Teesside University

School of Health & Life Sciences (SHLS)

MSc Dissertation

Course title: MSc Bioinformatics (With Advanced Practice)

Student name: ***Abiodun Joseph Akinade***

Student ID: ***A0413291***

Module title: Life Science Research Project

Module code: SCI4013-N-BF1-2023

Module leader: Dr. Mohammad Dadashipour (DADASHI)

Project title: A Genome-Wide Transcriptomics Analysis of *Mycobacterium tuberculosis*

Project supervisor: Dr. Shweta Kuba

Date of submission: 11/01/2024

Word count = 8,472

Semester & year: Semester 1, 2023 - 2024

Author Biography

Abiodun Joseph Akinade is the author of this research project. My main research interest is in the One Health approach for the control and prevention of Infectious diseases, epidemiology, and management of antimicrobial use and antimicrobial resistance surveillance. I hold a B.Tech in Animal Production and Health and MSc. in Public Health. My quest for knowledge and to further broaden my research and statistical analysis perspective made me apply for another masters (MSc in Bioinformatics), which is due to be completed in January 2024. I am particularly fascinated by the interplay between animals, humans, and the environment and interested in exploring the connection between them through the knowledge of bioinformatic analysis I have acquired during this Master's programme.

ABSTRACT

Mycobacterium tuberculosis (*Mtb*), the etiological agent of tuberculosis, is an infectious disease of bacteria that poses major concerns for public health worldwide because of its threat to health for many years. The challenges in diagnosis, prolonged treatment and drug resistance mechanism of tuberculosis necessitate the need to investigate the genetic factors that play a significant role in the pathogenesis of *Mtb*. This research aimed at utilising genome-wide transcriptomics analysis to identify the molecular signature of *Mtb*.

The gene expression profile datasets of active tuberculosis and healthy control were retrieved from the Gene Expression Omnibus (GEO) database. The four selected microarray datasets were analysed using Limma package in R to identify the Differentially Expressed Genes (DEGs). The functional enrichments were conducted using WebGestalt. The protein-protein interaction (PPI) network was constructed using STRING and Cytoscape software was used to visualise the hub genes. Finally, the Drug-target interactions were identified in DrugBank using the DGidb database.

The analysis produced a total of 36 common DEGs associated with *Mtb*. The functional analysis revealed that the genes were majorly enriched in biological regulation, membrane and protein binding. The pathway enrichment of the genes was mainly in immune responses. The high expression of GBP5, GBP1, BATF2 and other 10 hub genes may play a crucial role in the pathogenesis of *Mtb* which may be a biomarker for early diagnosis while SERPING1, LAP3, ADM, CACNA1I and BMX may be helpful in the development of novel therapy for tuberculosis disease in the future.

Keywords: *Mycobacterium tuberculosis or M. tuberculosis, Genome-wide analysis, transcriptomics, Molecular signature, Biomarker, Differentially Expressed Genes, Core genes.*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Adj-*P* value – Adjusted *P* value

ATB – Active tuberculosis

BATF2 – Basic Leucine Zipper Transcription Factor 2

BCG – Bacillus Calmette-Guerin

BP – Biological Process

CC – Cellular Component

CFP-10 – Culture Filtrate Protein, 10 kDa

COVID-19 – Coronavirus Disease

DAVID – Database for Annotation, Visualization and Integrated Discovery

DEGs – Differentially Expressed Genes

DGidb **–** Drug Interaction Database

EPTB – Extrapulmonary Tuberculosis

ESAT-6 – Early Secreted Antigenic Target, 6 kDa Protein

FBA – Flux Balance Analysis

FC – Fold Change

FDR – False Discovery Rate

GBPS – Guanylate Binding Proteins

GEO – Gene Expression Omnibus

GO – Gene ontology

GSEA - Gene Set Enrichment Analysis

IBD – Inflammatory Bowel Disease

KEGG – Kyolo Encyclopaedia of Gene and Genomes

Limma Packages – Linear Models for Microarray Analysis packages

LTBI – Latent Tuberculosis Infection

M. tuberculosis – *Mycobacterium tuberculosis*

MAP – Mycolic Acid Pathway

MCC – Maximal Clique Centrality

MF – Molecular Function

Mtb – *M. tuberculosis*

MTBC – *Mycobacterium tuberculosis* complex

NAA – Nucleic-Acid Amplification

NCBI – National Center for Biotechnology Information

NSAID – Nonsteroidal Anti-Inflammatory Drug

NTA – Network Topology-based Analysis

ORA – Over-Representation Analysis

PBMC – Peripheral Blood Mononuclear Cell

PPI – Protein-Protein Interaction

PTB – Pulmonary tuberculosis PTB

PZP – Pregnancy Zone Protein

STRING – Search Tool for the Retrieval of Interacting Genes

TB – Tuberculosis

WebGestalt – WEB-based Gene SeT AnaLysis Toolkit

WHO – World Health Organisation

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

Mycobacterium tuberculosis (*M. tuberculosis*), the etiological agent of tuberculosis is an infectious disease of bacterium that is characterised by the progressive development of certain granulomatous lesions or tubercles in the lung tissues (pulmonary tuberculosis), lymph nodes and other body parts (extrapulmonary tuberculosis) (Ahmad, 2011; Alam *et al*., 2019). It is estimated that *M. tuberculosis* latently infects about one-third of the world's population and causes between 8 and 10 million newly diagnosed cases of active tuberculosis (ATB) annually (Ottenhoff *et al*., 2012). Tuberculosis is one of the key health issues worldwide, especially in many developing countries. However, with more than one billion migrants all over the world (Dhavan *et al*., 2017), developed countries are not exempted from the threat of tuberculosis and so has become a major concern of public health in many countries. A better and clearer understanding and knowledge of the behaviour of transmission, its implications, possible future predictions about the method of transmission, tuberculosis diagnosis, therapeutic approach and survival rate of patients has been understood with the application of bioinformatics tools and system biology analysis.

This study uses genome-wide transcriptomics analysis to identify the molecular signature of *M. tuberculosis (Mtb)*. In this study, differential expression analyses were carried out to identifying Differentially Expressed Genes (DEGs) of *Mtb* from datasets downloaded from the Genes Expression Omnibus (GEO) database, functional enrichment and pathway, Protein-Protein Interactions (PPI) as well as drug interaction of the genes. The DEGs were identified using Linear Models for Microarray Analysis (Limma) packages in R programming language, functional enrichment and pathway were done using WebGestalt database, PPI network was computed by the Search Tool for the Retrieval of Interacting Genes (STRING) and constructed using Cytoscape software. The drug interactions were obtained from the DGidb database.

1.2 Epidemiology of *Mycobacterium tuberculosis*

Although tuberculosis is a largely preventable and curable disease. However, in 2022, it was the second most common infectious agent-related cause of death worldwide after coronavirus disease (COVID-19), tuberculosis caused nearly twice as many deaths as HIV/AIDS. Each year, more than 10 million people still contract tuberculosis (WHO, 2023). It is estimated that a quarter of the world's population has tuberculosis infection (Raman and Chandra, 2011; Houben and Dodd, 2016). After infection, the first two years have the highest chance of developing tuberculosis (about 5%), shortly after which the risk is significantly reduced (Menzies *et al.,* 2018) and some individuals will recover from the illness. Approximately 90% of all cases of tuberculosis occur in adults, with a higher incidence rate in men than in women (WHO, 2023).

In 2022, 7.5 million people worldwide were tuberculosis diagnosed newly and were formally reported as tuberculosis incidents. This was higher than the pre-COVID level (7.1 million in 2019), 16% higher than the level in 2021, 28% higher than the level in 2020, and the highest number in a single year since the WHO began tracking tuberculosis globally in the mid-1990s (Figure 1) (WHO, 2023). Over time, *Mtb* has adapted new subversion means to effectively evade the host's immune system and survive within the host, leading to the active or latent manifestation of disease (Ponnusamy and Arumugam, 2022).

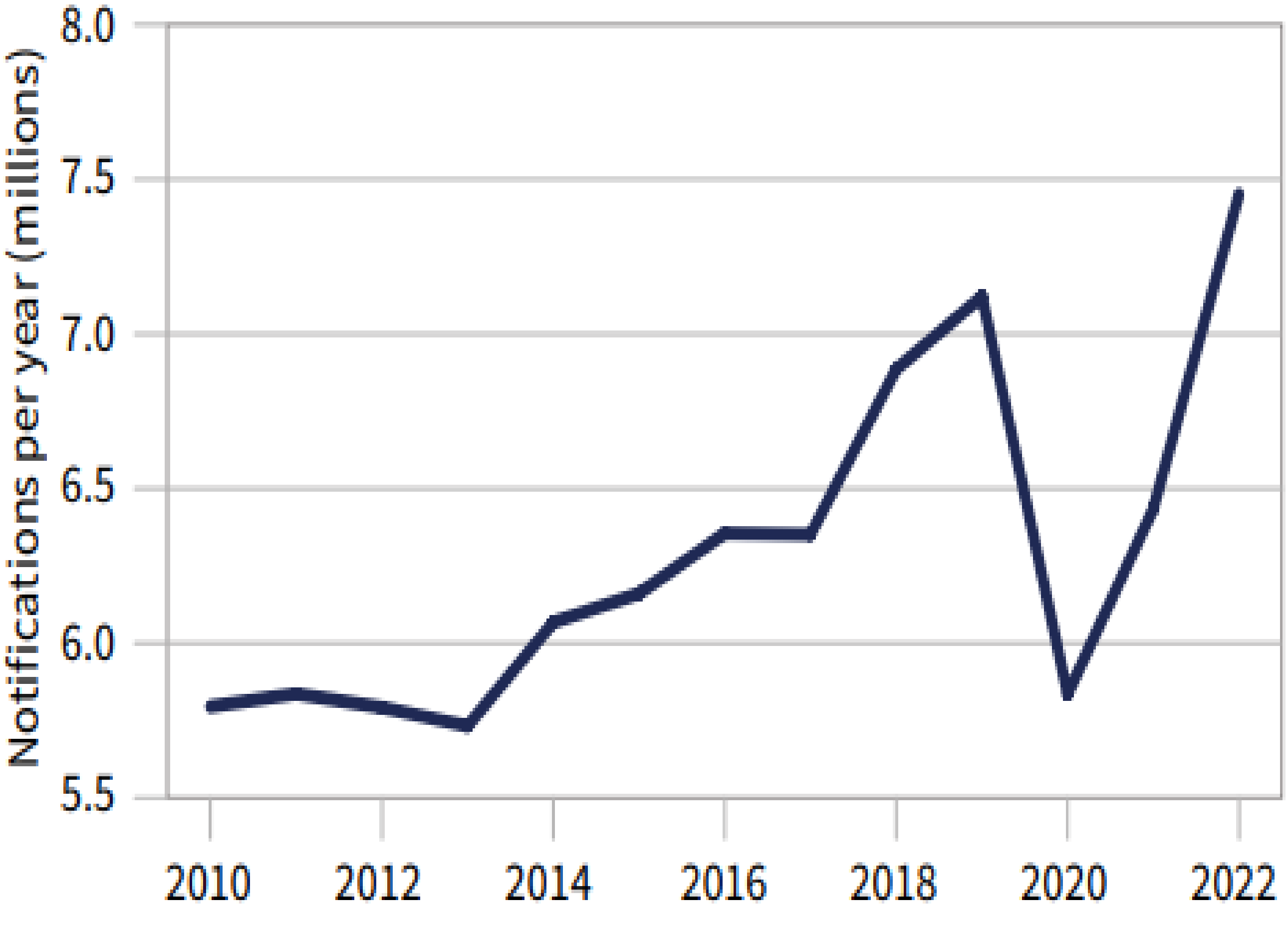


Figure 1: *Global Trend in Case of People Diagnosed with Tuberculosis, 2010 to 2022 (World Health Organisation (WHO), 2023)*

The mortality rate of tuberculosis disease is estimated to be high (approximately 50%) without treatment. However, approximately 85% of tuberculosis patients can be cured with the current treatments (a 4-to-6-month anti-tuberculosis medication regimen) recommended by WHO. Also, available are the 1-to-6-months regimens for tuberculosis infection treatment. It is also possible to lower the number of tuberculosis-related infections and illnesses (and consequently, the number of tuberculosis-related deaths) by implementing multisectoral initiatives to address tuberculosis determinants like undernourishment, poverty, HIV infection, diabetes and smoking, (WHO, 2023). Research advancements (such as identifying the core genes responsible for tuberculosis, development of a new vaccine and treatment) are required to quickly bring the number of new cases per year (i.e., tuberculosis incidence) down worldwide.

1.3 *Mycobacterium tuberculosis* Species

The Mycobacteriaceae family is made up of a wide range of bacteria that have different characteristics that make them pathogenic to humans and animals, as well as different host reservoirs and growth dynamics in culture (Kanabalan *et al*., 2021). These bacteria are primarily aerobic, Gram-positive, non-spore-forming, acid-fast bacilli, non-motile species that have a slightly curved shape and may branch out from their cell wall made of mycolic acid (Fong, 2020). The complex cell wall envelope of members of the *Mycobacterium* genus causes the cells to have low permeability. Additionally, the Zhiel-Neelsen acid-fast stain differential staining method can be used to distinguish the genus from another bacterial genus (Kanabalan *et al*., 2021). Based on their rates of growth, the *Mycobacterium* genus can be divided into two main groups: fast-growing and slow-growing *Mycobacteria*. For example, the fast-growing *Mycobacteria* consists of *Mycobacterium smegmatis,* a non-pathogenic or opportunistic bacteria in general whereas the slow-growing *Mycobacteria* such as *M. tuberculosis, Mycobacterium bovis* (*M. bovis*) *and Mycobacterium leprae* (*M. leprae*) which causes human tuberculosis (H. tuberculosis), bovine tuberculosis (B. tuberculosis) and leprosy, respectively (Forrellad *et al.,* 2013; Kanabalan *et al.,* 2021).

Mycobacterium tuberculosis complex (MTBC) comprises a group of genetically related *Mycobacteria* such as *M. tuberculosis, M. bovis, M. africanum, M. canettii, M. caprae, M. pinnipedii* and *M. microti* (Forrellad *et al.,* 2013)*.* In addition to the seven common species listed above, two novel species, *Mycobacterium orygis* and *Mycobacterium mungi*, are also referred to as MTBC (Pfyffer, 2015). A subgroup of the pathogenic species are animal-adapted strains that infect various mammalian species. Among them are *M. bovis* (found in cows), *M. caprae* (found in goats and sheep), *M. orygis* (found in oryxes), *M. microti* (found in voles) and *M. pinipedii* (found in seals or sea lions) (Smith *et al*., 2006; van Ingen *et al*., 2012; Kanabalan *et al.,* 2021). A summary of the fundamental traits of the MTBC members is shown in Table 1.

Table 1: *A Summary of the Fundamental Traits of the Mycobacterium tuberculosis complex's (MTBC) members (Forrellad et al., 2013; Pfyffer, 2015; Kanabalan et al., 2021)*

Species of MTBC	Summary of the Fundamental Traits
M. tuberculosis	Human tuberculosis. Most widely known species of the MTBC. Affecting more than one-third of the world's population. Causes between 8 and 10 million newly diagnosed cases of ATB It can be transmitted from humans to animals.
M. bovis	Bovine tuberculosis. Exhibits the broadest range of host infections. This affects cattle, goats, both domestic and wild and humans. Used to develop laboratory-selected mutant "*M. bovis var BCG*" early childhood sole vaccine used for the prevention of tuberculosis.
M. africanum and *M. canettii*	A strong connection to *M. tuberculosis*. Affecting humans. Typical isolates of patients from Africa.
M. caprae	An infection of goats.
M. orygis	An infection of large mammals on the continent of Africa, such as oryxes, antelopes, gazelles, and waterbucks.
M. microti	An infection from rodents
M. pinnipedii	Pathogens from seals or sea lions
M. mungi	Etiological agent of banded mongoose TB (*Mungo mungo).*

1.4 Molecular Signature of *Mycobacterium tuberculosis*

The virulence determinant of *Mtb* is enhanced by a secretion system encoded by the region of difference 1 locus (Volkman *et al.,* 2004), which is the primary molecular mechanism responsible for BCG attenuation (Pym *et al.,* 2002; Sgaragli and Frosini, 2016), identified as early secreted antigenic target, 6 kDa protein (ESAT-6) secreted by bacteria. Enhancing the recruitment of macrophages interacts with the host epithelium to induce matrix metalloproteinase-9 (MMP-9), which in turn promotes the maturation of nascent granulomas and bacterial growth (Volkman *et al.,* 2010). The culture filtrate protein, 10 kDa (CFP-10) family of mycobacterial secreted proteins is another group of proteins that ether-dimerizes with ESAT-6 (Renshaw *et al.,* 2002). Strong T-cell antigens such as CFP-10 and ESAT-6, are detected by serum from tuberculosis patients (Ulrichs *et al.,* 1998). The macrophage signaling pathway is modulated by ESAT6, CFP-10, and their complex, specifically the ERK 1/2 MAP kinase pathway (Ganguly *et al.,* 2007) by significantly reducing the phosphorylation of extracellular signal-regulated kinases 1/2 (ERK1/2) in the nucleus and then activating them. Increased phosphatase activity in the nucleus mediates this inhibition, leading to the dephosphorylation of pERK1/2 originating from the cytoplasm. The expression of the LPS-inducible c-myc gene is down-regulated because of the restriction of ERK1/2 activation, which also impacts the expression of c-Myc, a crucial component in macrophage activation (Sgaragli and Frosini, 2016). T-cell-derived lymphokines orchestrate CMI, which is executed by TDM-activated macrophage effector cells (Russell, 2007), is so effective that, during their lifetime, 90% of immunocompetent humans infected with *Mtb* can suppress the infection (LTB) and prevent the disease from progressing to clinical disease (ATB) (Sutherland, 1976; Sgaragli and Frosini, 2016).

The series of immunological events begins with alveolar macrophages engulfing *Mtb*. In turn, the bacilli may prevent macrophages from undergoing phago-lysosomal fusion. The macrophages secrete cytokines (IL1 [CXCL8], TNFα, IFN-γ, IL10, TGFβ, IL12, GM-CSF, RANTES [CCL5] and MCP1 [CCL2]). Major Histocompatibility Complex (MHC) is used to present *Mtb* antigens to CD4+, CD8+ T helper cells, CD1 and γδT cells. Antigens leave the lungs and are transported to draining lymph nodes after being presented to dendritic cells as well (Figure 2). T helper cells that are CD4+ and eventually CD8+ become activated in the tissues of lymph nodes. To increase the pool of lymphocytes specific for antigen, CD4+ cells produce IL2. The primed T cells return to the lung infection site and induce the development of granulomas (Sgaragli and Frosini, 2016).

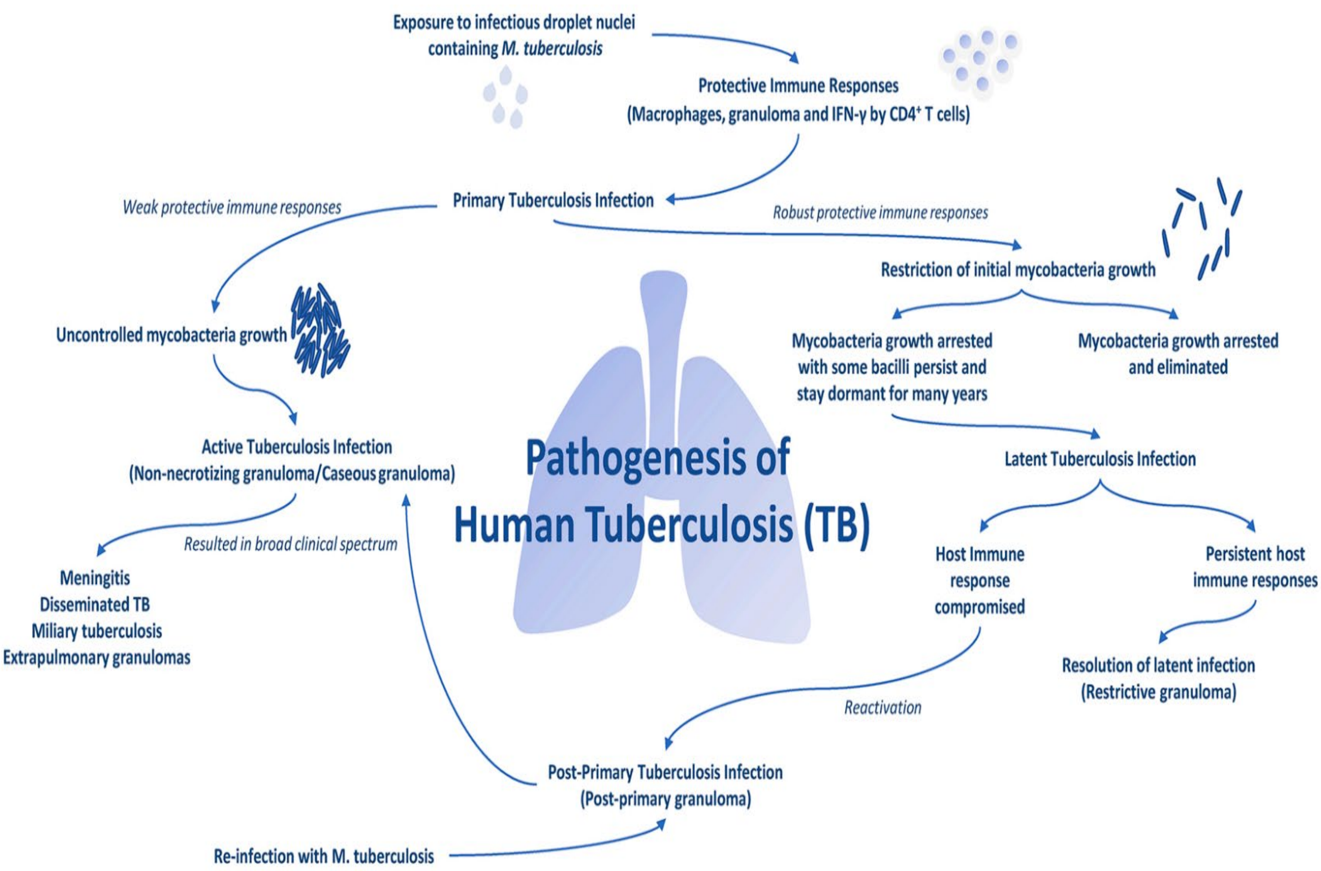

Exposure to infectious droplet nuclei containing M. tuberculosis
Protective Immune Responses (Macrophages, granuloma and IFN-γ by CD4+ T cells)
Primary Tuberculosis Infection
Weak protective immune responses
Robust protective immune responses
Restriction of initial mycobacteria growth
Uncontrolled mycobacteria growth
Mycobacteria growth arrested with some bacilli persist and stay dormant for many years
Mycobacteria growth arrested and eliminated
Active Tuberculosis Infection (Non-necrotizing granuloma/Caseous granuloma)
Pathogenesis of Human Tuberculosis (TB)
Latent Tuberculosis Infection
Resulted in broad clinical spectrum
Meningitis
Disseminated TB
Miliary tuberculosis
Extrapulmonary granulomas
Host Immune response compromised
Persistent host immune responses
Resolution of latent infection (Restrictive granuloma)
Reactivation
Post-Primary Tuberculosis Infection (Post-primary granuloma)
Re-infection with M. tuberculosis

Figure 2: *Pathogenesis of Human Tuberculosis (*Kanabalan *et al.,* 2021)

1.5 Genome-Wide Transcriptomic Analysis of *Mycobacterium tuberculosis*

Mycobacterium tuberculosis (*Mtb*), the bacterium that causes tuberculosis, continues to be a major global health concern, requiring a better understanding of the molecular mechanisms underlying its pathogenesis (Sgaragli *et al.,* 2016; Kanabalan *et al.,* 2021). Blood culture isolates provide a unique perspective on tuberculosis because they show how the bacterium behaves in the bloodstream, which is a vital site for the spread of infection. The investigation of *Mtb*'s genetic composition, transcriptional landscape, and protein expression patterns in this context is made possible by the integration of cutting-edge molecular biology techniques, including whole-genome sequencing, RNA-seq, and mass spectrometry-based proteomics (Li *et al.,* 2023). One essential tool for organising and understanding the enormous datasets produced by these high-throughput methods is bioinformatics. It enables the creation of complex gene regulatory networks, the identification of genetic variants, and differential gene expression patterns

(Mehmood *et al.,* 2014). Systems biology techniques support these analyses by offering a comprehensive comprehension of the intricate relationships between *Mtb* and the host, enabling researchers to go beyond studies of single genes and understand the larger biological context (Raman and Chandra, 2011). It is crucial to identify the important genes for several reasons. Firstly, it clarifies the particular genetic variables that allow *Mtb* to proliferate in the bloodstream and aid in its systemic spread. Secondly, these key genes might be crucial to the virulence of the bacteria, affecting the severity of the illness and its treatment results. Thirdly, comprehending the molecular aspects of the host-pathogen interaction provides opportunities for tailored therapeutic approaches, encompassing innovative pharmacological targets and plausible biomarkers for prognostic and diagnostic functions (Doran and Fulde, 2016).

Blood can reflect immunological and pathological changes in the parts of the body, Numerous studies have confirmed that human blood samples can be used to diagnose a variety of inflammatory and infectious diseases (Li *et al.,* 2023). For example, Shao *et al.,* (2021), found that the pregnancy zone protein (PZP) content of serum exosomes in patients with inflammatory bowel disease (IBD) can help with clinical therapy and may be a useful diagnostic biomarker. Li *et al.,* (2022) constructed a PPI network after studying the gene expression profiles in peripheral blood mononuclear cell (PBMC) samples. In addition, it has been discovered that three genes, HP, FUCA2 and SERPINA1 may be important in PTB. Validation of immune markers (for example, CXCL1, CXCL2, CXCL10, CCL1 and CCL3,) in plasma clearly distinguishes confirmed tuberculosis from unconfirmed tuberculosis in children, as demonstrated by a study that provided fresh evidence for the application of chemokines in plasma as markers of children tuberculosis (Kumar *et al.,* 2021).

1.6 Genomics and Discovery of Host Biomarkers

To treat tuberculosis, over 20 medications as well as the Bacillus Calmette-Guerin (BCG) vaccine are available. Even though the current medications are extremely valuable, they have several drawbacks. The most significant one is the development of drug resistance, which makes even the front-line medications ineffective (Raman and Chandra, 2011). Protease inhibitors, for instance, have been demonstrated to be incompatible with anti-tuberculosis regimens that contain rifampicin (Bonora and Di Perri 2008). Several obstacles in the fight against tuberculosis require the use of more advanced methods to research, comprehend, and develop strategies to combat tubercular infection (Raman and Chandra, 2011).

The fields of genomics and post-genomics are experiencing a rapid expansion in the types and quantity of information available, encompassing not only genome sequences and protein structures but also gene expression, regulation, and protein-protein interactions. This is due to the simultaneous advancements in high-throughput experimental methods and screening techniques to analyse whole genomes and proteomes (Raman and Chandra, 2011). There are now several in silico approaches to systematically address important questions in biology, with a clear impact on drug discovery, thanks to the availability of such data in publicly accessible databases and the advancements in computational power and methods for data mining and modeling (Apic *et al.*, 2005; Claus and Underwood 2002). Several steps in the drug discovery process benefit from systems-level approaches, especially target identification and determining the molecular cause of disease for sensible drug discovery (Raman and Chandra, 2011).

According to Raman *et al.,* (2005), the mycolic acid pathway (MAP) was stimulated and reconstructed for *Mtb* using Flux Balance Analysis (FBA), a method for analysing metabolic networks based on constraints. The biosynthesis of mycolic acids was mathematically abstracted, and the pathway was studied using fluorescence band alignment (FBA). This allowed for the identification of critical points in the pathway and the delineation of possible drug targets. Two genome-scale reconstructions of *Mtb* were published in 2007 (Beste *et al*., 2007; Jamshidi and Palsson 2007), through the examination of key genes and hard-coupled reaction sets, with applications in drug target identification (Raman and Chandra, 2011).

The identification of differential gene variation signatures in MTBC, specifically *Mtb* infection, coupled with a significant spike in host omics data has given rise to a valuable window of information regarding the role that genetic variation plays in tuberculosis diagnosis (Kanabalan *et al.,* 2021). For instance, Chang *et al*., (2018) analysed the genetic variations of the gene encoding IFN-induced protein-SP110 in a sizable patient cohort from Taiwan that included 68 latent tuberculosis infections, 301 active cases of tuberculosis, and 278 healthy controls. Among the five SNPs in the SP110 gene—rs7580912, rs7580900, rs9061, rs2241525, and rs11556887), the authors found that rs9061 is substantially associated with disease susceptibility to LTB1. The authors' additional analysis reveals that SP110 rs9061 SNP is linked to lower plasma TNF-α levels in patients with latent tuberculosis infection. This implies that genetic polymorphisms in SP110 could function as biomarkers for susceptibility to both latent and active tuberculosis infection in humans (Chang *et al*., 2018). However, a pilot study was conducted to investigate the possibility of epigenetic changes in immune cells serving as a tuberculosis biomarker (Esterhuyse *et al.,* 2015). The neutrophils and monocytes isolated from patients with latent and active tuberculosis infections were subjected to simultaneous transcriptome, proteome, and epigenome analyses by the authors. The authors emphasised the role that microRNAs and the epigenome (DNA methylation profiles) play in controlling function in both latent and active tuberculosis infections. Large-scale DNA methylome analysis based on age, gender, and cell types in tuberculosis infection is therefore required (Esterhuyse *et al.,* 2015; Kanabalan *et al.,* 2021).

1.7 Transcriptomics and Discovery of Host Biomarkers

Transcriptomics has been extensively used for the past decade to simplify the host-mycobacterial interaction and find likely host biomarkers for tuberculosis diagnosis. Examining blood transcriptomics profiles also helps us comprehend how host elements and the underlying molecular mechanism of *Mtb* infection are intertwined (Kanabalan *et al.,* 2021). Furthermore, although the primary immune response against *Mtb* is concentrated in the lungs, pathological events during tuberculosis infection are typically reflected in the peripheral blood by circulating host immune cells (Weiner *et al.,* 2013). Numerous research works have used blood transcriptomics analysis to discover host biomarkers. For example, a microarray study and quantitative polymerase chain reaction analysis identified several genes that were expressed differently in monocytes between peripheral blood mononuclear cells from tuberculosis patients and healthy donors infected with *Mtb*. These genes were primarily derived from monocytes (Jacobsen *et al.,* 2007). In another blood transcriptome study conducted by Berry *et al.,* (2010), 393 transcripts were associated with active tuberculosis. The authors also identified 86 distinct transcript signatures that distinguish tuberculosis from other inflammatory and infectious diseases. The authors emphasised that type-1 IFN-αβ signaling and IFN-γ are among the neutrophil-driven interferon-inducible genes that are primarily expressed in ATB (Berry *et al.,* 2010). Furthermore, Lee et al. (2016) showed that patients with LTBI greatly exhibit gene expression linked to natural killer cell activation and apoptosis, whereas the expression of innate immune-related genes is strongly correlated with ATB (Lee *et al*. 2016).

Microarray technology advancements have made it possible to analyse mRNA expression profiles at the genome-scale in a variety of organisms, including *Mtb*. Waddell *et al.,* (2007) reported a thorough examination of *Mtb*'s genome-scale expression analyses (Waddell *et al.,* 2007). A study conducted by Boshoff *et al.,* (2004) also reported an extensive examination of *Mtb*’s differential transcriptional reactions to growth-inhibitory conditions and medications (Boshoff *et al.,* 2004). In a different study, microarray analysis was used to determine *Mtb*'s response to the minimal inhibitory concentrations of six anti-microbials to clarify the mechanisms of *Mtb*'s innate resistance (Waddell *et al.,* 2004). Studies have also been conducted on the expression of *Mtb* genes in macrophages (Schnappinger *et al.,* 2003) by employing microarray technology to analyse RNA extracted from infected mouse macrophages. When comparing the macrophages to broth cultures, 454 induced and 147 repressed *Mtb* genes were found; these genes are referred to as the "differential intraphagosomal transcriptome." (Raman and Chandra, 2011). The integration of genome-scale transcriptional analyses can yield a multitude of data, which can contribute to a better understanding of the pathogenesis of TB disease.

Furthermore, there is increasing evidence that studying the impact of non-coding RNA at various stages of tuberculosis infection can be accomplished through RNA sequencing (Kanabalan *et al.,* 2021). The differential expression of a panel of microRNAs (miRNAs) between latent and ATB infections has been found in several studies, underscoring the potential of miRNAs as useful biomarkers in tuberculosis infection (Chakrabarty *et al.,* 2019; Lyu *et al.,* 2019; Kanabalan *et al.,* 2021). In addition to miRNA, other research by de Araujo *et al.,* (2019) showed that piRNA and small nucleolar RNA (snoRNA) may also be useful biomarkers to distinguish between latent and ATB infection (de Araujo *et al.,* 2019). However, the potential of circular RNA (circRNA) as a tuberculosis biomarker has been investigated by Fu *et al.,* (2019). The authors found that 171 deregulated circRNA were found in tuberculosis patients, with circRNA_101128, circRNA_103017, and circRNA_059914 being significantly up-regulated and circRNA_062400 being significantly down-regulated (Fu *et al.,* 2019). According to research by Lv *et al.,* (2017), there is a difference in the expression of exosomes between latent and ATB infections. This suggests that different stages of *Mtb* infection can cause different RNA cargoes to be packaged into exosomes. Further functional and pathway analysis showed that the immune system and the signaling pathway were downregulated, while the apoptotic and necrotic processes were upregulated (Lv *et al.,* 2017).

1.8 Functional Linkages in *Mycobacterium tuberculosis*

Protein-protein interactions play a crucial role in controlling cellular processes. They serve as the building blocks of numerous transcriptional regulatory networks and signal transduction pathways in the cell. The understanding of the structure and function of proteins has been a crucial motivator for biological research in the last few decades (Raman and Chandra, 2011). It is possible to deduce functional interactions between proteins across the genome through computational analyses or high-throughput experiments. There have been reported genome-wide functional linkages in *Mtb* by Eisenberg and colleagues (Strong *et al.,* 2003). Raman and Chandra, (2011) stated that it is possible to determine functionally linked gene clusters throughout the proteome and deduce the function of uncharacterized proteins by grouping proteins with comparable functional linkage profiles. These protein-protein interaction maps are also useful for identifying drug targets and in the resistance pathway analysis (Verkhedkar *et al*. 2007; Raman *et al.* 2008; Raman and Chandra 2008). It has been demonstrated that numerous highly connected proteins in protein interaction networks, also known as "hubs," are essential for cellular function; hub proteins like these could also be targets for medications (Jeong *et al.,* 2001; Verkhedkar *et al*. 2007).

False positives and negatives are other problems with computational approaches for predicting functional linkages, but these can be avoided by taking consensus predictions from several approaches into account. The STRING database gives each interaction a confidence score after taking into account predictions made using a variety of techniques and experimental data (Von Mering *et al.* 2007; Raman and Chandra, 2011). Future analyses will likely be more reliable due to the significant improvement in the quality of constructed interactomes, which will be made possible by advancements in both computational and experimental methods for defining protein-protein interactions (Raman and Chandra, 2011).

1.9 Significance of the Study

Pulmonary tuberculosis (PTB) and extrapulmonary tuberculosis (EPTB) are two subtypes of tuberculosis, a chronic infectious disease caused by *Mtb* (Harding, 2020; Li *et al.,* 2023). PTB is the most prevalent clinical form of tuberculosis among them, while tuberculosis affecting organs other than the lungs, such as the pleura, lymph nodes, bones, meninges, etc., is referred to as EPTB (Cukic and Ustamujic 2018; Holden *et al.*, 2019). A variety of symptoms, such as weight loss, cough and fever are experienced by people who harbour tuberculosis, and some people infected with *Mtb* may also develop latent tuberculosis infection (LTBI) (LoBue and Mermin, 2017). For the effective control of tuberculosis transmission, early diagnosis and appropriate therapy are therefore crucial. For example, sputum detection must come later because radiological diagnosis is not sufficient to reach a definitive independent diagnosis (Woodring *et al.,* 1986; Krysl *et al.,* 1994). The best sample for identifying lung diseases is sputum, and the most popular technique for tuberculosis diagnosis is sputum smear microscopy. Nevertheless, despite being simple to use, the technique requires a lot of time and has a low threshold for detection (Schaberg *et al.,* 1995; Steingart *et al.,* 2006). Although the sensitivity of nucleic-acid amplification (NAA) molecule detection is too low, it is a dependable technique for increasing diagnostic specificity, particularly in cases of smear-negative (paucibacillary) disease in which the clinical diagnosis is unclear (Greco *et al.,* 2006; Ling *et al.,* 2008). Early detection and therapy are delayed as a result of traditional methods' inability to detect *Mtb* quickly and effectively. More and more data shows that the population with tuberculosis is at an increased risk of long-term disability and death, with the percentage of mortality surpassing the total of all other common infectious diseases (Romanowski *et al.,* 2019; Li *et al.,* 2023). As a result, improving diagnostic techniques is crucial to increasing tuberculosis patients' chances of survival and this called for the importance of studying *Mtb* molecular signature through a genome-wide transcriptomics analysis for understanding pathogenesis, epidemiology and treatment of tuberculosis.

The findings from this study hold significant implications for both advancing academic knowledge and clinical applications. The knowledge gathered from these analyses can guide the creation of focused treatment plans and advance our understanding of the pathophysiology of tuberculosis. These are some main ideas emphasising how important this identification is:

Understanding Pathogenesis of Tuberculosis: one of the major significant of this study is to gain an understanding of the molecular processes that underlie *Mtb* infection and comprehend the genetic elements implicated in the adaptation to the host environment which makes *Mtb* thrive and survive in the host.

Discovery of Biomarker: identifying the core genes of *Mtb* may help in biomarker discovery for early detection, treatment and monitoring of tuberculosis which may help to reduce the spread of the disease

Understanding of Virulence and Host Interaction: some genes may be associated with the virulence of *Mtb*, thereby influencing its capacity to cause disease in the host. Knowledge of the interaction between such genes and the immune system of the host may help the development of interventions to modify the host response.

Drug Resistance Prediction: this can help to understand the genetic basis of resistance and predict drug resistance species for better development of therapeutic strategies.

Discovery of Drug Target: this can give insight into novel drug target discovery and may also increase the effectiveness of current medications thereby improving the treatment of tuberculosis.

Research and Development: this may serve as a basis for further tuberculosis research and development and can encourage continued efforts to improve treatment approaches, preventative measures, and diagnostic instruments.

1.10 Scope and Objectives of the Study

The scope of this study is to conduct a comprehensive genome-wide transcriptomics analysis of *Mtb* and identify the molecular signature associated with *Mtb*. The workflow of the research will involve analysing large-scale transcriptomics and genomics data to profile gene expression patterns in *Mtb*, to identify the differentially expressed genes. Achieving success in this study will involve the use of various bioinformatics tools and system biology analysis. This project aims to identify the core genes and drug target interactions of active tuberculosis using genome-wide transcriptomics analysis.

The specific objectives of this study are stated below:

- To identify the Differentially Expressed Genes (DEGs) in active tuberculosis.
- To carry out functional enrichment and KEGG pathway analysis on the identified genes to help in understanding the pathogenesis of *Mycobacterium tuberculosis*.
- To construct a Protein-Protein Interaction (PPI) network of genes to identify the potential hub genes in the pathogenesis of *M. tuberculosis*.
- To identify the list of drugs in the DrugBank targeting the genes

1.11 Organisation of the Study

Chapter one would be based on the introduction of the study, it would include the background of the study, exploring existing literature on *Mycobacterium tuberculosis*, significance of the study, scope and objectives of the study, and organisation of the study. Chapter two would be based on the research methodology, including research design, data collection and processing, statistical analysis and data validation. Chapter three will present the results of the analysis, including all analysis, explanations, figures and tables generated from the statistical analysis. Chapter four will focus on the discussion of the results and interpretations of the analysis's findings. Chapter five would be the final chapter and focus on presenting the study's conclusion and recommendations, and it would also cover its limitations.

CHAPTER TWO

METHODOLOGY

This chapter describes the study methodology and research design, clarifies the methods used in data collection, and also explains various techniques and tools used in analysing this study. The concept of this study is focused on quantitative research methodology because it aims at analysing broad gene profile datasets, calculating the DEGs in ATB patients and healthy controls and conducting precise measurements of the genes' interactions. Workflow of the Study with summary of methods are shown in (Figure 3).

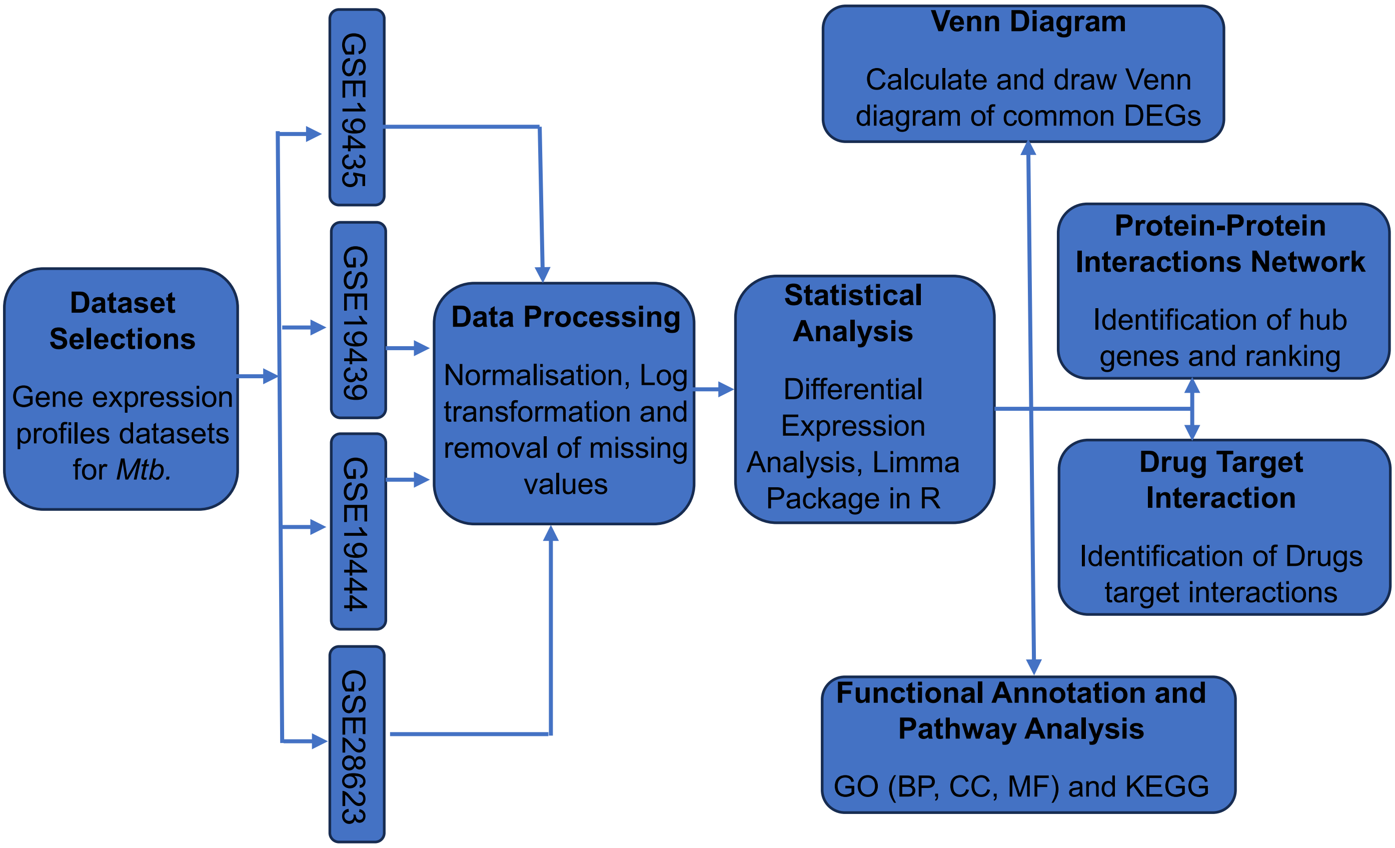

Dataset Selections
Gene expression profiles datasets for *Mtb.*
GSE19435
GSE19439
GSE19444
GSE28623
Data Processing
Normalisation, Log transformation and removal of missing values
Statistical Analysis
Differential Expression Analysis, Limma Package in R
Venn Diagram
Calculate and draw Venn diagram of common DEGs
Protein-Protein Interactions Network
Identification of hub genes and ranking
Drug Target Interaction
Identification of Drugs target interactions
Functional Annotation and Pathway Analysis
GO (BP, CC, MF) and KEGG

Figure 3: *Workflow of the Study with summary of methods.*

2.1 Research Design

Research design is a part of research methodology dealing with how the research is done (Goundar, 2012). Research design ensures that the data gathered, and the proof obtained can effectively and succinctly address the original research question. The process of gathering data or acquiring concrete evidence entails defining the kinds of data or evidence needed to address a research question, assess a programme, or verify a theory. Before beginning data collection or analysis, a structure or design must be developed for any research. This means that before beginning any research, we should honestly consider the following questions: concentrating on this research question, what kind of evidence or data is needed to provide a clear and compelling response? (Creswell, 2017). The research design that will be adopted for this study is a retrospective case-control study. This is because it permits the comparison of data and the evaluation of existing records.

According to Tasiou *et al.,* (2017), a retrospective case-control study compares people with a particular disease/condition (cases) to people without the disease/condition (controls). The research goes back in time to find and examine any possible risk factors that may have aided in the emergence of the disease/condition (Tasiou *et al.,* 2017).

This study will compare the genetic profiles of individuals with ATB (cases) to healthy individuals (controls). This will involve analysing whole blood isolate samples from previous ATB patients and healthy individuals. The study will analyse gene expression profile datasets of ATB and healthy controls. A series of bioinformatics tools will be employed to identify DEGs, biological processes, cellular components, molecular functions and pathways of the genes that are differentially expressed in ATB patients compared to the healthy control. This study will further construct the PPI of the DEGs, find the hub genes that may serve as markers for tuberculosis diagnosis and identify the list of drugs in the DrugBank targeting these DEGs, which may be helpful in the development of vaccines or drugs for tuberculosis.

2.2 Data Collection and Preparation

The gene expression profiles datasets of ATB and healthy control analysed in this study were retrieved from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/) built and maintained by the National Center for Biotechnology Information (NCBI). The GEO is a global public repository that the scientific community uses to submit functional genomic data sets from next-generation sequencing and high-throughput microarray technology. The database allows for the searchable, cross-linked, and indexed archiving of raw, processed, and metadata. All the data is freely downloadable in multiple formats. In addition, GEO offers several web-based techniques and tools to help users analyse, query and visualise data (Barrett *et al.,* 2012). The following selection criteria were used to further filter the datasets: (i) The gene expression profile datasets were from tuberculosis-affected patients with no current treatment,

(ii) only samples from healthy control and patients with active tuberculosis (ATB) were taken into consideration, (iii) each of the ATB and healthy control groups should have more than five samples, (iv) the datasets were from whole blood culture isolates of tuberculosis patients with no secondary diseases.

2.3 Dataset Features

At the time of the study (December 2023), 154 GEO microarray datasets for TB-related host response were found through database querying and filtering. However, based on further filtering according to the selection criteria for the study, a total of four (4) microarray datasets (GSE19435, GSE19439, GSE19444 (Berry *et al.,* 2010) and GSE28623 (Maertzdorf *et al.,* 2011)) which include active tuberculosis (ATB) patients and healthy control samples were selected for analysis (Table 2). The table also includes the two (2) RNA-Seq datasets (GSE107991 and GSE107994 (Singhania *et al.,* 2018)) used for validation. No experiment was conducted on any human or animal and all data retrieved was publicly available online.

Table 2: *GEO Datasets Features*

Dataset	Gene Expression	Platform	ATB	Healthy Control	Total Samples
GSE19435	Microarray	GPL6947 Illumina	21	12	33
GSE19439	Microarray	GPL6947 Illumina	13	12	25
GSE19444	Microarray	GPL6947 Illumina	12	12	24
GSE28623	Microarray	GPL4133 Agilent	46	37	83
GSE107991	RNA-Seq	GPL20301 Illumina	12	21	33
GSE107994	RNA-Seq	GPL20301 Illumina	53	50	103

2.4 Differentially Expressed Genes (DEGs) Analysis

The downloaded datasets that met the inclusion criteria were processed by using Linear Models for Microarray Analysis (Limma) packages in R programming language, the data were log-transformed, normalized and missing values were removed. Limma analyses microarray data by using linear models, it utilises a Bayesian framework to estimate differences in gene expression while accounting for the unpredictability present in microarray experiments (Symth, 2004; Ritchie *et al.,* 2015). The main concept is to apply empirical Bayes moderation to the standard errors of the estimated coefficients after fitting a linear model to the log-transformed gene expression data. To increase the differential expression estimates' accuracy. One important statistic that Limma uses to find genes with differential expression is the moderated t-statistic (Ritchie *et al.,* 2015). Therefore, the Limma package was used for this study because it a strong tool for reading, normalising, and analysing such data in this study.

The DEGs were identified between the ATB and healthy control calculating the Adjusted *P* value (Adj-*P* value) and $\mathrm{Log_2}$-Fold Change $|\mathrm{Log_2FC}|$ at the same time. The parameters for the DEGs were set at a level with Adj-*P* value < 0.05 and $|\mathrm{Log_2FC}|$ cut-off > 1.0. After the statistical analysis of each dataset, Boxplots and Volcano plots were drawn by R to have a picture representation of the dysregulated genes, upregulated genes and downregulated genes.

After the statistical analysis of each dataset, the common DEGs were calculated by using the Venn diagram (https://bioinformatics.psb.ugent.be/webtools/Venn/). Venn diagram is a free online web tool used to calculate the intersection(s) of the list of components. It produces a text output listing the components that are exclusive to a given list or that are in each intersection. It also produces a graphical output in a Venn/Euler diagram which can be downloaded as a figure in SVG and PNG format. The tool can calculate the intersections of a maximum of 30 lists.

2.5 Functional Annotations and Pathway Analysis

An essential step in interpreting gene lists obtained from extensive genetic, transcriptomic, and proteomic research is functional enrichment analysis (Wang *et al.,* 2013). Currently, there are many methods for functional annotations and pathway analysis. For instance, Database for Annotation, Visualization and Integrated Discovery (DAVID), is a frequently employed technique that utilises a modified Fisher's exact test to assess the significance of genes that are enriched in a particular pathway. One more well-known technique is gene set enrichment analysis (GSEA), which incorporates function enrichment analysis with the differential expression of genes. (Yang *et al.,* 2019). However, for this study, the functional annotations and pathway analysis were conducted on the genes using the WebGestalt (WEB-based Gene SeT AnaLysis Toolkit) database (https://www.webgestalt.org). This is because WebGestalt's statistical models are frequently predicated on tried-and-true techniques for gene set enrichment analysis and the hypergeometric test is one of the statistical technique frequently used by WebGestalt to determine whether a specific gene set is overrepresented in a list of differentially expressed genes relative to what would be predicted by chance. other techniques like the Fisher's exact test or the chi-squared test may also be used, depending on the type of data and analysis (Wang *et al.,* 2013). The gene functions were categorised by Gene Ontology (GO) annotation (Biological Processes (BPs), Cellular Components (CC), Molecular Functions (MF)) and Kyoto Encyclopaedia of Gene and Genomes (KEGG) Pathway enrichment analysis. The organism of interest was set at *Homo sapiens*; the method of interest was set at ORA and the parameters were left as default settings on WebGestalt.

2.6 Protein-Protein Interaction (PPI) Network Construction

The Search Tool for the Retrieval of Interacting Genes (STRING) database (http://string-db.org/) was used to complete PPI network construction with a 0.4 confidence level. The STRING database incorporates both predicted and known protein associations, including functional and physical interactions. Likewise, the pathogen proteins interact with the host DEGs (Ponnusamy and Arumugam, 2022). All the identified common DEGs were uploaded to the STRING database to assess and construct the potential human protein – *Mtb* protein interaction network.

2.7 Identification of Hub Genes

To identify and visualized the hub genes, Cytoscape software was used. Cytoscape is a bioinformatic platform that is free to use and can be enhanced with numerous plugins to increase visualisation options and network analysis power. It's simple to view a network's graphical representation using Cytoscape, and the interactome provides access to several levels of data, such as extensive genome-wide experiments and annotations of protein functions. Based on the Cytoscape API, the CytoHubba plugin is implemented in Java. Eleven node ranking techniques are implemented by the plugin to assess a node's significance in a biological network, including Degree. Out of the eleven techniques, the recently suggested method, Maximal Clique Centrality (MCC), performs better in terms of accuracy when predicting essential proteins from the yeast PPI network. (Chin *et al.,* 2014). The whole network was entered into the Cytoscape software (version 3.10.1) and using a plugin CytoHubba to evaluate the genes network, the hub genes were identified and ranked accordingly using MCC.

2.8 Drug Interactions

DGidb database (https://www.dgidb.org) was used for the Drug-target interactions and to identify the list of drugs in DrugBank targeting DEGs. DrugBank is an extensive database that contains details on drugs, including their targets, interactions, and mechanisms Ponnusamy and Arumugam, 2022). The common DEGs were entered into the DrugBank database, to query the

DEGs against the DrugBank to mine the drugs' interactions with the genes and to identify drugs with clinical and experimental evidence for direct interactions with the genes.

2.9 RNA-Seq Datasets

The statistical analysis was further validated using cross-validation by comparing the genes expressed in RNA-Seq datasets of ATB patients and healthy control samples. The RNA-Seq datasets were analysed using GEO2R and the results were compared with the results obtained from the statistical analysis, this is to obtain the strength of the results from the statistical analysis.

CHAPTER THREE

RESULTS

This chapter presents the results of the data analysis in this study. It presented the results as explained in the methodology and address the aim and objectives of the study stated in chapter one of this study.

3.1 Identification of DEGs

After the analysis of the datasets, a total of 259 DEGs, 109 upregulated and 150 downregulated genes were identified from GSE19435, a total of 211 DEGs, 133 upregulated and 78 downregulated genes were identified from GSE19439, a total of 211 DEGs, 133 upregulated and 78 downregulated genes were identified from GSE19444 and a total of 355 DEGs, 291 upregulated and 64 downregulated genes were identified from GSE28623 (Table 3). Figure 4 shows the Volcano plots of the DEGs of each dataset. The blue dots represent the downregulated genes, the red dots represent the upregulated genes, and the black dots represent not significant genes. The common genes among all 4 datasets were identified by the Venn diagram and the intersection of the DEGs was drawn by Venn analysis. 36 common DEGs, 26 upregulated genes and 10 downregulated were identified among the 4 datasets (Figure 5), indicating their status as high-confidence DEGs in the context of active tuberculosis versus healthy control samples.

Table 3: *DEGs, Upregulated and Downregulated Genes*

Dataset	Upregulated Genes	Downregulated Genes	DEGs
GSE19435	109	150	259
GSE19439	133	78	211
GSE19444	133	78	211
GSE28623	291	64	355
Total	**666**	**370**	**1,036**

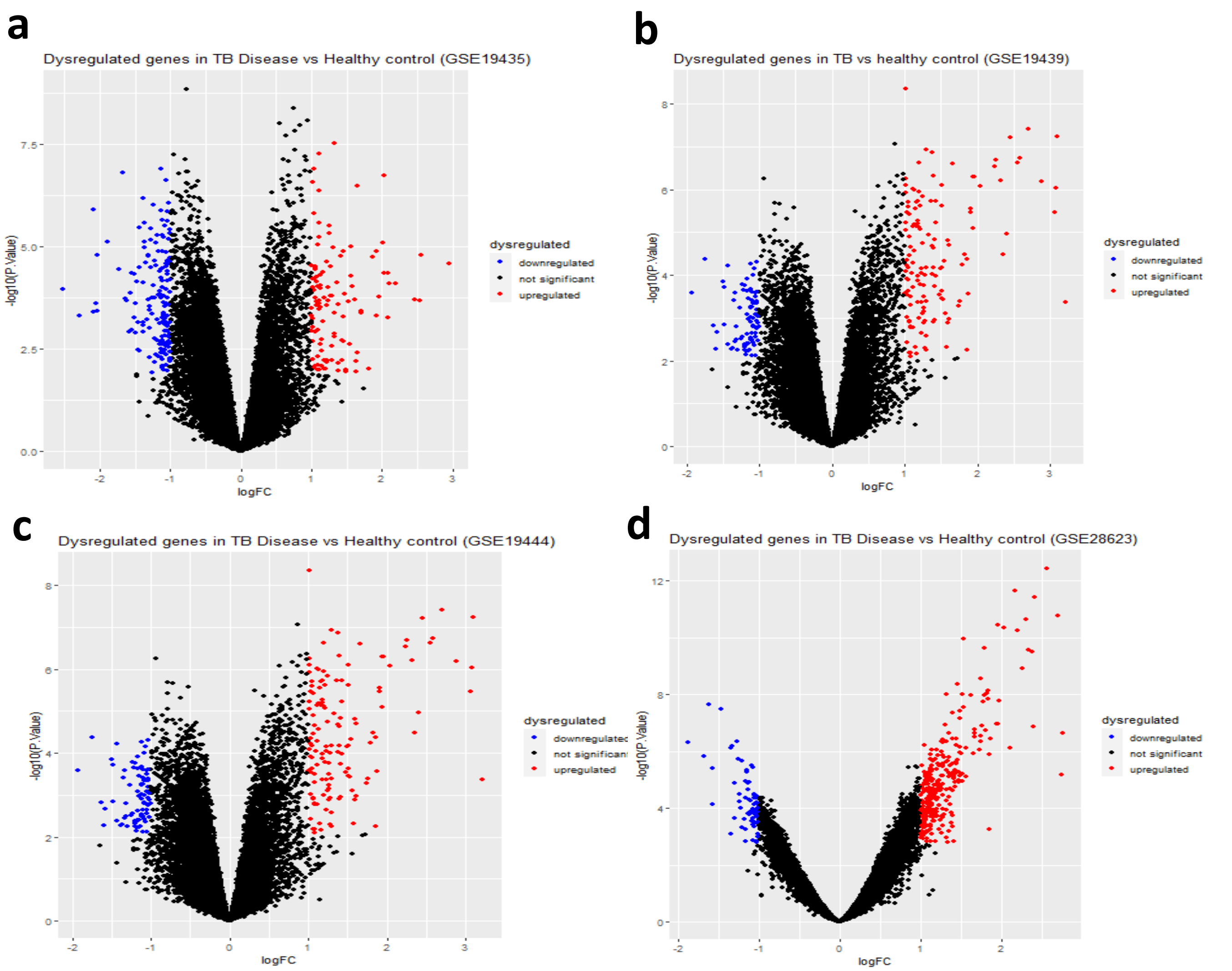


Figure 4: *Volcano plots. (a) represents the volcano plot for GSE19435, (b) is for GSE19439 (c) is for GSE19444 and (d) represents volcano plot of GSE28623.*

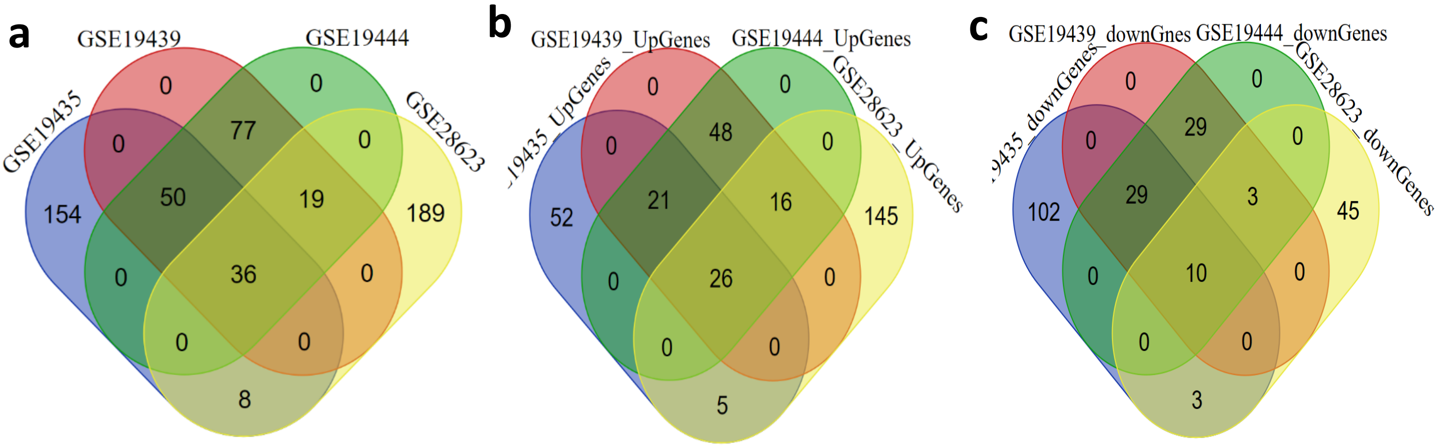


Figure 5: *Venn Diagram of DEGs common to the 4 datasets. (a) represents Venn diagram of total (36) (b) is for the upregulated genes (26) (c) is for the downregulated genes (10).*

3.2 Functional Enrichment Analysis

GO analysis discovered that upregulated DEGs were mainly relevant in BP, MF and CC. In BP, biological regulation was most enriched, followed by a response to stimulus, multicellular organismal process, localisation, metabolic process, development process, cell communication, multi-organism process, cellular component organisation, cell proliferation and others. MF showed the involvement in protein binding being the most enriched, followed by ion binding, nucleic acid binding, enzyme regulator activity, hydrolase activity, nucleotide binding, transferase activity and others. The CC analysis showed membrane involvement has been the most enriched, followed by the endomembrane system, vesicle, membrane-enclosed lumen, nucleus, extracellular space, cytosol, cell projection, Golgi apparatus, protein-containing complex and others (Figure 6a). GO analysis discovered that downregulated DEGs were particularly significant in BP and MF with no relevant difference in CC. BP showed biological regulation as the most enriched, followed by cell communication, response to stimulus, localisation, multicellular organismal process on the same level and metabolic process and developmental process on the same level as well. CC showed membrane as the most enriched with no significant difference in others. MF showed protein binding as the most enriched, followed by ion binding, transferase activity, nucleotide binding and others on the same levels (Figure 6b). The GO analysis identified that the total DEGs were significant in BP, MF and CC. In BP, biological regulation was most enriched, followed by a response to stimulus, multicellular organismal process, localisation and involvement in metabolic process, cell communication, and development process. MF showed the involvement in protein binding being the most enriched, followed by ion binding, nucleotide binding, transferase activity, nucleic acid binding, enzyme regulator activity, hydrolase activity and others. The CC analysis showed membrane involvement has been the most enriched, followed by a vesicle, endomembrane system, nucleus, extracellular space, membrane-enclosed lumen, cytosol, protein-containing complex, cell projection, Golgi apparatus and others (Figure 6c).

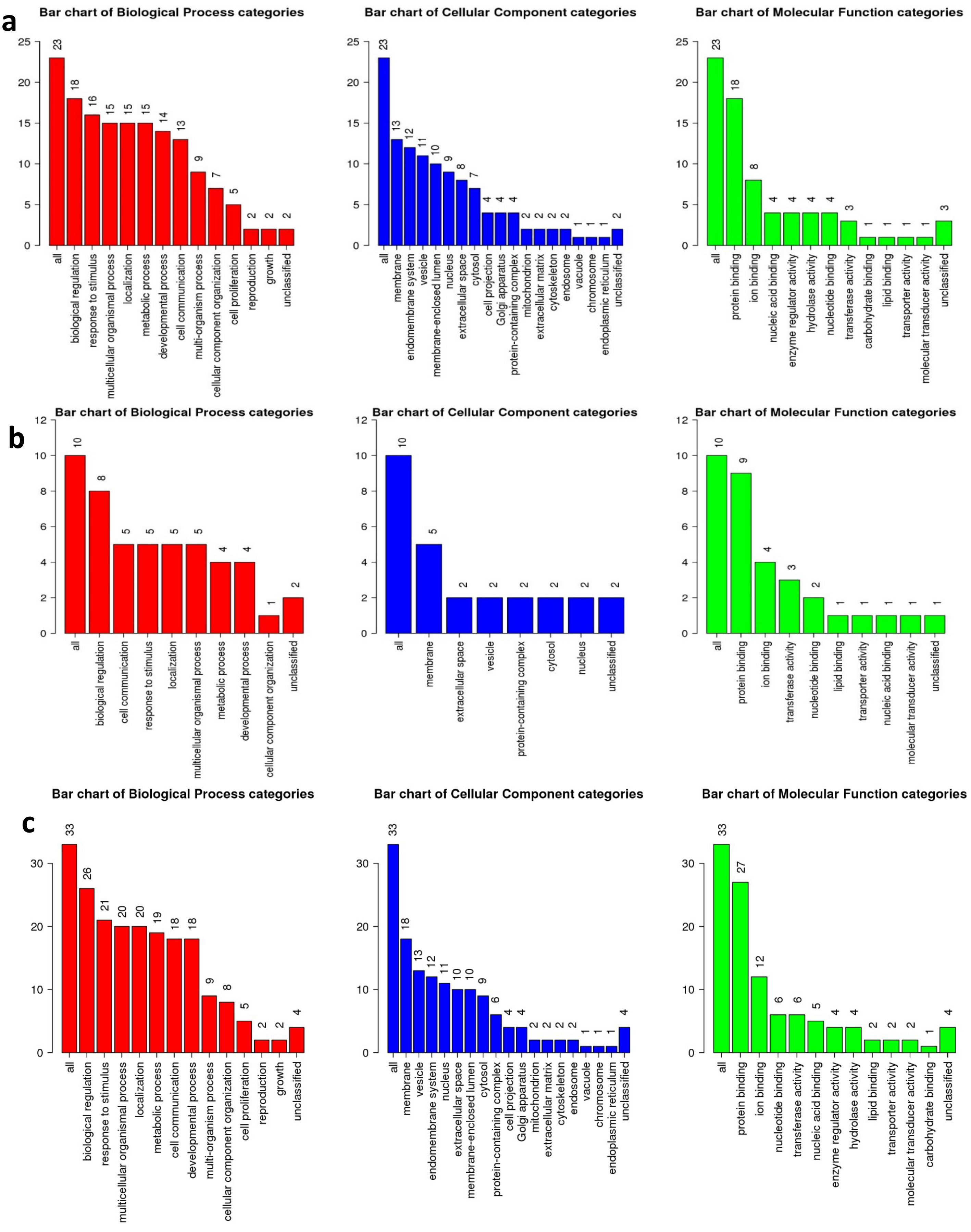


Figure 6: *Gene Oncology Analysis. (a) BP, CC and MF of the upregulated gene, (b) BP, CC and MF of the downregulated gene (c) BP, CC and MF of total DEGs.*

The KEGG pathway enrichment analysis was done for upregulated DEGs, downregulated DEGs and the total DEGs separately with WebGestalt. The analysis identified upregulated DEGs at a false discovery rate (FDR) of $\leq$ 0.05 and they are being enriched in pathways of regulation of innate immune response, positive regulation of immune response, regulation of immune response, response to biotic stimulus, innate immune response, immune effector process, defense response, secretion, secretion by cell and immune response (Figure 7a). The analysis identified downregulated DEGs and they are been enriched in pathways of paramethadione, negative regulation of glycogen biosynthetic process, negative regulation of glycogen metabolic process, C-X-C chemokine receptor activity, regulation of glucagon secretion, B cell receptor signaling pathways with PPI_BIOGRID_M162, PPI_BIOGRID_M74, antigen receptor-mediated signaling pathway and side of membrane as being least enriched (Figure 7b). The analysis identified total DEGs as being enriched in pathways of B cell receptor signaling pathway, antigen receptor-mediated signaling pathway, immune response-regulating cell surface receptor signaling pathway, immune response-activating cell surface receptor signaling pathway, regulation of innate immune response, activation of immune response, immune response-regulating signaling pathway, positive regulation of immune response, regulation of defense response, regulation of response to external stimulus, positive regulation of immune system process, response to biotic stimulus, innate immune response, secretion by cell, immune effector process, secretion and defense response, regulation of immune system process and immune response been the least enriched pathways (Figure 7c). The KEGG analysis also produced volcano plots for the upregulated DEGs, downregulated DEGs and the total DEGs shown in (Figure 8a-c). The top 10 significantly enriched GO for the upregulated DEGs and downregulated DEGs identified in the analysis are presented in (Table 4).

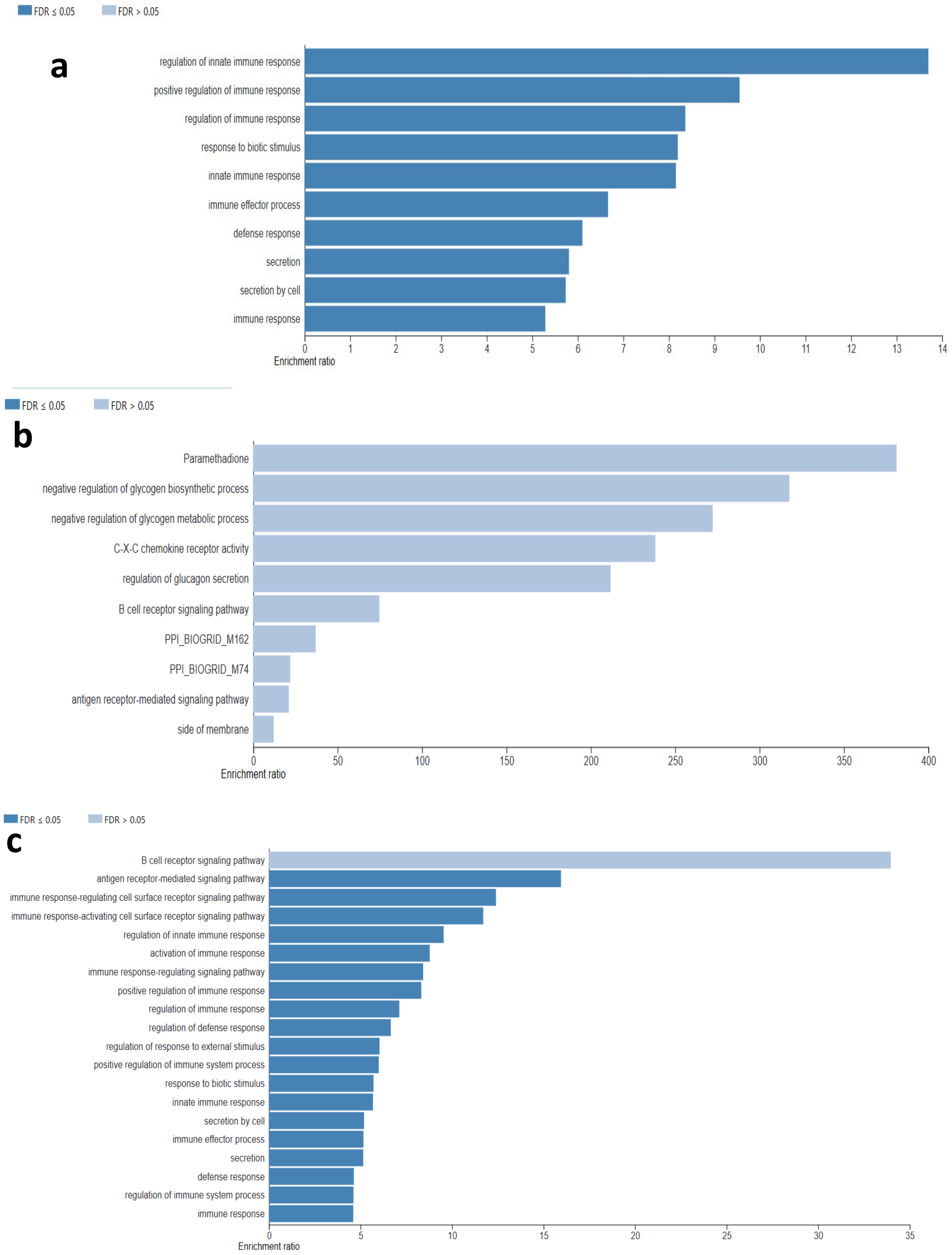


Figure 7: *KEGG Pathways. (a) Upregulated DEGs pathways (b) Downregulated DEGs Pathways (c) total DEGs pathways.*

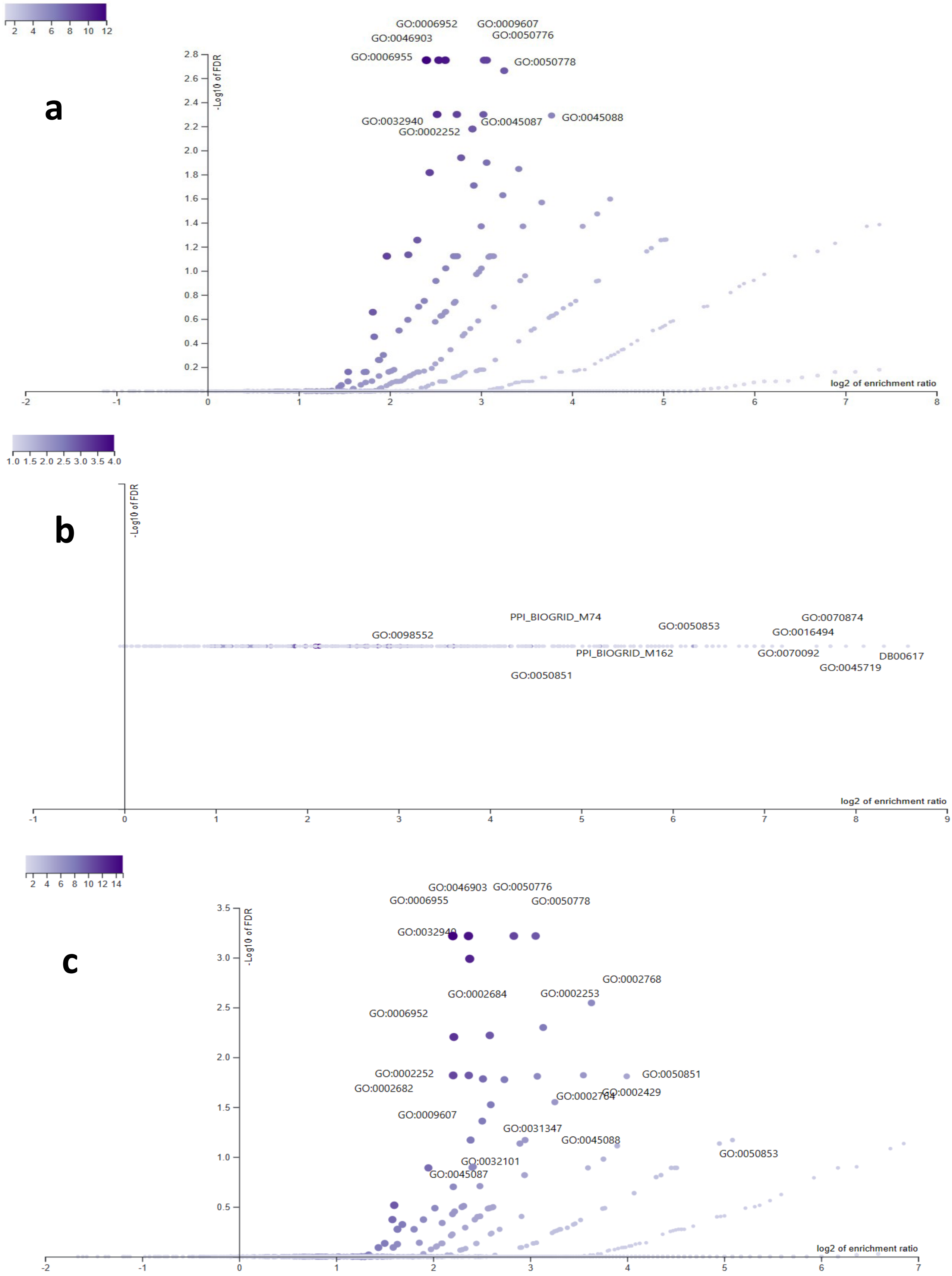


Figure 8: *Volcano plots of Pathways. (a) Upregulated DEGs pathways (b) Downregulated DEGs Pathways (c) total DEGs pathways.*

Table 4: *Top 10 significantly Enriched GO*

Sort	Gene Set	Description	Ratio	P value
Up-regulate	GO:0006952	Defense response	6.10	3.7022e-7
	GO:0006955	Immune response	5.29	3.9511e-7
	GO:0050776	Regulation of immune resp.	8.36	4.6746e-7
	GO:0009607	Response to biotic stimulus	8.19	5.5312e-7
	GO:0046903	Secretion	5.81	6.1072e-7
Down-regulate	GO:0050853	B cell receptor signaling pathway	74.7	0.000312
	GO:0045719	Negative regulation of glycogen biosynthetic process	318	0.003145
	GO:0070874	Negative regulation of glycogen metabolic process	272	0.003668
	GO:0050851	Antigen receptor-mediated signaling pathway	21.1	0.003839
	GO:0070092	Regulation of glucagon secretion	211	0.004714

3.3 Protein-Protein Interaction Network Construction

The PPI network of 34 nodes and 47 edges was identified after uploading the DEGs into the STRING database (Figure 9). about 24 genes were identified to show direct interactions with their neighbours. The top 10 hub genes were identified by using the CytoHubba plugin in Cytoscape software (Figure 10) and were ranked by MCC measures (Table 5). The results showed that GBP5 (guanylate binding protein 5) was the most significant gene with MCC score = 99, followed by GBP1 (MMC score = 86), BATF2 (MCC score = 73), EPSTI1 (MCC score = 48), PARP9 (MCC score = 48), FCGR1B (MCC score = 30), SERPING1 (MCC score = 24), LAP3 (MCC score = 24), ANKD22 (MCC score = 14) and ETV7 (MCC score = 12). All the identified hub genes are upregulated.

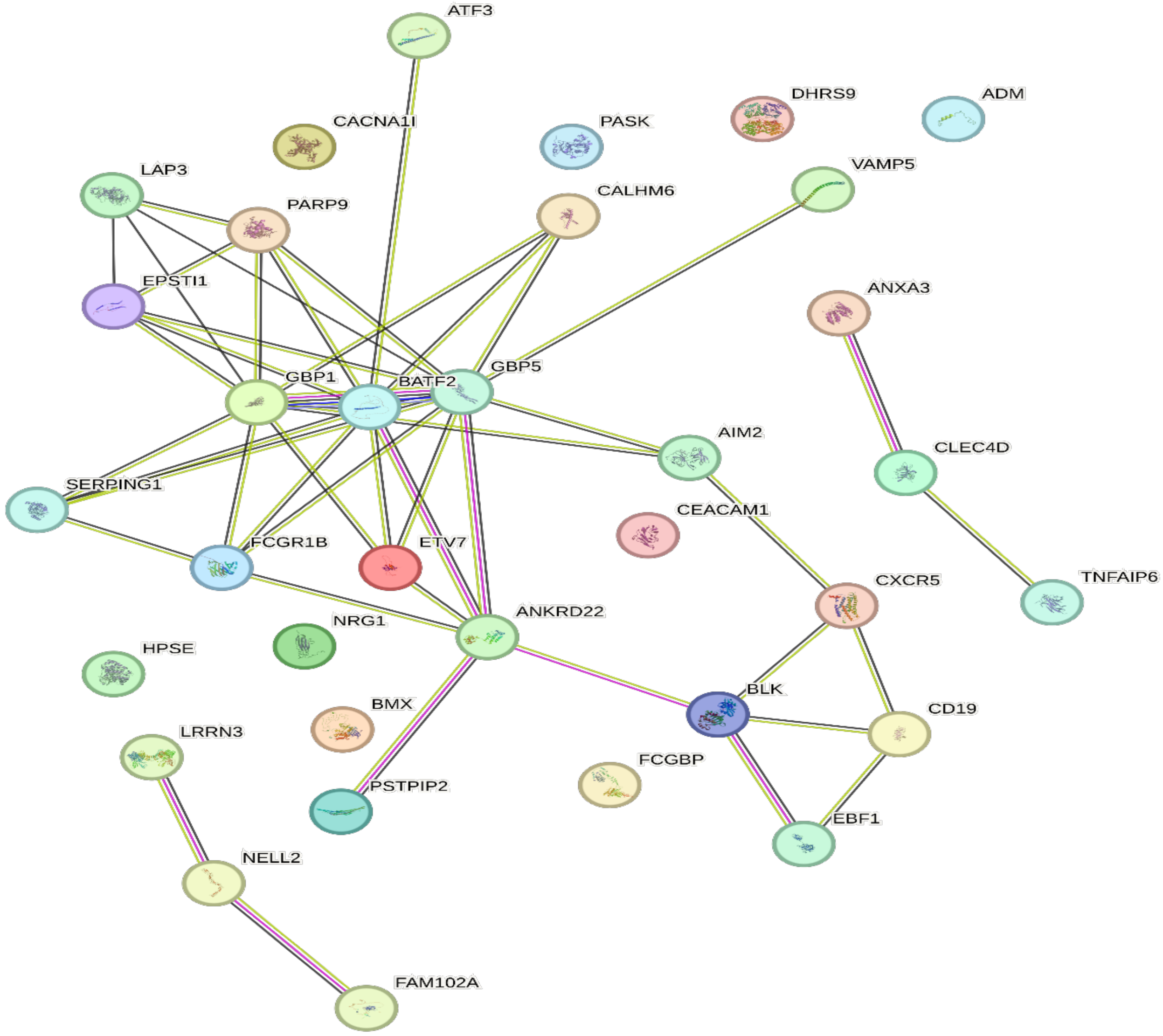


Figure 9: *Protein-Protein Interaction Network*

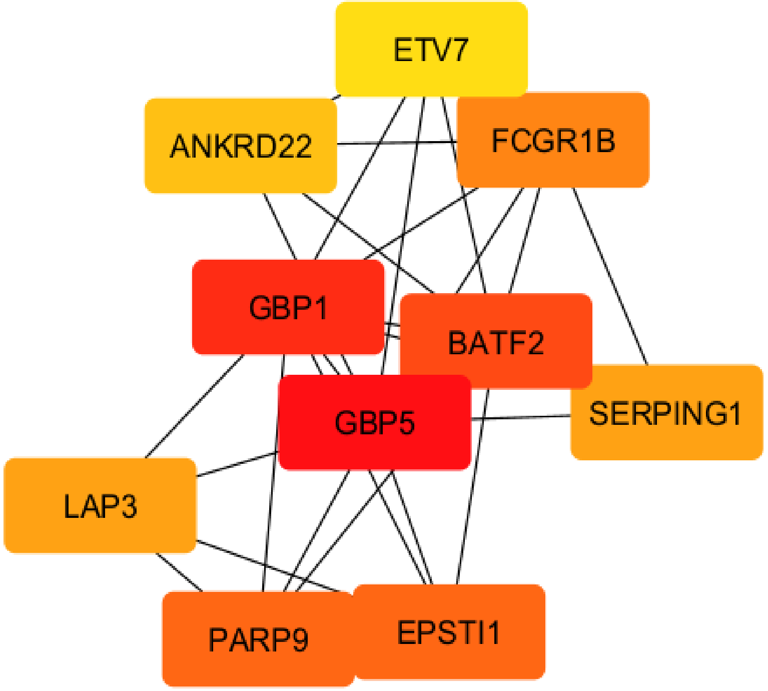


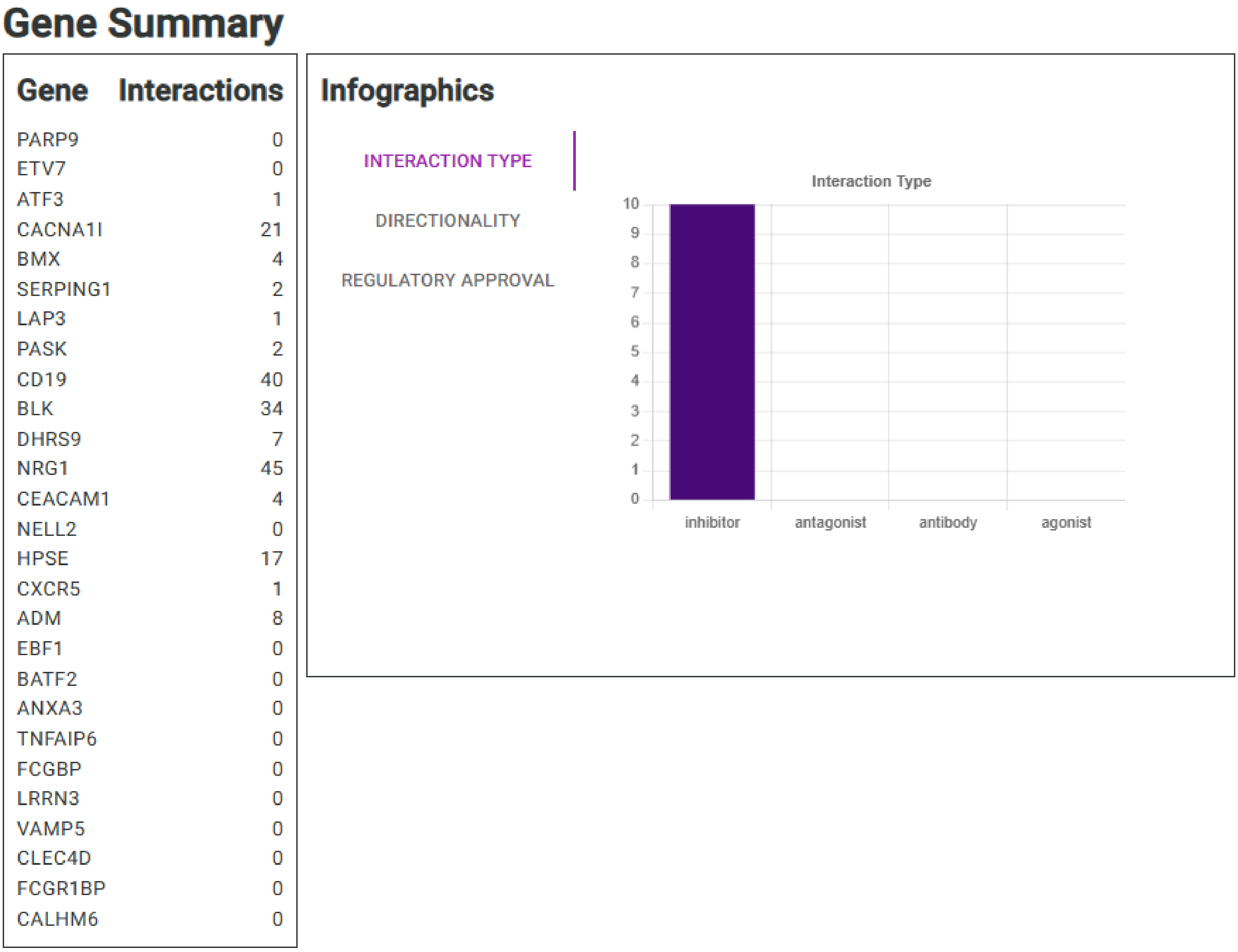
Figure 10: *PPI Network of Top 10 Hub Genes*

Table 5: *MCC Ranking of Top 10 Hub Genes*

Rank	Gene Name	MCC Score
1	GBP5	99
2	GBP1	86
3	BATF2	73
4	EPSTI1	48
4	PARP9	48
6	FCGR1B	30
7	SERPING1	24
7	LAP3	24
9	ANKD22	14
10	ETV7	12

3.4 Drug Interaction

DGidb database was used to query the DEGs against the DrugBank to mine drugs that target genes that could be utilized for treatment against tuberculosis. A total of 187 drugs targeting 14 DEGs were obtained from the process of the screening. Only the approved drugs that are used to treat a particular disease were selected (Appendix 1). Among them, each drug showed interactions with some of the target genes with NRG1 showing interaction with 45 drugs, CD19 having 40 drug interactions, BLK showing interaction with 34 drugs, CACNA1I with 21 drug interactions, HPSE with interactions with 17 drugs, ADM with 8 drugs interactions, DHRS9 have interactions with 7 drugs, CEACAM1 and BMX with 4 drugs interactions each, SERPING1 and PASK showed interactions with 2 drugs each, ATF3, LAP3 and CXCR5 have interaction with 1 drugs each, other genes does not show any drug interaction. all the identified drugs are inhibitors in their interaction type (Figure 11).



Gene Summary

Gene	Interactions
PARP9	0
ETV7	0
ATF3	1
CACNA1I	21
BMX	4
SERPING1	2
LAP3	1
PASK	2
CD19	40
BLK	34
DHRS9	7
NRG1	45
CEACAM1	4
NELL2	0
HPSE	17
CXCR5	1
ADM	8
EBF1	0
BATF2	0
ANXA3	0
TNFAIP6	0
FCGBP	0
LRRN3	0
VAMP5	0
CLEC4D	0
FCGR1BP	0
CALHM6	0

Infographics

Figure 11: *Summary of Drugs Interactions with Target Genes*

3.5 RNA-Seq Datasets Analysis

To validate the results and to assess the key genes more thoroughly, GEO2R was used to analysis 2 RNA-Seq datasets (GSE107991 and GSE107994), the results produced 316 DEGs including 280 upregulated and 36 downregulated, the volcano plot and Venn diagram are shown in figure 12 and figure 13 respectively. The GO and pathway enrichment analysis revealed enrichment in biological regulation, membrane and protein binding including innate immune response is the most enriched pathway (Figure 14 and 15) respectively. Through PPI construction and MCC ranking additional 10 hub genes were generated which include, IFIT3, STAT1, ISG15, OAS1, PARP9, GBP1, IFI35, IRF7, RTP4 and OASL (Figure 16 and Table 6) of which PARP9 and GBP1 were already in the 4 microarray datasets analysed.

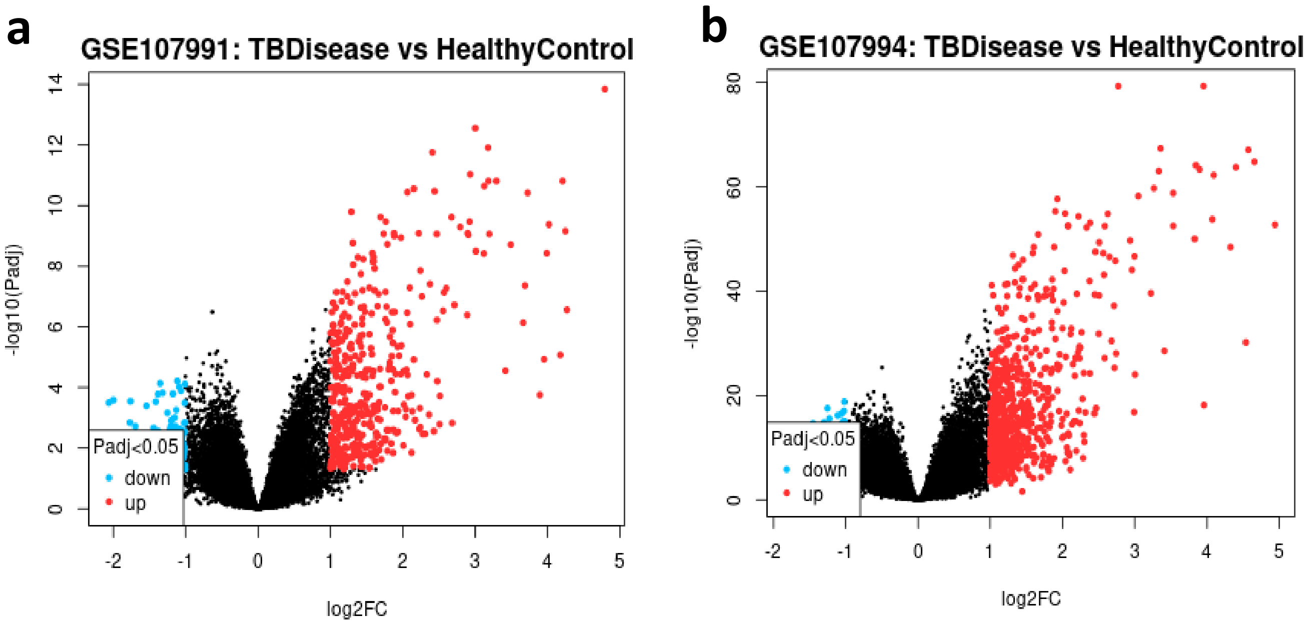


Figure 12: *Volcano plots of RNA-Seq Data. (a) represents the volcano plot for GSE107991, (b) is for GSE107994 (Red dots signify upregulated genes, blue dots for downregulated genes and black dots for not significant.*

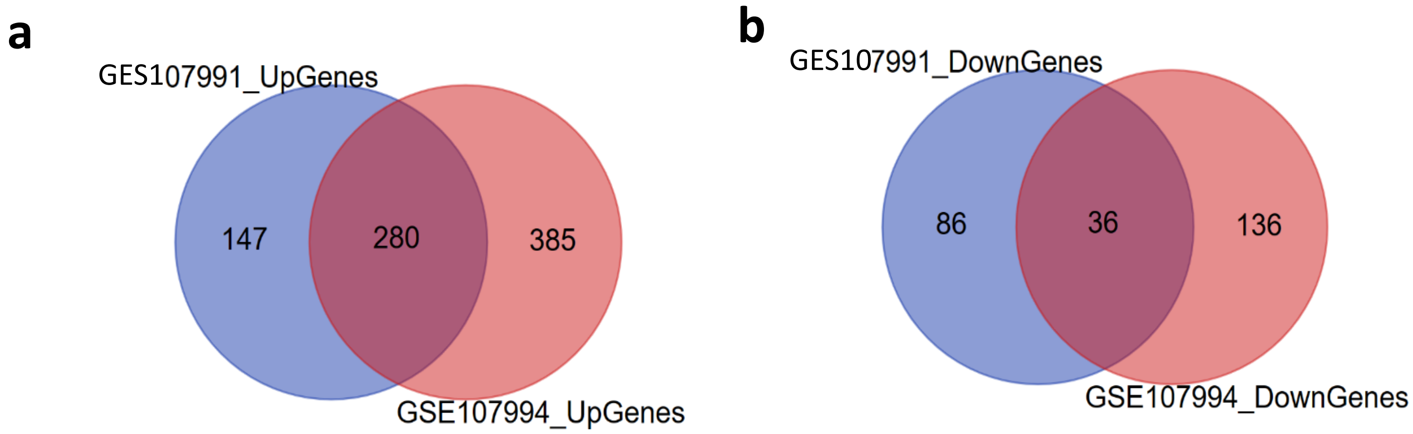


Figure 13: *Venn Diagram of DEGs common to the 2 RNA-Seq datasets. (a) represents Venn diagram of upregulated genes (b) is for the downregulated genes.*

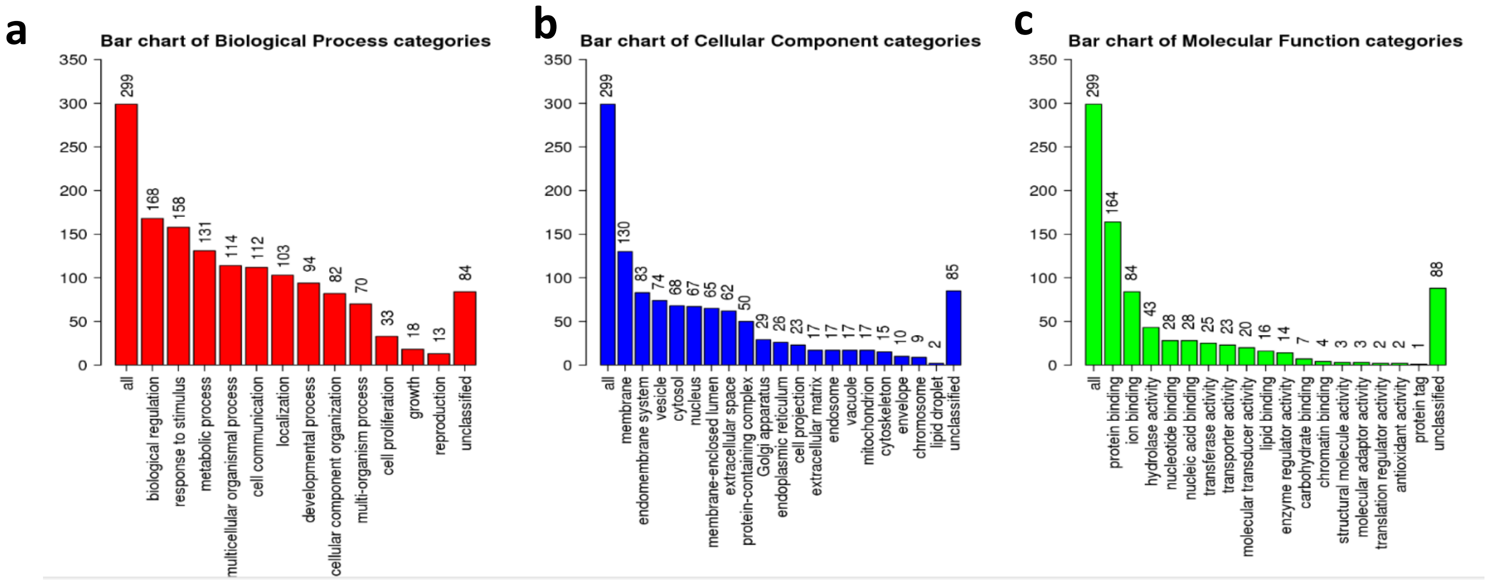


Figure 14: *Gene Oncology Analysis of RNA-Seq Datasets. (a) Biological Process (BP) (b) Cellular Component (CC) (c) Molecular Functions (MF).*

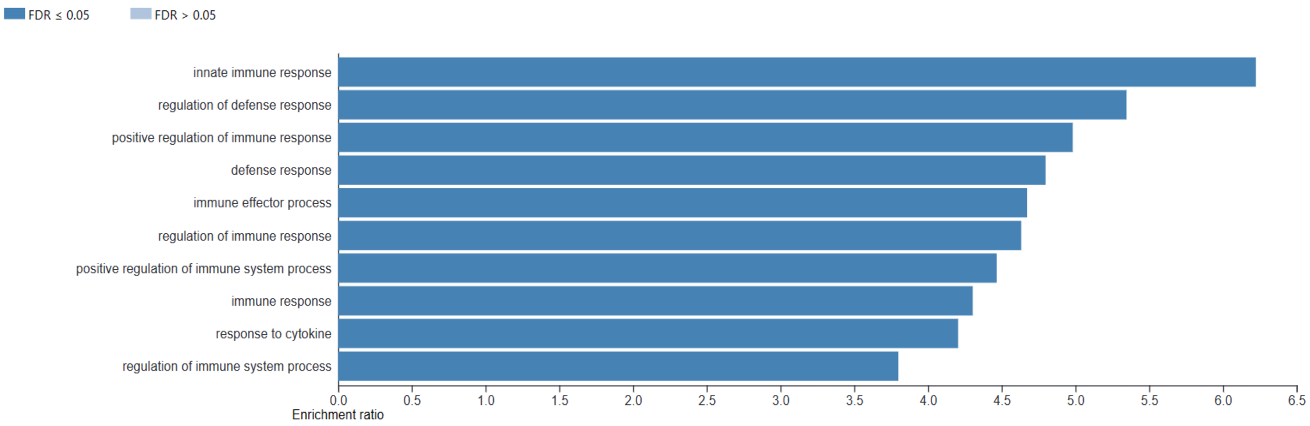


Figure 15: *KEGG Pathways of RNA-Seq Datasets (at FDR $\leq$ 0.05 which signifies true enrichment)*

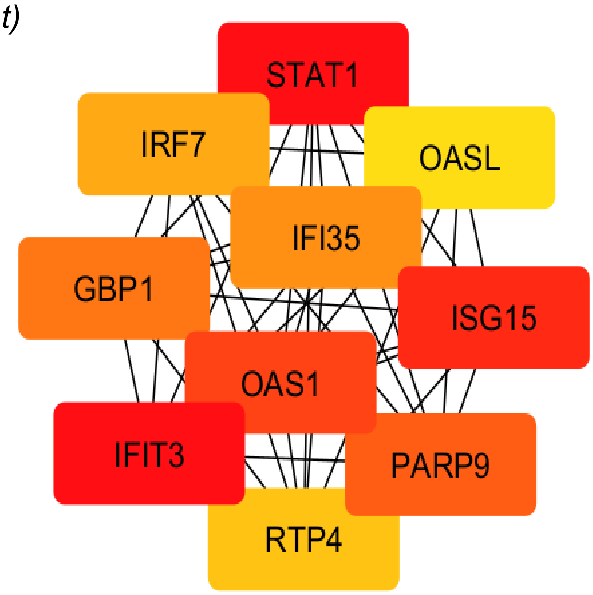


Figure 16: *PPI Network of Top 10 Hub Genes for the RNA-Seq Datasets*

Table 6: *MCC Ranking of Top 10 Hub Genes for RNA-Seq Datasets*

Rank	Gene Name	MCC Score
1	IFIT3	7.015
1	STAT1	7.015
3	ISG15	7.015
4	OAS1	7.015
5	PARP9	7.015
6	GBP1	7.015
7	IFI35	7.015
8	IRF7	7.015
9	RTP4	7.015
10	OASL	7.015

CHAPTER FOUR

DISCUSSION

This study included 4 microarray gene expression profile datasets and 2 RNA-Seq datasets as a validation cohort. All the included datasets are publicly available from the GEO database. All the datasets included active tuberculosis patients and healthy control groups in their samples, and they had more than 5 participants in each sample. The bioinformatics analysis was done to identify the molecular signature in *Mtb* which is the causative agent of tuberculosis.

Tuberculosis has been one of the major concerns for public health worldwide because of its threat to health for many years, its drug resistance ability and it is one of the leading causes of mortality worldwide. Therefore, identification of the core genes in *Mtb* could be a good development in the early diagnosis and treatment of tuberculosis, help to overcome the drug resistance of tuberculosis and improve advancement in the development of vaccines and new drugs targeting the core genes responsible for tuberculosis.

4.1 Summary of Key Findings

The findings in this study agreed with many studies such as a study conducted by Shi et al., (2022) which reported GBP1 to be upregulated in tuberculosis patients (Shi *et al.,* (2022). Yao *et al.,* (2022); Ponnusamy and Arumugam, (2022) in their studies identified GBP5 protein levels as being significantly upregulated in tuberculosis patients than non-tuberculosis patients (Yao *et al.,* 2022; Ponnusamy and Arumugam, 2022)). Chen *et al.,* (2019) also confirmed GPB5 and GBP1 as part of the top 10 hub genes in their studies, however, it was stated to be downregulated in ATB (Chen *et al.,* 2019). Guanylate binding proteins (GBPs) which include (GBP 1 – 7) belong to the GTPase subfamily, they are primarily induced by interferon gamma (IFN-γ). They are involved in numerous critical cellular processes, such as the activation of inflammasomes and innate immunity against a broad range of microbial pathogens (Li *et al.,* 2020). The high expression of GBP5 and GBP1 suggests that they may play a potential role as immune biomarkers for early detection and targeted TB treatment.

This study also identified BATF2 (Basic Leucine Zipper Transcription Factor 2) as one of the top 10 hub genes that are upregulated in ATB patients. In a blood transcriptomic study of pulmonary and extrapulmonary tuberculosis conducted by Roe *et al.,* (2016), it was also reported that BAFT2 levels were elevated in ATB compared to uninfected healthy individuals (Roe *et al.,* 2016). BATF2 was also reported by Ponnusamy and Arumugam, (2022) to be upregulated and one of the top 20 DEGs associated with *Mtb* (Ponnusamy and Arumugam, 2022). BATF2 is a transcription factor that is part of the activator protein 1 (AP-1) family. It is expressed in mononuclear phagocytic cells in response to IFN-stimulated innate immunity using lipopolysaccharide or *Mtb*. BATF2 mediates downstream proinflammatory responses through its interaction with IFN regulatory factor 1 (IRF1); some of these responses are also identified as elements of the host response to *Mtb* (Murphy *et al.,* 2013; Roy *et al.,* 2015). This research project proposes that BATF2 can also provide a sensitive biomarker to differentiate healthy individuals from tuberculosis-infected patients.

The findings of the drug interaction with the genes identified that Indomethacin and Celecoxib drugs, a nonsteroidal anti-inflammatory drug (NSAID) target ADM and CACNA1I genes respectively and Ibrutinib drug, an antineoplastic agent targets BMX gene. In a study published by Tonby *et al.,* (2016), it was reported that Indomethacin downregulates the fraction of FOXP3+T regulatory cells specific to *Mtb* significantly and *Mtb*-specific cytokine responses in ATB patients (Tony *et al.,* (2016). Naftalin *et al.,* (2018) in their study suggested that Celecoxib (a COX-2 inhibitor) may help treat tuberculosis through a variety of mechanisms, such as enhancing intracellular tuberculosis drug levels through efflux pump inhibition and having various effects on inflammation and the immune system (Naftalin *et al.,* 2018). Hu *et al.,* (2020) in their study conducted on mice reported that Ibrutinib inhibited the growth of intracellular *Mtb* in human macrophages. Also, Ibrutinib treatment dramatically reduced p62 and increased LC3b proteins in *Mtb*-infected macrophages, according to mechanisms studies. They finally verified that the administration of ibrutinib considerably decreased the amount of *Mtb* in the spleen and mediastinal node of mice infected with *Mtb* (Hu *et al.,* 2020). This research project suggests that the potential capacity of these drugs to treat tuberculosis can be further investigated to develop more and advanced treatments for tuberculosis.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

This study has employed genome-wide transcriptomic analysis to highlight the molecular signature, pathways, PPI and drug targets which are crucial for understanding tuberculosis disease and advancement in therapy. The findings have revealed that biological regulation, response to stimulus, membrane, endomembrane system, vesicle and protein binding, including regulation of innate immune response, positive regulation of immune response, response to biotic stimulus and innate immune response are common molecular signature in *Mycobacterium tuberculosis*. Further analysis also revealed involvement of paramethadione, negative regulation of glycogen biosynthetic process, negative regulation of glycogen metabolic process including C-X-C chemokine receptor activity, regulation of glucagon secretion, B cell receptor signaling pathway, antigen receptor-mediated signaling pathway, immune response-regulating cell surface receptor signaling pathway. It was discovered that GBP5, GBP1 and BATF2 were upregulated in active tuberculosis patients and they are at the top of the 10 hub genes an indication that high expression of GBP5 and GBP1 may play a crucial role in the pathogenesis of *Mycobacterium tuberculosis* which may be a biomarker for early diagnosis while SERPING1, LAP3, ADM, CACNA1I and BMX may be helpful in drugs development for the treatment of tuberculosis disease.

5.2 Recommendations

Based on the findings in this study, the following recommendations are therefore suggested both for clinical and academic applications for the early diagnosis and treatment of tuberculosis: Conducting further functional investigations, like overexpression or knockdown tests, to confirm the biological relevance of discovered genes. Analysing their effects on the spread of the disease, host response, and tuberculosis infection can add to the understanding of the disease mechanism. In addition, combining data from various omics platforms (genomics, transcriptomics, proteomics, and metabolomics) to obtain a thorough grasp of the molecular processes driving tuberculosis. This all-encompassing method can provide fresh perspectives on the dynamics of disease and possible treatment targets. Finally, investigating how the identified genes affect immune cell function, cytokine production, and overall immune regulation during infection can help in the early diagnosis and treatment of tuberculosis.

5.3 Limitations

One of the limitations faced during this study is the availability of well-annotated, high-quality datasets for tuberculosis which may impact the robustness of the analysis. Another limitation faced is the fold change threshold, genes with subtle but biologically significant changes may be overlooked in differential expression analysis if arbitrary fold change thresholds are set. In addition, functional enrichment analysis depends on pre-existing annotations, which might not include all biological pathways or functions, thus leaving important information out. Furthermore, gene expression data-based drug target prediction may miss off-target effects and other subtleties in drug action. Lastly, independent datasets were analysed using GEO2R as a validation cohort. While this is good for validating the findings, further experimental validation methods such as western blotting or qRT-PCR are necessary to confirm the biological significance and functional relevance of the genes identified.

5.4 Future Research

To confirm and build upon present findings in this study, future research on tuberculosis may pursue several directions. Potential study avenues and experiments include the following: To use animal models (guinea pigs, mice, etc.) to evaluate the in vivo significance of the discovered genes and therapeutic targets. Examine their roles in host-pathogen interactions, the pathophysiology of tuberculosis, and the effectiveness of possible treatment candidates. In addition, to work with medical professionals to obtain clinical samples to validate the expression patterns of the genes identified in tuberculosis patients and establish a relationship between the degree of gene expression and the course of the disease, its severity, and the response to therapy. Lastly, to further investigate if the identified drugs in this study can be used to treat tuberculosis, to determine their effectiveness against *M. tuberculosis* and to evaluate their safety profile through both in vitro and in vivo experiments.

ACKNOWLEDGEMENTS

I ascribe all the glory to the Almighty God who gave me the uncommon grace, unbeatable privilege, and extraordinary favour to experience a peaceful journey throughout my course of study at the University.

I am particularly grateful to my project supervisor, Dr. Shweta Kuba for her encouragement, constructive suggestions, inspiring discussion, and adequate supervision of this project work. Also, I would like to express my gratitude to Dr. Vasileios Panagiotis Lenis and Dr. Mengyuan Wang. Dr. Vasileios Panagiotis Lenis has been an incredible all-around mentor for my research project (especially getting me up to speed with the Limma package in R programming). I am indebted to him for all the invaluable feedback he has given me on this research project and for painstakingly going through my draft to the final edit.

I express my profound gratitude to my lovely partner, lover of my life, soul mate, and my best friend, Racheal Akinade, and my adorable daughters Testimony and Treasure for their unflinching support and their endurance when I am too busy with the project work to attend to them; you are more than anything to me in this world and I will always appreciate you. I also shower encomiums to my precious father Pa. J. O. Akinade and my late mother Mrs. D. O. Akinade who died shortly after I started this programme, your efforts to see me succeed in life will forever be in my memory. Also, my thanks would be incomplete without expressing my appreciation to my brothers and sisters, ASP. Samson Akinade, Pastor Timothy. Akinade, Mrs. Olubisi. E. Adelusi, Pastor. Moses Akinade, Bro. Olukola J. Akinade, and Odunayo Fashina, for their wonderful support; I say you are simply lovely and invaluable.

Finally, my expression goes to my colleagues in the Department of Bioinformatics, Teesside University whom we share knowledge for the success of this research project.

REFERENCES

Ahmad, S. (2011). Pathogenesis, immunology, and diagnosis of latent Mycobacterium tuberculosis infection. *Clinical and Developmental Immunology, 2011*

Alam, A., Imam, N., Ahmed, M. M., Tazyeen, S., Tamkeen, N., Farooqui, A., Malik, M. Z., & Ishrat, R. (2019). Identification and classification of differentially expressed genes and network meta-analysis reveals potential molecular signatures associated with tuberculosis. *Frontiers in Genetics, 10*, 932.

Apic, G., Ignjatovic, T., Boyer, S., & Russell, R. B. (2005). Illuminating drug discovery with biological pathways. *FEBS Letters, 579*(8), 1872-1877.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., & Holko, M. (2012). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research, 41*(D1), D991-D995.

Berry, M. P., Graham, C. M., McNab, F. W., Xu, Z., Bloch, S. A., Oni, T., Wilkinson, K. A., Banchereau, R., Skinner, J., & Wilkinson, R. J. (2010). An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature, 466*(7309), 973-977.

Beste, D. J., Hooper, T., Stewart, G., Bonde, B., Avignone-Rossa, C., Bushell, M. E., Wheeler, P., Klamt, S., Kierzek, A. M., & McFadden, J. (2007). GSMN-TB: a web-based genome-scale network model of Mycobacterium tuberculosismetabolism. *Genome Biology, 8*(5), 1-18.

Bonora, S., & Di Perri, G. (2008). Interactions between antiretroviral agents and those used to treat tuberculosis. *Current Opinion in HIV and AIDS, 3*(3), 306-312.

Boshoff, H. I., Myers, T. G., Copp, B. R., McNeil, M. R., Wilson, M. A., & Barry, C. E. (2004). The transcriptional responses of Mycobacterium tuberculosis to inhibitors of metabolism: novel insights into drug mechanisms of action. *Journal of Biological Chemistry, 279*(38), 40174-40184.

Chakrabarty, S., Kumar, A., Raviprasad, K., Mallya, S., Satyamoorthy, K., & Chawla, K. (2019). Host and MTB genome encoded miRNA markers for diagnosis of tuberculosis. *Tuberculosis, 116*, 37-43.

Chang, S., Chen, M., Lee, M., Liang, Y., Lu, T., Wang, J., & Yan, B. (2018). SP110 polymorphisms are genetic markers for vulnerability to latent and active tuberculosis infection in Taiwan. *Disease Markers, 2018*

Chen, J., Liu, C., Liang, T., Xu, G., Zhang, Z., Lu, Z., Jiang, J., Chen, T., Li, H., & Huang, S. (2021). Comprehensive analyses of potential key genes in active tuberculosis: A systematic review. *Medicine, 100*(30)

Chin, C., Chen, S., Wu, H., Ho, C., Ko, M., & Lin, C. (2014). cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Systems Biology, 8*(4), 1-7.

Claus, B. L., & Underwood, D. J. (2002). Discovery informatics: its evolving role in drug discovery. *Drug Discovery Today, 7*(18), 957-966.

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

Cukic, V., & Ustamujic, A. (2018). Extrapulmonary tuberculosis in Federation of Bosnia and Herzegovina. *Materia Socio-Medica, 30*(2), 153.

de Araujo, L. S., Ribeiro-Alves, M., Leal-Calvo, T., Leung, J., Durán, V., Samir, M., Talbot, S., Tallam, A., Mello, F. C. d. Q., & Geffers, R. (2019). Reprogramming of small noncoding RNA populations in peripheral blood reveals host biomarkers for latent and active Mycobacterium tuberculosis infection. *MBio, 10*(6), 10.1128/mbio. 01037-19.

Dhavan, P., Dias, H. M., Creswell, J., & Weil, D. (2017). An overview of tuberculosis and migration. *The International Journal of Tuberculosis and Lung Disease, 21*(6), 610-623.

Doran, K. S., Fulde, M., Gratz, N., Kim, B. J., Nau, R., Prasadarao, N., Schubert-Unkmeir, A., Tuomanen, E. I., & Valentin-Weigand, P. (2016). Host–pathogen interactions in bacterial meningitis. *Acta Neuropathologica, 131*, 185-209.

Esterhuyse, M. M., Weiner 3rd, J., Caron, E., Loxton, A. G., Iannaccone, M., Wagman, C., Saikali, P., Stanley, K., Wolski, W. E., & Mollenkopf, H. (2015). Epigenetics and proteomics join transcriptomics in the quest for tuberculosis biomarkers. *MBio, 6*(5), 10.1128/mbio. 01187-15.

Forrellad, M. A., Klepp, L. I., Gioffré, A., Sabio y Garcia, J., Morbidoni, H. R., Santangelo, M. D. L. P., Cataldi, A. A., & Bigi, F. (2013). Virulence factors of the Mycobacterium tuberculosis complex. *Virulence, 4*(1), 3-66.

Fu, Y., Wang, J., Qiao, J., & Yi, Z. (2019). Signature of circular RNAs in peripheral blood mononuclear cells from patients with active tuberculosis. *Journal of Cellular and Molecular Medicine, 23*(3), 1917-1925.

Ganguly, N., Giang, P. H., Basu, S. K., Mir, F. A., Siddiqui, I., & Sharma, P. (2007). Mycobacterium tuberculosis 6-kDa early secreted antigenic target (ESAT-6) protein downregulates lipopolysaccharide induced c-myc expression by modulating the extracellular signal regulated kinases 1/2. *BMC Immunology, 8*(1), 1-12.

Goundar, S. (2012). Research methodology and research method. *Victoria University of Wellington,* Greco, S., Girardi, E., Navarra, A., & Saltini, C. (2006). The current evidence on diagnostic accuracy of commercial based nucleic acid amplification tests for the diagnosis of pulmonary tuberculosis. *Thorax,*

Guler, R., Roy, S., Suzuki, H., & Brombacher, F. (2015). Targeting Batf2 for infectious diseases and cancer. *Oncotarget, 6*(29), 26575.

Harding, E. (2020). WHO global progress report on tuberculosis elimination. *The Lancet Respiratory Medicine, 8*(1), 19.

Hennink, M., Hutter, I., & Bailey, A. (2011). In-depth interviews. *Qualitative Research Methods.London: Sage,* , 108-134.

Holden, I. K., Lillebaek, T., Andersen, P. H., Bjerrum, S., Wejse, C., & Johansen, I. S. (2019). Extrapulmonary tuberculosis in Denmark from 2009 to 2014; characteristics and predictors for treatment outcome. Paper presented at the *Open Forum Infectious Diseases, , 6*(10) ofz388.

Houben, R. M., & Dodd, P. J. (2016). The global burden of latent tuberculosis infection: a re-estimation using mathematical modelling. *PLoS Medicine, 13*(10), e1002152.

Hu, Y., Wen, Z., Liu, S., Cai, Y., Guo, J., Xu, Y., Lin, D., Zhu, J., Li, D., & Chen, X. (2020). Ibrutinib suppresses intracellular mycobacterium tuberculosis growth by inducing macrophage autophagy. *Journal of Infection, 80*(6), e19-e26.

Jacobsen, M., Repsilber, D., Gutschmidt, A., Neher, A., Feldmann, K., Mollenkopf, H. J., Ziegler, A., & Kaufmann, S. H. (2007). Candidate biomarkers for discrimination between infection and disease caused by Mycobacterium tuberculosis. *Journal of Molecular Medicine, 85*, 613-621.

Jamshidi, N., & Palsson, B. Ø. (2007). Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ 661 and proposing alternative drug targets. *BMC Systems Biology, 1*, 1-20.

Jeong, H., Mason, S. P., Barabási, A., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature, 411*(6833), 41-42.

Kanabalan, R. D., Lee, L. J., Lee, T. Y., Chong, P. P., Hassan, L., Ismail, R., & Chin, V. K. (2021). Human tuberculosis and Mycobacterium tuberculosis complex: A review on genetic diversity, pathogenesis and omics approaches in host biomarkers discovery. *Microbiological Research, 246*, 126674.

Knopf, J. W. (2006). Doing a literature review. *PS: Political Science & Politics, 39*(1), 127-132.

Krysl, J., Korzeniewska-Kosela, M., Müller, N. L., & FitzGerald, J. M. (1994). Radiologic features of pulmonary tuberculosis: an assessment of 188 cases. *Canadian Association of*

Radiologists Journal= Journal L'Association Canadienne Des Radiologistes, 45(2), 101-107.

Kumar, N. P., Hissar, S., Thiruvengadam, K., Banurekha, V. V., Balaji, S., Elilarasi, S., Gomathi, N. S., Ganesh, J., Aravind, M. A., & Baskaran, D. (2021). Plasma chemokines as immune biomarkers for diagnosis of pediatric tuberculosis. *BMC Infectious Diseases, 21*(1), 1-11.

Lee, S., Wu, L. S., Huang, G., Huang, K., Lee, T., & Weng, J. T. (2016). Gene expression profiling identifies candidate biomarkers for active and latent tuberculosis. Paper presented at the *BMC Bioinformatics, , 17*(1) 27-39.

Li, B., Sun, L., Sun, Y., Zhen, L., Qi, Q., Mo, T., Wang, H., Qiu, M., & Cai, Q. (2023). Identification of the key genes of tuberculosis and construction of a diagnostic model via weighted gene co-expression network analysis. *Journal of Infection and Chemotherapy, 29*(11), 1046-1053.

Li, X., Liao, M., Guan, J., Zhou, L., Shen, R., Long, M., & Shao, J. (2022). Identification of key genes and pathways in peripheral blood mononuclear cells of type 1 diabetes mellitus by integrated bioinformatics analysis. *Diabetes & Metabolism Journal, 46*(3), 451-463.

Li, Z., Qu, X., Liu, X., Huan, C., Wang, H., Zhao, Z., Yang, X., Hua, S., & Zhang, W. (2020). GBP5 is an interferon-induced inhibitor of respiratory syncytial virus. *Journal of Virology, 94*(21), 10.1128/jvi. 01407-20.

Ling, D. I., Flores, L. L., Riley, L. W., & Pai, M. (2008). Commercial nucleic-acid amplification tests for diagnosis of pulmonary tuberculosis in respiratory specimens: meta-analysis and meta-regression. *PloS One, 3*(2), e1536.

Linnenluecke, M. K., Marrone, M., & Singh, A. K. (2020). Conducting systematic literature reviews and bibliometric analyses. *Australian Journal of Management, 45*(2), 175-194.

LoBue, P. A., & Mermin, J. H. (2017). Latent tuberculosis infection: the final frontier of tuberculosis elimination in the USA. *The Lancet Infectious Diseases, 17*(10), e327-e333.

Lv, L., Li, C., Zhang, X., Ding, N., Cao, T., Jia, X., Wang, J., Pan, L., Jia, H., & Li, Z. (2017). RNA profiling analysis of the serum exosomes derived from patients with active and latent Mycobacterium tuberculosis infection. *Frontiers in Microbiology, 8*, 1051.

Lyu, L., Zhang, X., Li, C., Yang, T., Wang, J., Pan, L., Jia, H., Li, Z., Sun, Q., & Yue, L. (2019). Small RNA profiles of serum exosomes derived from individuals with latent and active tuberculosis. *Frontiers in Microbiology, 10*, 1174.

Maertzdorf, J., Ota, M., Repsilber, D., Mollenkopf, H. J., & Weiner, J. (2011). Functional Correlations of Pathogenesis-Driven Gene Expression Signatures in.

Maertzdorf, J., Ota, M., Repsilber, D., Mollenkopf, H. J., Weiner, J., Hill, P. C., & Kaufmann, S. H. (2011). Functional correlations of pathogenesis-driven gene expression signatures in tuberculosis. *PloS One, 6*(10), e26938.

Mehmood, M. A., Sehar, U., & Ahmad, N. (2014). Use of bioinformatics tools in different spheres of life sciences. *Journal of Data Mining in Genomics & Proteomics, 5*(2), 1.

Menzies, N. A., Wolf, E., Connors, D., Bellerose, M., Sbarra, A. N., Cohen, T., Hill, A. N., Yaesoubi, R., Galer, K., & White, P. J. (2018). Progression from latent infection to active disease in dynamic tuberculosis transmission models: a systematic review of the validity of modelling assumptions. *The Lancet Infectious Diseases, 18*(8), e228-e238.

Murphy, T. L., Tussiwand, R., & Murphy, K. M. (2013). Specificity through cooperation: BATF–IRF interactions control immune-regulatory networks. *Nature Reviews Immunology, 13*(7), 499-509.

Naftalin, C. M., Verma, R., Gurumurthy, M., Hee, K. H., Lu, Q., Yeo, B. C. M., Tan, K. H., Lin, W., Yu, B., & Seng, K. Y. (2018). Adjunctive use of celecoxib with anti-tuberculosis drugs: evaluation in a whole-blood bactericidal activity model. *Scientific Reports, 8*(1), 13491.

Onwuegbuzie, A. J., & Leech, N. L. (2005). On becoming a pragmatic researcher: The importance of combining quantitative and qualitative research methodologies. *International Journal of Social Research Methodology, 8*(5), 375-387.

Ørngreen, R., & Levinsen, K. T. (2017). Workshops as a research methodology. *Electronic Journal of E-Learning, 15*(1), 70-81.

Ottenhoff, T. H., Dass, R. H., Yang, N., Zhang, M. M., Wong, H. E., Sahiratmadja, E., Khor, C. C., Alisjahbana, B., Van Crevel, R., & Marzuki, S. (2012). Genome-wide expression profiling identifies type 1 interferon response pathways in active tuberculosis.

Pandey, P., & Pandey, M. M. (2021). *Research methodology tools and techniques*. Bridge Center.

Pfyffer, G. E. (2015). Mycobacterium: general characteristics, laboratory detection, and staining procedures. *Manual of Clinical Microbiology,* , 536-569.

Ponnusamy, N., & Arumugam, M. (2022). Meta-analysis of active tuberculosis gene expression ascertains host directed drug targets. *Frontiers in Cellular and Infection Microbiology,* , 1512.

Pym, A. S., Brodin, P., Brosch, R., Huerre, M., & Cole, S. T. (2002). Loss of RD1 contributed to the attenuation of the live tuberculosis vaccines Mycobacterium bovis BCG and Mycobacterium microti. *Molecular Microbiology, 46*(3), 709-717.

Raman, K., & Chandra, N. (2008). Mycobacterium tuberculosis interactome analysis unravels potential pathways to drug resistance. *BMC Microbiology, 8*, 1-13.

Raman, K., & Chandra, N. (2011). Systems Biology of tuberculosis: Insights for drug discovery. *Understanding the Dynamics of Biological Systems: Lessons Learned from Integrative Systems Biology,* , 83-110.

Raman, K., Rajagopalan, P., & Chandra, N. (2005). Flux balance analysis of mycolic acid pathway: targets for anti-tubercular drugs. *PLoS Computational Biology, 1*(5), e46.

Raman, K., Yeturu, K., & Chandra, N. (2008). targetTB: a target identification pipeline for Mycobacterium tuberculosis through an interactome, reactome and genome-scale structural analysis. *BMC Systems Biology, 2*, 1-21.

Renshaw, P. S., Panagiotidou, P., Whelan, A., Gordon, S. V., Hewinson, R. G., Williamson, R. A., & Carr, M. D. (2002). Conclusive evidence that the major t-cell antigens of themycobacterium tuberculosis complex esat-6 and cfp-10 form a tight, 1: 1 complex and characterization of the structural properties of esat-6, cfp-10, and the esat-6· cfp-10 complex: Implications for pathogenesis and virulence. *Journal of Biological Chemistry, 277*(24), 21598-21603.

Ritchie, M. E., Phipson, B., Wu, D. I., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research, 43*(7), e47.

Roe, J. K., Thomas, N., Gil, E., Best, K., Tsaliki, E., Morris-Jones, S., Stafford, S., Simpson, N., Witt, K. D., & Chain, B. (2016). Blood transcriptomic diagnosis of pulmonary and extrapulmonary tuberculosis. *JCI Insight, 1*(16)

Romanowski, K., Baumann, B., Basham, C. A., Khan, F. A., Fox, G. J., & Johnston, J. C. (2019). Long-term all-cause mortality in people treated for tuberculosis: a systematic review and meta-analysis. *The Lancet Infectious Diseases, 19*(10), 1129-1137.

Roy, S., Guler, R., Parihar, S. P., Schmeier, S., Kaczkowski, B., Nishimura, H., Shin, J. W., Negishi, Y., Ozturk, M., & Hurdayal, R. (2015). Batf2/Irf1 induces inflammatory responses in classically activated macrophages, lipopolysaccharides, and mycobacterial infection. *The Journal of Immunology, 194*(12), 6035-6044.

Russell, D. G. (2007). Who puts the tubercle in tuberculosis? *Nature Reviews Microbiology, 5*(1), 39-47.

Schaberg, T., Reichert, B., Schulin, T., Lode, H., & Mauch, H. (1995). Rapid drug susceptibility testing of Mycobacterium tuberculosis using conventional solid media. *European Respiratory Journal, 8*(10), 1688-1693.

Schnappinger, D., Ehrt, S., Voskuil, M. I., Liu, Y., Mangan, J. A., Monahan, I. M., Dolganov, G., Efron, B., Butcher, P. D., & Nathan, C. (2003). Transcriptional adaptation of Mycobacterium tuberculosis within macrophages: insights into the phagosomal environment. *The Journal of Experimental Medicine, 198*(5), 693-704.

Sgaragli, G., & Frosini, M. (2016). Human tuberculosis I. Epidemiology, diagnosis and pathogenetic mechanisms. *Current Medicinal Chemistry, 23*(25), 2836-2873.

Sgaragli, G., Frosini, M., Saponara, S., & Corelli, F. (2016). Human tuberculosis. III. Current and prospective approaches in anti-tubercular therapy. *Current Medicinal Chemistry, 23*(21), 2245-2274.

Shao, J., Jin, Y., Shao, C., Fan, H., Wang, X., & Yang, G. (2021). Serum exosomal pregnancy zone protein as a promising biomarker in inflammatory bowel disease. *Cellular & Molecular Biology Letters, 26*(1), 1-13.

Shi, T., Huang, L., Zhou, Y., & Tian, J. (2022). Role of GBP1 in innate immunity and potential as a tuberculosis biomarker. *Scientific Reports, 12*(1), 11097.

Singhania, A., Verma, R., Graham, C. M., Lee, J., Tran, T., Richardson, M., Lecine, P., Leissner, P., Berry, M. P., & Wilkinson, R. J. (2018). A modular transcriptional signature identifies phenotypic heterogeneity of human tuberculosis infection. *Nature Communications, 9*(1), 2308.

Smith, N. H., Kremer, K., Inwald, J., Dale, J., Driscoll, J. R., Gordon, S. V., Van Soolingen, D., Hewinson, R. G., & Smith, J. M. (2006). Ecotypes of the Mycobacterium tuberculosis complex. *Journal of Theoretical Biology, 239*(2), 220-225.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology, 3*(1)

Steingart, K. R., Ng, V., Henry, M., Hopewell, P. C., Ramsay, A., Cunningham, J., Urbanczik, R., Perkins, M. D., Aziz, M. A., & Pai, M. (2006). Sputum processing methods to improve the sensitivity of smear microscopy for tuberculosis: a systematic review. *The Lancet Infectious Diseases, 6*(10), 664-674.

Strong, M., Graeber, T. G., Beeby, M., Pellegrini, M., Thompson, M. J., Yeates, T. O., & Eisenberg, D. (2003). Visualization and interpretation of protein networks in Mycobacterium tuberculosis based on hierarchical clustering of genome-wide functional linkage maps. *Nucleic Acids Research, 31*(24), 7099-7109.

Sudre, P., Ten Dam, G., & Kochi, A. (1992). Tuberculosis: a global overview of the situation today. *Bulletin of the World Health Organization, 70*(2), 149.

Sutherland, I. (1976). Recent studies in the epidemiology of tuberculosis, based on the risk of being infected with tubercle bacilli. *Advances in Tuberculosis Research.Fortschritte Der Tuberkuloseforschung.Progres De L'Exploration De La Tuberculose, 19*, 1-63.

Tasiou, A., Giannis, T., Brotis, A. G., Siasios, I., Georgiadis, I., Gatos, H., Tsianaka, E., Vagkopoulos, K., Paterakis, K., & Fountas, K. N. (2017). Anterior cervical spine surgery-associated complications in a retrospective case-control study. *Journal of Spine Surgery, 3*(3), 444.

Tonby, K., Wergeland, I., Lieske, N. V., Kvale, D., Tasken, K., & Dyrhol-Riise, A. M. (2016). The COX-inhibitor indomethacin reduces Th1 effector and T regulatory cells in vitro in Mycobacterium tuberculosis infection. *BMC Infectious Diseases, 16*(1), 1-12.

Ulrichs, T., Munk, M. E., Mollenkopf, H., Behr-Perst, S., Colangeli, R., Gennaro, M. L., & Kaufmann, S. H. (1998). Differential T cell responses to Mycobacterium tuberculosis ESAT6 in tuberculosis patients and healthy donors. *European Journal of Immunology, 28*(12), 3949-3958.

van Ingen, J., Rahim, Z., Mulder, A., Boeree, M. J., Simeone, R., Brosch, R., & Van Soolingen, D. (2012). Characterization of Mycobacterium orygis as M. tuberculosis complex subspecies. *Emerging Infectious Diseases, 18*(4), 653.

Verkhedkar, K. D., Raman, K., Chandra, N. R., & Vishveshwara, S. (2007). Metabolome Based Reaction Graphs of M. tuberculosis and M. leprae: A.

Volkman, H. E., Clay, H., Beery, D., Chang, J. C. W., Sherman, D. R., & Ramakrishnan, L. (2004). Tuberculous granuloma formation is enhanced by a mycobacterium virulence determinant. *PLoS Biology, 2*(11), e367.

Volkman, H. E., Pozos, T. C., Zheng, J., Davis, J. M., Rawls, J. F., & Ramakrishnan, L. (2010). Tuberculous granuloma induction via interaction of a bacterial secreted protein with host epithelium. *Science, 327*(5964), 466-469.

Von Mering, C., Jensen, L. J., Kuhn, M., Chaffron, S., Doerks, T., Krüger, B., Snel, B., & Bork, P. (2007). STRING 7—recent developments in the integration and prediction of protein interactions. *Nucleic Acids Research, 35*(suppl_1), D358-D362.

Waddell, S. J., Butcher, P. D., & Stoker, N. G. (2007). RNA profiling in host–pathogen interactions. *Current Opinion in Microbiology, 10*(3), 297-302.

Waddell, S. J., Stabler, R. A., Laing, K., Kremer, L., Reynolds, R. C., & Besra, G. S. (2004). The use of microarray analysis to determine the gene expression profiles of Mycobacterium tuberculosis in response to anti-bacterial compounds. *Tuberculosis, 84*(3-4), 263-274.

Wang, J., Duncan, D., Shi, Z., & Zhang, B. (2013). WEB-based gene set analysis toolkit (WebGestalt): update 2013. *Nucleic Acids Research, 41*(W1), W77-W83.

Weiner, J., Maertzdorf, J., & Kaufmann, S. H. (2013). The dual role of biomarkers for understanding basic principles and devising novel intervention strategies in tuberculosis. *Annals of the New York Academy of Sciences, 1283*(1), 22-29.

Woodring, J. H., Vandiviere, H. M., Fried, A. M., Dillon, M. L., Williams, T. D., & Melvin, I. G. (1986). Update: the radiographic features of pulmonary tuberculosis. *American Journal of Roentgenology, 146*(3), 497-506.

World Health Organization. (2023). Global tuberculosis report 2023. *Global tuberculosis report 2023* ()

Yang, Q., Wang, S., Dai, E., Zhou, S., Liu, D., Liu, H., Meng, Q., Jiang, B., & Jiang, W. (2019). Pathway enrichment analysis approach based on topological structure and updated annotation of pathway. *Briefings in Bioinformatics, 20*(1), 168-177.

Yao, X., Liu, W., Li, X., Deng, C., Li, T., Zhong, Z., Chen, S., Ge, Z., Zhang, X., & Zhang, S. (2022). Whole blood GBP5 protein levels in patients with and without active tuberculosis. *BMC Infectious Diseases, 22*(1), 1-8.

APPENDICES

Appendix 1: List of Drugs in the DrugBank and their Interactions with the DEGs

Gene	Drug	Disease Treated	Interaction score
SERPING1	C1 ESTERASE INHIBITOR	Hereditary angioedema	29.49
CEACAM1	ARCITUMOMAB	Diagnostic agent	9.83
CACNA1I	PARAMETHADIONE	Anticonvulsants	2.80
ATF3	PROGESTERONE	For reducing the risk of preterm birth for women with short cervix a mid-pregnancy, for prevention of preterm delivery, for symptomatic treatment of menopausal symptoms, neuroprotectant for stroke victims	2.22
CACNA1I	ETHOSUXIMIDE	Anticonvulsants	1.12
BMX	IBRUTINIB	Antineoplastic agent	0.86
CD19	BLINATUMOMAB	Antineoplastic agent	0.84
	TAFASITAMAB	Antineoplastic agent	0,73
CACNA1I	TRIMETHADIONE	Anticonvulsants	0.70
ADM	PAROXETINE HYDROCHLORIDE, HEMIHYDRATE	Antidepressant	0.64
HPSE	ASTEMIZOLE	Anti-Allergic Agents	0.53
HPSE	LABETALOL	Antihypertensive Agents	0.49
ADM	INDOMETHACIN	NSAID(Non-steroidal anti-inflammatory drugs)	0.42
CEACAM1	TRETINOIN	For treatment of acne	0.35
HPSE	THROMBIN	Topical tissue sealant	0.28
ADM	INSULIN, REGULAR, HUMAN	For the treatment of diabetic foot ulcers, antidiabetics	0.28
NRG1	PROGESTIN	Contraceptive	0.23
NRG1	AFATINIB	Antineoplastic agent	0.21
NRG1	PERTUZUMAB	Antineoplastic agent	0.18

NRG1	BUPIVACAINE	Local, Anesthetics, neuralgia, analgesic, local anestethic	0.17
CACNA1I	ZONISAMIDE	Anticonvulsant, antipsychotic agent, appetite suppressant	0.12
CACNA1I	PREGABALIN	For treatment of restless legs syndrome, neuropathic pain, analgessic	0.10
CACNA1I	GABAPENTIN ENACARBIL	For treatment of restless legs syndrome	0.10
BLK	IBRUTINIB	Antineoplastic agent	0.10
NRG1	LAPATINIB	Antineoplastic agent	0.10
HPSE	TINZAPARIN SODIUM	For treatment of cystic fibrosis, pelvic pain of bladder origin and interstitial cystitis, antithrombotic, anticoagulant, Anticoagulants	0.10
NRG1	ISOPROTERENOL	Bronchodilator Agents; Cardiotonic Agents	0.09
CACNA1I	CELECOXIB	NSAID	0.08
CACNA1I	VERAPAMIL	Antihypertensive agent	0.07
CACNA1I	GABAPENTIN	Analgesic, for the treatment of neuropathic pain	0.07
NRG1	CETUXIMAB	Antineoplastic agent	0.07
NRG1	PANITUMUMAB	Antineoplastic agent	0.06
NRG1	NICOTINE POLACRILEX	Central Nervous System Stimulants	0.06
NRG1	COLCHICINE	For treatment of gout	0.05
NRG1	PROGESTERONE	For reducing the risk of preterm birth for women with short cervix a mid-pregnancy, for prevention of preterm delivery, for symptomatic treatment of menopausal symptoms, neuroprotectant for stroke victims	0.04
NRG1	TAMOXIFEN	Hormonal, Antineoplastic Agents	0.04
BLK	ERLOTINIB	Antineoplastic agent	0.03
NRG1	VINCRISTINE	Antineoplastic agent	0.03

NRG1	DEXAMETHASONE	For the treatment of Meniere’s disease, glucocorticoid, an anti-inflammatory agent	0.03
NRG1	ASPIRIN	NSAID	0.03
NRG1	CYTARABINE	Antineoplastic agent	0.03
BLK	DASATINIB ANHYDROUS	Antineoplastic agent	0.02
BLK	GEFITINIB	Antineoplastic agent	0.02
NRG1	ERLOTINIB	Antineoplastic agent	0.02
BLK	SORAFENIB	Antineoplastic agent	0.01
NRG1	GEMCITABINE	Antineoplastic agent	0.01
NRG1	PACLITAXEL	For treatment of peripheral arterial disease (PAD), DMARD, anti-inflammatory agent, antineoplastic agent	0.01