Bioinformatic Analysis of RNA Sequencing Data to Identify Significant Genes for Tuberculosis in Diabetes Mellitus Comorbidity

Submitted by

HARISANKAR D

(Reg. No. 21378016)

In partial fulfilment for the award of the degree of

Master of Science in Bioinformatics

Under the Supervision of

Dr. R Krishna

Professor

DEPARTMENT OF BIOINFORMATICS

PONDICHERRY UNIVERSITY

PUDUCHERRY

May 2023

DEPARTMENT OF BIOINFORMATICS
PONDICHERRY UNIVERSITY
PUDUCHERRY

Bonafide Certificate

Certified that this Project titled **“Bioinformatic Analysis of RNA Sequencing Data to Identify Significant Genes for Tuberculosis in Diabetes Mellitus Comorbidity”** is the bonafide work of **HARISANKAR D (Roll No. 21378016),** who carried out the work under the supervision of **Dr. R Krishna**. Certified further that to the best of my knowledge, the work reported herein does not form part of any other dissertation or dissertation on the basis of which a degree or award was conferred on an earlier occasion onthis or any other candidate.

Dr. P.T.V. LAKSHMI
Professor and Head
Department of Bioinformatics
Pondicherry University
Puducherry

Dr. R KRISHNA
Professor
Department of Bioinformatics
Pondicherry University
Puducherry

Declaration

I hereby declare that the dissertation entitled **"Bioinformatic Analysis of RNA Sequencing Data to Identify Significant Genes for Tuberculosis in Diabetes Mellitus Comorbidity"** is submitted for partial fulfillment for completion of my project for **Master of Science** in **Bioinformatics**. This work is entirely an original done by me at the Department of Bioinformatics, Pondicherry University, Puducherry. This work has not been submitted elsewhere for any other degree, diploma, associateship, or fellowship.

Place: Puducherry

Date: 19/05/2023

Harisankar D

MSc Bioinformatics

Pondicherry University

Acknowledgement

I would like to express my heartfelt gratitude to **Dr. R Krishna**, Professor at the Department of Bioinformatics, Pondicherry University, for providing me with the opportunity to conduct this research project and for his invaluable guidance and unwavering support throughout.

I would also like to extend my sincere thanks to **Dr. P. T. V. Lakshmi**, Head of the Department of Bioinformatics, Pondicherry University, and all the faculty members of the Department for their support and encouragement throughout my study.

Furthermore, I am deeply grateful to **Ms. Subhashree S Nayak**, Research Scholar at the Department of Bioinformatics, for her invaluable guidance and constant supervision during the research period.

I would like to express my appreciation to my family for their unwavering support and encouragement during my academic period and research work. I would also like to thank my friends and classmates for their support and encouragement.

.

Harisankar D

Table of Contents

1. List of Tables, Figures and Abbreviations

Tables

Figures

Abbreviations

ATB/TB	Active Tuberculosis
LTBI	Latent Tuberculosis Infection
DM/T2DM	Type 2 Diabetes Mellitus
DEG	Differentially Expressed Genes
TBDM	Tuberculosis - Diabetes Mellitus
STRING	Search Tool for the Retrieval of Interacting Genes

2. Abstract

Background : Tuberculosis (TB) and Diabetes Mellitus (DM) are two of the world's most significant health problems, and their comorbidity in a single individual, has become a significant health concern. Early detection of Tuberculosis in T2DM patients is hence necessary. This study aimed to identify significant genes for Active Tuberculosis from analysing RNA sequencing data of whole blood samples from the NCBI SRA database.

Methods : RNAseq samples of TB only, DM only, TBDM comorbid groups and HCs from the SRA NCBI database were pre-processed in Galaxy server. Common Differentially Expressed Genes (DEGs) between TB only and TBDM comorbid groups were found out. Functional enrichment analyses of the DEGs were performed using the DAVID database. Then, the STRING database and Cytoscape were used to construct protein-protein interaction network and to identify hub genes. The expression levels of these hub genes were investigated in the DM only group (under the thresholds of p-value < 0.01 and | fold change (FC) | >1. mirWalk2.0 was used to construct a gene-miRNA network.

Results : Analysis of RNA sequencing samples revealed 105 common DEGs between TB only and TBDM comorbid sets. The functional enrichment was made, which showed a few important biological activities in which the identified DEGs were involved in. A PPI network consisting of 96 nodes and 98 edges was constructed, and the top 10 hub genes were identified based on MCC. Out of these hub genes, the most significant genes for ATB were speculated.

Conclusion : The hub genes EOMES and TNFSF10 have been suggested as potential biomarkers for Active Tuberculosis, along with the identification of ten significant miRNAs that target the ten hub genes.

Keywords : diabetes mellitus, tuberculosis, differentially expressed genes, bioinformatics

2. Introduction

Tuberculosis (TB) is a highly contagious infectious disease caused by bacteria belonging to the genus Mycobacterium, with *Mycobacterium tuberculosis* (MTB) being the primary causative agent [1]. Aside from TB, this particular bacterial genus is also known to cause other severe diseases in mammals such as Leprosy [2]. It is considered one of the deadliest bacterial infections globally, with a substantial number of people contracting it each year, and a significant portion of those succumbing to the disease. This is particularly true for African countries, where the number of cases continues to rise annually. In 2021 alone, there were approximately 1.5 million cases, a significant increase from the 1.4 million cases reported in the previous year [3].

TB primarily exists in two forms or stages, **Latent** and **Active** [4]. Latent TB (LTBI) is when an individual, although infected with MTB, doesn't show symptoms of TB and cannot spread the disease. LTBI affects at least one-fourth of the world's population [4]. Although the infection remains dormant in most, in an estimated 5-10 percent of individuals, LTBI progresses to a higher level of severity, known as the Active stage (ATB). This is the actual infectious phase of TB [4]. Symptoms of low-grade TB are generally vague and non-specific, like a low-grade fever, general fatigue, etc. Although the primary location of TB is the Lungs (Pulmonary TB), there is a high chance of it spreading to other parts of the body depending on its progression, including the brain [5]. As mentioned earlier, TB is highly prevalent in its latent form throughout the world. The progression of TB from the Latent to the Active stage may depend upon a lot of factors, many of which are still not fully understood. The compromised immune system is one such important factor, especially in people with immunosuppressive conditions [6] like HIV/AIDS and DM. Therefore, there is an urgent need to better understand the underlying molecular mechanisms that support the perpetuation of TB, especially in such comorbid cases.

Diabetes is a chronic condition brought on by either insufficient insulin production by the pancreas or inefficient insulin utilization by the body [7]. Insulin controls blood sugar levels. Uncontrolled diabetes frequently causes **hyperglycaemia**, also known as high blood glucose or raised blood sugar, which over time can seriously harm many different bodily systems, including the neurons and blood vessels, and in turn the organs [7]. There are two types of Diabetes Mellitus - Type 1 and Type 2.

Type 1 Diabetes occurs when the body's immune system attacks and destroys insulin-producing cells in the pancreas. Type 2 Diabetes (T2DM) occurs when the body becomes resistant to insulin or doesn't produce enough insulin. Symptoms of diabetes include increased thirst and hunger, frequent urination, fatigue, blurred vision, slow-healing wounds, and numbness in the hands or feet. Diabetes can lead to serious health complications such as cardiovascular disease, kidney disease, nerve damage, and blindness [7]. According to a study by the World Diabetes Federation, diabetes incidence is on a sharp upward trend. If left unchecked, there could nearly be 700 million T2DM patients worldwide by 2045 [8].

Risk factors for T2DM include being overweight or obese, having a family history of diabetes, being physically inactive, and having a poor diet high in sugar and saturated fats. Symptoms of type 2 diabetes include frequent urination, increased thirst, hunger, fatigue, blurred vision, and slow healing of wounds [9]. Almost 8.5% of persons who were 18 years of age and older had diabetes in 2014. A total of 1.5 million deaths were directly related to diabetes in 2019, and 48% of these deaths occurred in those under the age of 70. Diabetes contributed to an additional 460,000 kidney-related fatalities, while high blood sugar is responsible for approximately 20% of all cardiovascular-related deaths [7]. Type 2 Diabetes is complex, in the sense, it is more of a lifestyle disease than a hereditary one, thereby its underlying mechanisms are difficult to understand, especially the genetics aspect of it, as the living environment of people varies drastically across populations and within populations [10]. Hence, more research is needed to fully understand this disease at its core.

Research has demonstrated that diabetes and tuberculosis (TB) have a multifaceted relationship in which diabetes increases the risk of TB and worsens TB outcomes [11]. People with diabetes are more vulnerable to TB infection due to their compromised immune system. Furthermore, diabetes can negatively affect the ability of phagocytes and lymphocytes to contain the TB bacteria, impairing chemotaxis, phagocytosis, activation, and antigen presentation by phagocytes in response to M tuberculosis [11].

Additionally, diabetes might adversely affect T-cell production of interferon γ, and T-cell growth, function, and proliferation. These diabetes-related differences in the immune response to TB might predispose patients to infections, and decreased phagocyte and T-cell function may be contributing factors [11]. The co-occurrence of TB and diabetes poses significant obstacles to the global plan of reducing the annual incidence of ATB to less than one case per million population by 2050 [12]. A study has found that DM patients have a 3.11-fold greater risk of ATB compared to non-DM individuals [13]. Consequently, periodic screening tests for ATB are essential for patients with diabetes.

Blood-based tests for TB include interferon-gamma release assays (IGRAs) and serological tests. IGRAs are blood tests that measure the release of interferon-gamma (IFN-γ) by T-cells in response to TB antigens. However, IGRAs are limited in their sensitivity, particularly in immunocompromised individuals [14]. Serological tests measure the presence of TB-specific antibodies in the blood. While these tests are quick and easy to perform, they are not recommended for the diagnosis of TB due to their poor accuracy and lack of standardization [15]. It is hence clear that we are still lacking a fast and accurate screening method for TB, especially in TB-DM comorbid individuals.

Early diagnosis of TB is therefore essential for successful treatment and containment of the disease. Significant genes can be used as biomarkers for early detection of TB or in monitoring the progression of the disease. This is particularly important in resource-limited settings where diagnostic tools are not readily available, and the disease burden is high. The main objective of this project is to analyze the RNA sequencing data extracted from whole blood samples of patients with the aim of identifying genes that may play a vital role in the diagnosis of Active Tuberculosis in comorbid cases. The insights gained from this study can potentially pave the way for the development of innovative and enhanced diagnostic techniques for TB in the near future.

3. Review of Literature

3.1 Tuberculosis : *An Overview*

Tuberculosis (TB) is a chronic bacterial infection caused by MTB. It primarily affects the lungs but can also spread to other organs in the body [3]. According to the World Health Organization (WHO), TB is one of the top 10 causes of death worldwide and the leading cause of death from a single infectious agent [3]. The transmission of TB occurs through the inhalation of airborne droplets containing the bacteria, usually from an infected person who is coughing, sneezing or speaking [3]. Once inhaled, the bacteria can infect the lungs and replicate, leading to the formation of granulomas (small nodules) and eventually to the destruction of lung tissue[3][5]. TB can present with a wide range of symptoms, which may include persistent coughing, chest pain, fatigue, fever, night sweats, and weight loss. However, not all infected individuals will develop symptoms, and in some cases, the infection can remain dormant for years before becoming active [3][5].

3.2.1 Phases of Tuberculosis

Tuberculosis may broadly be classified into two phases: Active and Latent. Latent tuberculosis infection (LTBI) is characterized as an ongoing immune response triggered by MTB antigens, but without any clinical manifestation of Active Tuberculosis disease. Those with LTBI serve as a source for cases of Active TB.

a) Latent TB

The definition of LTBI by the World Health Organization (WHO) involves a persistent immune response to M. tuberculosis antigens but no clinically apparent Active TB. About 25% of the global population is estimated to have LTBI, and the duration of the latent period varies. While healthy individuals can carry LTBI for their whole lives, a small percentage may reactivate within 2 to 5 years of infection, making people with LTBI a significant source of new ATB cases. Factors influencing LTBI reactivation include bacterial, host, and environmental factors [16]. Although healthy individuals with documented LTBI have a lifetime risk of reactivation of about 5% to 15% [17][18], certain comorbidities and risk factors can increase the likelihood of developing ATB.

HIV infection is the most influential risk factor, increasing the probability of developing ATB by over 100-fold [19]. Moderate-risk individuals include those treated with tumour necrosis factor alpha (TNF-α) inhibitors (used for many autoimmune and inflammatory conditions) [20] or glucocorticoids [20], those with diabetes [13]. The majority of these conditions leading to an increased risk of LTBI reactivation have one common feature, which is **suppressed immunity**.

b) Active TB

ATB is the actual infectious phase of the disease and is a leading cause of death from infectious diseases worldwide [21]. In 2017, an estimated 10 million people, including 5.8 million men, 3.2 million women, and 1 million children, developed TB, with 4 million people going undiagnosed and untreated [22]. The majority of cases were concentrated in eight countries, with India having the highest incidence at 27%, followed by China, Indonesia, the Philippines, Pakistan, Nigeria, Bangladesh, and **South Africa** [22]. While TB is found in all age groups and countries, about 90% of cases occur in adults aged 15 years or older. The report also revealed that 464,633 HIV-positive individuals had TB, with 72% being from Africa [22]. Multidrug-resistant TB (MDR-TB) is another concerning issue, with approximately 558,000 people developing rifampicin-resistant TB, of whom an estimated 458,000 had MDR-TB [22]. Unfortunately, only 50% of patients with MDR-TB are cured after treatment with WHO-approved regimens [23], [24]. TB can affect any part of the body, with 5% to 45% of cases having extrapulmonary TB (EPTB) [25], [26], [27] affecting all organs of the body. Common sites of EPTB include lymph nodes, pleura, bones, meninges, and the urogenital tract [28].

3.2.3 Active Tuberculosis – Type 2 Diabetes Mellitus Comorbidity

The World Health Organization (WHO) states that individuals with diabetes mellitus (DM) are at a significantly higher risk of developing tuberculosis (TB), with a risk that is two to three times greater than that of people without DM [29]. The underlying cause of this heightened risk is that DM can compromise the immune system, thus increasing an individual's susceptibility to contracting infections such as TB [30]. Furthermore, TB infection can exacerbate DM by raising blood glucose levels, resulting in hyperglycaemia, in individuals with DM [31]. This is because TB infection may cause insulin resistance and a decrease in insulin secretion, resulting in suboptimal glucose control in those with DM [32]. The interaction between TB and DM can lead to negative health outcomes, including a greater likelihood of treatment failure, relapse, and mortality compared to those with TB alone [33]. DM can also complicate TB diagnosis and treatment, resulting in delayed diagnosis and suboptimal treatment [34]. Screening for TB in those with DM and vice versa is recommended to address the comorbidity between TB and DM [34]. Integrated care models that address both diseases simultaneously have been demonstrated to improve treatment outcomes and decrease mortality rates [34].

3.3 Significant Genes for Tuberculosis

Identifying significant genes for ATB could be useful for early diagnosis, monitoring disease progression, and evaluating the efficacy of therapies in ATB. In this study, terms such as “Blood based biomarkers for Tuberculosis”, “Significant genes for Tuberculosis”, “Hub genes” etc, were used to retrieve all the literature. Several significant genes have been proposed for Tuberculosis and also there have been studies in which certain differentially expressed genes have been speculated to be potential biomarkers for Tuberculosis, including cases where ATB is comorbid with immunosuppressive conditions like T2DM and HIV AIDS.

One study [35] utilized an integrative bioinformatics approach to analyze gene expression profiles of individuals with Active TB and Latent TB infection in two distinct populations of the same country (South Africa). By comparing the gene expression patterns, the researchers identified a set of seven genes that showed differential expression between Active TB and Latent TB infection. Furthermore, the functional annotation and pathway enrichment analysis revealed that the differentially expressed genes were involved in key pathways related to immune response and signalling. Specifically, the upregulated genes (FCGR1B, ANKRD22, CARD17, IFITM3, TNFAIP6) were associated with interferon signalling and other immune-related pathways, while the downregulated genes (FCGBP and KLF12) were linked to Wnt and B-cell signalling pathways.

Another study [36] discusses the use of bioinformatics tools to analyze the differential gene expression between patients with tuberculosis (TB) and healthy individuals to unravel the pathogenesis of TB. The analysis identified 190 differentially expressed genes (DEGs) including 36 up-regulated genes and 154 down-regulated genes. The majority of the DEGs were involved in molecular function, specifically protein binding. GO enrichment analysis and KEGG pathway enrichment analysis to determine the significant enrichments in the DEGs, such as T cell activation, chemotaxis, leukocyte activation involved in immune response, cytokine secretion, head development, etc. The establishment of a protein-protein interaction network identified six key genes including LRRK2, FOXC1, CCR7, FYN, CXCR5, and Fas/Fas-ligand system. The study also discusses the role of transcription factors (TFs) such as FOXC1, which is essential for mesenchymal lineage specification and organ development during the pathogenic processes of TB.

A comparative study similar to this experiment was done [37] in which gene expression datasets for healthy controls (HCs), TB patients, DM patients, TB+DM patients, and metformin-treated cells were obtained from the Gene Expression Omnibus (GEO) database and differentially expressed genes (DEGs) were identified from pairwise dataset comparisons. DEGs were verified by comparing them to DEGs for TB+DM vs HCs. Enrichment analysis of DEGs common to all three dataset comparisons was conducted using DAVID. The protein–protein interaction (PPI) network was established via STRING and visualised in Cytoscape. Hub genes were identified using the Cytoscape plug-in CytoHubba and then were verified using reverse transcription-quantitative polymerase chain reaction (RT-qPCR) analysis. Targeted miRNA prediction analysis identified metformin treatment-induced gene expression changes in peripheral blood mononuclear cells. A total of 422 DEGs were common to all three dataset comparisons. Functional enrichment analysis revealed these DEGs were enriched for functional terms of type I interferon signalling pathway, innate immune response, inflammatory response, and infectious diseases. Ten hub genes identified using PPI network analysis were screened for interactions with metformin target gene INS using CytoHubba based on maximal clique centrality (MCC) score. Subsequently, five hub genes were predicted to functionally interact with INS, including STAT1, IFIT3, RSAD2, IFI44L, and XAF1, as verified by RT-qPCR. Meanwhile, seven miRNAs (miR-3680-3p, miR-3059-5p, miR-629-3p, miR-29b-2-5p, miR514b-5p, miR-4755-5p, miR-4691-3p) were associated with regulation of hub genes. Notably, six hub genes (STAT1, IFIT3, RSAD2, ISG15, IFI44, IFI6) were down-regulated in cells exposed to both metformin and MTB antigens.

In another study conducted on two datasets [38], researchers analyzed the data of 113 HIV/TB specimens and 109 HIV/LTBI specimens to identify key genes involved in HIV/TB progression. A total of 83 differentially expressed genes (DEGs) were identified, with 64 up-regulated and 19 down-regulated DEGs common to both datasets. GO and KEGG pathway analyses were conducted on the DEGs, and the results showed that the up-regulated DEGs were particularly enriched in various biological processes and molecular functions such as innate immune response, extracellular exosome, serine-type endopeptidase activity, and protein tyrosine kinase binding. On the other hand, down-regulated DEGs were enriched in various pathways related to T cell receptor V(D)J recombination, leukocyte chemotaxis, beta-catenin binding, among others. A significant module of 12 vital genes (CAMP, CTSG, DEFA1, DEFA1B, DEFA3, DEFA4, ELANE, HP, HPSE, OLFM4, PGLYRP1, TCN1) was constructed from the PPI network through Cytoscape MCODE analysis.

4. Objective

- To use Bioinformatic and Statistical tools to identify and characterize significant genes for Active Tuberculosis in Type 2 Diabetes using RNAseq Data Analysis of whole blood samples
- To find Micro-RNAs that target the significant genes

It involves :

1. Pre-processing the SRA data
2. Identification of Differentially Expressed Genes (DEGs)
3. Functional Enrichment Analysis of the DEGs
4. Protein-Protein Interaction (PPI) Network Construction
5. Hub Gene Identification
6. miRNA Interaction Network Construction

4.1 Workflow

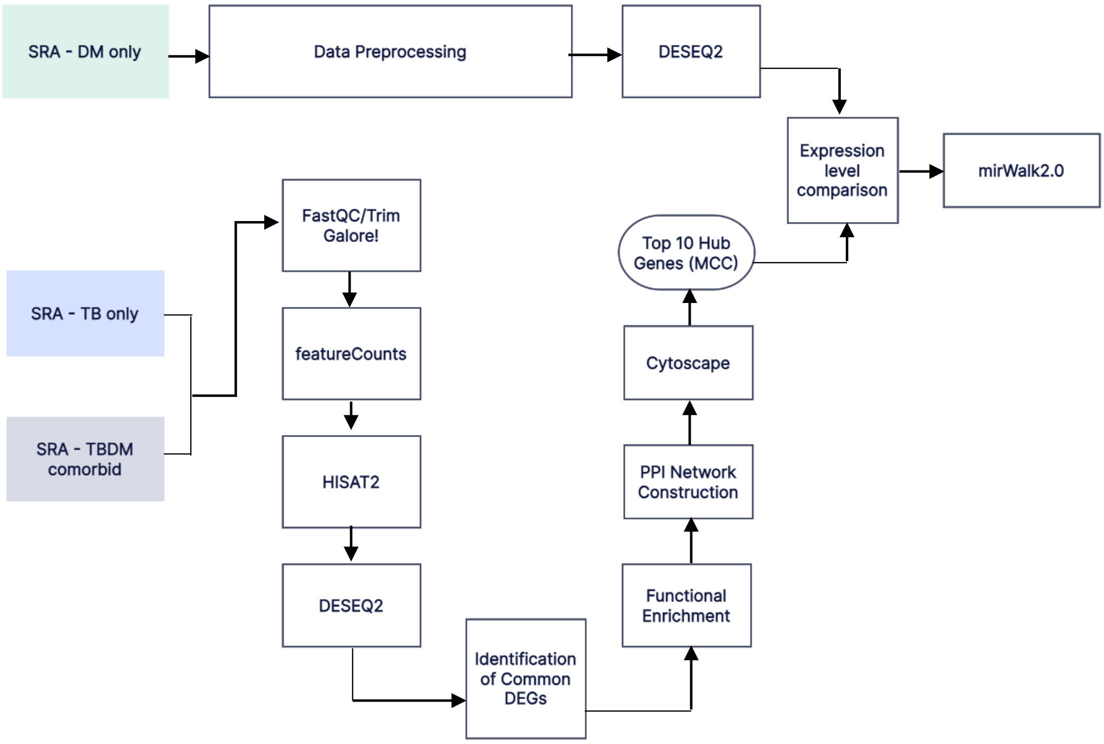


Fig 1. Workflow of the experiment

5. Materials and Methods

5.1 Data Collection

The NCBI Sequence Read Archive (SRA) [39] is a public repository for storing raw sequencing data from a variety of high-throughput sequencing platforms, such as Illumina, PacBio, and Oxford Nanopore. The SRA provides a centralized location for researchers to deposit and access raw sequencing data, which can then be used for a wide range of downstream analyses, such as transcriptome profiling, metagenomics, and genome assembly [39].

The primary objective was to get a particular Illumina dataset containing sequence reads of TB patients alone, DM patients alone, TB-DM comorbid patients and Healthy Controls (HC). South Africa was the preferred geographical region.

NCBI SRA (Sequence Read Achieve), a public repository of High Throughput Next Generation Sequencing Data, was hence searched using the keywords : "**Tuberculosis**", "**Diabetes Mellitus**", "**Africa**" and "**Illumina**". The project PRJNA470512 was chosen. It had the SRA dataset **GSE114192**. It is based on the GPL18573 Platform (Illumina NextSeq 500). This dataset was based on a landmark study conducted by Clare Eckold and team [40], in which the aforementioned categories were found out in four different countries. South Africa was among them, the rest three being Indonesia, Romania and Peru. From within the South African cohort, 7 samples each of DM, TB, TBDM and HC were chosen and uploaded into the Galaxy server [41] for further analysis, separately, in the form of SRR accession text files (Table 1).

Sample

	TB	DM	TB-DM	HC
SRR IDs	SRR7134839	SRR7134902	SRR7134876	SRR7134907
	SRR7134887	SRR7134891	SRR7134874	SRR7134906
	SRR7134859	SRR7134882	SRR7134865	SRR7134896
	SRR7134833	SRR7134877	SRR7134853	SRR7134895
	SRR7134816	SRR7134864	SRR7134804	SRR7134893
	SRR7134802	SRR7134861	SRR7134794	SRR7134885
	SRR7134793	SRR7134847	SRR7134792	SRR7134852

Table 1. Accession numbers of all the 28 samples

Reference Genome File

The Ensembl Human Reference Genome was chosen for reference. It was downloaded from the EBI-Ensembl website, and was uploaded to the Galaxy server.

The File : Homo_sapiens.GRCh38.dna_sm.primary_assembly.fa.gz

Sequence Annotation File

The Ensembl Sequence Annotation File was downloaded from ENSEMBL home page and was uploaded to the Galaxy server.

The File : Homo_sapiens.GRCh38.109.gtf.gz

5.2 Data Pre-processing

5.2.1 Quality Check

The quality of the reads were checked using the FastQC [42] tool.

FastQC is a tool that is commonly used in bioinformatics to assess the quality of DNA sequence data. It works by analyzing the raw data generated by high-throughput sequencing technologies and producing a set of diagnostic graphs and tables that allow researchers to quickly identify potential issues with the data. One of the key features of FastQC is its ability to provide a visual representation of the sequence quality scores, which are used to measure the accuracy of the base calls made by the sequencing machine. By examining these scores, researchers can identify regions of the sequence that may be of lower quality and may need to be filtered or trimmed before downstream analysis [42]. **MultiQC** [43] is a tool that takes in FastQC results and produces a single, integrated and interactive output.

Some of the reads were found to have slight quality issues related to per base sequence quality and presence of adaptor sequences. These reads were Trimmed using the Trim Galore! tool [44], following which the quality was again rechecked using FastQC. Trim Galore! is a bioinformatics tool used for trimming adapters and low-quality regions in Next-Generation Sequencing data.

5.2.2 Sequence Alignment and Mapping

The high-quality RNA sequencing data were aligned to the human reference genome (GRCh38) file, from EBI-ENSEMBL. The HISAT2 [45] software was used for the alignment. The HISAT2 alignment is quick and sensitive for mapping next-generation sequencing reads to a population of human genomes and a single reference genome. Compared to other alignment tools, the HISAT2 is more precise and faster [45]. The RNA sequence file is the input that the HISAT2 tool accepts.

HISAT2 utilizes a clever algorithm called Hierarchical Indexing for Spliced Alignment of Transcripts (HISAT) to align high-throughput sequencing reads to a reference genome. This algorithm takes advantage of a hierarchical indexing scheme that makes it fast and memory-efficient, allowing it to handle large-scale sequencing data. [45].

After the alignment, the output files are generated in BAM format, which is the binary form of the Sequence Alignment Map (SAM) format file. Because BAM files are compressed, they take up less storage space and are easier to manage [45].

The Output Summary File option was in order to visualize the percentage of alignment for all the reads.

5.2.3 Quantification

Quantification was done using featureCounts [46]. A read is a cluster sequence produced by the sequencing process that represents an element of a unique fragment. The number of reads that cover a certain feature, such as a gene, is quantified using the read count quantitation technique, which is widely used and may be changed based on a variety of criteria to allow for comparison with other data sets. The featureCounts tool is used to get such read counts from RNA sequencing data. It is more efficient than earlier approaches, requiring less computer memory and generating tabular output. [46].

The output are the Gene ID and their related counts, indicating the number of copies of the gene. This method provides a quantitative view of the reading distribution across all probes and is a good initial approach to analyzing the data. The BAM files created by the HISAT2 alignment tool are used as input and the output has a summary file and read counts files.

5.3 Differential Expression Analysis

DESeq2 [47] (accessed via Galaxy) was the tool of choice for performing Differential Gene Expression Analysis between the 3 test sets and 1 set of healthy controls, all carried out separately (as TB vs HC, DM vs HC, TB-DM vs HC).

DESeq2 is a powerful bioinformatics tool used for differential gene expression analysis in RNA-Seq experiments. It is designed to detect changes in gene expression between different experimental conditions or treatments. DESeq2 uses statistical methods to normalize raw RNA-Seq read counts and identify differentially expressed genes (DEGs) with high accuracy and sensitivity [47].

Volcano plots [48] were generated using the ggplot2 tool in R, after specifying the necessary inputs (FDR, p-value, logFC).

A Volcano Plot is a commonly used data visualization tool for displaying the results of differential gene expression analysis [48]. A volcano plot typically shows the log-fold change (LFC) on the x-axis and the negative log10-transformed p-value on the y-axis for each gene in a dataset. Genes with a large LogFC and a small p-value (i.e., highly significant) are located in the upper left or upper right corner of the plot, while genes with a small LFC and a large p-value (i.e., not significant) are located in the lower left or lower right corner [48].

5.3.1 Identification of Common Differentially Expressed Genes between TB-HC set and TBDM-HC Comorbid set

The Venn Online [49] tool was used to identify common upregulated and downregulated DEGs between TB-HC and TBDM-HC sets.

5.4 Functional Enrichment Analysis

DAVID (Database for Annotation, Visualization and Integrated Discovery) [50][51] combines biological data with analytical tools to provide systematic and thorough functional annotation data on biological systems. Gene Ontology (GO), which provides terms related to functions, and Kyoto Encyclopaedia of Genes and Genomes (KEGG), which provides information on functional pathways, were the tools used in this study to perform functional enrichment analysis of DEGs. The GO terms found here were used to annotate activities falling into the biological Processes (BP), cellular component (CC), or molecular function categories (MF). p-values lower than 0.05 were regarded as statistically significant.

5.5 PPI Network Construction and Hub Gene Identification

The STRING database [52] (version 11.5) aims to integrate all known and predicted associations between proteins, including both physical interactions as well as functional associations. To achieve this, STRING collects and scores evidence from several sources. In addition to the protein-protein interaction data, the STRING database also provides functional annotations for proteins, including Gene Ontology (GO) terms, protein domains, and pathways. These annotations can be used to interpret the biological significance of the protein-protein interactions and to identify potential functional modules or pathways [52].

To determine the interrelationships between differentially expressed genes (DEGs) and proteins in the STRING database, a minimum interaction score of 0.40 was set. The resulting protein-protein interaction (PPI) networks were visualized using **Cytoscape** (version 3.9.1). **Cytoscape** [53] is an open-source bioinformatics software platform for visualizing molecular interaction networks and integrating with gene expression profiles and other data.

Top 10 Hub genes were identified using **CytoHubba** (version 0.1), under the maximal clique centrality (**MCC**) [54] algorithm and degree. **CytoHubba** [54] is a plugin for Cytoscape. CytoHubba provides a collection of network analysis algorithms and tools for identifying important nodes or subnetworks in a protein-protein interaction (PPI) network. CytoHubba includes several different methods for identifying important nodes in a PPI network, including degree centrality, betweenness centrality, and closeness centrality. It also includes algorithms for identifying network motifs, topological overlaps, and clusters within a PPI network. The plugin can also be used to prioritize genes or proteins based on their relevance to a specific disease or phenotype [54].

5.6 Prediction of miRNAs and Construction of a Gene–miRNA Interaction Network

miRWalk [55] is a website that allows users to predict and analyze interactions between microRNAs (miRNAs) and their target genes. It combines information from various sources, including both computational predictions and verified miRNA-target interactions, to create a comprehensive tool for studying how miRNAs regulate gene expression. These tiny RNA molecules can bind to the 3' untranslated region (UTR) of target mRNA and control their expression by reducing translation or accelerating degradation. With miRWalk, researchers have access to a database of predicted and verified miRNA-target interactions and various tools to analyze miRNA-target networks [55].

The identified hub genes were uploaded to this database to predict the miRNAs that could target them. The filter was set to a score of 0.90, the target gene binding region was the 3′ untranslated region (UTR), and miRDB was used for intersection with other databases. The results were then analyzed using Cytoscape.

6. RESULTS

6.1 Data Pre-processing

6.1.1 Quality Check

FastQC was used to assess the quality of the data. The report for all the samples were summarized using MultiQC tool **(**Table 2).

Sample Name	% Dups	% GC	M Seqs
SRR7134792	22.5%	53%	20.4
SRR7134793	22.0%	52%	18.1
SRR7134794	25.5%	53%	13.9
SRR7134802	45.2%	53%	20.0
SRR7134804	34.6%	56%	19.3
SRR7134816	22.9%	52%	18.4
SRR7134833	24.6%	54%	18.7
SRR7134839	24.6%	54%	19.2
SRR7134847	26.7%	54%	16.3
SRR7134852	24.3%	54%	27.7
SRR7134853	24.6%	56%	16.5
SRR7134859	26.3%	54%	17.6
SRR7134861	22.9%	56%	18.6

Sample Name	% Dups	% GC	M Seqs
SRR7134864	21.4%	54%	16.7
SRR7134865	38.3%	55%	15.1
SRR7134874	22.2%	51%	15.4
SRR7134876	55.9%	56%	16.8
SRR7134877	26.0%	54%	15.4
SRR7134882	40.7%	59%	18.6
SRR7134885	26.0%	56%	15.9
SRR7134887	22.5%	56%	18.3
SRR7134891	52.8%	55%	17.2
SRR7134893	26.0%	57%	14.9
SRR7134895	31.8%	55%	25.6
SRR7134896	45.1%	56%	19.8
SRR7134902	30.4%	56%	20.1
SRR7134906	17.2%	53%	20.4
SRR7134907	21.2%	53%	17.5

Table 2. General quality statistics generated using MultiQC**.** The FastQC raw data for all 28 samples were given as the input. The average GC content in human genomes range from 30 to 60 percent [56].

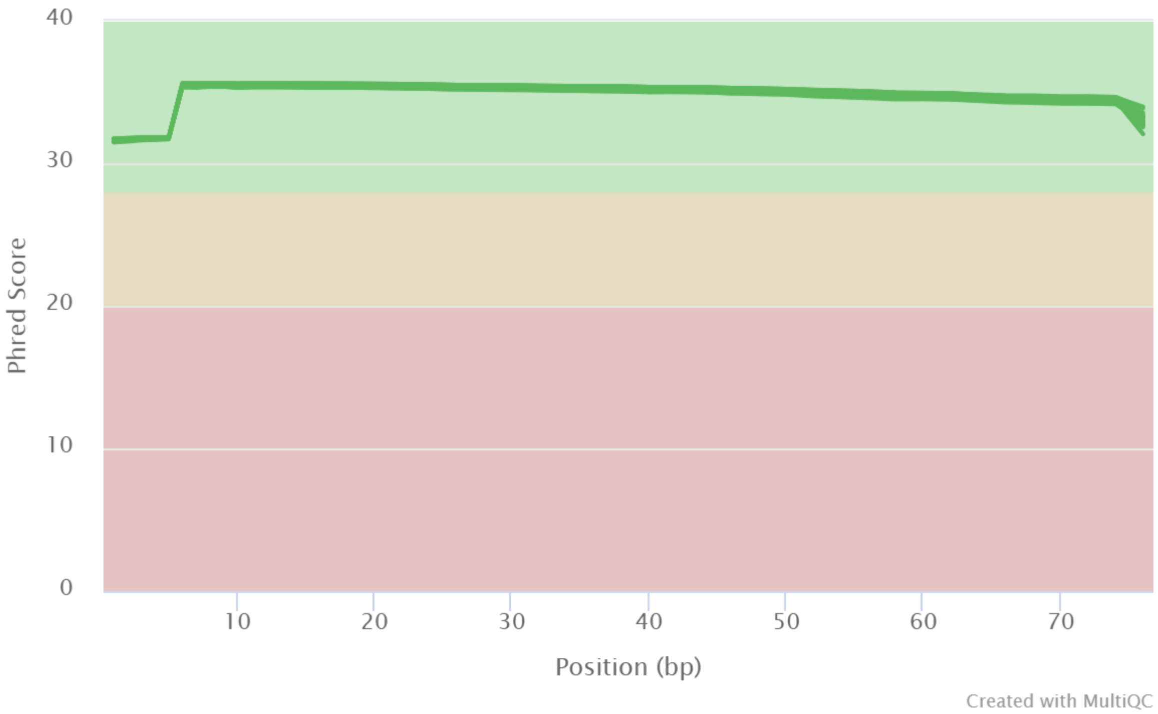
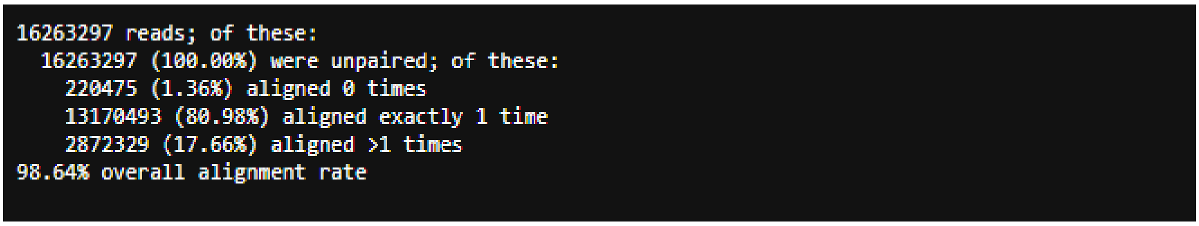


Fig 2. Mean Quality Scores. MultiQC integrates the quality scores for all individual reads into a single interactive output. The Phred score [57] is a logarithmic scale used to express the probability that a base call is incorrect, where a higher Phred score indicates a higher probability that the base call is correct.

6.1.2 Sequence Alignment and Mapping

Alignment was performed using HISAT2. The average alignment rate for all the 28 reads were **97.5%.**

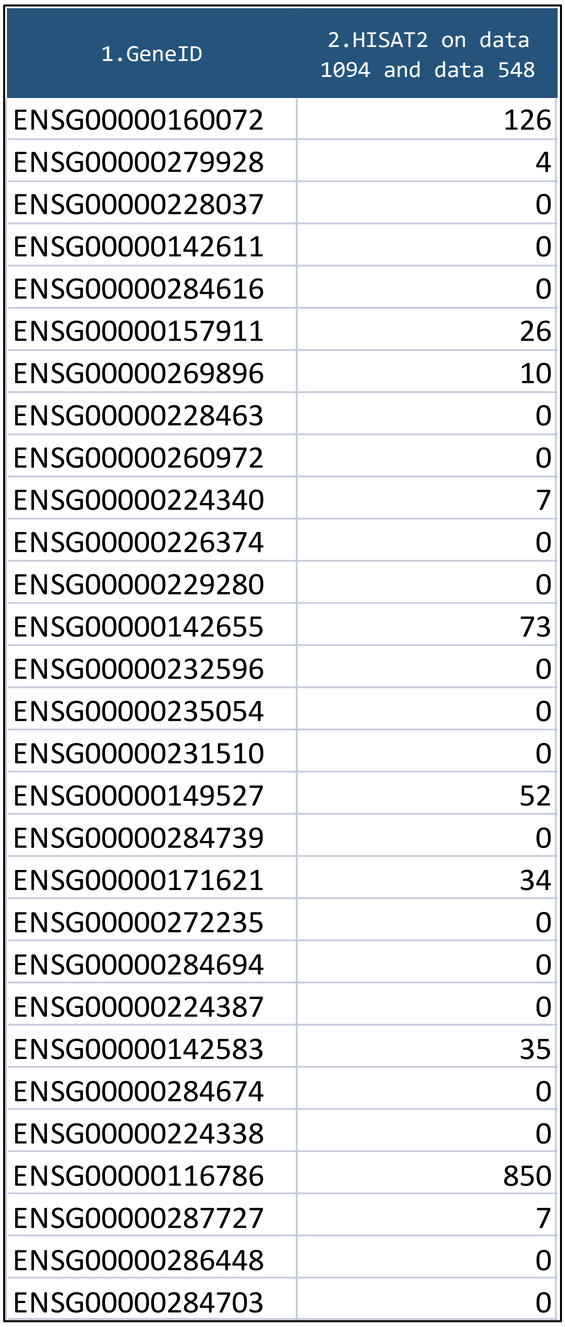


```
16263297 reads; of these:
  16263297 (100.00%) were unpaired; of these:
    220475 (1.36%) aligned 0 times
    13170493 (80.98%) aligned exactly 1 time
    2872329 (17.66%) aligned >1 times
98.64% overall alignment rate
```

Fig 3. HISAT2 alignment rate for one read (98.6%). This is the output summary file for one alignment. The overall alignment rate in HISAT2 refers to the percentage of reads that were successfully aligned to the reference genome. A higher overall alignment rate indicates that more reads were successfully mapped to the reference genome, which can be indicative of a high-quality RNA sequencing experiment.

6.1.3 Quantification

Quantification of the reads were performed using featureCounts tool. The result is a table that matches the gene IDs to their corresponding gene counts. RNA-Seq quantifies the level of gene abundance by tallying the reads that correspond to that gene.



1.GeneID	2.HISAT2 on data 1094 and data 548
ENSG00000160072	126
ENSG00000279928	4
ENSG00000228037	0
ENSG00000142611	0
ENSG00000284616	0
ENSG00000157911	26
ENSG00000269896	10
ENSG00000228463	0
ENSG00000260972	0
ENSG00000224340	7
ENSG00000226374	0
ENSG00000229280	0
ENSG00000142655	73
ENSG00000232596	0
ENSG00000235054	0
ENSG00000231510	0
ENSG00000149527	52
ENSG00000284739	0
ENSG00000171621	34
ENSG00000272235	0
ENSG00000284694	0
ENSG00000224387	0
ENSG00000142583	35
ENSG00000284674	0
ENSG00000224338	0
ENSG00000116786	850
ENSG00000287727	7
ENSG00000286448	0
ENSG00000284703	0

Fig 4. Quantification of a read based on the given annotation file

6.2 Differential Expression Analysis

DESeq2 tool was used to identify which all genes were getting differentially expressed between the tests and healthy controls. The annotated tabular files (Figure 5b) were downloaded and visualized in Microsoft Excel.

6.2.1 Identification of Differentially Expressed Genes

Differentially Expressed Genes were identified to be those with a | fold change (FC) | >1 0.5 and -0.5 and a p-value of less than 0.01, among the expressed genes. The DEGs were either upregulated or downregulated. Those with a log fold change less than -0.5 were classified as down-regulated DEGs, whereas those with a log fold change greater than 0.5 were classified as up-regulated DEGs.

A total of 230 differentially expressed genes were obtained in the case of TB vs HC, out of which 86 are downregulated and the remaining 144 are upregulated.

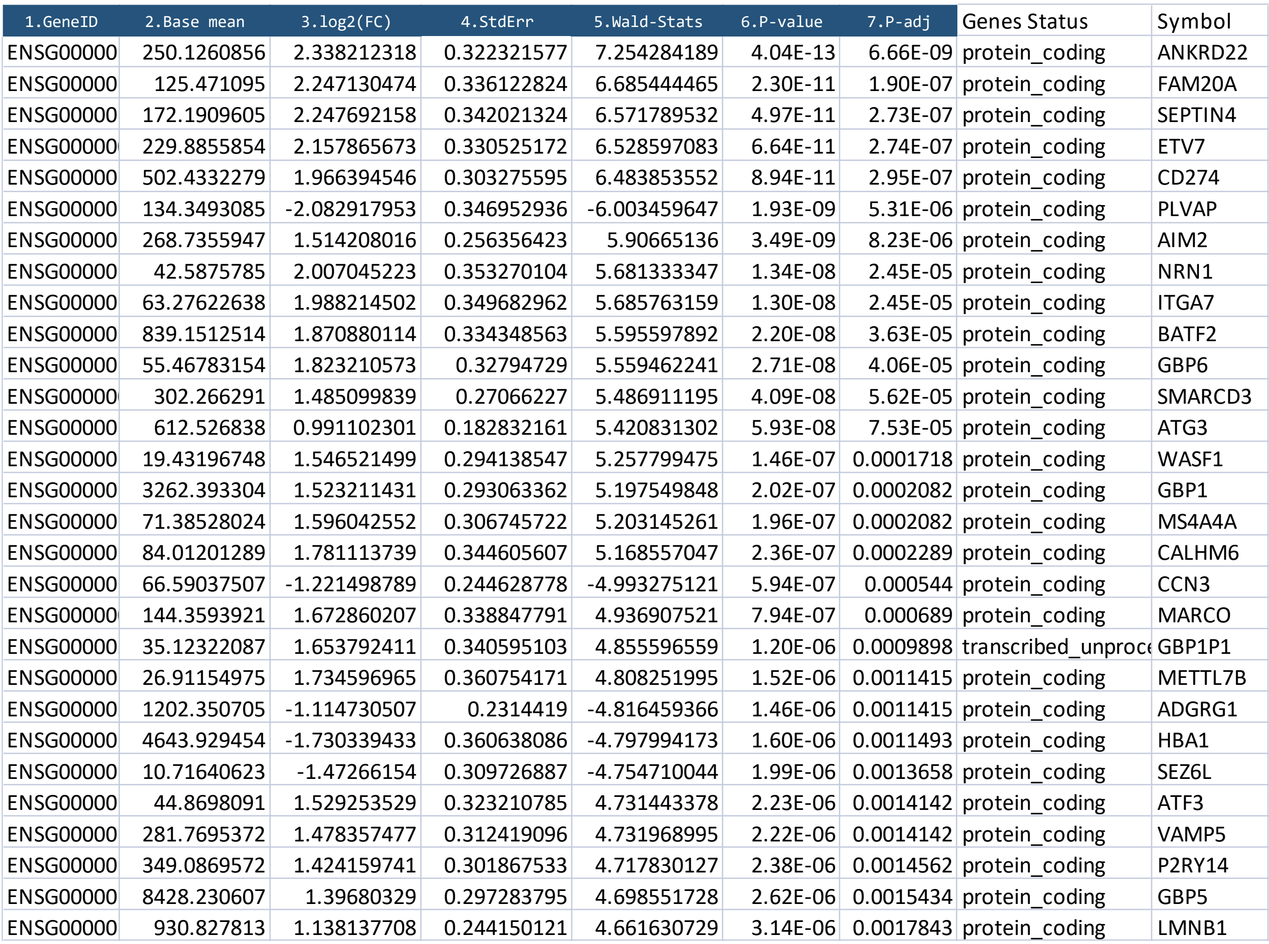
A total of 684 DEGs were observed for TBDM vs HC. Out of 684, 249 were downregulated and the remaining 435 DEGs were upregulated.

A total of 506 DEGs were identified in case of DM vs HC, out of which 217 were downregulated and 289 genes were upregulated.

The Venn Online tool was used to identify common upregulated and downregulated DEGs between TB-HC and TBDM-HC sets. A total of 105 common DEGs were identified between the two sets and these common DEGs were considered for further functional enrichment, pathway, and interaction analysis as well as for PPI network construction.

Volcano Plots

Volcano Plots for the respective cases have also been constructed (Fig. 7).



1.GeneID	2.Base mean	3.log2(FC)	4.StdErr	5.Wald-Stats	6.P-value	7.P-adj	Genes Status	Symbol
ENSG00000	250.1260856	2.338212318	0.322321577	7.254284189	4.04E-13	6.66E-09	protein_coding	ANKRD22
ENSG00000	125.471095	2.247130474	0.336122824	6.685444465	2.30E-11	1.90E-07	protein_coding	FAM20A
ENSG00000	172.1909605	2.247692158	0.342021324	6.571789532	4.97E-11	2.73E-07	protein_coding	SEPTIN4
ENSG00000	229.8855854	2.157865673	0.330525172	6.528597083	6.64E-11	2.74E-07	protein_coding	ETV7
ENSG00000	502.4332279	1.966394546	0.303275595	6.483853552	8.94E-11	2.95E-07	protein_coding	CD274
ENSG00000	134.3493085	-2.082917953	0.346952936	-6.003459647	1.93E-09	5.31E-06	protein_coding	PLVAP
ENSG00000	268.7355947	1.514208016	0.256356423	5.90665136	3.49E-09	8.23E-06	protein_coding	AIM2
ENSG00000	42.5875785	2.007045223	0.353270104	5.681333347	1.34E-08	2.45E-05	protein_coding	NRN1
ENSG00000	63.27622638	1.988214502	0.349682962	5.685763159	1.30E-08	2.45E-05	protein_coding	ITGA7
ENSG00000	839.1512514	1.870880114	0.334348563	5.595597892	2.20E-08	3.63E-05	protein_coding	BATF2
ENSG00000	55.46783154	1.823210573	0.32794729	5.559462241	2.71E-08	4.06E-05	protein_coding	GBP6
ENSG00000	302.266291	1.485099839	0.27066227	5.486911195	4.09E-08	5.62E-05	protein_coding	SMARCD3
ENSG00000	612.526838	0.991102301	0.182832161	5.420831302	5.93E-08	7.53E-05	protein_coding	ATG3
ENSG00000	19.43196748	1.546521499	0.294138547	5.257799475	1.46E-07	0.0001718	protein_coding	WASF1
ENSG00000	3262.393304	1.523211431	0.293063362	5.197549848	2.02E-07	0.0002082	protein_coding	GBP1
ENSG00000	71.38528024	1.596042552	0.306745722	5.203145261	1.96E-07	0.0002082	protein_coding	MS4A4A
ENSG00000	84.01201289	1.781113739	0.344605607	5.168557047	2.36E-07	0.0002289	protein_coding	CALHM6
ENSG00000	66.59037507	-1.221498789	0.244628778	-4.993275121	5.94E-07	0.000544	protein_coding	CCN3
ENSG00000	144.3593921	1.672860207	0.338847791	4.936907521	7.94E-07	0.000689	protein_coding	MARCO
ENSG00000	35.12322087	1.653792411	0.340595103	4.855596559	1.20E-06	0.0009898	transcribed_unproc	GBP1P1
ENSG00000	26.91154975	1.734596965	0.360754171	4.808251995	1.52E-06	0.0011415	protein_coding	METTL7B
ENSG00000	1202.350705	-1.114730507	0.2314419	-4.816459366	1.46E-06	0.0011415	protein_coding	ADGRG1
ENSG00000	4643.929454	-1.730339433	0.360638086	-4.797994173	1.60E-06	0.0011493	protein_coding	HBA1
ENSG00000	10.71640623	-1.47266154	0.309726887	-4.754710044	1.99E-06	0.0013658	protein_coding	SEZ6L
ENSG00000	44.8698091	1.529253529	0.323210785	4.731443378	2.23E-06	0.0014142	protein_coding	ATF3
ENSG00000	281.7695372	1.478357477	0.312419096	4.731968995	2.22E-06	0.0014142	protein_coding	VAMP5
ENSG00000	349.0869572	1.424159741	0.301867533	4.717830127	2.38E-06	0.0014562	protein_coding	P2RY14
ENSG00000	8428.230607	1.39680329	0.297283795	4.698551728	2.62E-06	0.0015434	protein_coding	GBP5
ENSG00000	930.827813	1.138137708	0.244150121	4.661630729	3.14E-06	0.0017843	protein_coding	LMNB1

Table 3. Table showing DEGs between TB and HC after annotation

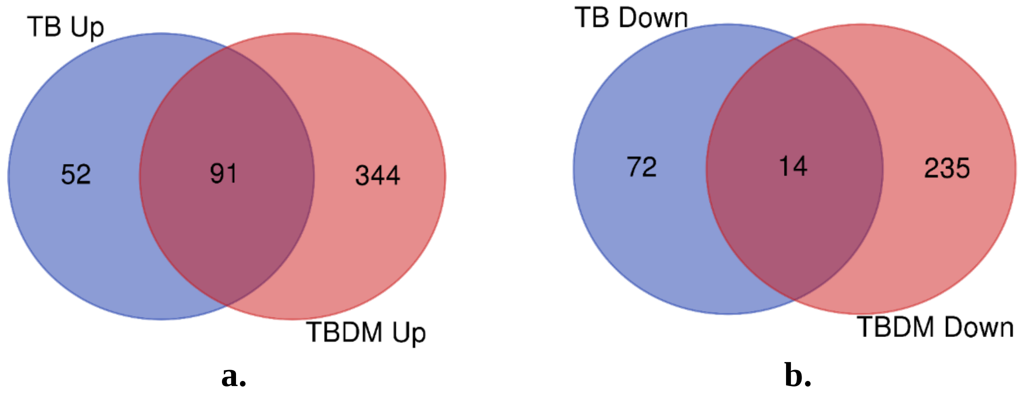


Fig 5. Common upregulated **(a)** and downregulated **(b)** DEGs between TB HC and TB-DM HC groups. A total of 91 genes were commonly upregulated between TB only and TBDM groups while 14 DEGs were commonly downregulated between the two groups.

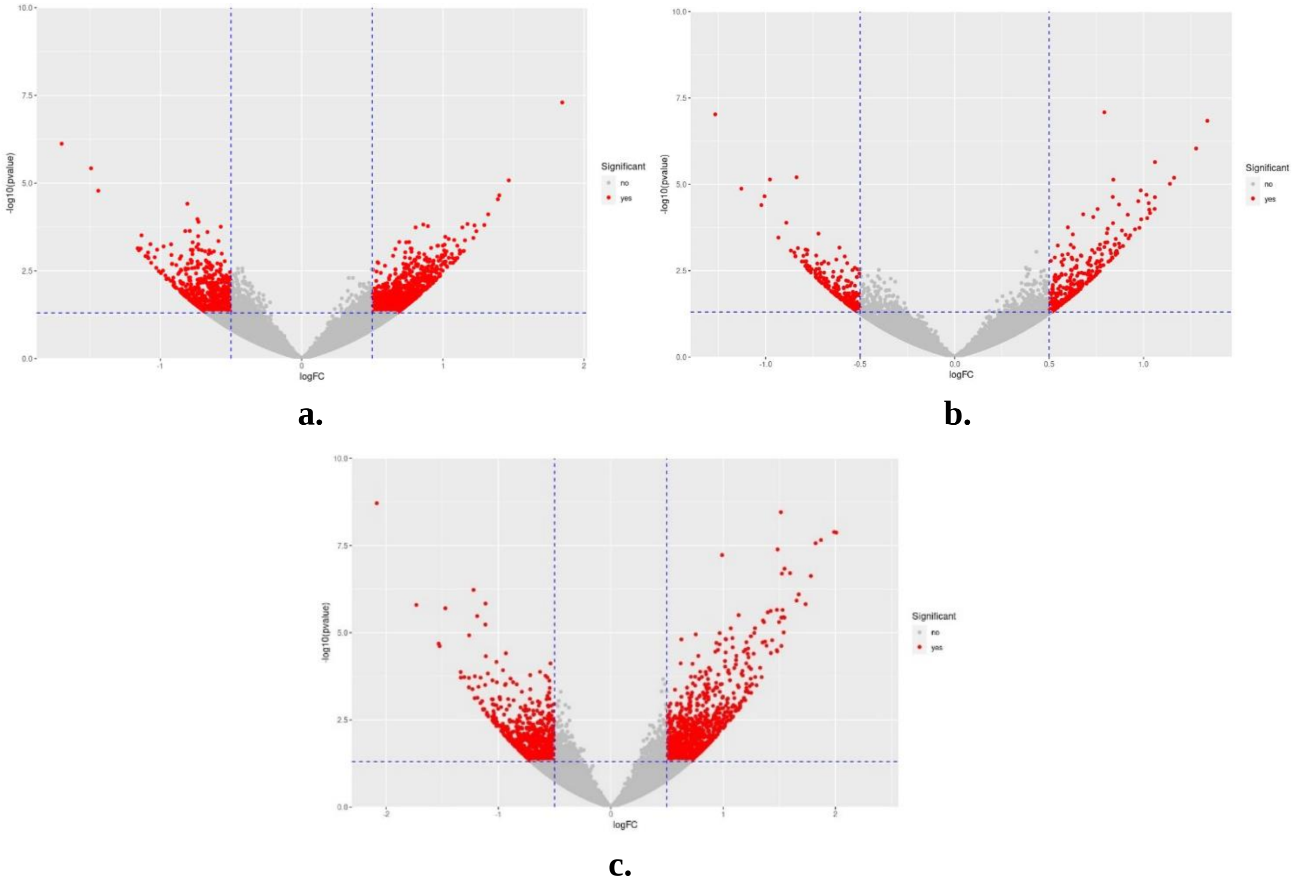


Fig 6. Volcano Plots for TB only vs HC **(a),** DM only vs HC **(b)** and TBDM vs HC **(c).** The genes that lie towards the left of the plot (logFC < - 0.5) are the significant downregulated DEGs while those that lie towards the right (logFC > 0.5) are the significant upregulated DEGs

6.3 Functional Enrichment Analysis

The 105 common DEGs were uploaded to the DAVID server to gain insight into the Biological Processes, Molecular Functions and Cellular Components (Fig 7). The genes were uploaded as Official Gene Symbols and the cut off for p-value was < 0.05.

In the Biological Process (BP) term, the results demonstrated that the DEGs were mainly involved in :

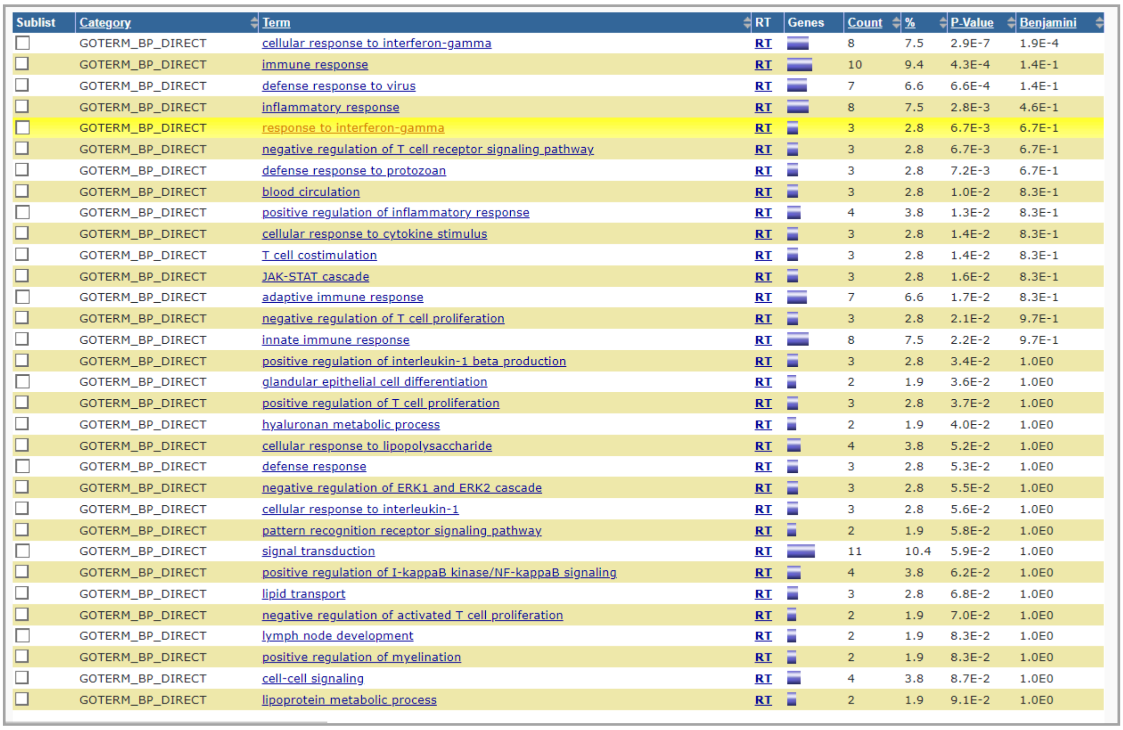
- Cellular response to interferon-gamma
- Immune response
- Defense response to virus
- Inflammatory response
- Response to interferon gamma
- Negative regulation of T-Cell receptor signaling pathway
- Defense response to protozoan

In the Cellular Component (CC) term, DEGs were involved mainly in these components:

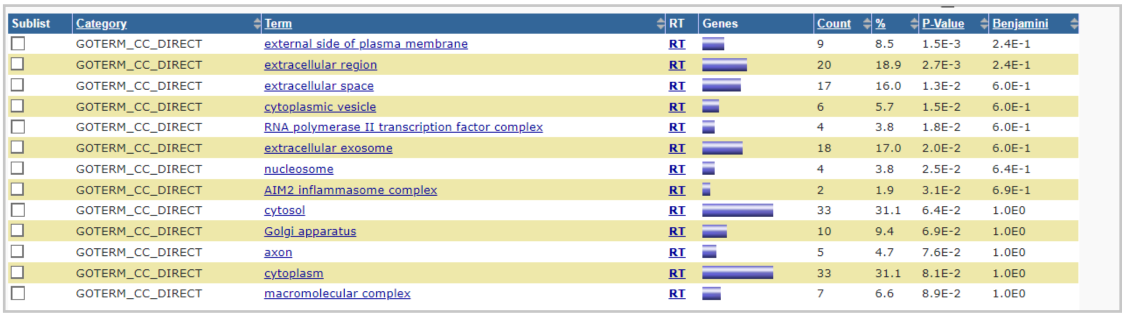
- External side of plasma membrane
- Cytoplasmic vesicle
- RNA polymerase 2 transcription factor complex
- Nucleosome
- Extracellular Region
- Extracellular Exosome
- Extracellular Space

In the Molecular Function (MF) term, DEGs were mainly associated with :

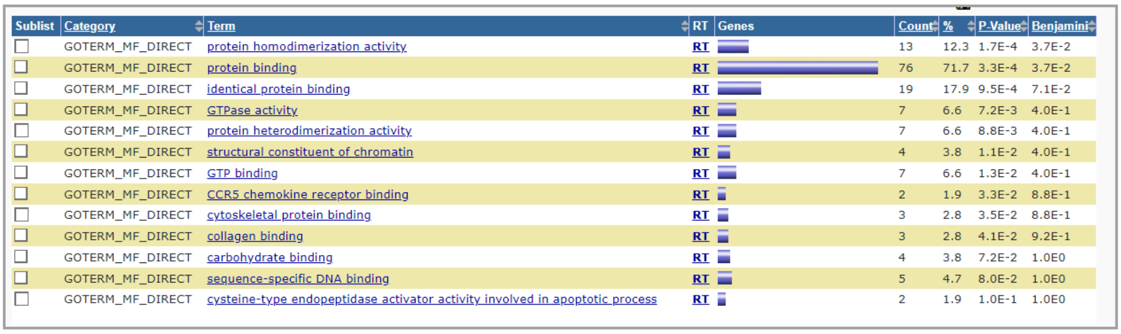
- Protein Binding
- Identical Protein Binding
- Protein Homodimerization Activity
- GTPase Activity



Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	GOTERM_BP_DIRECT	cellular response to interferon-gamma	RT		8	7.5	2.9E-7	1.9E-4
□	GOTERM_BP_DIRECT	immune response	RT		10	9.4	4.3E-4	1.4E-1
□	GOTERM_BP_DIRECT	defense response to virus	RT		7	6.6	6.6E-4	1.4E-1
□	GOTERM_BP_DIRECT	inflammatory response	RT		8	7.5	2.8E-3	4.6E-1
□	GOTERM_BP_DIRECT	response to interferon-gamma	RT		3	2.8	6.7E-3	6.7E-1
□	GOTERM_BP_DIRECT	negative regulation of T cell receptor signaling pathway	RT		3	2.8	6.7E-3	6.7E-1
□	GOTERM_BP_DIRECT	defense response to protozoan	RT		3	2.8	7.2E-3	6.7E-1
□	GOTERM_BP_DIRECT	blood circulation	RT		3	2.8	1.0E-2	8.3E-1
□	GOTERM_BP_DIRECT	positive regulation of inflammatory response	RT		4	3.8	1.3E-2	8.3E-1
□	GOTERM_BP_DIRECT	cellular response to cytokine stimulus	RT		3	2.8	1.4E-2	8.3E-1
□	GOTERM_BP_DIRECT	T cell costimulation	RT		3	2.8	1.4E-2	8.3E-1
□	GOTERM_BP_DIRECT	JAK-STAT cascade	RT		3	2.8	1.6E-2	8.3E-1
□	GOTERM_BP_DIRECT	adaptive immune response	RT		7	6.6	1.7E-2	8.3E-1
□	GOTERM_BP_DIRECT	negative regulation of T cell proliferation	RT		3	2.8	2.1E-2	9.7E-1
□	GOTERM_BP_DIRECT	innate immune response	RT		8	7.5	2.2E-2	9.7E-1
□	GOTERM_BP_DIRECT	positive regulation of interleukin-1 beta production	RT		3	2.8	3.4E-2	1.0E0
□	GOTERM_BP_DIRECT	glandular epithelial cell differentiation	RT		2	1.9	3.6E-2	1.0E0
□	GOTERM_BP_DIRECT	positive regulation of T cell proliferation	RT		3	2.8	3.7E-2	1.0E0
□	GOTERM_BP_DIRECT	hyaluronan metabolic process	RT		2	1.9	4.0E-2	1.0E0
□	GOTERM_BP_DIRECT	cellular response to lipopolysaccharide	RT		4	3.8	5.2E-2	1.0E0
□	GOTERM_BP_DIRECT	defense response	RT		3	2.8	5.3E-2	1.0E0
□	GOTERM_BP_DIRECT	negative regulation of ERK1 and ERK2 cascade	RT		3	2.8	5.5E-2	1.0E0
□	GOTERM_BP_DIRECT	cellular response to interleukin-1	RT		3	2.8	5.6E-2	1.0E0
□	GOTERM_BP_DIRECT	pattern recognition receptor signaling pathway	RT		2	1.9	5.8E-2	1.0E0
□	GOTERM_BP_DIRECT	signal transduction	RT		11	10.4	5.9E-2	1.0E0
□	GOTERM_BP_DIRECT	positive regulation of I-kappaB kinase/NF-kappaB signaling	RT		4	3.8	6.2E-2	1.0E0
□	GOTERM_BP_DIRECT	lipid transport	RT		3	2.8	6.8E-2	1.0E0
□	GOTERM_BP_DIRECT	negative regulation of activated T cell proliferation	RT		2	1.9	7.0E-2	1.0E0
□	GOTERM_BP_DIRECT	lymph node development	RT		2	1.9	8.3E-2	1.0E0
□	GOTERM_BP_DIRECT	positive regulation of myelination	RT		2	1.9	8.3E-2	1.0E0
□	GOTERM_BP_DIRECT	cell-cell signaling	RT		4	3.8	8.7E-2	1.0E0
□	GOTERM_BP_DIRECT	lipoprotein metabolic process	RT		2	1.9	9.1E-2	1.0E0

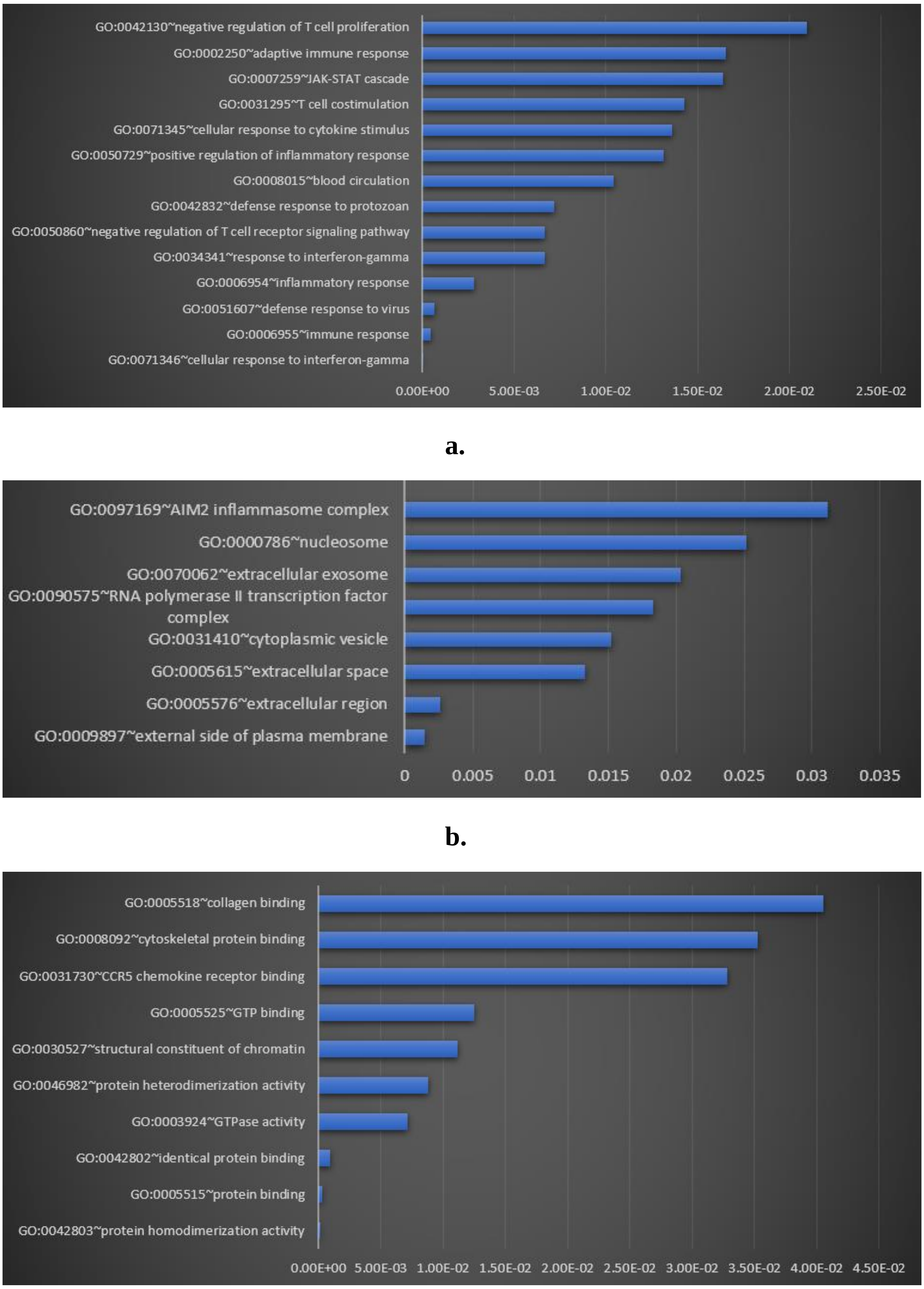


Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	GOTERM_CC_DIRECT	external side of plasma membrane	RT		9	8.5	1.5E-3	2.4E-1
□	GOTERM_CC_DIRECT	extracellular region	RT		20	18.9	2.7E-3	2.4E-1
□	GOTERM_CC_DIRECT	extracellular space	RT		17	16.0	1.3E-2	6.0E-1
□	GOTERM_CC_DIRECT	cytoplasmic vesicle	RT		6	5.7	1.5E-2	6.0E-1
□	GOTERM_CC_DIRECT	RNA polymerase II transcription factor complex	RT		4	3.8	1.8E-2	6.0E-1
□	GOTERM_CC_DIRECT	extracellular exosome	RT		18	17.0	2.0E-2	6.0E-1
□	GOTERM_CC_DIRECT	nucleosome	RT		4	3.8	2.5E-2	6.4E-1
□	GOTERM_CC_DIRECT	AIM2 inflammasome complex	RT		2	1.9	3.1E-2	6.9E-1
□	GOTERM_CC_DIRECT	cytosol	RT		33	31.1	6.4E-2	1.0E0
□	GOTERM_CC_DIRECT	Golgi apparatus	RT		10	9.4	6.9E-2	1.0E0
□	GOTERM_CC_DIRECT	axon	RT		5	4.7	7.6E-2	1.0E0
□	GOTERM_CC_DIRECT	cytoplasm	RT		33	31.1	8.1E-2	1.0E0
□	GOTERM_CC_DIRECT	macromolecular complex	RT		7	6.6	8.9E-2	1.0E0



Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	GOTERM_MF_DIRECT	protein homodimerization activity	RT		13	12.3	1.7E-4	3.7E-2
□	GOTERM_MF_DIRECT	protein binding	RT		76	71.7	3.3E-4	3.7E-2
□	GOTERM_MF_DIRECT	identical protein binding	RT		19	17.9	9.5E-4	7.1E-2
□	GOTERM_MF_DIRECT	GTPase activity	RT		7	6.6	7.2E-3	4.0E-1
□	GOTERM_MF_DIRECT	protein heterodimerization activity	RT		7	6.6	8.8E-3	4.0E-1
□	GOTERM_MF_DIRECT	structural constituent of chromatin	RT		4	3.8	1.1E-2	4.0E-1
□	GOTERM_MF_DIRECT	GTP binding	RT		7	6.6	1.3E-2	4.0E-1
□	GOTERM_MF_DIRECT	CCR5 chemokine receptor binding	RT		2	1.9	3.3E-2	8.8E-1
□	GOTERM_MF_DIRECT	cytoskeletal protein binding	RT		3	2.8	3.5E-2	8.8E-1
□	GOTERM_MF_DIRECT	collagen binding	RT		3	2.8	4.1E-2	9.2E-1
□	GOTERM_MF_DIRECT	carbohydrate binding	RT		4	3.8	7.2E-2	1.0E0
□	GOTERM_MF_DIRECT	sequence-specific DNA binding	RT		5	4.7	8.0E-2	1.0E0
□	GOTERM_MF_DIRECT	cysteine-type endopeptidase activator activity involved in apoptotic process	RT		2	1.9	1.0E-1	1.0E0

Fig 7. Functional Enrichment Analysis of common DEGs for Biological Process **(a),** Cellular Component **(b)** and Molecular Function **(c)** terms. The count represents the number of genes involved in the respective functions/components.



a.

b.

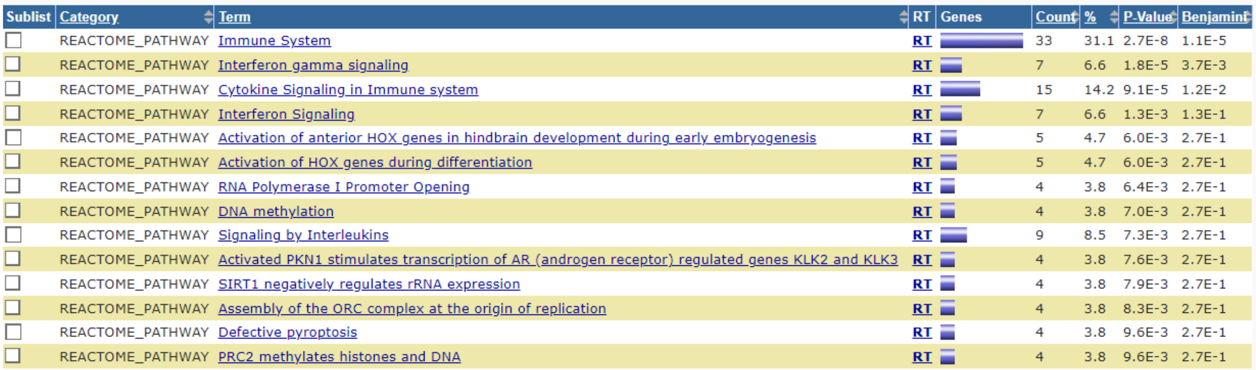
c.

Fig 8. Graphical representation of major Biological Processes **(a),** Cellular Components **(b)** and Molecular Functions **(c)** for the given list of genes. This is based p-value and the cut off was set to < 0.05.

6.3.1 Pathway Enrichment Analysis

Pathway enrichment was performed in DAVID. From the options, “Reactome Pathways” was chosen. The cut off set for p-value was < 0.05.

The DEGs were mainly enriched in pathways related to Immune system, Interferon gamma signalling, Cytokine signalling in Immune system and Interferon Signalling.



Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	REACTOME_PATHWAY	Immune System	RT		33	31.1	2.7E-8	1.1E-5
□	REACTOME_PATHWAY	Interferon gamma signaling	RT		7	6.6	1.8E-5	3.7E-3
□	REACTOME_PATHWAY	Cytokine Signaling in Immune system	RT		15	14.2	9.1E-5	1.2E-2
□	REACTOME_PATHWAY	Interferon Signaling	RT		7	6.6	1.3E-3	1.3E-1
□	REACTOME_PATHWAY	Activation of anterior HOX genes in hindbrain development during early embryogenesis	RT		5	4.7	6.0E-3	2.7E-1
□	REACTOME_PATHWAY	Activation of HOX genes during differentiation	RT		5	4.7	6.0E-3	2.7E-1
□	REACTOME_PATHWAY	RNA Polymerase I Promoter Opening	RT		4	3.8	6.4E-3	2.7E-1
□	REACTOME_PATHWAY	DNA methylation	RT		4	3.8	7.0E-3	2.7E-1
□	REACTOME_PATHWAY	Signaling by Interleukins	RT		9	8.5	7.3E-3	2.7E-1
□	REACTOME_PATHWAY	Activated PKN1 stimulates transcription of AR (androgen receptor) regulated genes KLK2 and KLK3	RT		4	3.8	7.6E-3	2.7E-1
□	REACTOME_PATHWAY	SIRT1 negatively regulates rRNA expression	RT		4	3.8	7.9E-3	2.7E-1
□	REACTOME_PATHWAY	Assembly of the ORC complex at the origin of replication	RT		4	3.8	8.3E-3	2.7E-1
□	REACTOME_PATHWAY	Defective pyroptosis	RT		4	3.8	9.6E-3	2.7E-1
□	REACTOME_PATHWAY	PRC2 methylates histones and DNA	RT		4	3.8	9.6E-3	2.7E-1

Fig 9. Pathway Enrichment Analysis using Reactome (accessed from DAVID).

6.4 Protein – Protein Interaction Network

All 105 common DEGs were uploaded to STRING database as Official Gene Symbols. A cluster of 96 genes were obtained on setting the minimum required interaction score to 0.4 (Figure 10). The PPI network had 96 nodes and 97 edges, with an average node degree of 2.02.

The network was downloaded and visualized in Cytoscape (Figure 11). In order to view the hub genes, the CytoHubba plugin was installed in Cytoscape. The top 10 hub genes were visualized and ranked based on the MCC (Maximal Clique Centrality) and Degree (Figure 12) (Table 4).

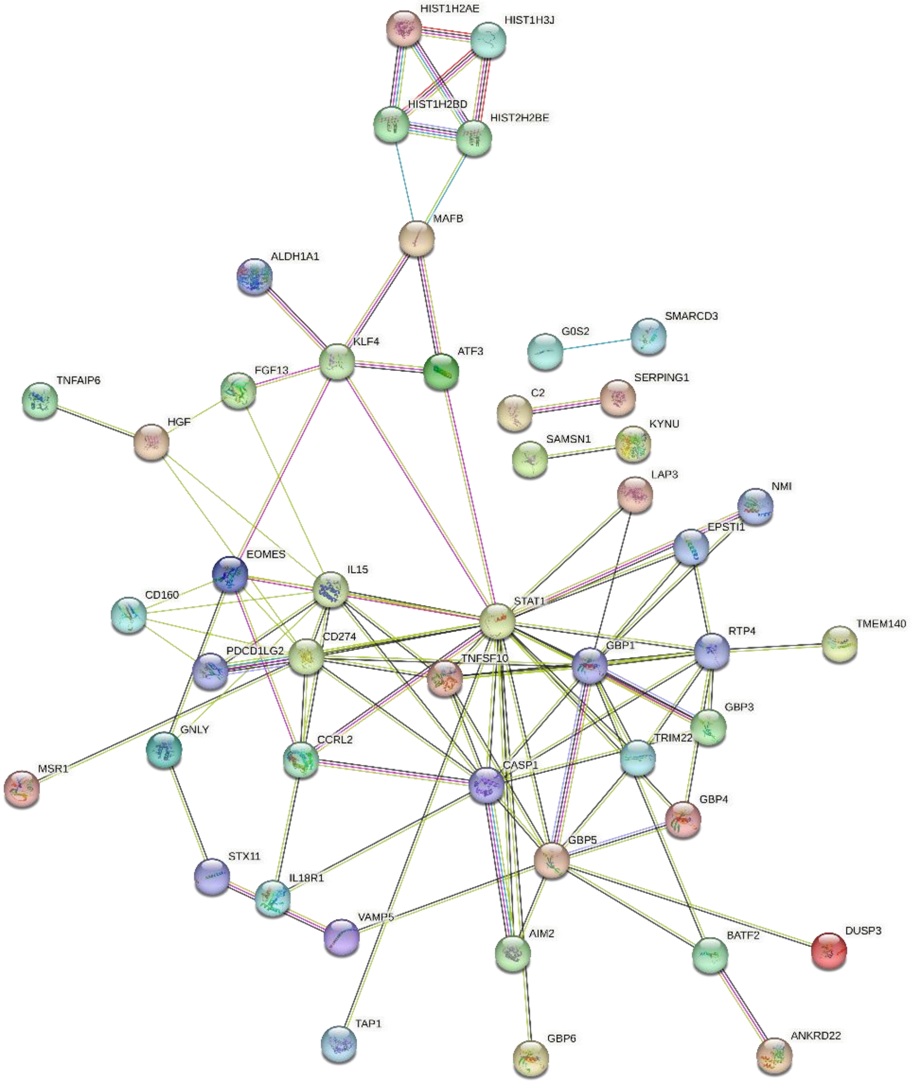


Fig 10. PPI network obtained from STRING DB

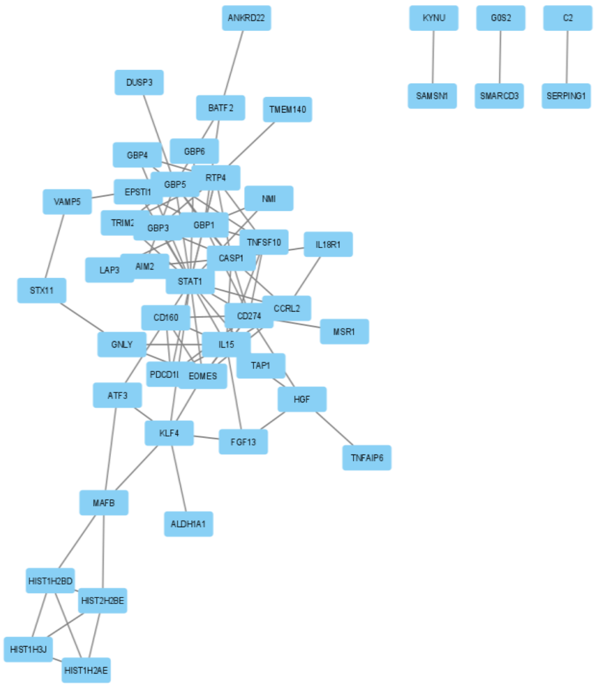


Fig 11. PPI Network as visualized in Cytoscape

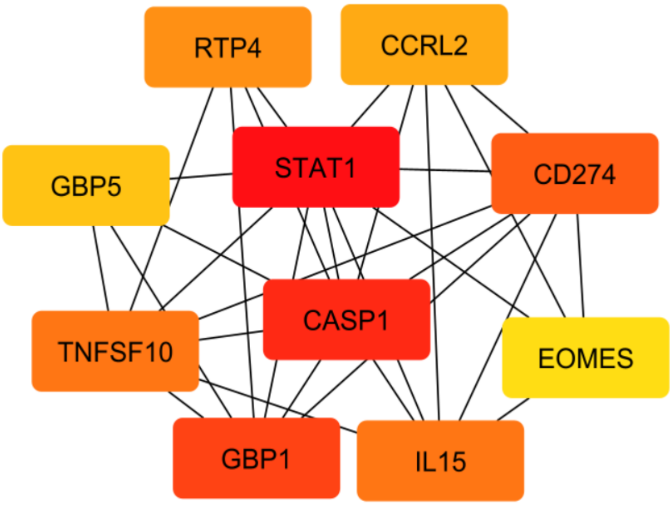


Fig.12 Top 10 Hub Genes

Sl. No.	Hub Genes	Full Name
1.	STAT1	Signal transducer and activator of transcription 1
2.	CASP1	Caspase 1
3.	GBP1	Guanylate-binding protein 1
4.	CD274	Cluster of differentiation 274
5.	TNFSF10	Tumor necrosis factor ligand superfamily member 10
6.	IL15	Interleukin 15
7.	RTP4	Receptor transporter protein 4
8.	CCLR2	C-C motif chemokine receptor 2
9.	GBP5	Guanylate-binding protein 5
10.	EOMES	Eomesodermin

Table 4. List of Hub Genes, ranked based on *MCC*

In total, 9 genes were upregulated and 1 gene was downregulated in both TB Only and TBDM groups. The expression levels of these genes in all three groups are given below (cut-off for significance : logFC value between -0.5 and 0.5 and p-value < 0.01) :

Sl.no	Genes	p-value	Log$_2$FC
1.	STAT1	0.00067	0.63
2.	CASP1	0.00017	0.60
3.	GBP1	2.27E-06	1.06
4.	CD274	2.02E-05	1.01
5.	TNFSF10	0.00218	0.57
6.	IL15	0.00299	0.61

7.	RTP4	0.00066	0.90
8.	CCRL2	0.00160	0.71
9.	GBP5	3.09E-05	0.97
10.	EOMES	0.00919	-0.56

Table 5. Expression levels of hub genes in TB only group

Sl.no	Genes	p-value	Log$_2$FC
1.	STAT1	0.930	-0.01
2.	CASP1	0.684	0.06
3.	GBP1	0.662	0.12
4.	CD274	0.999	0.0002
5.	TNFSF10	0.405	-0.19
6.	IL15	0.177	0.35
7.	RTP4	0.115	0.46
8.	CCRL2	0.075	0.45
9.	GBP5	0.633	-0.13
10.	EOMES	0.344	-0.20

Table 6. Expression levels of hub genes in DM only group (**statistically insignificant**)

Sl.no	Genes	p-value	Log_2FC
1.	STAT1	0.00067	0.78
2.	CASP1	4.64E-05	0.77
3.	GBP1	2.02E-07	1.52
4.	CD274	8.94E-11	1.96
5.	TNFSF10	0.00065	0.85
6.	IL15	0.00027	0.93
7.	RTP4	0.00020	1.24
8.	CCRL2	0.00146	0.99
9.	GBP5	2.62E-06	1.39
10.	EOMES	0.00086	-0.86

Table 7. Expression levels of hub genes in TBDM comorbid group

It is to be mentioned that 5 common genes were found to be expressed in all three groups, **within the range of statistical significance**. The genes and their expression levels are represented below (Table 8) :

Genes	logFC TB only	logFC TBDM	logFC DM only
ANKRD22	0.83	2.33	0.90
IL31RA	0.83	1.25	1.22
VAMP5	0.70	1.5	0.80
ALDH1A1	0.65	1.1	0.80

Table 8. Genes common to all three sets and their expression levels

- None of the hub genes were observed / expressed in the DM only group under the given thresholds of p-value < 0.01 and | fold change (FC) | >1
- Moreover, the expression levels of these genes were noticeably higher in TBDM comorbid group when compared to TB only group

6.5 miRNA Interaction Network

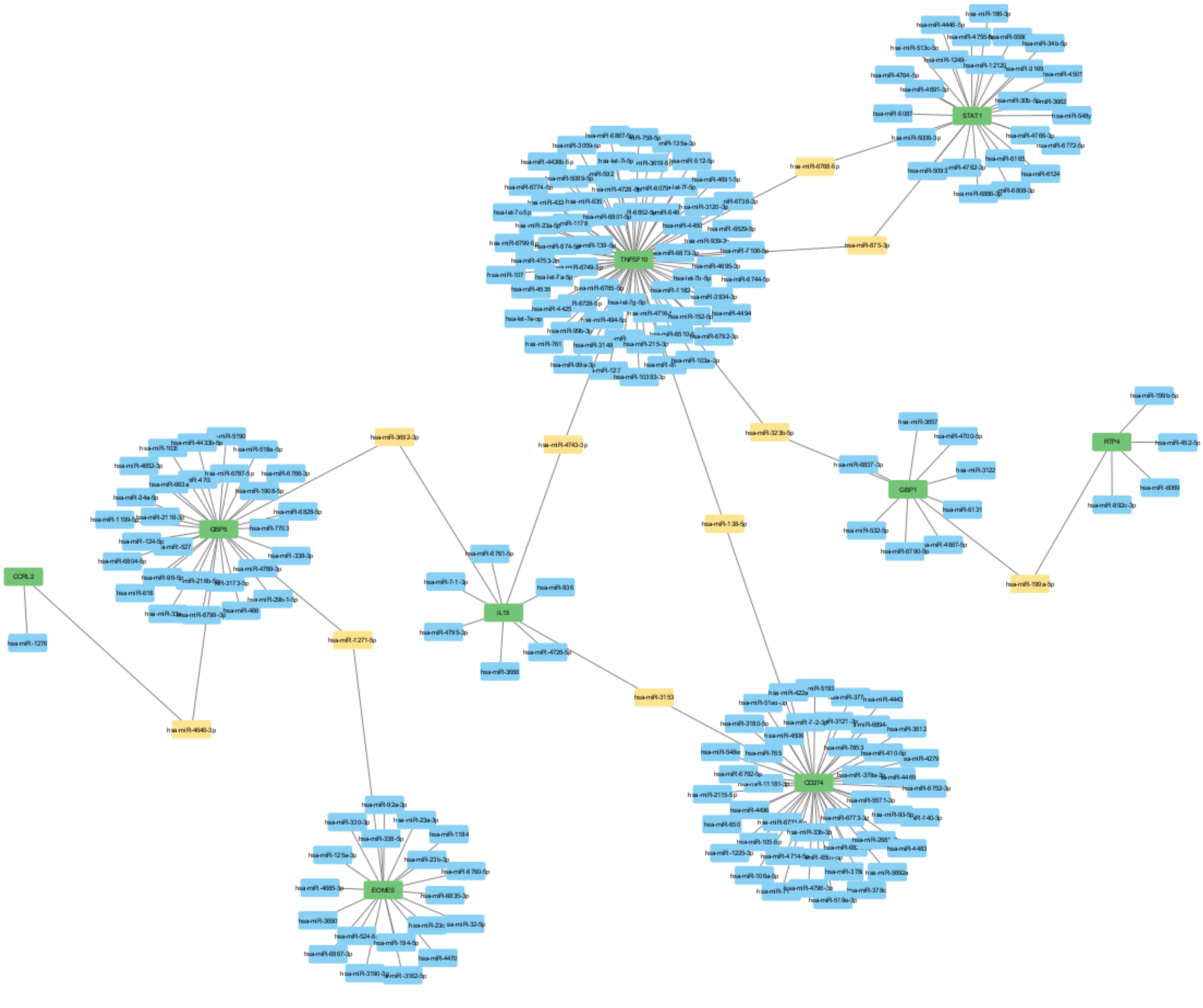


Fig 13. Network of miRNAs and their target genes. **GREEN** boxes represent the hub genes, while the **BLUE** and **YELLOW** boxes represent miRNAs and significant miRNAs respectively. When an miRNA targets 2 genes, it means that the miRNA has the ability to bind to the mRNA of both genes, leading to the suppression of their expression. This suggests that the miRNA plays a regulatory role in the biological processes that involve the two targeted genes.

miRNA(s)	Target Genes
hsa-mir-1271-5p	GBP5, EOMES
hsa-mir-4646-3p	GBP1, CCRL2
hsa-mir-3692-3p	GBP5, IL15
hsa-mir-3153	CD274, IL15
hsa-mir-4743-3p	IL15, TNFSF10
hsa-mir-138-5p	TNFSF10, CD274
hsa-mir-323b-5p	TNFSF10, GBP1
hsa-mir-199a-5p	GBP1, RTP4
hsa-mir-6768-5p,	TNFSF10, STAT1
hsa-mir-875-3p	

Table 9. miRNAs targeting at least 2 genes

The **mirWalk2.0** tool was used to construct the network. The network was downloaded and visualized in Cytoscape. Significant miRNAs were assumed to be those that bind to at least 2 genes. In total, 10 such significant miRNAs (Figure 13) (Table 9) were identified.

7. Discussion

Both tuberculosis and type 2 diabetes mellitus (T2DM) are widespread diseases that affect a significant portion of the global population, and they can impose a substantial burden on patients as well as healthcare systems. It is well-established that individuals with T2DM are at a higher risk of developing tuberculosis compared to those without T2DM, and numerous studies have confirmed this association [58] [59]. However, the exact mechanisms underlying this link are not yet fully understood.

In this project, a series of Bioinformatic and Statistical analysis were conducted on RNAseq data of individuals with ATB only, DM only and TBDM comorbidity so as to find genes that could potentially act as biomarkers for Tuberculosis in DM patients.

In the present study, the gene expression dataset **GSE114192** was screened, following which 105 DEGs shared by TB Only and TBDM Comorbid groups were identified, that included 91 up- and 14 downregulated genes. Functional enrichment analysis was performed in DAVID server. Under the BP term, the DEGs were mainly involved in Immune response, Defense response to virus, Signal transduction, Cellular response to interferon-gamma, Inflammatory response, Adaptive immune response and Innate immune response. Under the CC term, the DEGs were mainly involved in Cytosol, Cytoplasm, Extracellular Region, Extracellular Exosome and Extracellular Space. Under the MF term, the DEGs were mainly involved in functions like Protein Binding, Identical Protein Binding, Protein Homodimerization Activity, GTPase Activity, and GTP Binding.

Furthermore, analysis of the PPI network in STRING database among shared DEGs revealed ten hub genes (STAT1, CASP1, GBP1, CD274, TNFSF10, IL15, RTP4, CCRL2, GBP5, EOMES) that were subsequently identified using the plug-in CytoHubba of Cytoscape based on the Maximum Clique Centrality (MCC) scores. Also, four genes were found to be regulated in all three (TB, DM and TBDM comorbid) cases (ANKRD22, IL31RA, VAMP5, ALDH1A1).

Out of the ten hub genes identified, 9 genes were upregulated in both TB and TBDM groups, while the gene EOMES was downregulated in both.

The genes STAT1, CASP1, GBP1, GBP5, CD274, IL15, CCRL2 and RTP4 have already been studied with regards to Tuberculosis, and their significance as potential biomarkers in the same have been looked into.

STAT1, a crucial gene in host defense against tuberculosis (TB) infection, has been shown to increase susceptibility to MTB infection when mutated [60, 61]. Phosphorylated STAT1 levels have been found to drive the expression of pro-apoptotic genes resulting in anti-TB effects. Conversely, unphosphorylated STAT1 represses macrophage apoptosis, facilitating immune evasion by MTB, leading to continued infection [62]. TNFSF10, also known as TRAIL, is a pro-apoptotic cytokine gene that induces apoptosis in transformed and tumour cells but not in normal cells [63]. While it is associated with diseases such as colorectal cancer and pancreatic cancer, it is also involved in related pathways such as MIF-mediated glucocorticoid regulation and procaspase-8 dimerization [63]. Upregulation of STAT1 may drive the expression of TRAIL. Further research into the potential diagnostic viability of this gene is warranted.

CASP1 (Caspase - 1) has been seen to be associated with a number of biological pathways including cytokine signalling in immune system, Signalling by interleukins and Innate immune response and in processes like positive regulation of IL-1 Beta. The most well-known function of active caspase-1 is to cleave the pro-forms of inflammatory cytokines IL-1Beta and -18 into their active forms in response to inflammatory stimuli in immune cells [64]. The upregulation of CASP1 is likely due to the activation of inflammasomes, multi-protein complexes that mediates activation of caspase-1 which promotes the secretion of IL-1β and IL-18 as well as pyroptosis, a form of cell death induced by bacterial pathogens [65].

GBP1 (Guanylate Binding Protein 1) has been found to be associated with a wide range of BP terms including cellular response to interferon-gamma, defense response to virus, and negative regulation of T cell receptor signalling pathway . It has also been reported in pathways including that of Interferon gamma signalling, Cytokine Signalling in Immune system and Interferon Signalling. A previous study [66] found that GBP1 expression was generally upregulated in TB patients, suggesting that high expression of GBP1 may play a role in fighting MTB infection. Furthermore, ISGs (interferonstimulated genes) and inflammasome-activation genes were highly expressed in both the TB and GBP1-high groups, suggesting that GBP1 is associated with the interferon signalling pathway and inflammasome activation in TB. The study concluded that GBP1 could act as an optimal diagnostic biomarker for TB. This correlates with the present study in that GBP1 was upregulated in both TB and TBDM groups compared to HCs.

GBP5 (Guanylate Binding Protein 5)is an interferon-inducible gene, meaning that its expression can be initiated by interferons, which are cytokines produced by the immune system in response to viral infections and other pathogens [67]. GBP5 has been found to be associated with BP terms like Interferon gamma signalling, Cytokine signalling in Immune system and Interferon signalling. A recent study concluded that GBP5 protein in whole blood is a potential biomarker for differentiating ATB from non-TB group [68]. The study also proposed that whole blood GBP5 protein assay has similar performance for ATB diagnosis as IGRA, and the combination of the GBP5 assay and IGRA results in approximately 90% accuracy for diagnosing ATB in more than half of the suspected patients [68]. The upregulation of GBP5 in both the groups further confirms this.

CD274 has been observed to be involved in negative regulation of T cell proliferation. CD274, also known as programmed cell death ligand 1 (**PD-L1**), is a cell surface protein that plays a role in regulating the immune response [69]. CD274 expression can be induced by interferon-gamma (IFN-γ) and other pro-inflammatory cytokines, and it is known to interact with its receptor, programmed cell death protein 1 (PD-1), on T cells to inhibit their activity [69]. It has been found to be upregulated in both the groups. This upregulation might be mediated by IFN-γ and other cytokines produced by T-cells and other immune cells during the immune response to MTB infection. The upregulation of CD274 in TB might hence play a role in immune evasion by MTB, as it can inhibit T-cell activity and promote the survival of MTB-infected cells.

Interleukin 15 is a cytokine that plays a major role in the development of inflammatory and protective immune responses to microbial invaders and parasites by modulating immune cells of both the innate and adaptive immune systems [70]. It stimulates the proliferation of natural killer cells, T-cells and B-cells and promotes the secretion of several cytokines [71][72]. It is associated with BP terms like positive regulation of T cell proliferation and inflammatory response. A previous study has shown that IL15 is able to enhance the survivability of MTB infected mice, compared to other interleukins such as IL2 or IL4, when delivered as a treatment post the infection [73].

CCRL2 (Chemokine CC Motif Receptor Like 2) is a chemokine receptor that is involved in the recruitment of immune cells to sites of inflammation. Studies have shown that CCRL2 is upregulated in macrophages and dendritic cells in response to MTB infection, suggesting a potential role in the host immune response to TB [74].

EOMES is a transcription factor that belongs to the T-box family, which is closely related to T-bet. Its role in CD8 T cell and natural killer cell differentiation is well known [75]. EOMES is involved in biological processes such as the adaptive immune response[75]. Numerous studies have examined the function of EOMES in human CD4 T-cell differentiation. One study [76] suggests that EOMES is critical in regulating the phenotype shift of Th17 cells towards non-classic Th1 cells, indicating its involvement in helper T cell (Th) plasticity. Another study [77] proposes that EOMES acts as a lineage-defining transcription factor for regulatory T-cells (Tr1 cells) that produce both IL-10 and IFN-γ. Both studies demonstrate that EOMES drives the secretion of IFN-γ and conveys a "cytotoxic" signature, while also suppressing Th17 characteristics [75]. An important characteristic of EOMES+ CD4 T cells is their tendency to accumulate in inflamed tissues in patients with chronic inflammatory disorders [75], such as Tuberculosis. However, EOMES was downregulated in both TB only and TBDM groups. Upregulating it might enhance the immune response of the host towards MTB. EOMES could potentially serve as a therapeutic biomarker for ATB, although further research is needed to validate this statement.

RTP4 (receptor-transporting protein 4), is a putative chaperone protein that facilitates the trafficking and functional cell surface expression of certain G-protein coupled receptors (GPCRs), including the bitter taste receptor TAS2R16 [78]. A very recent study examined the expression levels of RTP4 before and after anti-TB therapy [79]. The findings revealed that RTP4 gene expression was upregulated in individuals with

ATB compared to healthy controls and downregulated after antituberculosis therapy. These results suggest that RTP4 may serve as a promising biomarker for the diagnosis of ATB, and the upregulation of RTP4 in this study further validates this point.

The analysis also revealed **ten miRNAs** (hsa-mir-1271-5p, hsa-mir-4646-3p, hsa-mir-3692-3p, hsa-mir-3153, hsa-mir-4743-3p, hsa-mir-138-5p, hsa-mir-323b-5p, hsa-mir-199a-5p, hsa-mir-6768-5p, and hsa-mir-875-3p) that potentially target two or more genes and could exert post-transcriptional effects on them. These miRNAs have the potential to serve as biomarkers for ATB, but their clinical significance needs validation through further research.

8. Conclusion

Active Tuberculosis is a serious infectious disease caused by the bacterium MTB. The co-occurrence of TB and Diabetes Mellitus (DM) is a growing concern, as diabetes can increase the risk of TB infection and worsen TB outcomes. People with diabetes are more susceptible to TB infection and are more likely to have TB reactivation. Early diagnosis and treatment of Active TB is crucial in order to prevent the spread of TB to others and to improve treatment outcomes, especially in countries with high TB burden.

In this study, a dataset from NCBI SRA was analyzed to investigate the gene expression differences between TB only, DM only, TBDM comorbid, and healthy control groups from South Africa. The analysis revealed 105 common differentially expressed genes (DEGs), out of which 91 were upregulated and 14 were downregulated in TBDM comorbid and TB only groups. Functional enrichment analysis and pathway analysis showed that these DEGs were enriched in processes related to immune responses and cytokine signalling in immune system pathways. Ten hub genes were identified through network construction and analysis, namely *STAT1, CASP1, GBP1, CD274, TNFSF10, IL15*, *RTP4*, *CCRL2*, *GBP5,* and *EOMES*. These genes were significantly expressed in the TB only and TBDM groups, but not in the DM only group. **EOMES** and **TNFSF10** were identified as potential novel therapeutic and diagnostic biomarkers, respectively, from literature, but further validation in a larger dataset is required.

Additionally, the miRNA-hub gene interaction network identified **10 miRNAs** (hsa-mir-1271-5p, hsa-mir-4646-3p, hsa-mir-3692-3p, hsa-mir-3153, hsa-mir-4743-3p, hsa-mir-138-5p, hsa-mir-323b-5p, hsa-mir-199a-5p, hsa-mir-6768-5p, and hsa-mir-875-3p) that target at least 2 hub genes and could be potential markers for TB, although more research is necessary to confirm this hypothesis.

9. References

1. Division of Tuberculosis Elimination, National Center for HIV, Viral Hepatitis, STD, and TB Prevention, Centres for Disease Control and Prevention
2. Mycobacterium leprae–host-cell interactions and genetic determinants in leprosy: an overview Roberta Olmo Pinheiro,1 Jorgenilce de Souza Salles,1 Euzenir Nunes Sarno,1 and Elizabeth Pereira Sampaio1,2,†

3. World Health Organization. Global Tuberculosis Report 2020; World Health Organization: Geneva, Switzerland, 2020. Available online: https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2022
4. World Health Organization. Latent tuberculosis infection: updated and consolidated guidelines for programmatic management. Geneva, Switzerland: World Health Organization, 2018. Licence: CC BY-NC-SA 3.0 IGO.
5. Bennett JE, et al. Mycobacterium tuberculosis. In: Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases. 9th ed. Elsevier; 2020. https://www.clinicalkey.com
6. Screening and prevention for latent tuberculosis in immunosuppressed patients at risk for tuberculosis: a systematic review of clinical practice guidelines Tasnim Hasan,1 Eric Au,2 Sharon Chen,1,3 Allison Tong,4,5 and Germaine Wong 2,5
7. Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2019. Results. Institute for Health Metrics and Evaluation. 2020 (https://vizhub.healthdata.org/gbd-results/).
8. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th editionPouya Saeedi 1, Inga Petersohn 2, Paraskevi Salpea 2, Belma Malanda 2, Suvi Karuranga 2, Nigel Unwin 3, Stephen Colagiuri 4, Leonor Guariguata 5, Ayesha A Motala 6, Katherine Ogurtsova 7, Jonathan E Shaw 8, Dominic Bright 9, Rhys Williams 9; IDF Diabetes Atlas Committee
9. American Diabetes Association Professional Practice Committee. Standards of medical care in diabetes—2022. Diabetes Care. 2022;45(suppl 1):S17–S38. doi:10.2337/dc22-S002
10. Cook, Curtiss B et al. "Geoenvironmental diabetology." Journal of diabetes science and technology vol. 5,4 834-42. 1 Jul. 2011, doi:10.1177/193229681100500402
11. Dooley KE, Chaisson RE. Tuberculosis and diabetes mellitus: convergence of two epidemics. Lancet Infect Dis. 2009 Dec;9(12):737-46. doi: 10.1016/S1473-3099(09)70282-8. PMID: 19926034; PMCID: PMC2945809.

12. Dye C, Glaziou P, Floyd K, Raviglione M. Prospects for tuberculosis elimination. Annu Rev Public Health. 2013;34:271-86. doi: 10.1146/annurev-publhealth-031912-114431. Epub 2012 Dec 14. PMID: 23244049.
13. Jeon CY, Murray MB. Diabetes mellitus increases the risk of active tuberculosis: a systematic review of 13 observational studies. PLoS Med. 2008 Jul 15;5(7):e152. doi: SS10.1371/journal.pmed.0050152. PMID: 18630984; PMCID: PMC2459204.
14. Pai M, Denkinger CM, Kik SV, Rangaka MX, Zwerling A, Oxlade O, Metcalfe JZ, Cattamanchi A, Dowdy DW, Dheda K, Banaei N. Gamma interferon release assays for detection of Mycobacterium tuberculosis infection. Clin Microbiol Rev. 2014 Jan;27(1):3-20. doi: 10.1128/CMR.00034-13. PMID: 24396134; PMCID: PMC3910908.
15. Steingart KR, Ramsay A, Dowdy DW, Pai M. Serological tests for the diagnosis of active tuberculosis: relevance for India. Indian J Med Res. 2012 May;135(5):695-702. PMID: 22771604; PMCID: PMC3401705.
16. Getahun H, Matteelli A, Chaisson RE, Raviglione M. Latent Mycobacterium tuberculosis infection. N Engl J Med. 2015;372:2127-35. 10.1056/NEJMra1405427 [PubMed] [CrossRef] [Google Scholar]
17. Comstock GW, Livesay VT, Woolpert SF. The prognosis of a positive tuberculin reaction in childhood and adolescence. Am J Epidemiol. 1974;99:131-8. 10.1093/oxfordjournals.aje.a121593 [PubMed] [CrossRef] [Google Scholar]
18. Vynnycky E, Fine PE. The natural history of tuberculosis: the implications of age-dependent risks of disease and the role of reinfection. Epidemiol Infect. 1997;119:183-201. 10.1017/S0950268897007917 [PMC free article] [PubMed] [CrossRef] [Google Scholar]
19. Selwyn PA, Hartel D, Lewis VA, Schoenbaum EE, Vermund SH, Klein RS, Walker AT, Friedland GH. A prospective study of the risk of tuberculosis among intravenous drug users with human immunodeficiency virus infection. N Engl J Med. 1989;320:545-50. 10.1056/NEJM198903023200901 [PubMed] [CrossRef] [Google Scholar]
20. Moore D, Liechty C, Ekwaru P, Were W, Mwima G, Solberg P, Rutherford R, Mermin J. Prevalence, incidence and mortality associated with tuberculosis in HIV-infected patients initiating antiretroviral therapy in rural Uganda. AIDS. 2007;21:713-9. 10.1097/QAD.0b013e328013f632 [PubMed] [CrossRef] [Google Scholar]
21. WHO. The top 10 causes of death. WHO https://www. who.int/news-room/fact-sheets/detail/the-top10-causes-of-death (2019).
22. WHO. WHO Global Tuberculosis report 2018. WHO http://who.int/tb/publications/global_report/en/ (2018).

23. Floyd, K., Glaziou, P., Zumla, A. & Raviglione, M. The global tuberculosis epidemic and progress in care, prevention, and research: an overview in year 3 of the End TB era. Lancet Respir. Med. 6, 299–314 (2018).
24. WHO. Guidelines for the treatment of drug-susceptible tuberculosis and patient care. 2017 update. WHO http://apps.who.int/iris/bitstream/10665/255052/1/9789241550000-eng.pdf?ua=1 (2017).
25. Lawn, S. D. & Zumla, A. I. Tuberculosis (Seminar). Lancet 378, 57–72 (2011).
26. Furin, J., Cox, H. & Pai, M. Tuberculosis. Lancet 393, 1642–1656 (2019).
27. Kulchavenya, E. Extrapulmonary tuberculosis: are statistical reports accurate? Ther. Adv. Infect. Dis. 2, 61–70 (2014).

28. Kulchavenya, E., Naber, K. & Bjerklund Johansen, T. E. Urogenital tuberculosis: classification, diagnosis, and treatment. Eur. Urol. 15, 112–121 (2016).

29. World Health Organization. Global tuberculosis report 2021. https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2021 (Accessed 5 May 2023).
30. Restrepo BI, Fisher-Hoch SP, Pino PA, et al. Tuberculosis in poorly controlled type 2 diabetes: altered cytokine expression in peripheral white blood cells. Clin Infect Dis. 2008;47(5):634-641. doi:10.1086/590557
31. Marais BJ, Lönnroth K, Lawn SD, et al. Tuberculosis comorbidity with communicable and non-communicable diseases: integrating health services and control efforts. Lancet Infect Dis. 2013;13(5):436-448. doi:10.1016/S1473-3099(13)70015-X
32. Viswanathan V, Kumpatla S, Aravindalochanan V, et al. Prevalence of diabetes and pre-diabetes and associated risk factors among tuberculosis patients in India. PLoS One. 2012;7(7):e41367. doi:10.1371/journal.pone.0041367
33. Jeon CY, Murray MB. Diabetes mellitus increases the risk of active tuberculosis: a systematic review of 13 observational studies. PLoS Med. 2008;5(7):e152. doi:10.1371/journal.pmed.0050152
34. Workneh MH, Bjune GA, Yimer SA. Prevalence and associated factors of tuberculosis and diabetes mellitus comorbidity: A systematic review. PLoS One. 2017;12(4):e0175925. doi:10.1371/journal.pone

35. Natarajan S, Ranganathan M, Hanna LE, Tripathy S. Transcriptional Profiling and Deriving a Seven-Gene Signature That Discriminates Active and Latent Tuberculosis: An Integrative Bioinformatics Approach. Genes (Basel). 2022 Mar 29;13(4):616. doi: 10.3390/genes13040616. PMID: 35456421; PMCID: PMC9032611.

36. Xie L, Chao X, Teng T, Li Q, Xie J. Identification of Potential Biomarkers and Related Transcription Factors in Peripheral Blood of Tuberculosis Patients. Int J Environ Res Public Health. 2020 Sep 24;17(19):6993. doi: 10.3390/ijerph17196993. PMID: 32987825; PMCID: PMC7579196.

37. Liu S, Ren W, Yu J, Li C, Tang S. Identification of Hub Genes Associated with Diabetes Mellitus and Tuberculosis Using Bioinformatic Analysis. Int J Gen Med. 2021 Jul 30;14:4061-4072. doi: 10.2147/IJGM.S318071. PMID: 34354368; PMCID: PMC8331204.

38. Li X, Xie Y, Zhang D, Wang Y, Wu J, Sun X, Wu Q. Identification of Significant Genes in HIV/TB via Bioinformatics Analysis. Ann Clin Lab Sci. 2020 Sep;50(5):600-610. PMID: 33067206.

39. Leinonen R, Sugawara H, Shumway M; International Nucleotide Sequence Database Collaboration. The sequence read archive. Nucleic Acids Res. 2011 Jan;39(Database issue):D19-21. doi: 10.1093/nar/gkq1019. Epub 2010 Nov 9. PMID: 21062823; PMCID: PMC3013647.
40. Eckold C, Kumar V, Weiner J, Alisjahbana B et al. Impact of Intermediate Hyperglycemia and Diabetes on Immune Dysfunction in Tuberculosis. Clin Infect Dis 2021 Jan 23;72(1):69-78. PMID: 32533832
41. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Grüning BA, Guerler A, Hillman-Jackson J, Hiltemann S, Jalili V, Rasche H, Soranzo N, Goecks J, Taylor J, Nekrutenko A, Blankenberg D. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 2018 Jul 2;46(W1):W537-W544. doi: 10.1093/nar/gky379. PMID: 29790989; PMCID: PMC6030816.
42. Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online].
43. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016 Oct 1;32(19):3047-8. doi: 10.1093/bioinformatics/btw354. Epub 2016 Jun 16. PMID: 27312411; PMCID: PMC5039924.
44. Krueger, F. (2021). Trim Galore. In *GitHub repository*. GitHub. https://https://github.com/FelixKrueger/TrimGalore.com/fenderglass/Flye
45. Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357â€ "360. https://doi.org/10.1038/nmeth.3317
46. Liao, Y., Smyth, G. K., & Shi, W. (2013). featureCounts: an efficient general purpose program for assigning sequence reads to genomic

features. *Bioinformatics*, *30*(7), 923â€ “930. https://doi.org/10.1093/bioinformatics/btt656

47. Love MI, Huber W, Anders S (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” Genome Biology, 15, 550. doi: 10.1186/s13059-014-0550-8.
48. Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.
49. https://bioinformatics.psb.ugent.be/webtools/Venn/
50. Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols, 4(1), 44â€ “57. https://doi.org/10.1038/nprot.2008.211
51. - Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., & Lempicki, R. A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biology, 4(5). https://doi.org/10.1186/gb-2003-4-5-p3
52. Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, Lars J Jensen, Christian von Mering, STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets, *Nucleic Acids Research*, Volume 47, Issue D1, 08 January 2019, Pages D607–D613, https://doi.org/10.1093/nar/gky1131
53. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003 Nov;13(11):2498-504. doi: 10.1101/gr.1239303. PMID: 14597658; PMCID: PMC403769.
54. Chin, CH., Chen, SH., Wu, HH. et al. cytoHubba: identifying hub objects and sub-networks from complex interactome. BMC Syst Biol 8 (Suppl 4), S11 (2014). https://doi.org/10.1186/1752-0509-8-S4-S11
55. Sticht C, De La Torre C, Parveen A, Gretz N (2018) miRWalk: An online resource for prediction of microRNA binding sites. PLoS ONE 13(10): e0206239. https://doi.org/10.1371/journal.pone.0206239
56. Piovesan A, Pelleri MC, Antonaros F, Strippoli P, Caracausi M, Vitale L. On the length, weight and GC content of the human genome. BMC Res Notes. 2019 Feb 27;12(1):106. doi: 10.1186/s13104-019-4137-z. PMID: 30813969; PMCID: PMC6391780.
57. Zhang S, Wang B, Wan L, Li LM. Estimating Phred scores of Illumina base calls by logistic regression and sparse modeling. BMC Bioinformatics. 2017 Jul

11;18(1):335. doi: 10.1186/s12859-017-1743-4. PMID: 28697757; PMCID: PMC5504792.

58. Association of diabetes and tuberculosis: impact on treatment and post-treatment outcomes - María Eugenia Jiménez-Corona1, Luis Pablo Cruz-Hervert2, Lourdes García-García2, Leticia Ferreyra-Reyes2, Guadalupe Delgado-Sánchez2, Miriam Bobadilla-del-Valle3, Sergio Canizales-Quintero2 - Instituto Nacional de Salud Pública, Av. Universidad # 655, Col. Sta. María Ahuacatitlán, Cuernavaca, Morelos C.P. 62100, México; garcigarml@gmail.com

59. Restrepo BI. Diabetes and Tuberculosis. Microbiol Spectr. 2016 Dec;4(6):10.1128/microbiolspec.TNMI7-0023-2016. doi: 10.1128/microbiolspec.TNMI7-0023-2016. PMID: 28084206; PMCID: PMC5240796.

60. . Pedraza-Sanchez S, Lezana-Fernandez JL, Gonzalez Y, et al. Disseminated tuberculosis and chronic mucocutaneous candidiasis in a patient with a gain-of-function mutation in signal transduction and activator of transcription 1. Front Immunol. 2017;8:1651. doi:10.3389/fimmu.2017.01651

61. Yi X-H, Zhang B, Fu Y-R, et al. STAT1 and its related molecules as potential biomarkers in Mycobacterium tuberculosis infection. J Cell Mol Med. 2020;24(5):2866–2878. doi:10.1111/jcmm.14856

62. Yao K, Chen Q, Wu Y, Liu F, Chen X, Zhang Y. Unphosphorylated STAT1 represses apoptosis in macrophages during Mycobacterium tuberculosis infection. J Cell Sci. 2017;130:1740–1751.

63. Hernandez, A., Wang, Q.D., Schwartz, S.A., and Evers, B.M. (2001) Sensitization of human colon cancer cells to TRAILmediated apoptosis. J Gastrointest Surg 5: 56–65.

64. Sun, Q. , Gao, W. , Loughran, P. , Shapiro, R. , Fan, J. , Billiar, T. R. , Scott, M. J. (2013) Caspase 1 activation is protective against hepatocyte cell death by up-regulating beclin 1 protein and mitochondrial autophagy in the setting of redox stress. *J. Biol. Chem.* 288, 15947–15958. [PMC free article] [PubMed] [Google Scholar]

65. Franchi L, Eigenbrod T, Muñoz-Planillo R, Nuñez G. The inflammasome: a caspase-1-activation platform that regulates immune responses and disease pathogenesis. Nat Immunol. 2009 Mar;10(3):241-7. doi: 10.1038/ni.1703. PMID: 19221555; PMCID: PMC2820724.

66. Shi, T., Huang, L., Zhou, Y. *et al.* Role of GBP1 in innate immunity and potential as a tuberculosis biomarker. *Sci Rep* **12**, 11097 (2022). https://doi.org/10.1038/s41598-022-15482-2

67. Taylor MW. Interferons. Viruses and Man: A History of Interactions. 2014 Jul 22:101–19. doi: 10.1007/978-3-319-07758-1_7. PMCID: PMC7123835.

68. Yao, X., Liu, W., Li, X. et al. Whole blood GBP5 protein levels in patients with and without active tuberculosis. BMC Infect Dis 22, 328 (2022). https://doi.org/10.1186/s12879-022-07214-8.

69. Francisco, L. M., Sage, P. T. & Sharpe, A. H. The PD-1 pathway in tolerance and autoimmunity. *Immunol Rev* **236**, 219–242, https://doi.org/10.1111/j.1600-065X.2010.00923.x (2010).

70. Ratthé C, Girard D. Interleukin-15 enhances human neutrophil phagocytosis by a Syk-dependent mechanism: importance of the IL-15Ralpha chain. J Leukoc Biol. 2004 Jul;76(1):162-8. doi: 10.1189/jlb.0605298. Epub 2004 May 3. PMID: 15123770.

71. Grabstein KH, Eisenman J, Shanebeck K, Rauch C, Srinivasan S, Fung V, Beers C, Richardson J, Schoenborn MA, Ahdieh M, et al. Cloning of a T cell growth factor that interacts with the beta chain of the interleukin-2 receptor. Science. 1994 May 13;264(5161):965-8. doi: 10.1126/science.8178155. PMID: 8178155.

72. Badolato R, Ponzi AN, Millesimo M, Notarangelo LD, Musso T. Interleukin-15 (IL-15) induces IL-8 and monocyte chemotactic protein 1 production in human monocytes. Blood. 1997 Oct 1;90(7):2804-9. PMID: 9326248.

73. Maeurer MJ, Trinder P, Hommel G, Walter W, Freitag K, Atkins D, Störkel S. Interleukin-7 or interleukin-15 enhances survival of Mycobacterium tuberculosis-infected mice. Infect Immun. 2000 May;68(5):2962-70. doi: 10.1128/IAI.68.5.2962-2970.2000. PMID: 10768995; PMCID: PMC97510.

74. Petrilli, J.D., Araújo, L.E., da Silva, L.S. *et al.* Whole blood mRNA expression-based targets to discriminate active tuberculosis from latent infection and other pulmonary diseases. *Sci Rep* **10**, 22072 (2020). https://doi.org/10.1038/s41598-020-78793-2.

75. Dejean AS, Joulia E, Walzer T. The role of Eomes in human CD4 T cell differentiation: A question of context. Eur J Immunol. 2019 Jan;49(1):38-41. doi: 10.1002/eji.201848000. PMID: 30536524.

76. Mazzoni, A., Maggi, L., Siracusa, F., Ramazzotti, M., Rossi, M. C., Santarlasci, V., Montaini, G. et al., Eomes controls the development of Th17-derived (non-classic) Th1 cells during chronic inflammation. *Eur. J. Immunol.* 2018.

77. Gruarin, P., Maglie, S., De Simone, M., Haringer, B., Vasco, C., Ranzani, V., Bosotti, R. et al., Eomesodermin controls a unique differentiation program in human IL-10 and IFN-gamma coproducing regulatory T cells. *Eur. J. Immunol.* 2018.

78. Behrens M, Bartelt J, Reichling C, Winnig M, Kuhn C, Meyerhof W. Members of RTP and REEP gene families influence functional bitter taste receptor expression. J Biol Chem. 2006 Jul 21;281(29):20650-9. doi: 10.1074/jbc.M513637200. Epub 2006 May 23. PMID: 16720576.

79. Li H, Zhou Q, Ding Z, Wang Q. RTP4, a Biomarker Associated with Diagnosing Pulmonary Tuberculosis and Pan-Cancer Analysis. Mediators Inflamm. 2023 Apr 26;2023:2318473. doi: 10.1155/2023/2318473. PMID: 37152371; PMCID: PMC10156460.