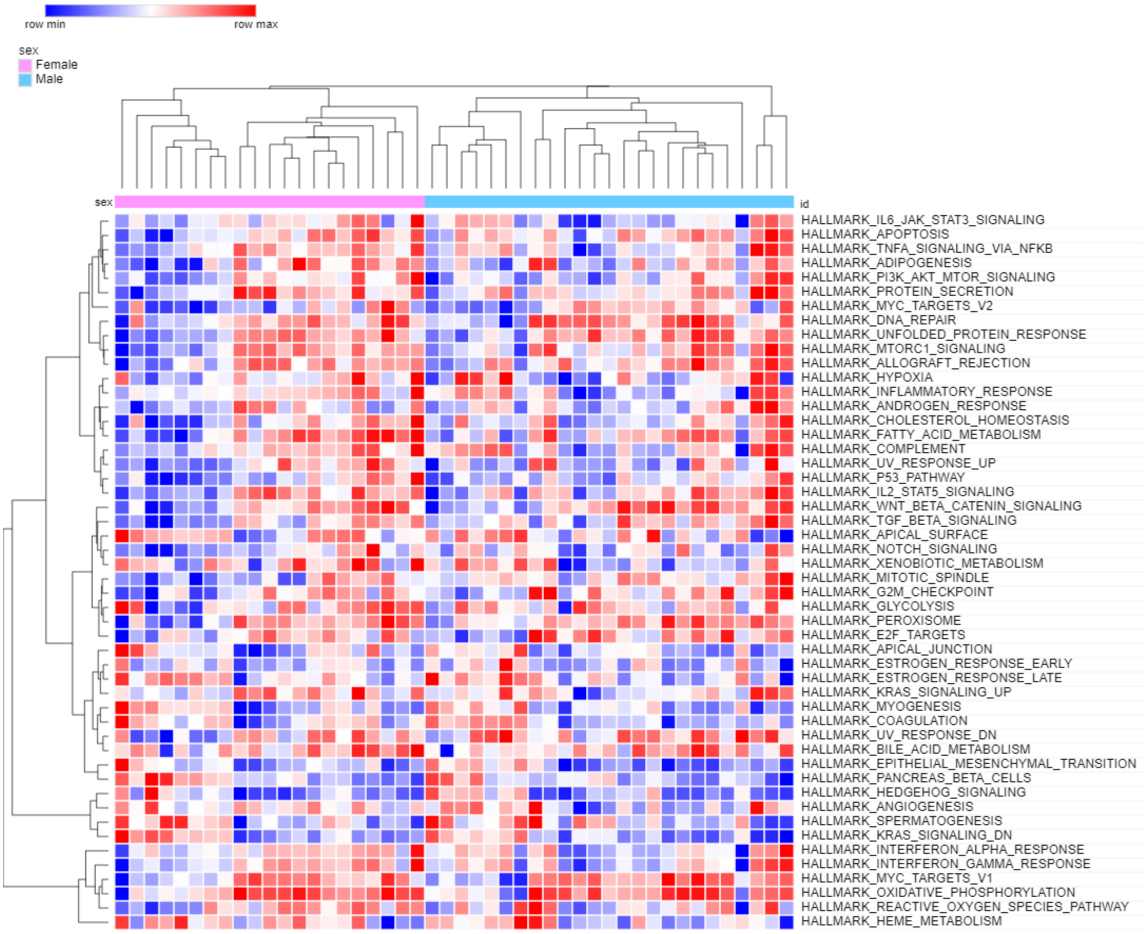


Figure 27: **S**ingle-sample gene set enrichment analysis of hallmark pathways for GSE28623.

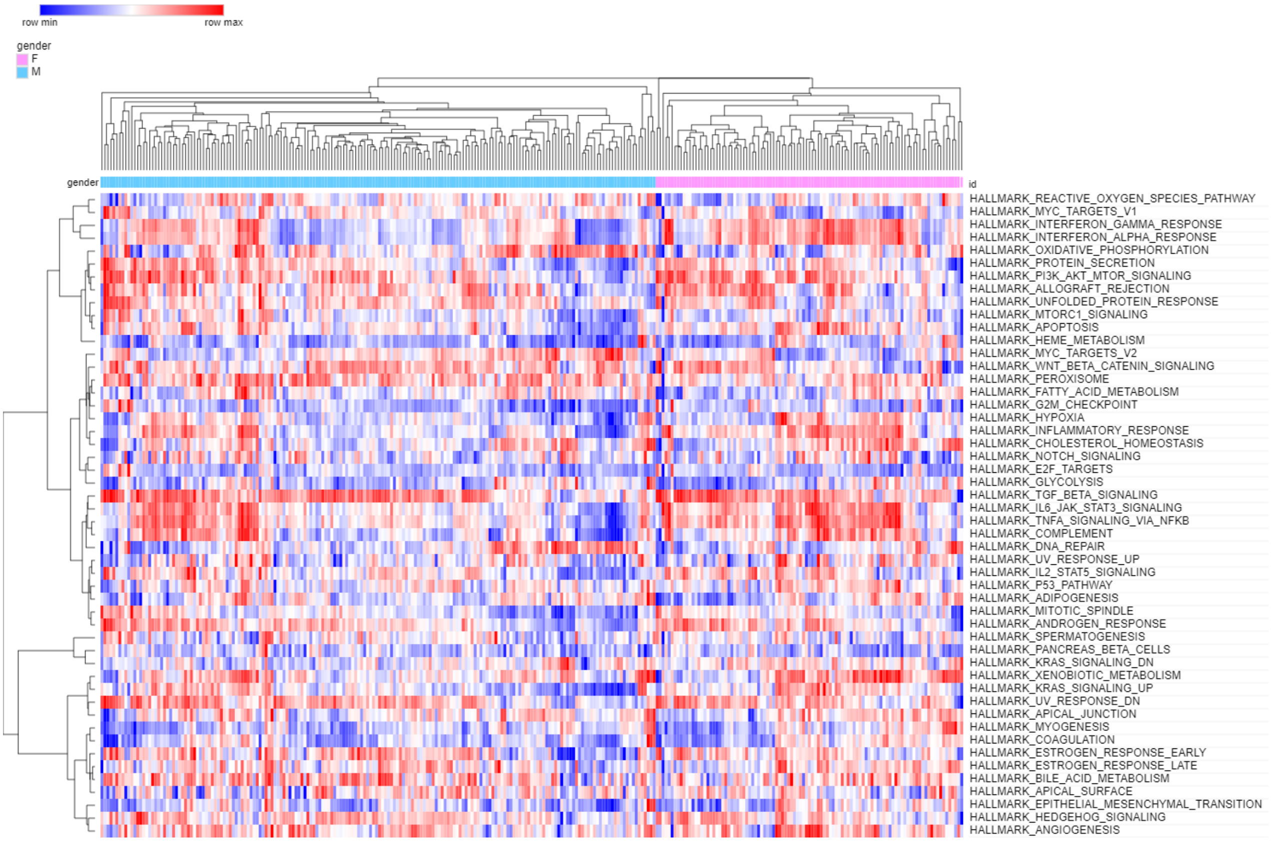


Figure 28: Single-sample gene set enrichment analysis of hallmark pathways for GSE157657.

CHAPTER 5: DISCUSSION

Despite the gender-based differences in TB incidence being heavily attributed to lifestyle factor such as smoking, alcohol use and movement patterns, there exist significant differences in Tuberculosis immune responses that need to be investigated further, i.e. by integrating other forms of –**omics** data (including methylation profiling), more datasets and better inclusion of confounders that impact TB progression. Previous studies have already shown that there exists heterogeneity in immune cell populations across cohorts of TB patients in different areas, therefore, ethnicity-specific differences need to be looked at in more detail [36].

The study also raises questions about the validity of transcriptomic signatures sets that are being proposed for TB diagnosis as very little data from Indian TB patients is available and has been analysed. But, India has the biggest TB burden across the globe.

CHAPTER 6: EXPERIMENTAL WORK

6.1. EXPERIMENT 1: Media Preparation and Sterilization

Chemicals Required: Luria Broth Powder, LB Agar Powder, Nutrient Broth Powder, Nutrient Agar, Distilled water

Equipment Required: Conical flask, weighing balance, spatula, measuring cylinder, autoclave, cotton plugs, test tubes

Procedure followed:

1. Weighed the media according to the instructions provided and the volume needed, using the weighing balance.
2. Transferred the media powder to conical flasks.
3. Added distilled water using measuring cylinder to make up the appropriate volumes.
4. Dissolved the media by stirring.
5. Added 10 ml liquid media (broth) to the test tubes
6. All test tubes and conical flasks were capped using tight cotton plugs.
7. Glassware containing the media put in the autoclave and sterilized at 121º C for 15-20 minutes.

Results:

Sterilized solid and liquid media was prepared. The media was used 12-18 hours after preparation to check for any unwanted microbial contamination.

6.2. EXPERIMENT 2: Culturing Microorganisms in solid and liquid media

Chemicals Required: Luria Broth, Nutrient Broth, LB Agar, Nutrient Agar (autoclaved), bacterial culture without contamination (*E. coli*), Test tubes and Flasks containing media

Equipment Required: Ethanol, inoculating loop, burner, cotton, sterilized petri plates, Laminar Air Flow chamber, parafilm, and incubator

Procedure followed:

Culturing in solid media:

1. The surface of the LAF was wiped clean with ethanol and cotton.
2. Placed all required material inside the LAF and the UV light was turned for 10-15 minutes.
3. Heated the agar media to melt it.
4. Turned off the UV light, switched on the fan and the light inside the LAF
5. With the burner lit, the petri plates were opened and media carefully poured into each of them (20-25 ml per plate).
6. The plates were allowed to set for 25-35 minutes.
7. Using an inoculating loop, the culture was taken and streaked onto the plates. Some plates were also streaked in quadrants.
8. Plates were covered in parafilm and incubated overnight.

Culturing in liquid media:

1. The surface of the LAF was wiped clean with ethanol and cotton.
2. Placed all required material inside the LAF and the UV light was turned for 10-15 minutes.
3. Heated the agar media to melt it.
4. Turned off the UV light, switched on the fan and the light inside the LAF
5. With the burner lit, an inoculating loop was used to take the culture and inoculate the liquid media with *E. coli.*
6. Tubes and flasks were covered with cotton plugs and incubated overnight.

Results:

After 12-18 hours of incubation, isolated colonies could be seen on the streaked plates and the liquid media turned opaque indicating bacterial growth.

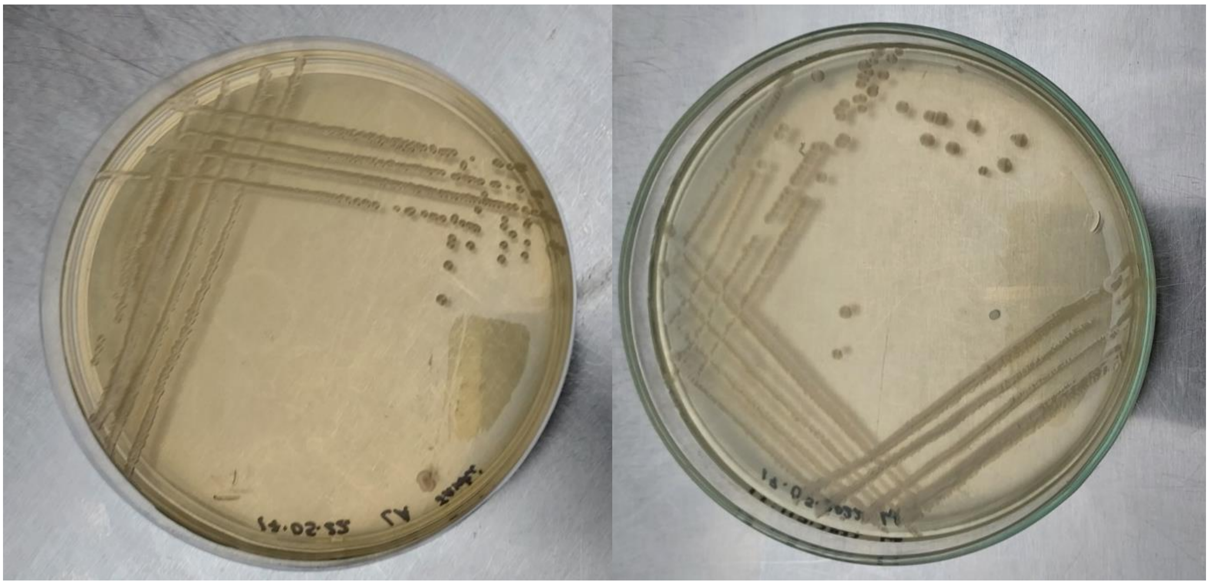


Figure 29: Quadrant streaking of E. coli

6.3. EXPERIMENT 3: Gram Staining

Chemicals Required: Primary stain (crystal violet), Gram's Iodine, 95% alcohol or acetone, Safranin (counterstain)

Equipment Required: Glass slides, microscope, inoculating loop, distilled water, immersion oil, bacterial culture, dropper, burner

Procedure followed:

1. Using the inoculating loop, a small amount of bacterial inoculum was put onto a drop of distilled water on a clean slide and mixed.
2. The smear was heat-fixed.
3. Added crystal violet and kept it for 30 seconds to 1 minute.
4. Rinsed with water.
5. Added gram's iodine and kept it for 1 minute.
6. Rinsed with water.
7. Washed with 95% alcohol or acetone for 10-20 seconds.
8. Added Safranin for 1 minute.
9. Washed with water.
10. Air-dried and observed the slide under the microscope.

Results:

Pink *E. coli* cells were observed under the microscope at 100X, as shown in the Figure below.

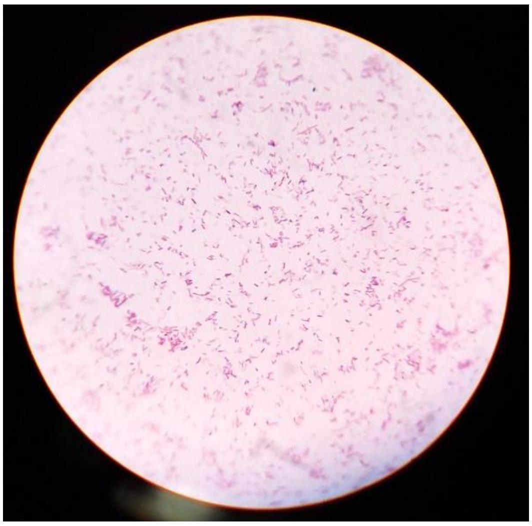


Figure 30: *E.coli* under microscope after Gram Staining

6.4. EXPERIMENT 4: Acid-fast staining

Chemicals Required: Mycobacterial culture, Carbol Fuschin dye, Acid-alcohol (3% HCl in 95% alcohol), Malachite green

Equipment Required: Glass slides, inoculating needle, heating plate, dropper, water (distilled and tap water), immersion oil, and microscope

Procedure followed:

1. Prepared a bacterial smear on a clean glass slide by adding a needle of bacterial growth on a glass slide on a drop of water.
2. Mixed the specimen with water and left to dry.
3. Air-dried smear and heat fixed.
4. Covered the smear with carbol fuschin dye.
5. Heated the stain until vapours began to rise.
6. Allowed heated stain to remain on the slide for 5 minutes.
7. Washed off with tap water
8. Flooded the slide with acid-alcohol for 30 seconds.
9. Rinsed with water.
10. Covered the smear with malachite green and kept for 1-2 minutes.
11. Rinsed with water, wiped the back of the slide and allowed the stained smear to air-dry.
12. Examined the slide under the microscope under 100X objective.

Results:

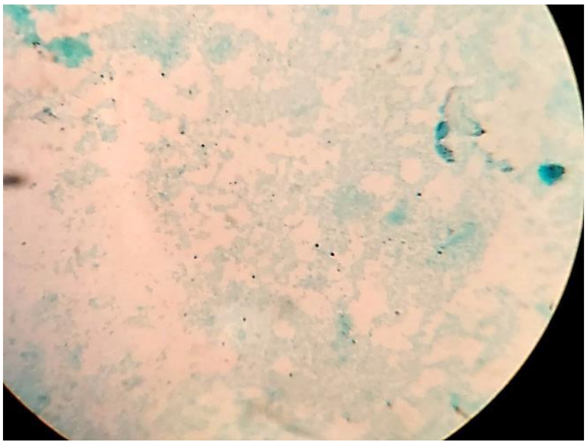


Figure 31: Non acid-fast bacteria

6.5. EXPERIMENT 5: Genomic DNA Isolation using phenol-chloroform method

Chemicals Required: Tris base, Tris-Cl, Phenol: Chloroform (1:1), absolute ethanol, SDS, Lysis Buffer, 1X TE Buffer, chilled isopropanol

Equipment Required: *E. coli* culture, Luria broth containing flask , tarson tubes (50 ml), microfuge tubes (2 ml), Glass slides, centrifuge, Agarose gel electrophoresis buffers and apparatus, pipettes and tips, incubator (37 ° C), ice bucket with ice, -20° C refrigerator, 4 ° C refrigerator

Procedure followed:

1. A fresh flask containing LB Broth was inoculated and incubated overnight at 37 ° C.
2. From a fresh overnight culture, 10 ml culture was transferred to tarson tubes.
3. To get the cell pellet, the tubes were spun at 7,000 rpm for 10 minutes.
4. The supernatant was discarded, and the pellet resuspended in 4ml Lysis buffer.
5. Completely vortexed the tubes to allow proper mixing.
6. Incubated the tubes at 37 ° C for 1 hour.
7. Added equal volumes of phenol: chloroform (2ml each) and mixed properly. The phenol used has saturated at pH 8.0 using Tris-Cl.
8. Centrifuged at 10,000 rpm for 10 minutes. After centrifugation a white layer could be seen at the interface of organic and aqueous layers.
9. Carefully transferred the aqueous phase with a pipette to 2ml microfuge tubes.
10. Repeated the phenol: chloroform step, but with 05.ml of each and spun for 10,000 rpm for 5 minutes.
11. Transferred the aqueous layer to a new tube and added 1 ml isopropanol for precipitation.
12. Incubated the tubes at -20 ° C overnight.
13. Spun the tubes for 15 minutes at 4 ° C.
14. Discarded the supernatant and rinsed the pellet with 1 ml of 70% ethanol.
15. Repeated the alcohol washing step
16. Resuspended the DNA in TE buffer
17. Agarose gel was loaded with DNA samples and electrophoresis carried out to visualize the bands of genomic DNA.

Results:

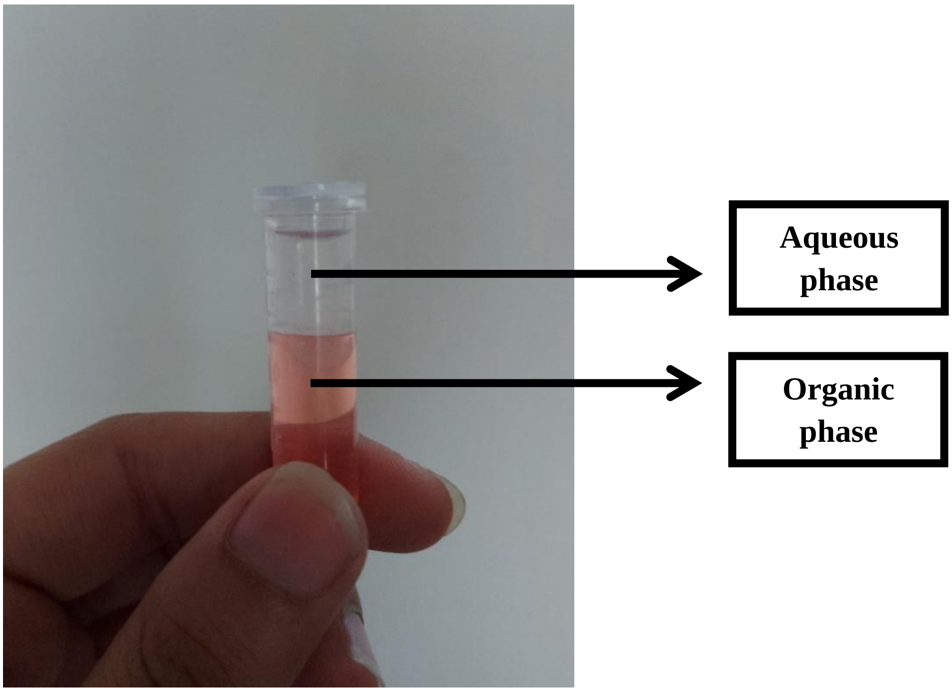


Figure 32: Separated aqueous and organic phases. The upper layer was transferred to a fresh vial.

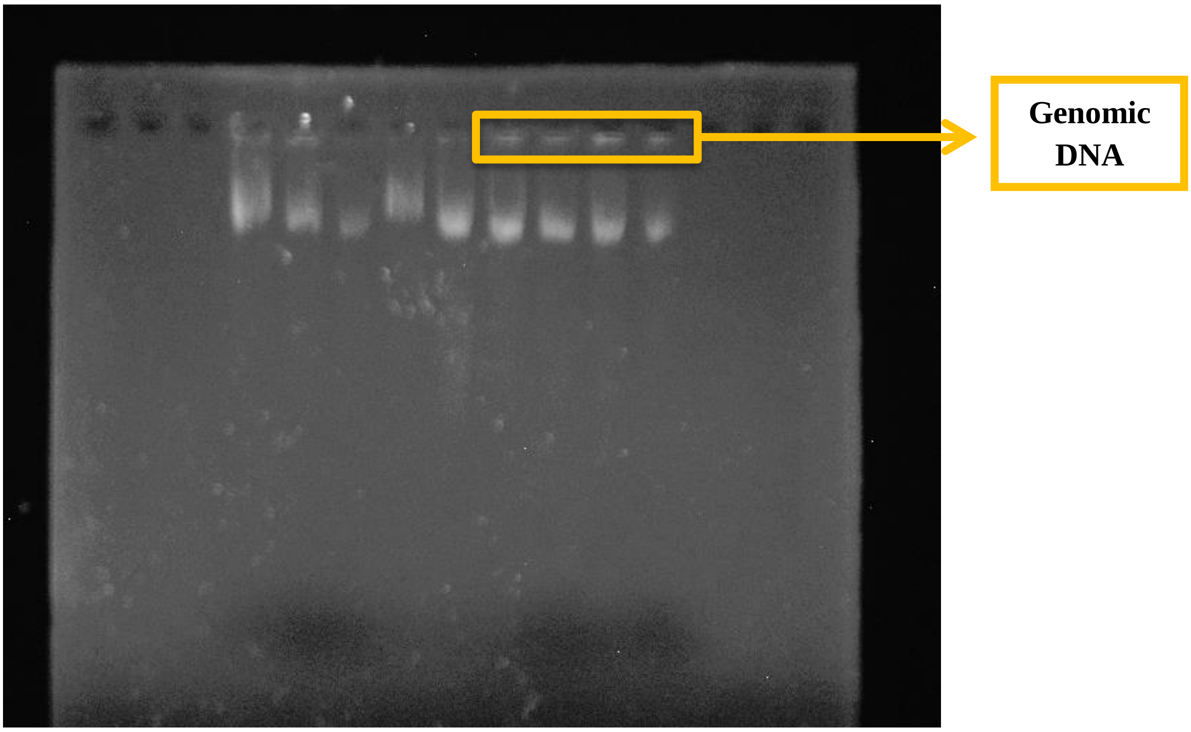


Figure 33: DNA visualized on 0.8% Agarose gel.

6.6. EXPERIMENT 6: Plasmid DNA Isolation using phenol-chloroform method

Chemicals Required: Alkaline Lysis Buffers I, II and III, 96% ethanol, 70% ethanol, chilled isopropanol, DNA loading dye (6X), EtBr

Equipment Required: *E. coli* pUC19 culture, Luria broth containing flask , tarson tubes (50 ml), microfuge tubes (2 ml), Glass slides, centrifuge, Agarose gel electrophoresis buffers and apparatus, pipettes and tips, incubator (37 º C), ice bucket with ice, -20º C refrigerator, 4 º C refrigerator

Procedure followed:

1. A fresh flask containing LB Broth was inoculated and incubated overnight at 37 º C.
2. From a fresh overnight culture, 45 ml culture was transferred to tarson tubes.
3. To get the cell pellet, the tubes were spun at 7,000 rpm for 10 minutes at 4 º C.
4. Added 0.600 ml of ALS-I (GTE).
5. Completely vortexed the tubes to allow proper mixing.
6. Added 1.2 ml of SDS-NaOH solution (ALS-II) and inverted the tubes rapidly.
7. Incubated at 37 º C for 5 minutes, this gave rise to a slimy texture.
8. Carefully added 0.450 ml of acetate solution (ALS-III) to allow renaturation of circular DNA.
9. Mixed gently 5-6 times.
10. Incubated in ice for 30 minutes.
11. Centrifuged at 7,000 rpm for 20 minutes at 4 º C.
12. Transferred the supernatant to fresh tubes
13. Added equal volume (~1ml) of chilled isopropanol (1:1).
14. Incubated the tubes at -20 º C overnight.
15. Spun the tubes for 10,000 rpm for 15 minutes at 4 º C.
16. Discarded the supernatant and rinsed the pellet with 0.750 ml of 96% ethanol.
17. Centrifuged at 10,000 rpm for 15 minutes.
18. Discarded the supernatant and rinsed the pellet with 0.750 ml of 70% ethanol.
19. Centrifuged at 10,000 rpm for 10 minutes.
20. Discarded the supernatant
21. Air-dried the pellet by allowing the ethanol to evaporate.
22. Suspended the pellet in autoclaved distilled water (10 microliters) and mixed by tapping.
23. Briefly spun the tubes and then pooled plasmid DNA into one tube.

24. Casted 0.8% Agarose gel with 3 microliters of EtBr (50 ml) in an 8-well tank.

25. 5 microliters sample and 1 microliter loading dye were mixed and loaded into the wells.

26. Agarose gel electrophoresis was carried out at 100Vto visualize the bands of plasmid DNA.

Results:

Plasmid DNA was visualized.

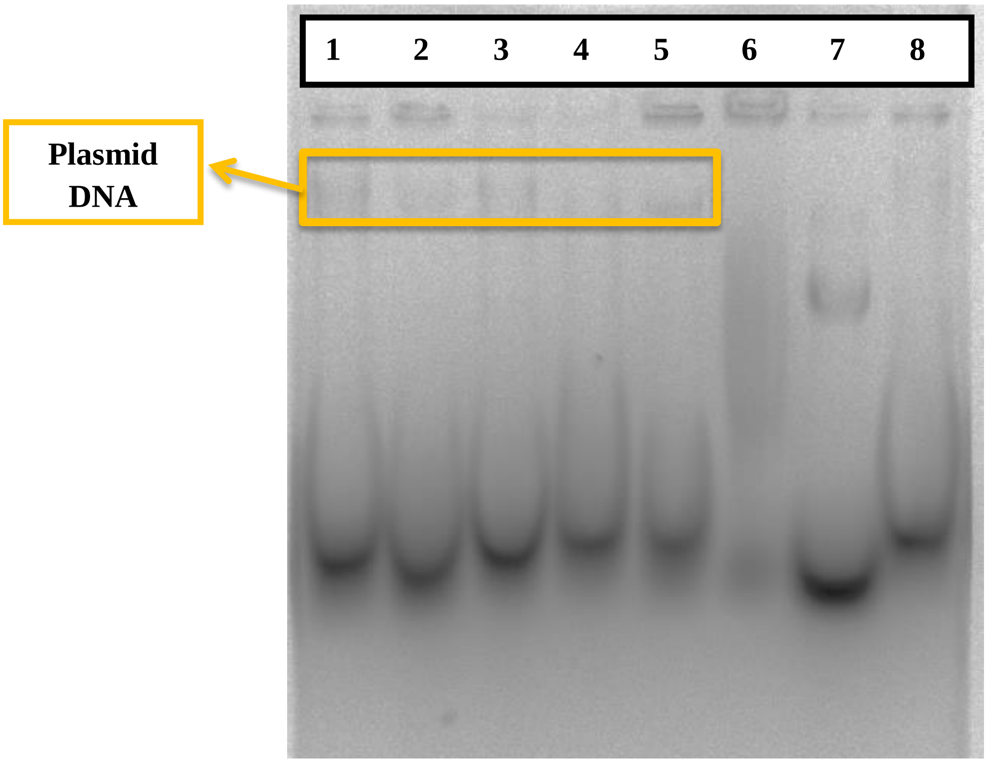


Figure 34: Plasmid DNA on 0.8% Agarose gel, the first five lanes are all pUC-19, the sixth is the pRT vector.

6.7. EXPERIMENT 7: Antimicrobial Susceptibility test/ Disc Diffusion

Chemicals required: MHA Media, antibiotic discs, distilled water, 2% agar, liquid culture

Equipment required: distilled water, flasks, pouring plates, burner, ethanol, spreader, forceps, pipettes, tips , LAF

Procedure followed:

1. The MHA media was prepared (2% Agar+ MHA media) following the instructions given on the box. Water was added to make up the required volume.
2. After the autoclave was performed and the media checked for contamination, pouring was done.
3. After the media solidified, 1ml of liquid culture (of *S. aureus*) was pipette and poured inside the LAF
4. Using a spreader, the culture was carefully spread all over the media plate till it was completely dried.
5. Now the antibiotics disc were taken and placed across the plate using forceps.
6. Plates were sealed with a parafilm and incubated overnight for the culture to grow.

Results:

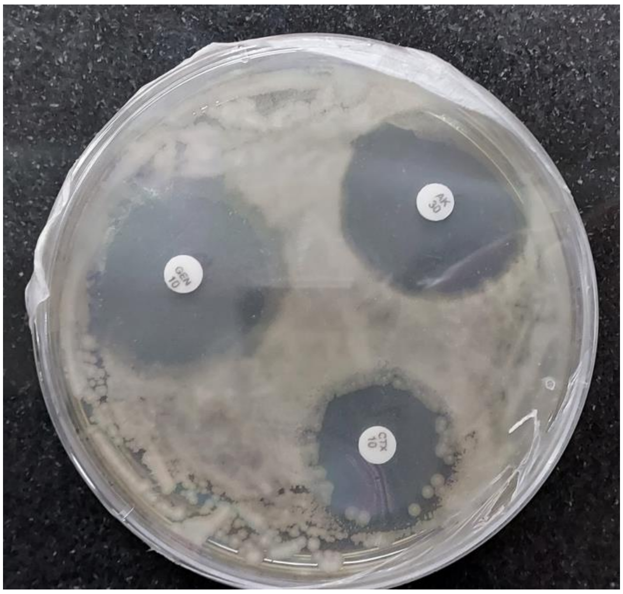


Figure 35: AST by Disk Diffusion Results

Table 4: Observed zones of inhibition against *S. aureus*.

Antibiotic	**Concentration (mcg/disk)**	**Observed Diameter (in mm)**
Gentamycin	10.00	30.0
Amikacin	10.00	32.0
Cephotaxime	30.00	10.0

REFERENCES

[1] R. M. G. J. Houben and P. J. Dodd, ‘The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling’, *PLOS Medicine*, vol. 13, no. 10, p. e1002152, Oct. 2016, doi: 10.1371/journal.pmed.1002152.

[2] WHO Team, ‘Global tuberculosis report 2021’, World Health Organization, Geneva, Oct. 2021. [Online]. Available: https://www.who.int/publications/i/item/9789240037021

[3] WHO Team, ‘Impact of the COVID-19 pandemic on TB detection and mortality in 2020’. WHO, Mar. 21, 2021. [Online]. Available: https://www.who.int/publications/m/item/impact-of-the-covid-19-pandemic-on-tb-detection-and-mortality-in-2020

[4] J. G. Burel *et al.*, ‘Host Transcriptomics as a Tool to Identify Diagnostic and Mechanistic Immune Signatures of Tuberculosis’, *Front Immunol*, vol. 10, pp. 221–221, Feb. 2019, doi: 10.3389/fimmu.2019.00221.

[5] A. Singhania, R. J. Wilkinson, M. Rodrigue, P. Haldar, and A. O’Garra, ‘The value of transcriptomics in advancing knowledge of the immune response and diagnosis in tuberculosis’, *Nature Immunology*, vol. 19, no. 11, pp. 1159–1168, Nov. 2018, doi: 10.1038/s41590-018-0225-9.

[6] P. Narasimhan, J. Wood, C. R. Macintyre, and D. Mathai, ‘Risk factors for tuberculosis’, *Pulm Med*, vol. 2013, pp. 828939–828939, 2013, doi: 10.1155/2013/828939.

[7] M. Bates, B. J. Marais, and A. Zumla, ‘Tuberculosis Comorbidity with Communicable and Noncommunicable Diseases’, *Cold Spring Harb Perspect Med*, vol. 5, no. 11, p. a017889, Feb. 2015, doi: 10.1101/cshperspect.a017889.

[8] Drain Paul K. *et al.*, ‘Incipient and Subclinical Tuberculosis: a Clinical Review of Early Stages and Progression of Infection’, *Clinical Microbiology Reviews*, vol. 31, no. 4, pp. e00021-18, doi: 10.1128/CMR.00021-18.

[9] L. E. Connolly, P. H. Edelstein, and L. Ramakrishnan, ‘Why is long-term therapy required to cure tuberculosis?’, *PLoS Med*, vol. 4, no. 3, pp. e120–e120, Mar. 2007, doi: 10.1371/journal.pmed.0040120.

[10] M. P. R. Berry *et al.*, ‘An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis’, *Nature*, vol. 466, no. 7309, pp. 973–977, Aug. 2010, doi: 10.1038/nature09247.

[11] M. W. Borgdorff, N. J. Nagelkerke, C. Dye, and P. Nunn, ‘Gender and tuberculosis: a comparison of prevalence surveys with notification data to explore sex differences in case detection’, *Int J Tuberc Lung Dis*, vol. 4, no. 2, pp. 123–132, Feb. 2000.

[12] K. C. Horton, P. MacPherson, R. M. G. J. Houben, R. G. White, and E. L. Corbett, ‘Sex Differences in Tuberculosis Burden and Notifications in Low- and Middle-Income Countries: A Systematic Review and Meta-analysis’, *PLoS Med*, vol. 13, no. 9, p. e1002119, Sep. 2016, doi: 10.1371/journal.pmed.1002119.

[13] O. Neyrolles and L. Quintana-Murci, ‘Sexual inequality in tuberculosis’, *PLoS Med*, vol. 6, no. 12, pp. e1000199–e1000199, Dec. 2009, doi: 10.1371/journal.pmed.1000199.

[14] E. I. Bini *et al.*, ‘The influence of sex steroid hormones in the immunopathology of experimental pulmonary tuberculosis’, *PLoS One*, vol. 9, no. 4, pp. e93831–e93831, Apr. 2014, doi: 10.1371/journal.pone.0093831.

[15] C.-Y. Lin *et al.*, ‘Effects of Gender and Age on Development of Concurrent Extrapulmonary Tuberculosis in Patients with Pulmonary Tuberculosis: A Population Based Study’, *PLOS ONE*, vol. 8, no. 5, p. e63936, May 2013, doi: 10.1371/journal.pone.0063936.

[16] S. Nhamoyebonde and A. Leslie, ‘Biological differences between the sexes and susceptibility to tuberculosis.’, *J Infect Dis*, vol. 209 Suppl 3, pp. S100-106, Jul. 2014, doi: 10.1093/infdis/jiu147.

[17] Jackie Marchildon, ‘Gender Inequality Is Seriously Harming the Global Fight Against Tuberculosis’. https://www.globalcitizen.org/en/content/gender-equality-and-tuberculosis/

[18] Hadley Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. [Online]. Available: https://ggplot2.tidyverse.org

[19] Kamil Slowikowski, *ggrepel: Automatically Position Non-Overlapping Text Labels with ‘ggplot2’*. [Online]. Available: https://github.com/slowkow/ggrepel

[20] Hadley Wickham and Dana Seidel, *scales: Scale Functions for Visualization*. 2022. [Online]. Available: https://scales.r-lib.org

[21] Garnier *et al.*, *viridis - Colorblind-Friendly Color Maps for R*. 2021. [Online]. Available: https://sjmgarnier.github.io/viridis/

[22] J. Maertzdorf *et al.*, ‘Human gene expression profiles of susceptibility and resistance in tuberculosis.’, *Genes Immun*, vol. 12, no. 1, pp. 15–22, Jan. 2011, doi: 10.1038/gene.2010.51.

[23] J. Maertzdorf *et al.*, ‘Functional correlations of pathogenesis-driven gene expression signatures in tuberculosis.’, *PLoS One*, vol. 6, no. 10, p. e26938, 2011, doi: 10.1371/journal.pone.0026938.

[24] T. O. J. P. Elliott *et al.*, ‘Dysregulation of Apoptosis Is a Risk Factor for Tuberculosis Disease Progression.’, *J Infect Dis*, vol. 212, no. 9, pp. 1469–1479, Nov. 2015, doi: 10.1093/infdis/jiv238.

[25] J. M. Cliff *et al.*, ‘Distinct phases of blood gene expression pattern through tuberculosis treatment reflect modulation of the humoral immune response.’, *J Infect Dis*, vol. 207, no. 1, pp. 18–29, Jan. 2013, doi: 10.1093/infdis/jis499.

[26] M. Kaforou *et al.*, ‘Detection of tuberculosis in HIV-infected and -uninfected African adults using whole blood RNA expression signatures: a case-control study’, *PLoS Med*, vol. 10, no. 10, pp. e1001538–e1001538, Oct. 2013, doi: 10.1371/journal.pmed.1001538.

[27] D. E. Zak *et al.*, ‘A blood RNA signature for tuberculosis disease risk: a prospective cohort study’, *Lancet*, vol. 387, no. 10035, pp. 2312–2322, Jun. 2016, doi: 10.1016/S0140-6736(15)01316-1.

[28] W. E. Johnson *et al.*, ‘Comparing tuberculosis gene signatures in malnourished individuals using the TBSignatureProfiler’, *BMC Infect Dis*, vol. 21, no. 1, pp. 106–106, Jan. 2021, doi: 10.1186/s12879-020-05598-z.

[29] O. Tabone *et al.*, ‘Blood transcriptomics reveal the evolution and resolution of the immune response in tuberculosis’, *J Exp Med*, vol. 218, no. 10, p. e20210915, Oct. 2021, doi: 10.1084/jem.20210915.

[30] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2020. [Online]. Available: https://www.R-project.org/

[31] Hadley Wickham and Maximilian Girlich, *tidyr: Tidy Messy Data*. 2022. [Online]. Available: https://tidyr.tidyverse.org

[32] A. Piovesan, F. Antonaros, L. Vitale, P. Strippoli, M. C. Pelleri, and M. Caracausi, ‘Human protein-coding genes and gene feature statistics in 2019’, *BMC Research Notes*, vol. 12, no. 1, p. 315, Jun. 2019, doi: 10.1186/s13104-019-4343-8.

[33] W. A. Haynes *et al.*, ‘EMPOWERING MULTI-COHORT GENE EXPRESSION ANALYSIS TO INCREASE REPRODUCIBILITY’, *Pac Symp Biocomput*, vol. 22, pp. 144–153, 2017, doi: 10.1142/9789813207813_0015.

[34] A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, and P. Tamayo, ‘The Molecular Signatures Database (MSigDB) hallmark gene set collection’, *Cell Syst*, vol. 1, no. 6, pp. 417–425, Dec. 2015, doi: 10.1016/j.cels.2015.12.004.

[35] S. Hänzelmann, R. Castelo, and J. Guinney, ‘GSVA: gene set variation analysis for microarray and RNA-Seq data’, *BMC Bioinformatics*, vol. 14, no. 1, p. 7, Jan. 2013, doi: 10.1186/1471-2105-14-7.

[36] R. Roy Chowdhury *et al.*, ‘A multi-cohort study of the immune factors associated with M. tuberculosis infection outcomes’, *Nature*, vol. 560, no. 7720, pp. 644–648, Aug. 2018, doi: 10.1038/s41586-018-0439-x.

APPENDICES

APPENDIX 1: Code used for pre-processing, normalization and cleaning

```
#libraries
library(tidyr)
library(MetaIntegrator)
library(GSVA)
library(readxl)
library(ggplot2)

#read dataset
setwd("E:/ACADEMICS/Major project/data/")
if(!(dir.exists("GSE157657"))dir.create("GSE157657")
#==========================================================================
#GSE1576657 Cleaning
GSE157657_exprFile<- "GSE157657_norm.data.txt"
GSE157657_exprData_all <- read.delim(GSE157657_exprFile, header =TRUE,
row.names=1)

#read series matrix
GSE157657_sampleData<- read.delim("GSE157657_series_matrix.txt", skip=25,
header=FALSE)

#selecting the features we want
#filtering all samples with the parameters we need:
GSE157657_metaData<- as.data.frame(t( apply(GSE157657_sampleData[,-1],
MARGIN=2, function(x){
  if (grep("gender:", x) && grep("subgroup:", x) && grep("ethnicity:", x)
&& grep("days_from_att:", x)){
    return(x[c(1, 10, 12,  grep("subgroup:", x), grep("gender:",
x),grep("ethnicity:", x), grep("days_from_att:", x), grep("age:", x))])
  }
  else
    return(c(NA,0,0,0,0))
})))

colnames(GSE157657_metaData)<- c("sample", "patient_id", "group", "sub-
group", "sex", "ethnicity", "days_att", "age")

#remove the strings
GSE157657_metaData$group <- sapply(GSE157657_metaData$group, func-
tion(x){unlist(strsplit(x, split = 'group: '))})[2,]
GSE157657_metaData$subgroup <- sapply(GSE157657_metaData$subgroup, func-
tion(x){unlist(strsplit(x, split = 'subgroup: '))})[2,]
GSE157657_metaData$ethnicity <- sapply(GSE157657_metaData$ethnicity, func-
tion(x){unlist(strsplit(x, split = 'ethnicity: '))})[2,]
GSE157657_metaData$sex <- sapply(GSE157657_metaData$sex, func-
tion(x){unlist(strsplit(x, split = 'gender: '))})[2,]
GSE157657_metaData$days_att <-
as.numeric(sapply(GSE157657_metaData$days_att, func-
tion(x){unlist(strsplit(x, split = 'days_from_att: '))})[2,])
GSE157657_metaData$patient_id <-
as.numeric(sapply(GSE157657_metaData$patient_id, func-
tion(x){unlist(strsplit(x, split = 'patient id: '))})[2,])
GSE157657_metaData$age <- as.numeric(sapply(GSE157657_metaData$age, func-
tion(x){unlist(strsplit(x, split = 'age: '))})[2,])

#filter only PTB samples
```

```
GSE157657_PTB_samples <- sapply(GSE157657_metaData, function(x){grep(
"PTB", x)})
GSE157657_LTB_samples <- sapply(GSE157657_metaData, function(x){grep(
"PTB", x)})
GSE157657_metaData_PTB <- GSE157657_metaData[GSE157657_PTB_samples$group,]
#filter only days from att<=60 (2 months)
GSE157657_metaData_PTB <-GSE157657_metaData_PTB
[which(GSE157657_metaData_PTB$days_att<=60),]
#filter only those samples from the expression dataset
samplesPTB_match<- match(GSE157657_metaData_PTB$sample,
colnames(GSE157657_exprData_all))
GSE157657_exprData_PTB<-GSE157657_exprData_all[, c(1,samplesPTB_match)]

#modify GSE157657_metaData_PTB to have continent specific ethnicity values!
GSE157657_metaData_PTB$ethnicity<- sapply(GSE157657_metaData_PTB$ethnicity, function(x){
  if(length(grep( "South Asia", x))>0){
    return("South Asia")
  }
  else if(length(grep("Africa", x))>0){
    return("African")
  }
  else if(length(grep("Europe", x))>0||x=="British"){
    return("European")
  } else if (x=="British Indian"){
    return("British Indian")
  }
})

#create comparison lists for the PTB ethnicity
GSE157657_compList <- list(c("South Asia", "African"),
                           c("African", "European"),
                           c("European","British Indian"),
                           c("South Asia", "British Indian"))

#==========================================================================
======
#GSE19491 Analysis
GSE19491_exprData_all <- read.delim("GSE19491_series_matrix.txt",
                                    skip=81,
                                    header =TRUE,
                                    row.names=1)
#read sample data from series matrix
GSE19491_sampleData<- read.delim("GSE19491_series_matrix.txt",
                                 skip=33,
                                 nrows=30,
                                 header=FALSE)

#removing samples which are not whole-blood, come from healthy control
#also removing samples which were taken at end of treatment and were not
TB
cond_Tb<-c("Whole blood from patient with active TB before treatment",
           "Whole blood from patient with active TB 2 months after treatment started",
           "Whole Blood from patient with Latent TB",
           "Whole Blood from patient with Active TB")
TB_samples_idx<- unlist(sapply(cond_Tb, function(x){
  grep(x, GSE19491_sampleData[8,])
}))
```

```
GSE19491_sampleData<- GSE19491_sampleData[, TB_samples_idx]

#selecting the features we want
#filtering all samples with the parameters we need:
GSE19491_metaData<- as.data.frame(t(GSE19491_sampleData[c(2,10:13),-1]))
colnames(GSE19491_metaData)<- c("sample", "age", "sex", "ethnicity", "illness")

#remove the strings
GSE19491_metaData$illness <- sapply(GSE19491_metaData$illness, function(x){unlist(strsplit(x, split = 'illness:'))})[2,]
GSE19491_metaData$ethnicity <- sapply(GSE19491_metaData$ethnicity, function(x){unlist(strsplit(x, split = 'ethnicity: '))})[2,]
GSE19491_metaData$sex <- sapply(GSE19491_metaData$sex, function(x){unlist(strsplit(x, split = 'gender: '))})[2,]

#keep only the numbers for age??
GSE19491_metaData$age <- as.numeric(sapply(GSE19491_metaData$age, function(x){
  if(length(grep("years", x))>0){
    x<- unlist(strsplit(x, split = 'years'))[1]}
  unlist(strsplit(x, split = 'age: '))})[2,])

#filter only active TB, whole blood samples, which are not end of treatment
GSE19491_PTB_samples <- sapply(GSE19491_metaData, function(x){grep( "PTB",
x)})
GSE19491_metaData_PTB <- GSE19491_metaData[GSE19491_PTB_samples$illness,]

#filter only those samples from the expression dataset
samplesPTB_match<- match(GSE19491_metaData_PTB$sample,
colnames(GSE19491_exprData_all))

#removing any duplicated rows
#remove the last row, which is empty and NAs
GSE19491_exprData_PTB<-GSE19491_exprData_all[-48804, samplesPTB_match]
#remove genes/probes with NAs
GSE19491_exprData_PTB<- GSE19491_exprData_PTB %>% drop_na()

#scaling values to mean
GSE19491_exprData_PTB_scaled<-scale(GSE19491_exprData_PTB,
                                    scale=apply(GSE19491_exprData_PTB, 2,
mean))

#add gene names to first column
#read platform file
GPLplatform_data <- read.delim("GPL6947-13512.txt", skip=30)
GPLplatform_data <- GPLplatform_data[which(GPLplatform_data$Symbol!=""),]

#matching Probe IDs
genesMatch <- match(GPLplatform_data$ID,
row.names(GSE19491_exprData_PTB_scaled))
GSE19491_exprData_PTB<- cbind.data.frame(GPLplatform_data$Symbol[-
which(is.na(genesMatch))],
                                         GSE19491_exprData_PTB_scaled[
na.omit(genesMatch),])
colnames(GSE19491_exprData_PTB)[1]<- "Gene_name"

#modify GSE19491_metaData_PTB to have continent specific ethnicity values!
```

```
GSE19491_metaData_PTB$ethnicity[which(GSE19491_metaData_PTB$ethnicity=="As
ian other")]<- "Asian Other"

#create comparison lists for the PTB ethnicity
GSE19491_compList <- list(c("South Asian",  "Asian Other"),
                          c("South Asian","White" ),
                          c("White","Black"),
                          c("Black","Other"),
                          c("Asian Other","Other"))
##========================================================================
======
#GSE28623 Loading and Preprocessing
GSE28623_exprData<- read.delim("GSE28623_series_matrix.txt",
                               header =TRUE,
                               skip=60)
GSE28623_sampleData<- read.delim("GSE28623_series_matrix.txt",
                               header =TRUE,
                               skip=30,
                               nrows=28)

#remove healthy controls
expr_id_ref<- GSE28623_exprData[-45016,1]
controls_idx<- grep("healthy non-infected donors",
GSE28623_sampleData[9,])
GSE28623_sampleData<- GSE28623_sampleData[, -controls_idx]

GSE28623_metaData<- as.data.frame(t(GSE28623_sampleData[c(1, 9, 10), -1]))
colnames(GSE28623_metaData)<- c("Sample", "group", "sex")

#remove the strings
GSE28623_metaData$sex <- sapply(GSE28623_metaData$sex, func-
tion(x){unlist(strsplit(x, split = 'gender: '))})[2,]
GSE28623_metaData$group <- sapply(GSE28623_metaData$group, func-
tion(x){unlist(strsplit(x, split = 'group: '))})[2,]

#make a new dataset only for active TB patients
GSE28623_PTB_samples <- grep("tuberculosis patients",
GSE28623_metaData$group)
GSE28623_metaData_PTB <- GSE28623_metaData[GSE28623_PTB_samples,]

#matching samples with the expression dataset
samplesPTB_match<- match(GSE28623_metaData_PTB$Sample,
                         colnames(GSE28623_exprData))

#remove the last row, which is empty and NAs
GSE28623_exprData<-GSE28623_exprData[-45016, samplesPTB_match]
#remove genes/probes with NAs
GSE28623_exprData<- GSE28623_exprData %>% drop_na()

#log2 normalization
#adding a constant>minimum value to ensure no NaNs or -Inf's are produced
min(GSE28623_exprData) #min value=10.57
boxplot(GSE28623_exprData)
GSE28623_exprData_norm<-apply(GSE28623_exprData, 2,
                                  function(x){log2(x+10.6)})
boxplot(GSE28623_exprData_norm)
GSE28623_exprData_norm<- cbind.data.frame(expr_id_ref,
GSE28623_exprData_norm)

#add gene names to first column
```

```
#read platform file
GPLplatform_data <- read.delim("GPL4133-12599.txt", skip=22)
GPLplatform_data <- GPLplatform_data[which(GPLplatform_data$GENE_SYMBOL!=""),]

#matching Probe IDs
genesMatch <- match(GPLplatform_data$ID, GSE28623_exprData_norm[,1])
GSE28623_exprData_norm<- cbind.data.frame(GPLplatform_data[,10],
                                          GSE28623_exprData_norm[
na.omit(genesMatch),-1])
colnames(GSE28623_exprData_norm)[1]<- "Gene_name"

#==========================================================================
======
#FILTER PROTEIN-CODING GENES AND KEEP UNIQUE GENES WITH THE HIGHEST VARIANCE

coding_genes<- readxl::read_xlsx("C:/Users/hp/Downloads/Genes.xlsx")

rmGenes<- function (exprData){
  exprData<- dplyr::distinct(exprData)

  #in case of genes with more than one rows
  #keeping expression values with higher variance
  #getting unique genes
  uniqueGenes<- as.vector(unique(exprData[,1]))
  rowsKeep<- as.numeric()
  for(gene in uniqueGenes){
    rowGene <- which(exprData[,1]==gene)
    if (length(rowGene)==1){
      rowsKeep<- c(rowGene, rowsKeep)
    }
    else if(length(rowGene)>1){
      sd<- sapply(rowGene, function(x){sd(as.numeric(exprData[x,-1]))})
      idx<- which(sd==max(sd))
      rowsKeep<- c(rowGene[idx], rowsKeep)
    }
  }
  exprData<- exprData[unique(rowsKeep),]
  head(exprData)


  mat_gene<- na.omit(match(coding_genes$Gene_Symbol, exprData[,1]))
  exprData_coding<- exprData[mat_gene, ]
  return(exprData_coding)
}


GSE157657_exprData_coding_PTB<- rmGenes(exprData=GSE157657_exprData_PTB)
GSE19491_exprData_coding_PTB<- rmGenes(exprData=GSE19491_exprData_PTB)
GSE28623_exprData_coding_PTB<- rmGenes(exprData=GSE28623_exprData_norm)

#changing illness label to group quickly!
colnames(GSE19491_metaData_PTB)[5]<- "group"
```

APPENDIX 2: Code used for MetaIntegrator Analysis

```
#==========================================================================
#MetaIntegrator
#getting gene lists to use as keys
GSE157657_genes_exprData_coding<- GSE157657_exprData_coding_PTB[,1]
names(GSE157657_genes_exprData_coding)<-
rownames(GSE157657_exprData_coding_PTB)
GSE157657_exprData_coding_PTB<- apply(GSE157657_exprData_coding_PTB[,-1],
2, as.numeric)
row.names(GSE157657_exprData_coding_PTB)<-
names(GSE157657_genes_exprData_coding)

#getting gene lists
GSE19491_genes_exprData_coding<- GSE19491_exprData_coding_PTB[,1]
names(GSE19491_genes_exprData_coding)<-
rownames(GSE19491_exprData_coding_PTB)
GSE19491_exprData_coding_PTB<- apply(GSE19491_exprData_coding_PTB[,-1], 2,
as.numeric)
row.names(GSE19491_exprData_coding_PTB)<-
names(GSE19491_genes_exprData_coding)

#getting gene lists
GSE28623_genes_exprData_coding<- GSE28623_exprData_coding_PTB[,1]
names(GSE28623_genes_exprData_coding)<-
rownames(GSE28623_exprData_coding_PTB)
GSE28623_exprData_coding_PTB<- apply(GSE28623_exprData_coding_PTB[,-1], 2,
as.numeric)
row.names(GSE28623_exprData_coding_PTB)<-
names(GSE28623_genes_exprData_coding)
#=========================================================
#changing metaData to fit pheno format
GSE157657_pheno<- GSE157657_metaData_PTB[,-1]
row.names(GSE157657_pheno)<- GSE157657_metaData_PTB[,1]
GSE19491_pheno<- GSE19491_metaData_PTB[,-1]
row.names(GSE19491_pheno)<- GSE19491_metaData_PTB[,1]
GSE28623_pheno<- GSE28623_metaData_PTB[,-1]
row.names(GSE28623_pheno)<- GSE28623_metaData_PTB[,1]

#running MetaIntegrator
#making discovery and validation objects
PTB_object1<- list(expr=GSE157657_exprData_coding_PTB,
                   keys=GSE157657_genes_exprData_coding,
                   pheno=GSE157657_pheno,
                   formattedName="GSE157657_PTB_discovery")
PTB_object2<- list(expr=GSE19491_exprData_coding_PTB,
                   keys=GSE19491_genes_exprData_coding,
                   pheno=GSE19491_pheno,
                   formattedName="GSE19491_PTB_discovery")
PTB_object1_validation<- list(expr=GSE28623_exprData_coding_PTB,
                   keys=GSE28623_genes_exprData_coding,
                   pheno=GSE28623_pheno,
                   formattedName="GSE28623_PTB_validation")

#write data to files
write.table(GSE157657_exprData_coding_PTB, paste0(Sys.Date(), "_expr-
data_GSE157657_coding_PTB.txt"), sep="\t")
write.table(GSE19491_exprData_coding_PTB, paste0(Sys.Date(), "_expr-
data_GSE19491_coding_PTB.txt"), sep="\t")
```

```
write.table(GSE28623_exprData_coding_PTB, paste0(Sys.Date(), "_expr-
data_GSE28623_coding_PTB.txt"), sep="\t")

#add class as sex
#discovery dataset 1
PTB_object1$class <- sapply(GSE157657_pheno$sex, function(x){
  if (x=='M'){
    return (0)}
  else{
    return (1)}})
names(PTB_object1$class)<- as.character(row.names(GSE157657_pheno))

#discovery dataset 2
PTB_object2$class <- sapply(GSE19491_pheno$sex, function(x){
  if (x=='Male'){
    return (0)
  } else{
    return (1)
  }
})
names(PTB_object2$class)<- as.character(row.names(GSE19491_pheno))

#validation dataset
PTB_object1_validation$class <- sapply(GSE28623_pheno$sex, function(x){
  if (x=='Male'){
    return (0)}
  else{
    return (1)}})
names(PTB_object1_validation$class)<-
as.character(row.names(GSE28623_pheno))

#checking the datasetObjects
checkDataObject(PTB_object1, "Dataset")
checkDataObject(PTB_object2, "Dataset")
checkDataObject(PTB_object1_validation, "Dataset")

discovery_datasets <- list(PTB_object1, PTB_object2)
names(discovery_datasets) = c(PTB_object1$formattedName,
PTB_object2$formattedName)
PTBMetaObj=list()
PTBMetaObj$originalData <- discovery_datasets
checkDataObject(PTBMetaObj, "Meta", "Pre-Analysis")

#running the analysis
PTBMetaObj <- runMetaAnalysis(PTBMetaObj, maxCores=1)
PTBMetaObj<- filterGenes(PTBMetaObj, FDRThresh = 0.001)
PTBMetaObj<- filterGenes(PTBMetaObj, FDRThresh = 0.05)
PTBMetaObj<- filterGenes(PTBMetaObj, FDRThresh = 0.0001)


#writing positive and negative genes to a file!
write.table(PTBMetaObj[["filterResults"]][["FDR0.001_es0_nStudies1_looaFAL
SE_hetero0"]][["posGeneNames"]], file=paste0(Sys.Date(), "_pos_genes_Meta-
Integrator.csv"), sep=",", row.names=FALSE)
write.table(PTBMetaObj[["filterResults"]][["FDR0.001_es0_nStudies1_looaFAL
SE_hetero0"]][["negGeneNames"]], file=paste0(Sys.Date(),"_neg_genes_Meta-
Integrator.csv"), sep=",",  row.names=FALSE)
```

```
write.table(PTBMetaObj[["filterResults"]][["FDR0.05_es0_nStudies1_looaFALS
E_hetero0"]][["posGeneNames"]], file=paste0(Sys.Date(),
"_FDR=0.05_pos_genes_Meta-Integrator.csv"), sep=",", row.names=FALSE)
write.table(PTBMetaObj[["filterResults"]][["FDR0.05_es0_nStudies1_looaFALS
E_hetero0"]][["negGeneNames"]],
file=paste0(Sys.Date(),"_FDR=0.05_neg_genes_Meta-Integrator.csv"),
sep=",",  row.names=FALSE)
#summarize filter results
summary_results_FDR0.05<-summarizeFilterResults(PTBMetaObj,
"FDR0.05_es0_nStudies1_looaFALSE_hetero0")
summary_results_FDR0.001<-summarizeFilterResults(PTBMetaObj,
"FDR0.001_es0_nStudies1_looaFALSE_hetero0")
summary_results_FDR0.0001<-summarizeFilterResults(PTBMetaObj, "FDR1e-
04_es0_nStudies1_looaFALSE_hetero0")

#write summarized results to files
write.table(summary_results_FDR0.001[["pos"]], "sum-
mary/summary_results_FDR0.001_pos.txt", sep="\t")
write.table(summary_results_FDR0.001[["neg"]], "sum-
mary/summary__results_FDR0.001_neg.txt", sep="\t")

#calculate z-scores
z_scores_dat1<- calculateScore(PTBMetaObj$filterResults[[3]], PTB_object1,
suppressMessages=FALSE)
z_scores_dat2<- calculateScore(PTBMetaObj$filterResults[[3]], PTB_object2,
suppressMessages=FALSE)

#plots
violinPlot(PTBMetaObj$filterResults[["FDR1e-
04_es0_nStudies1_looaFALSE_hetero0"]], PTB_object1_validation, labelColumn
= 'sex')


rocPlot(PTBMetaObj$filterResults[["FDR1e-
04_es0_nStudies1_looaFALSE_hetero0"]], PTB_object1_validation,
        title = "ROC plot for Validation Dataset, FDR: 0.0001")

#heatmap
heatmapPlot(PTBMetaObj, PTBMetaObj$filterResults[[3]])
```

APPENDIX 3: Code used for ssGSEA

```
#ssGSEA
#read the hallmarks gene sets
gene.sets <- read.delim("h.all.v7.4.symbols.gmt", header=FALSE)
gene.sets.list <- lapply(1:50, function(x){
  idxGenes<- which(gene.sets[x,3:202]!="")
  return(as.character(gene.sets[x,idxGenes+2]))
})
names(gene.sets.list)<- gene.sets[,1]

#GSVA
#ssGSEA
set.seed(9837)
row.names(GSE157657_exprData_coding_PTB)<-GSE157657_genes_exprData_coding
GSE157657_gsva.ssgsea<- gsva(GSE157657_exprData_coding_PTB,
                             gene.sets.list,
                             method = "ssgsea",
                             verbose=FALSE)
GSE157657_gsva.ssgsea<-rbind(gender=GSE157657_pheno$sex,
GSE157657_gsva.ssgsea)

row.names(GSE19491_exprData_coding_PTB)<-GSE19491_genes_exprData_coding
GSE194941_gsva.ssgsea<- gsva(GSE19491_exprData_coding_PTB,
                             gene.sets.list,
                             method = "ssgsea",
                             verbose=FALSE)
GSE194941_gsva.ssgsea<-
rbind(sex=GSE19491_pheno$sex,GSE194941_gsva.ssgsea)

row.names(GSE28623_exprData_coding_PTB)<-GSE28623_genes_exprData_coding
GSE28623_gsva.ssgsea<- gsva(GSE28623_exprData_coding_PTB,
                             gene.sets.list,
                             method = "ssgsea",
                             verbose=FALSE)
GSE28623_gsva.ssgsea<- rbind(sex=GSE28623_pheno$sex,GSE28623_gsva.ssgsea)


#writing ssGSEA tables to files
write.table(GSE194941_gsva.ssgsea, file= "GSE194941_gsva.ssgsea.txt",
sep="\t")
write.table(GSE157657_gsva.ssgsea, file= "GSE157657_gsva.ssgsea.txt",
sep="\t")
write.table(GSE28623_gsva.ssgsea, file= "GSE28623_gsva.ssgsea.txt",
sep="\t")
```

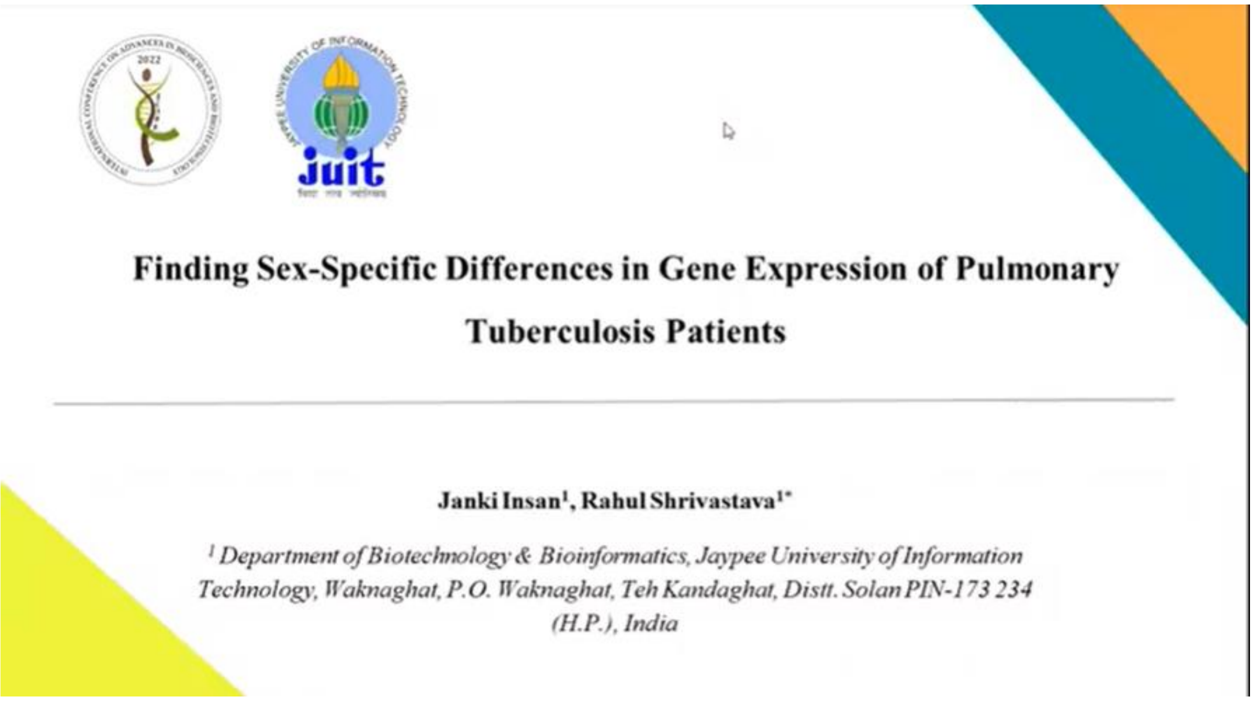
PUBLICATIONS

1. **Sex-Specific Transcriptomic Differences in Pulmonary Tuberculosis Patients**

Janki Insan, Rahul Shrivastava

E-short Presentation at: International Conference on Advances in Biosciences and Biotechnology at Jaypee Institute of Information Technology, Noida, India, Held in Online mode from 20th-22nd January, 2022

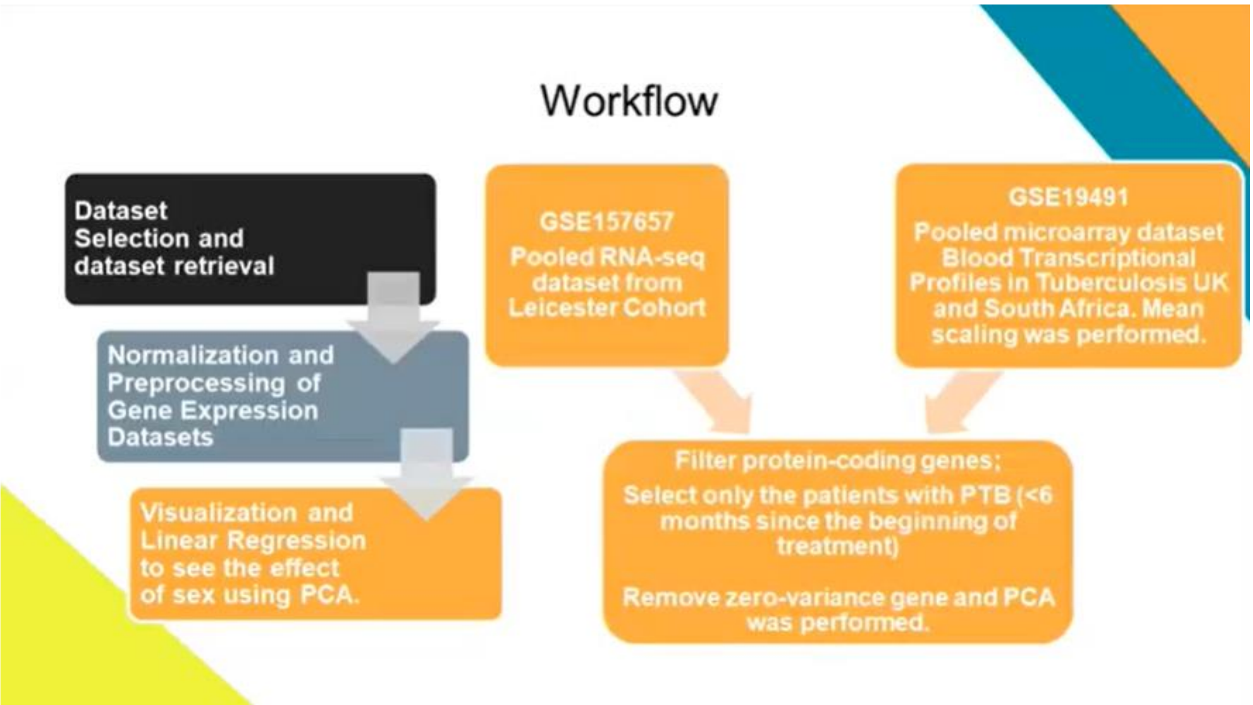
Slides:



Finding Sex-Specific Differences in Gene Expression of Pulmonary Tuberculosis Patients

Janki Insan[1], Rahul Shrivastava[1*]

[1] *Department of Biotechnology & Bioinformatics, Jaypee University of Information Technology, Waknaghat, P.O. Waknaghat, Teh Kandaghat, Distt. Solan PIN-173 234 (H.P.), India*



Workflow

Dataset Selection and dataset retrieval

Normalization and Preprocessing of Gene Expression Datasets

Visualization and Linear Regression to see the effect of sex using PCA.

GSE157657 Pooled RNA-seq dataset from Leicester Cohort

GSE19491 Pooled microarray dataset Blood Transcriptional Profiles in Tuberculosis UK and South Africa. Mean scaling was performed.

Filter protein-coding genes; Select only the patients with PTB (<6 months since the beginning of treatment)

Remove zero-variance gene and PCA was performed.