Investigating the Role of Phylogenetic Approach in Predicting Prognostic Biomarkers for Multiple Myeloma

Submitted By

Ms. Jawaria
(57-FOC/MSBI/F22)

Supervised by

Ms. Tahira Noor
Lecturer

Department of Bioinformatics
Faculty of Computing & Information Technology
International Islamic University Islamabad

2025

Investigating the Role of Phylogenetic Approach in Predicting Prognostic Biomarkers for Multiple Myeloma

MS Research Thesis

Submitted By

Ms. Jawaria
(57-FOC/MSBI/F22)

Supervised by

Ms. Tahira Noor
Lecturer

Final year project report submitted to the Department of Bioinformatics as a part of course of studies of Master's degree in Bioinformatics of the International Islamic University, Islamabad.

Department of Bioinformatics
Faculty of Computing & Information Technology
International Islamic University Islamabad
2025

Department of Bioinformatics
Faculty of Computing & Information Technology
International Islamic University, Islamabad

Date: _____

Final Approval

This is to certify that we have read the thesis of **'Investigating the Role of Phylogenetic Approach in Predicting Prognostic Biomarkers for Multiple Myeloma'** submitted by Ms. Jawaria, 57-FOC/MSBI/F22. It is our judgment that this project is of sufficient standard to warrant its acceptance by the International Islamic University Islamabad for the Master Degree in Bioinformatics.

Committee:

External Examiner

Dr. Sajid Rashid
Professor
National Center for Bioinformatics
Quaid-i-Azam University,
Islamabad

Internal Examiner

Dr. Attiya Kanwal
Assistant Professor
Head of Department
FCIT
IIUI, Islamabad

Supervisor

Ms. Tahira Noor
Lecturer
In charge DBI Female Program
FCIT
IIUI, Islamabad

DISSERTATION

A dissertation submitted to Department of Bioinformatics, Faculty of Computing & Information Technology, International Islamic University Islamabad as a partial fulfillment of the requirements for the award of the degree of Master in Bioinformatics (MSBI).

DEDICATION

I would like to dedicate this thesis to Allah Almighty our creator, Strong pillar and source of aspiration, wisdom, knowledge and understanding has been the foundation of our strength to this project. I also dedicate this work to our parents, family, friends and teachers who have encouraged us all the way. I want to pay special thanks to my supervisor Ms. Tahira Noor who not only helped me throughout the course of this time to solve every query but also gave me moral support and guidance to fulfill this task.

DECLARATION

I hereby declare that this thesis "Investigating the Role of Phylogenetic Approach in Predicting Prognostic Biomarkers for Multiple Myeloma" neither as a whole nor as a part has been copied out from any source. It is further declared that we have done this research with the accompanied report entirely on the basis of our personal efforts, under the proficient guidance of our teachers especially our supervisor Ms. Tahira Noor. If any part of the system is proved to be copied out from any source or found to be reproduction of any project from any of the training institute or educational institute, I shall stand by the consequences.

_____

Jawaria

57-FOC/MSBI/F22

ACKNOWLEDGEMENTS

My deepest gratitude to ALLAH Almighty, who helped me in every step of life. It has been a strenuous journey, but I have finally reached the finish line with my final thesis.

I owe an immense amount of gratitude to all those who contributed in one way or another; without you, this would not be possible. My special regards to all the faculty members of the Department of Bioinformatics, especially all my teachers and supervisors; who always encourages me and always have tried to develop my interest in research. I am thankful that they provided great opportunities to me even in hard times like pandemic.

Moreover, all the assistance provided by my supervisor, **Ms. Tahira Noor,** is greatly appreciated. I am thankful for her patience, guidance, encouragement and useful critiques for this project.

Finally, I want to extend my thanks towards my husband, parents, family and friends for always being my support and making all this possible.

Thank You All!

_____

Jawaria

57-FOC/MSBI/F22

PROJECT IN BRIEF

Project Title:	Investigating the Role of Phylogenetic Approach in Predicting Prognostic Biomarkers for Multiple Myeloma
Organization:	Department of Bioinformatics Faculty of Computing and Information Technology International Islamic University H-10, Islamabad
Undertaken By:	Jawaria 57-FOC/MSBI/F22
Supervised By:	Ms. Tahira Noor
Start Date:	2023
Completion Date:	2025
Objective:	The main goal of this thesis is to identify novel prognostic biomarkers for MM progression by conducting comprehensive bioinformatics analysis and confirm the clinical significance and application of discovered prognostic biomarkers
Tools & Technologies:	RStudio, GEO2R, DAVID, STRING, Cytoscape, ClustalW, MEGA11, KMplotter.
Documentation Tool:	MS Word, MS PowerPoint
Operation System:	Windows 10
System Used:	Lenovo core i5

Table of Contents

List of Abbreviations

Acronym	Abbreviations	Acronym	Abbreviations
MM	Multiple Myeloma	AID	Activation-Induced Cytidine Deaminase
MGUS	Monoclonal Gammopathy of Unknown Significance	RPS19	Ribosomal Protein S19
SMM	Smoldering Multiple Myeloma	RPS17	Ribosomal Protein S17
ISS	International Staging System	RPS25	Ribosomal Protein S25
R-ISS	Revised International Staging System	RPL35A	Ribosomal Protein S35A
LDH	Lactate Dehydrogenase	FISH	Fluorescence In Situ Hybridization
NGS	Next Generation Sequencing	KEGG	Kyoto Encyclopaedia of Genes and Genomes
WGCNA	Weighted Gene Co-Expression Network Analysis	DAVID	Database for Annotation, Visualization and Integrated
DEGs	Differentially Expressed Genes	PPI	Protein-Protein Interaction
ME	Module Eigengenes	DMNC	Density of Maximum Neighborhood Component
GO	Gene Ontology	EPC	Edge Percolated Component
CC	Cellular Component	MNC	Maximum Neighborhood Component
BP	Biological Process	MCC	Maximal Clique Centrality
MF	Molecular Function	MSA	Multiple Sequence Alignment
KM	Kaplan-Meier	MEGA11	Molecular Evolutionary Genetics Analysis
ROC	Receiver Operator Characteristic	AUC	Area Under the Curve

List of Figures

List of Tables

ABSTRACT

The prognosis and survival rate of Multiple Myeloma (MM) has notably enhanced by the introduction of novel therapeutic agents such as proteasome inhibitors, immunomodulatory drugs and monoclonal antibodies. However, it remains untreatable and relapse-prone. Previous literature highlights substantial progress in understanding MM pathogenesis and therapeutic development, yet biomarker identification through comprehensive bioinformatics approaches remains a critical gap. Genomic-based biological targets may offer enhanced support for rational disease intervention. Therefore, this study modified existing methodological framework by adding phylogenetic approach for the identification of biomarkers involved in the progression of MM. This study employed integrated bioinformatics and phylogenetic analysis with expression datasets (GSE6477 and E-GEOD-5900) downloaded from GEO and ArrayExpress. The GSE6477dataset was used to screen DEGs and E-GEOD-5900 dataset was used to generate a co-expression network. The Datasets GSE13591 and GSE24080 were used for validating the hub genes, while the PPI network identified hub nodes. Moreover, functional enrichment and phylogenetic analyses were carried out to further explore DEGs and key genes for MM progression. All the potential hub genes were assessed using Kaplan-Meier Plotter for survival-analysis. There were total 1505 (DEGs) identified. GO annotation and KEGG pathway mapping revealed a connection of DEGs with the ribosome biogenesis process. It is proved that RPS17, RPS25, RPS19 and RPL35A have clinical prognostic significance for MM. New biomarkers RPS17, RPS25, RPS19, and RPL35A have been discovered which can impact the outcome of prognosis in MM patients. The hub genes identified are prospective biomarkers for disease progression. Further treatment-biomarker interactions and clinical studies are required to make them potential targets for clinical treatment.

Chapter 1

Introduction

1 Introduction

1.1. Multiple Myeloma (MM)

Multiple myeloma (MM) represents blood neoplasm which involves abnormal division of antibody producing plasma cells. It is a type of immune cell that play a key function in producing antibodies [3]–[5]. MM progression several stages, from Monoclonal gammopathy of unknown significance (MGUS), the precursor stage of smouldering multiple myeloma (SMM), and to the final stage of active multiple myeloma (MM) [6], [7]. Active MM is further subdivided into three stages (Stage I, Stage II and Stage III) based on the degree of abnormal proteins in the blood and the extent of organ damage. This staging is based on the Revised International Staging System (R-ISS) which incorporates both biomedical markers and cytogenetic features to asses disease severity and prognosis. Stage I includes patients with serum β2-microglobulin levels less than 3.5 mg/L, albumin levels equal to or greater than 3.5 g/dL, normal lactate dehydrogenase (LDH) levels and absence of high-risk chromosomal abnormalities such as deletion 17p, t(4;14), or t(14;16). This stage is associated with a favorable prognosis. Stage II is an intermediate category defined by values that do not meet the criteria for either Stage I or Stage III. It represents intermediate risk group with variable prognosis. Stage III is characterized by β2 microglobulin levels equal to or exceeding 5.5 mg/L. It is accompanied by either elevated LDH levels or the presence of high-risk cytogenetic abnormalities. It indicates more aggressive disease and a poorer clinical outcome. This staging framework plays a crucial role in guiding therapeutic decisions and predicting patient outcomes. MM comprises nearly 10% of every hematologic malignancies and 1% of neoplasms globally. As therefore, it is a rare cancer type [8]. It has been more common in elderly adults with an average diagnosis age of 69 years [4], [5].

1.1.1. Clinical Symptoms of Multiple Myeloma

In MM, abnormal plasma cells accumulate in the bone marrow, displacing healthy blood cells and producing abnormal antibodies. This disturbance causes various problems such as anemia, bone pain, and increased susceptibility to infections. Bone fracture, fatigue, recurring infections and unexpected weight loss are typical symptoms of MM [7], [9]. There is no new bone development in the osteolytic bone lesions of MM, unlike other cancers that spread to the bone marrow. The primary cause of morbidity is bone disease which is best identified by magnetic resonance imaging (MRI), Fluoro-deoxyglucose (FDG) positron emission tomography/computed tomographic scans (PET/CT), or low dose whole body computed tomography(WB-CT) [10]. 8% of patients develop

extramedullary disease (EMD) later in the course of illness, whereas only 1% to 2% of patients had EMD at the time of initial diagnosis [11]. These clinical features underscore the importance of early detection in MM management.

1.1.2. Molecular and Cellular basis of Multiple Myeloma

A complex interaction of molecular and cellular processes that originates within the B-cell developmental route leads to MM. Fundamentally, the disease is caused by random mutagenesis, a natural and pervasive process in replicating tissue that arises from enzymatic changes DNA bases. Exogenous factors that arise mutation rates may make these mutational events even worse. Although the majority of mutations are neutral or harmful, but rare variants arise that give impacted plasma cells a proliferative or survival benefit. The initial clonal proliferation, often clinically silent can progressively accumulate more genetic changes. This makes the shifts from benign precursor state to overt malignancy easier. In a dynamic and complex multicellular microenvironment, the competition for nutrients, oxygen and growth stimulant along with critical requirement for immune evasion influences the natural selection of clones. Tumor plasticity is further enhanced by loss of genomic stability due to compromised DNA repair and the acquisition of hypermutator phenotypes which allow MM cells to adapt and flourish under selection pressures. All of these cellular and molecular processes work together to support the pathophysiology of MM. Finally, determines its clinical heterogeneity and variable progression form asymptomatic phases to aggressive, symptomatic myeloma [12]. This complexity poses significant challenges in prognosis and treatment planning.

1.2 Epidemiological Background

MM is a highly heterogeneous disease with significantly varied clinical presentations and outcomes. It is incurable with most patients relapsing or becoming resistant to drugs over time. Despite significant advancement in treatments in the last two decades involving immunomodulatory drugs, monoclonal antibodies and proteasome inhibitors, it is incurable with most patients relapsing or becoming resistant to drugs over time [13]–[16]. In recent years, overall survival has improved to more than five years, but more effective and personalized treatment options are still required [17], [18]. Therefore, identifying reliable biomarkers remains essential to improve early diagnosis, monitor disease progression and guide targeted therapy.

1.2.1 Incidence Trends in Different Populations

MM incidence rates have gradually increased globally over the past two decades globally. This increase is seen in a wide range of age groups, geographical regions and urbanization levels. Different parts of the world have been shown to have varying incidence patters. Some areas have consistently higher rates, while others claim lower incidence. The highest incidence of MM was continuously seen in non-Hispanic Black populations across age, U.S. census regions and urbanization levels. On the other hand, Asian American and Pacific Islander (AAPI) and non-Hispanic American Indian/Alaska Native (AIAN) population has lowest incidence rates [73]. Men experienced higher MM incidence compared to women, reflecting gender disparities. A major impact of this disease is disturbing the quality of life of a patient. [8], [9]. However, the latest trends seem to indicate that there is an alarming global trend for MM incidence and mortality. [19] predict that between 2018 to 2043, there will be an increase of 14.9% in incidence per year. It implies the increasing burden of this disease especially in the elderly population [18], [19]. By 2023, approximately 160,000 new cases of this disease will be diagnosed annually all over the world along with 106,000 deaths. It reflects the gravity and severity of MM [20]. Genetic susceptibility, environment exposures, healthcare access, diagnostic procedures and lifestyle-related risk factors including obesity can all contribute to differences in incidence. Additionally, the distribution of precursor illnesses such as MGUS, varies in prevalence between populations. It influences geographical inequalities in MM incidence. The significance of more research into the underlying biological, sociodemographic and environmental factors contributing to the trends highlighted by the geographic clustering of higher MM incidence in particular areas [73].

1.2.2 Prevalence and Global Disease Burden of Multiple Myeloma

The prevalence of MM has dramatically increased globally over the past three decades. The prevalence, incidence, mortality and disability-adjusted life years (DALYs) related to MM have significantly increased between 1990 and 2021. In certain measures, these increases have nearly doubled. Age standardized rates such as DALYs (ASDR), mortality (ASDR), incidence (ASIR) and prevalence (ASPR) showed increasing trends particularly among males and population in middle-SDI regions. While mortality and disability rates in women did not decrease. Their projected annual percentage changes (EAPCs) were marginally negative and indicates the growth of the disease burden that may have outpaced due to female population. Furthermore, the

percentage of MM cases linked to high body mass index (BMI) increased from 6.40% in 1990 to 7.96% in 2021 on global scale. Over the next 15 years, the global burden of MM is expected to rise continuously according to the ARIMA model projections. Several factors influence the rising trends in disease such as global population aging and improvements in healthcare and diagnostics [74]. To improve patient outcomes and address healthcare disparities, the growing prevalence of MM especially among older individuals in the 70-74 age range calls for continued research and focused therapies [19]. Recent research indicates that, with median age of 69 upon diagnosis, the disease primarily affects older individuals [21]. Black individuals are almost twice as likely as white person to be diagnosed with MM, likely due to a higher prevalence of MGUS, genetic predisposition and socioeconomic disparities. These factors collectively contribute to increased disease susceptibility and delayed diagnosis. This emphasizes the need for focused health interventions [22], [23]. The actual burden of MM may be underestimated in areas with less developed healthcare systems, where underreporting is still a problem. These global trends highlight the intricate interactions among healthcare infrastructure, diagnostic procedures and demographic shifts that shape the trends in MM burden that have been documented [74].

1.2.3 Mortality Rates and Survival Outcomes of Multiple Myeloma

Despite these alarming statistics, MM death rates have dropped dramatically over the last decade about 27.5% due to better treatment outcomes. An estimated 12,540 fatalities in 2024, roughly 3.0 deaths per 100,000 individuals occurs in the cause of MM [19]. Therapeutic strategies, such as alkylating agents, corticosteroids immunomodulatory medications, proteasome inhibitors and monoclonal antibodies for MM evolved rapidly. This evolution led to more patients attaining deep remission and favorable prognoses. Over the past 10 years, the median survival span has increased from 3-5 or 8-10 years [24]. However, five years of survival have improved to only 54% suggesting advancements in therapeutic approaches and early diagnosis techniques [19], [22] as mortality is one of the predominant reasons of inadequate diagnosis. For instance, MM is more likely to go undetected or misdiagnosed in the early stage, which might worsen patient's chances of survival due to lack of timely diagnosis. Since young patients are more likely to have a better prognosis due to their improved physical health rather than older patients. Therefore, stem cell transplantation and novel medication therapy are offered ot them preferentially. Meanwhile, it was stated that the current treatments were ineffective for elder populations. As a result, new regimens

with lower toxicity and frailty-adapted therapy are needed to potentially address the age disparity in MM management [24]. To significantly improve survival rates and address persistent difficulties faced by people with this complex disease, more research and clinical trials are needed.

1.2.4 Current Drug Treatment and Limitations

Based on thorough understanding of mechanism and importance of the MM microenvironment, Proteasome inhibitors like bortezomib have been developed in conjunction with immunomodulatory drugs (IMiDs) and steroids. This has resulted in a significant improvement in treatment response and survival for patients with MM. However, MM is still incurable because of different medication resistance mechanism. These include clonal evolution, including PSMD4 hyperexpression linked to chromosome 1q21 amplification and t(4;14) unbalanced translocation, MDR gene polymorsphism [25]. Current prognostic techniques that rely on clinical and laboratory characteristic like the ISS and Revised ISS, may not fully account for the underlying genetic and molecular heterogeneity of MM [18], [21]. Recent research has identified a number of high risk cytogenetic abnormalities including t(4;14), t(14;16) and del(17p). These abnormalities are linked to worse outcomes. Yet, there is still a lack of incorporation of these abnormalities into routine clinical practice [26] To overcome these limitations, researchers are looking into emerging techniques such as CAR-T cells, checkpoint inhibitors, mTOR inhibitors, anti-cytokine drugs, second-generation proteasome inhibitors, and monoclonal antibodies [25]. Finding prognostic and predictive biomarkers remains crucial for optimizing therapeutic strategies and overcome resistance in MM.

1.3. Novel Biomarker Discovery in Multiple Myeloma

1.3.1 Relapse and Drug Resistance in Multiple Myeloma

Despite advances in treatment, patients with MM are still at high risk of relapse and disease progression. This necessitates early treatments to reduce morbidity and mortality. Patients who shows progression during undergoing treatment and fails to achieve at least a minimal response to initial therapy are classified as “primary refractory”. The term “double refractory” refers to MM that progresses during or after treatment with both immunomodulatory agent and a proteasome inhibitor. “triple class refractory disease is linked to a particularly poor prognosis. It is defined by resistance to monoclonal antibodies in addition to these therapies [27]. The International Myeloma Working Group (IMWG) defines relapsed and refractory multiple myeloma (RRMM) as the

absence of at least minimal response, biochemical or clinical signs of progression. It progresses within 60 days of the most recent therapy in patients who had previously responded. Even in the absence of biochemical advancement, IMWG defines progressive disease as having at least as 25% increase in serum or urine paraprotein from the lowest response level, elevated free light chains or new CRAB symptoms. Patients with MM frequently go through several cycles of remission and relapse during the course of the disease. This necessitates multiple lines of combination therapy. Treatment planning is made more difficult by the complex and poor understanding of etiology of medication resistance and recurrence. Short durations of previous remission, high-risk cytogenetic profiles, plasma cell leukaemia, immune system malfunction and insufficient response to previous treatments are all contributing factors. Del(17p), gain(1q)/del(1p), t(4;14) and t(14;16) are high risk cytogenetic characteristics that are present in about 20% of patients who relapse aggressively [21]. Patients are classified as having "double hit" myeloma if they have two of these anomalies and "triple hit" status if they have three or more. Similarly, patients with trisomies or translocations such as t(11;14) and t(6;14) are classified as standard-risk. Furthermore, toxicity form aggressive combination treatments and poor performance status can cause therapy to be interrupted or delayed. This can result in early relapse, reduced overall survival and diminished quality of life for patients as well as highlighting the need for more tolerable and targeted approaches [27].

1.3.2 Gaps in Existing Biomarker-Based Prognostic Tools

There are significant limitations with the current biomarker based prognostic methods for MM, despite improvements in genomic and transcriptome profiling. Firstly, patients in the newly discovered MDMS8 subgroup have the same high risk biomarkers like 1q gain, del17p and t(4;14) and shows worse progression free and overall survival outcomes. This suggests that using traditional biomarkers alone might not be enough to accurately stratify patient prognosis. Secondly, compared to traditional gene expression (GE) or copy number (CN) based approaches indicates a need for more comprehensive bioinformatics frameworks. Thirdly, despite their wide usage, GE based classifiers such as EMC92 and UAMS70 frequently classify patients with different biological profiles and clinical outcomes under the same risk category. As seen with MDMS8 raises the possibility of misclassification [28]. All in all, results demonstrate the necessity of more thorough molecular stratification beyond existing methods and supports the integration of phylogenetic analysis as a complementary approach. Phylogenetic approaches have the potential

to improve disease classification and prognosis accuracy in MM by fine-tuning the mapping of clonal evolution and tumor heterogeneity.

1.4 Overview of Research Approach

1.4.1 Computational Framework for Biomarker Discovery

Recent advancements in computational biology has enabled the development of diverse framework for predicting drug responses and identifying biomarkers using large-scale pharmacogenomics datasets. Numerous computational strategies like linear regression models, Bayesian inference methods and matrix factorization techniques have been developed with the availability of extensive datasets. These models seek to infer drug sensitivity patterns across various cancer types by utilizing omics data, namely gene expression profiles. By learning nonlinear drug response functions from similarity matrices of both pharmaceuticals and cell lines, similarity-regularized matrix factorization (SRMF) has superior performance [29]. While some indirect approaches like SRMF that model nonlinear relations through similarity matrices conceptually align with phylogenetic analysis in their shared goal of uncovering pattern within biological data. Integrative techniques like DualNets integrate drug structures and protein interaction networks demonstrate the promise of merging several omics layers to increase prediction accuracy. While gene expression is still the most informative data modality. Although deep learning models like CDRscan are gaining popularity but their advantage over traditional methods is not yet definitive. This suggests that model design and data integration strategies remain critical areas for biomarker discovery. Lastly, there is ongoing need for robust, interpretable and accurate computational methods tailored to cancer-specific drug response prediction and biomarker identification [29].

1.4.2 Role of Multi-Omics Integration in MM Analysis

Multiple Myeloma is a perfect candidate for multi-omics analysis due to its genetic heterogeneity and rapid evolution. Research on MM has been revolutionized by high-throughput omics technologies such transcriptomics, genomics, proteomics, metabolomics. It enabled comprehensive profiling of molecular changes that promote disease progression, relapse and drug resistance [30]. Genomic studies using NGS, WES and WGS have unveiled variety of somatic mutations, chromosomal translocation and copy number changes that cause MM heterogeneity. Notably, gene mutations have become critical biomarkers for risk assessment and treatment response. Furthermore, cis-regulatory elements and non-coding areas have been found as

recurrently mutated and functionally significant. This expands the scope of genetic biomarkers [31]. Since transcriptomics profiling can identify patterns of gene expression, fusion events and signaling pathways activations that are not possible form DNA data alone. Therefore, it has emerged as a vital complement to genomics in MM [32]. Epigenomics has further expanded the understanding of MM by revealing abnormal DNA methylation and histone modification patterns that regulate gene expression, B-cell differentiation, disease development, and epigenomic profiling. A number of epigenetic enzymes have been identified as potential therapeutic targets [33]. Proteomics adds functional layer to MM analysis through identification of protein-level changes, treatment response biomarkers and resistance mechanism. Sensitive monitoring of minimal residue disease has been made possible by the discovery of key dysregulated pathways including kinase signaling by mass spectrometry and phosphoproteomics. In addition to genomic and transcriptomic data, proteomic profiling of serum, bone marrow and extracellular vesicles improves the precision of diagnosis, prognosis and therapeutic targeting [34]. Additionally, clonal progression from precursor states (MGUS and SMM) to overt MM had been illuminated by integrative multi-omics approaches. It also demonstrated how treatment pressure alters tumor subclonal architecture. Collectively, multi-omics integration is changing the landscape of MM by guiding personalized therapy, influencing precision medicine strategies and enhancing patient outcomes through molecularly guided risk assessment and treatment adaptation [30].

1.5 Research Gap

The use of phylogenetic analysis to predict prognostic biomarkers by integrating high-dimensional data from many databases has not received much attention. Moreover, the creation of customized treatment plans based on patient's risk profile is further complicated by the absence of trustworthy biomarkers for forecasting treatment relapse and disease progression. Therefore, more research is required to bridge these gaps and identify novel prognostic biomarkers using bioinformatics analysis. The aim of filling these gaps is to create more effective treatment strategies that can overcome medication resistance and enhance long-term outcomes for MM patients.

1.6 Problem Statement

According to previous studies, there hasn't been much exploration into the use of bioinformatics techniques to identify multiple myeloma prognostic indicators. The use of bioinformatics analysis

in discovering prognostic biomarkers must be investigated in order to bridge this gap and offer valuable data for improved prognostic assessments in MM.

1.7 Research Questions

The exploration of predictive biomarkers in MM has been guided by many key concerns that aim to address significant elements of the disease. These questions aim to explore the predictive power of identified biomarkers, the role of phylogenetic analysis in proving these biomarkers and the relative effectiveness of various computational methods used in the study. The specific research questions are as followed:

- What novel prognostic biomarkers can be identified for multiple myeloma prognosis through comprehensive bioinformatics analysis?
- Do the identified biomarkers predict MM prognosis?
- Can phylogenetic analysis support the prediction of novel prognostic biomarkers in MM?

1.8 Aim and Objectives

Accurate identification of prognostic biomarkers play role in improving early diagnosis, risk stratification and targeted treatment. The increasing accessibility of transcriptomic data from public repositories presents a strong case for using advance bioinformatics methodologies to find and confirm reliable biomarkers. Significant challenges are presented by the inherent heterogeneity of MM and the lack of integrated multi-level analyses. Bridging these gaps requires a systematic approach that combines evolutionary insights with common bioinformatics approaches. The emphasis is placed on discovering biologically significant genes with high diagnostic and prognostic potential utilizing existing computational pipelines, thereby contributing to precision medicine for MM.

Aim:

To identify novel prognostic biomarkers for MM progression by conducting comprehensive bioinformatics analysis and confirm the clinical significance and application of discovered prognostic biomarkers. We achieved our aim by accomplishing the objectives listed below:

Objectives:

- To identify novel prognostic biomarker in MM by performing comprehensive bioinformatics analysis
- To perform statistical and survival analyses to evaluate the prognostic potential of the identified biomarkers
- To investigate the role of phylogenetic approach in predicting novel biomarkers
- To bridge current knowledge gap and provide valuable insights for improved prognostic assessments and therapeutics

1.9 Proposed Solution

An important aspect of our research involves proving hub genes identified through bioinformatics analyses to determine their evolutionary importance and possible role in MM pathogenesis. This investigation ultimately will contribute to personalized therapeutic approaches and bridge the existing gaps in the literature about relapse prediction customized treatment options for MM patients.

1.10 Proposed Methodology

This study uses a comprehensive multi-step methodology designed to find potential prognostic biomarkers, in order to better understand the processes and course of MM. Firstly, Differential Gene Expression analysis was performed in order to identify genes that show notable variations in expression between healthy and MM samples. Weighted Gene Co-Expression Network Analysis was used to find important modules associated with the progression of the disease. Essential hub genes to the etiology and development of MM are found using network analysis. The molecular pathways implicated are further clarified by enrichment and protein-protein interaction (PPI) analysis. ROC analysis, survival analysis and phylogenetic analysis was utilized to assess important hub genes. Lastly, this comprehensive approach which is described below seeks to enhance patient outcomes by filling in current gaps in the literature and laying the groundwork for future developments in the treatment of MM. A detail elaboration of this comprehensive approach is provided below.

The collection of reliable and comprehensive datasets is the foundation of any genomic analysis. For this study, gene expression data was obtained from publically accessible databases like the Gene Expression Omnibus (GEO) and ArrayExpress. These platforms are curated databases that host high-throughput gene expression data form various studies involving different diseases, conditions and tissue types [35]. The selection of datasets was based on the term Multiple

Myeloma that contains both disease and control samples, sufficient sample size and clear annotation. By ensuring the usage of standardized and peer-reviewed data, these databases offer a solid foundation for downstream bioinformatics analysis. Figure 1.1 shows the schematic diagram of data collection from GEO and ArrayExpress.

Afterwards, preprocessing of the data was done to ensure the quality and comparability. This step involved normalization, filtering of low quality or irrelevant probes and correction of batch effects. For this, NCBI's online tool GEO2R was utilized. GEO2R allows users to compare two or more groups of samples in a GEO series and find DEGs through Limma package in R [36]. Quick yet reliable preprocessing is supported by its statistical tools, visual output and user-friendly interfaces. To reduce the noise and boost confidence in downstream analysis, it is crucial to ensure data quality at this stage.

DEGs represent those genes that exhibit a substantial difference in expression levels between samples and healthy controls. Identification of DEGs aids in pinpointing key players in the onset and progression of disease. Once again, GEO2R was used for DEG analysis by filtering statistically significant genes using log fold-change and adjusted p-value thresholds. This approach guarantees that only the most significant changes in gene expression are recorded. DEGs serve as the foundation for later enrichment, network and biomarker identification analyses.

The Database for Annotation, Visualization and Integrated Discovery (DAVID) was used to conduct functional annotation and pathway enrichment analyses in order to assess the biological significance of DEGs. DAVID offers extensive tools for Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway mapping and Gene Ontology (GO) term analysis (biological processes, molecular functions, cellular components) [37]. These enrichments suggest potential therapeutic targets and diseases mechanisms by revealing the biological roles, molecular activities and pathways most impacted in MM.

A system biology method called Weighted Gene Co-Expression Network Analysis (WGCNA) was used to describe the patterns of gene correlation across microarray samples. It ensures the possibility to identify gene modules that are highly associated and relates these modules to external samples traits such as clinical symptoms. The current study utilized this method to identify gene modules linked to multiple myeloma which shed light on the intricate molecular networks that underlie the disease [38]. WGCNA was employed due to its ability to

simplify data and identify functionally significant gene sets. Thus, this approach facilitates the identification of possible targets and biomarkers with biological significance.

Using the STRING database, a protein-protein interaction (PPI) network was built in order to obtain insight into the protein level interactions of DEGs. STRING compiles known and predicted interactions derived from experimental data, computational predictions and literature [75]. This network offers an analytical and visual framework for comprehending biological processes and molecular interactions. To ensure reliability in downstream hub gene identification, interactions with high confidence scores were selected. The PPI network built with STRING is shown in Figure 1.2.

In the PPI network, hub genes are the nodes with the high connection and potentially the most influential. This network was examined using Cytoscape software, a power tool for network visualization and analysis [39]. These core genes were found using key algorithms with cytoscape like Degree, Density of Maximum Neighborhood Component (DMNC), Edge Percolated Component (EPC), Maximal Clique Centrality (MCC) and Maximum Neighborhood Component (MNC). Each algorithm provides a different perspective on node centrality, ensuring comprehensive hub gene identification procedure.

By the combination of DEGs, WGCNA modules and hub genes, the identification of crossover candidate genes becomes easier. These are the crucial genes that have central role in pathophysiology of MM and are the excellent candidates for further research and therapeutic targeting. The biological relevance and reliability of certain biomarkers are improved by this integrative approach.

Phylogenetic analysis was used to evaluate the functional and sequence similarity among hub genes. It provides insights about their evolutionary conservation and possible roles that are shared across different species. ClustalW, a widely used tool for aligning proteins or nucleotide sequences was utilized to perform multiple sequence alignments (MSA) [40]. MEGA 11 (Molecular Evolutionary Genetics Analysis) offers tools for sequence alignment, model selection and tree visualization was used to generate phylogenetic tree from the alignments [41]. This analysis aids in placing gene function with the context of evolutionary framework. In Figure 1.3, (a) and (b) illustrates the workflow of ClustalW and MEGA11 used for sequence alignment and phylogenetic tree construction.

A detailed literature and database-based validation was conducted to further confirm the relevance of genes to MM progression and pathogenesis. To make sure that every candidate gene is associated to disease biology, expression patterns and clinical importance, several specialized databases were reviewed. Online Mendelian Inheritance in Man (OMIM) is a comprehensive and reliable database of human genes and genetic traits. It provides comprehensive reviews of genetic findings and their clinical implication with a primary focus on hereditary disorders and gene-disease relationships. GeneCards database integrates gene-centric information from more than 150 sources, such as genomic, transcriptomic, proteomic, genetic and functional data. It offers a comprehensive gene profile that covers gene function, diseases, pathways and tissue-specific expression patterns. DisGeNET gathers information on gene-disease connections from variety of sources such as scientific publications and databases that have been carefully selected by experts by providing useful score and evidence levels for each association. It facilitates the systematic investigation of the genetic basis of human diseases. For manual literature mining, PubMed was used. It was utilized to cross-reference each gene with studies directly connected to MM. it contains millions of peer-reviewed biomedical papers. The Human Protein Atlas offers comprehensive protein expression profiles in both healthy and malignant tissues that is supported by transcriptomic data. It helps in confirming if there is differential expression of potential genes in tissues relevant to MM. Open Targets uses publically accessible datasets to link genes, pathways and diseases. Gene expression, molecular interactions and clinical trial relevance are among the integrated evidence that it uses to assess potential genes as a therapeutic target. A comprehensive database of somatic mutations linked to human cancer, the Catalogue of Somatic Mutations in Cancer (COSMIC). It was employed to determine whether hematologic cancers such as MM had known mutation patterns in the hub genes. IntOGen finds driver genes and mutations implicated in carcinogenesis by combining data from cancer genome sequencing projects. It aided in determining of genes that are known to be involved in MM or other types of cancers. Using these databases makes it possible to verify the gene's biological significance and expression in MM while providing a more comprehensive understanding of their potential as therapeutic target and prognostic biomarkers.

ROC curve analysis was performed using pROC package in R to assess the diagnostic capability of hub genes. ROC analysis measures the sensitivity and specificity of a gene in distinguishing MM from normal samples. Excellent diagnostic performance is indicated by an area

under the curve (AUC) near to 1 [42]. To evaluate the biomarker potential of candidate genes in clinical applications, this approach is essential.

A vital statistical analysis was used that assesses time-to-event data. Survival analysis usually focuses on patient survival times in relation to specific clinical or genetic traits. In the context of cancer research, it aids in determining whether differences in patient survival outcomes are correlated to the expression levels of particular genes. Prognostic biomarker or genes whose expression can forecast the course or outcome of the disease can be identified using this analysis. The KMplotter online tool, which combines gene expression data with clinical survival data from several publically accessible datasets was utilized in this study for survival analysis [43]. Kaplan-Meier survival curves are generated in KMplotter to compare the overall survival (OS) of groups with high and low expression levels. Hazard ratios (HRs) provides the risk estimates, while log-rank test evaluated the statistical significance. This allows to rank the candidate genes and prioritize those most closely associated to patient outcomes in MM by assessing their prognostic ability.

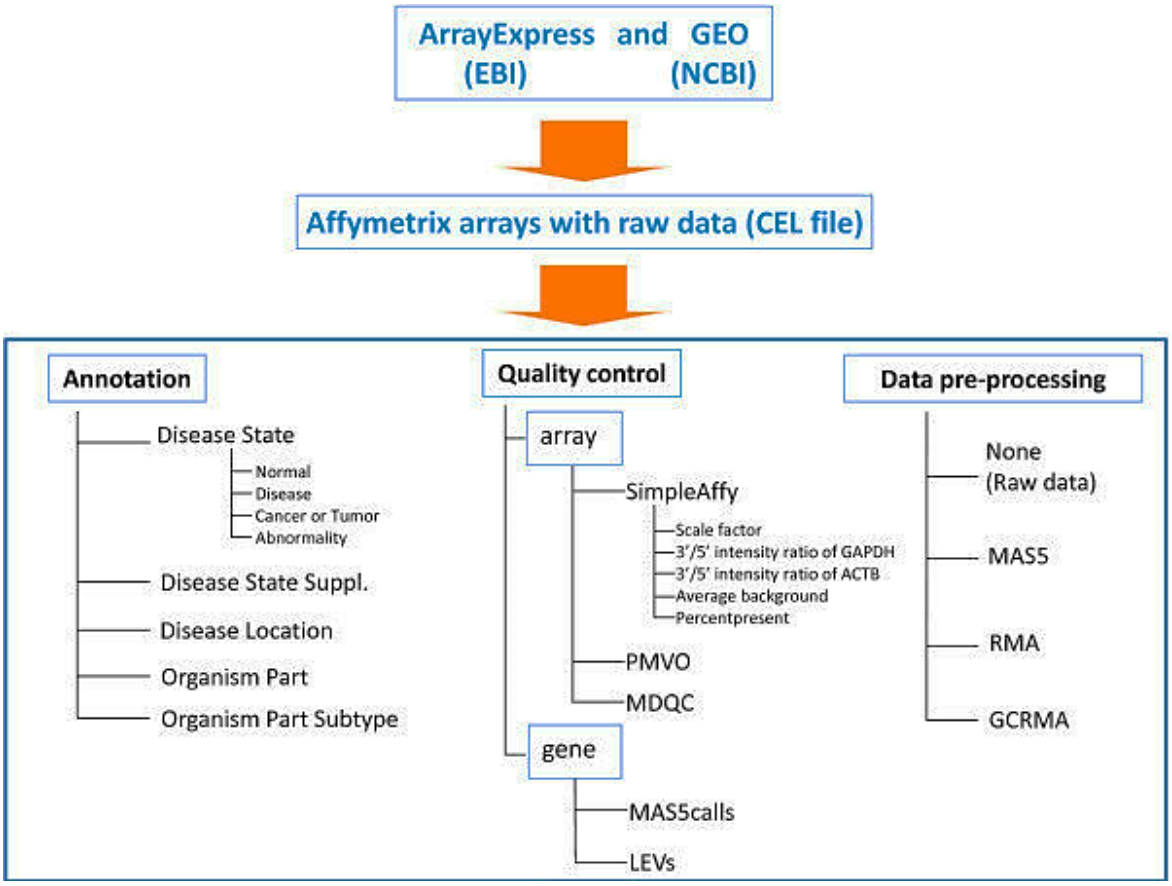


Figure 1.1. The schematic diagram of data collection from GEO and ArrayExpress [44].

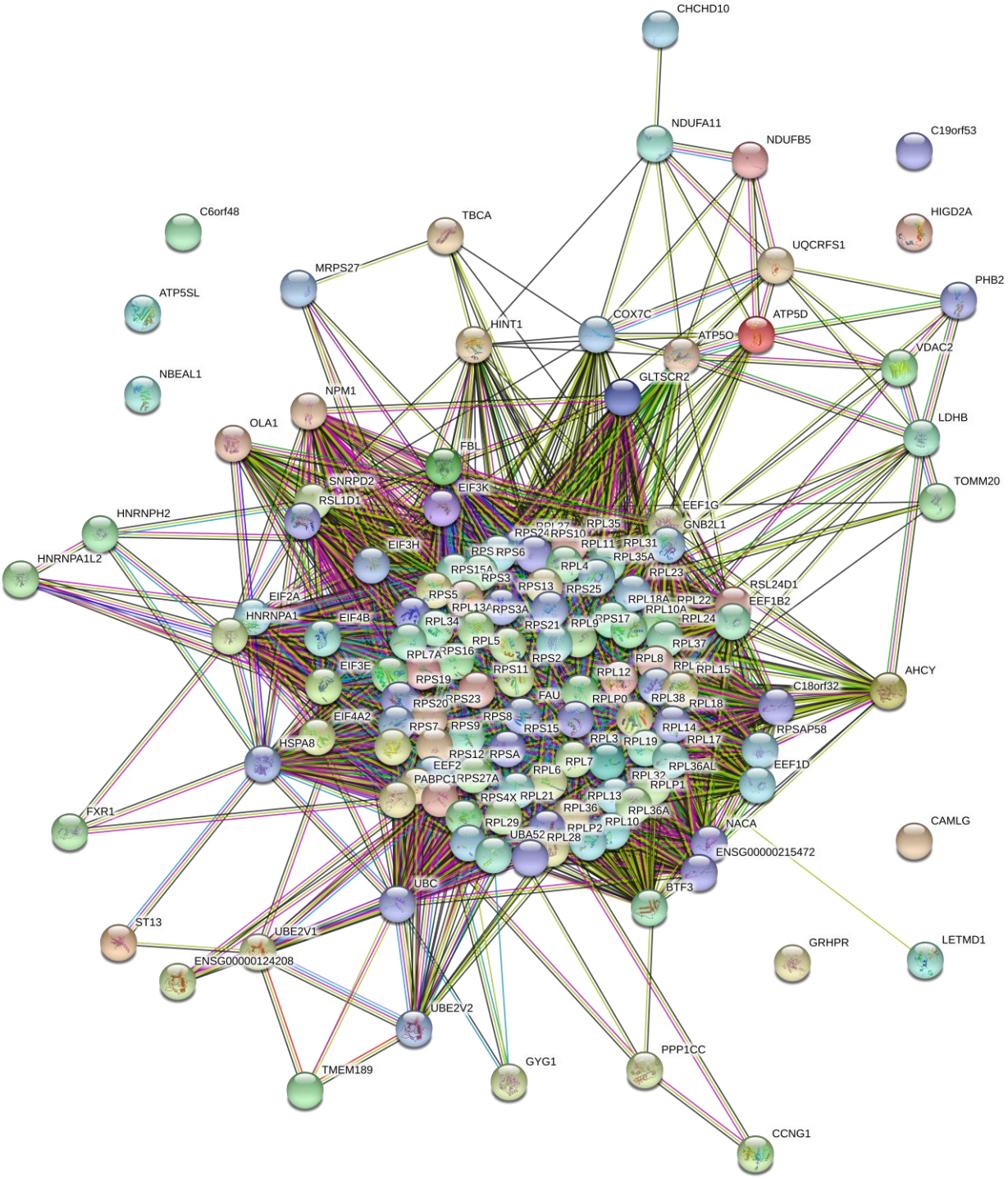
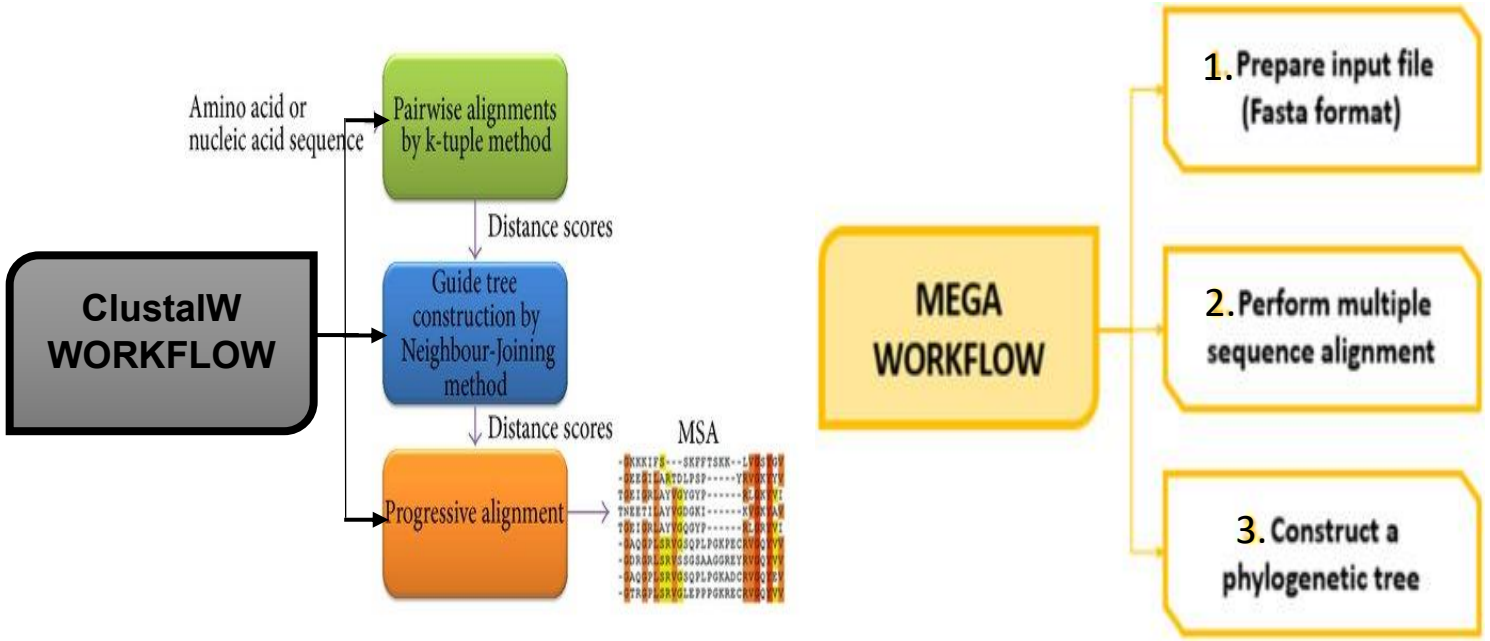


Figure 1.2. The Protein-Protein Interaction network built with STRING.



(A) (B)

Figure 1.3. (A) and (B) illustrates the workflow of ClustalW and MEGA11 used for sequence alignment and phylogenetic tree construction [1], [2].

1.11 Scope and Limitations

The discovery of new prognostic biomarkers that can help with early relapse diagnosis and guide effective treatment plans for advanced stages of MM could significantly change the way of disease management. Moreover, finding important regulatory networks and using phylogenetic method for the hub gene validation, especially linked to disease progression could help us better understand the molecular mechanism of MM. This would allow clinicians to personalize treatment based on the molecular profile of each patient. The development of personalized therapies that target specific genetic alteration could improve patient's outcomes and quality of life with the help of reliable biomarkers.

Furthermore, our results have application in clinical settings, where use of customized treatment approaches could significantly affect patient survival and treatment planning. This research may help identify potential treatment targets and gene-drug interactions for drug development efforts aimed at improving response rates and overcoming resistance in MM patients. To a significant extent, the work supports the continuous endeavours in cancer research to create precision medicine strategies that minimize side effects while simultaneously improving treatment effectiveness, thereby promoting a more patient-centered healthcare paradigm.

However, our study may have a number of limitations. First, despite the fact that samples from the various datasets were rigorously quality controlled, it is indisputable that the impacts of inherent variability might not have been totally eradicated. The in-silico nature of the study necessitates additional experimental validation in laboratory or clinical settings to validate the therapeutic applications of the biomarkers. The complexity of drug resistance or off-target effects required thorough experimental trials to assess treatment-biomarker interactions and their effectiveness. Despite these limitations, this study presents insightful information about MM and lays a solid basis for further investigation.

Chapter 2

Literature Review

2 Literature Review

The literature review offers a thorough overview of prior research related to the present study. It synthesized key findings, methodological frameworks and methods of previous studies that facilitate the identification and prediction of prognostic biomarkers for MM

2.1 Introduction to the Literature on Multiple Myeloma

This review seeks to understand the existing and emerging prognostic biomarkers in MM regarding their contribution to risk-based classification of patients and personalized treatment modalities. While prior research has largely depended on DEGs analysis, WGCNA and related approaches for biomarker identification. However, our approach will take a different track by integrating data from combination of databases and new method such as phylogenetic analysis. We aim to uncover new biomarkers while confirming the predictive value of existing ones. We hope to contribute to expand the literature of research on MM risk assessment, ultimately improving clinical decision making and patient outcomes.

MM is a heterogeneous malignancy of plasma cells with considerable differences in clinical behavior among patients. It includes variation in symptom onset, therapeutic response and long-term survival outcomes. Although treatment options have expanded in recent years with newer drugs like proteasome inhibitors and immunomodulatory agents with improving outcomes. However, the disease remains incurable for most patients. A majority of patients eventually relapse or stop responding to therapies [45], [46]. Early and precise risk stratification is critical, as high risk cases often show significantly worse survival despite aggressive treatments [14].

The historical development of MM research has progressed through a series of significant milestones that have shaped current understanding of the disease. The first clinical case was documented by Samuel Solly in 1844, followed by the introduction of the term “multiple myeloma” by von Rustizky in 1873. A key diagnostic breakthrough came in 1845 when Henry Bence Jones identified a unique urinary protein; later names Bence Jones protein. It is now recognized as a hallmark of MM [47]. Advance in protein electrophoresis and immunoelectrophoresis throughout the 20th century greatly enhanced the detection of monoclonal proteins, contributing to improved diagnostic accuracy. The identification of MGUS further enriched understanding of MM progression, recognizing it as a precursor state. In terms of staging, the development of the Durie-Salmon System in 1975 provided a structured clinical framework,

which was later refined into International Staging System (ISS) and the Revised ISS (R-ISS). This system incorporated biomarkers such as β2-microglobulin, albumin, cytogenetic abnormalities and LDH levels for more precise risk stratification. The discovery of MGUS further enhanced the MM progression understanding, attributing it to a precursor stage. Staging wise, the introduction of the Durie-Salmon System in 1975 offered a systematic clinical approach. It eventually classified into ISS and revised ISS. The system included biomarkers like β2-micorglobulin, albumin, cytogenetic aberration and LDH levels for accurate risk stratification [76]. Recently, genomic advances have enabled researchers to probe the evolutionary trajectory of MM. Techniques such as NGS have shed light on the accumulation of key driver mutations and their effects on the different stages of disease progression form precursor conditions like MGUS and SMM to over MM [48]. Studies revealed that MM arises not as malignancy driven by single event, rather due to clonal evolution where early and late mutational events shape the disease phenotype and treatment response. Thus, the historical timeline of MM research reflects a gradual transition from clinical observation and refinement of diagnosis to a deep molecular dissection of the biology of the disease. It indicated the complexity of MM and its associated clinical challenges.

2.2 Pathogenesis and Molecular Mechanisms of Multiple Myeloma

MM is derived from B cells, typically originating from the germinal center where malignant transformation occurs during somatic hypermutation and antibody isotype switching. These are frequently mediated by activation-induced cytidine deaminase (AID). Progressing usually through identifiable precursor stages, MGUS followed by SMM and leading to clinically active condition. The first step along the progression usually involves initial genetic events such as chromosomal abnormalities (t(14;16), t(4;14), translocation or hyperdiploidy). It is then transformed into secondary eventualities such as deletions in chromosome 17p, amplifications of 1q and mutations in critical genes such as KRAS, NRAS and TP53. These genetic alterations trigger several crucial cellular pathways particularly NF-κB, MAPK, JAK/STAT and PI3K/Akt cascades. These pathways drive tumor cell growth, enhance survival mechanisms and confer resistance to therapies. The bone marrow niche plays a supportive role through multiple mechanisms such as cytokine release stimulation of new blood vessel formation. MM also disrupts normal bone remodeling by tipping the balance between bone destruction (via elevated RANKL and MIP-1 α stimulating osteoclasts) and bone formation (via inhibition of osteoblasts by DKK1 and other Wnt pathway blockers). Finally, resulting in the distinctive lytic bone lesions seen in patients [12], [49].

2.3 Clinical Characteristics and Advancements in Therapeutics of Multiple Myeloma

MM presents varied clinical features, most notably the classic CRAB symptoms; hypercalcemia, renal impairment, anemia and bone lesions [14], [22]. Diagnosis requires meeting IMWG criteria which includes either ≥ 10% clonal plasma cells in bone marrow or biopsy-proven plasmacytoma along with CRAB features or specific high-risk markers like serum free light chain ratio exceeding 100 or multiple MRI detected bone lesions [17]. Despite advancements in imaging, diagnostic delays remain a significant issue [50]. Treatment has evolved from conventional chemotherapy to include targeted approaches such as proteasome inhibitors (bortezomib, carfilzomib), immunomodulators (lenalidomide), monoclonal antibodies (daratumumab) and breakthrough BCMA-directed CAR-T cell therapies. These all contributing to better survival outcomes [21], [51]. However, challenges remain such as treatment resistance, extramedullary progression and high-risk genetic abnormalities like del (17p) and t(4;14). A new therapeutic direction focuses on MYC inhibition and ribosome-targeting agents (CX-5461). It promises better outcomes in refractory cases [52].

2.4 Integrative Bioinformatics Approaches for Biomarker Discovery in Multiple Myeloma

Bioinformatics has become indispensable in the identification of prognostic biomarkers for MM, allowing research to integrate large-scale gene expression data with clinical parameters. A wide range of computational approaches have been employed including differential gene expression analysis, WGCNA, PPI networks and functional enrichment methods such as Gene Ontology (GO) and Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway analysis. Studies using GE) database have consistently identified differentially expressed genes (DEGs) with potential clinical value [53], [54]. For instance, WGCNA was used for the identification of FCER1G as a hub gene significantly associated with MM pathogenesis [55] While SSBP1 was found to have strong prognostic significance [4]. In another analysis, PCNA and CDC7 were highlighted as promising prognostic markers in high-risk MM though DEG profiling of the GSE87900 dataset. [56] identified 1,759 DEGs across multiple microarray datasets and uncover key genes such as RPN1, SEC61A1 and SPCS1 through WGCNA reinforcing their diagnostic potential. Similarly, [57] used WGCNA to correlate gene modules with clinical traits, identifying FGFR3, MMP9 and SNORA71A as potential biomarkers for early detection and targeted therapy. Integrating GO and KEGG analyses, [54] discovered hub genes including FH, TSTA3 and POLR3G, whose predictive

power was further supported by ROC curve analyses. [53] combined bioinformatics with experimental validation to confirm RRM2 as a survival-associated biomarker that demonstrates the value of integrating in-silico and clinical approaches. [58] focused on cell-adhesion related DEGs and identified ITGA9 and LAMB1 as a novel biomarkers linked to disease stage and β2-microglobulin levels. It provides insights into MM progression via bone marrow microenvironment. Collectively, these studies exemplify the power of integrative bioinformatics in uncovering molecular mechanisms, predicting clinical outcomes and proposing new therapeutic targets in MM.

2.5 Comparative Review: Traditional Biomarker Methods vs. Phylogenetic Approaches in Biomarker Prediction

Traditional biomarker techniques in MM like serum protein electrophoresis, β2-microglobulin measurements and fluorescence in situ hybridization (FISH) have long been used as markers for risk assessment, diagnosis and disease monitoring. These IMWG supported clinically validated tools provide standardized and easily available evaluations of cytogenetic risk and disease burden [17], [21]. However, these methods frequently miss the temporal and spatial variability of developing MM clones and only offer a static picture of tumor genetics. In contrast, phylogenetic approaches utilize whole-exome sequencing and single-cell analysis to reconstruct the clone architecture and evolutionary history of MM. Studies have demonstrated that mutations in genes like KRAS, NRAS, BRAF and TP53 are often found in minor subclones at diagnosis and supporting a branching evolution model rather than a linear one [59]. Phylogenetic analysis also enables detection of genome-wide duplication, deletion events and subclonal dynamics that are critical for early relapse prediction and therapy resistance. Importantly, phylogenetic-guided biomarker discovery has revealed immune related candidates such MPO which negatively correlated with CD8+ T-cell infiltration and has been casually linked to MM risk through Mendelian randomization [73]. While traditional biomarkers remain valuable due to their clinical practicality and widespread use and phylogenetic methods offer a dynamic, real-time perspective of tumor evolution. This enables personalized therapeutic strategies especially in the context of relapse, resistance and emerging immunotherapies such as CAR-T cells and bispecific antibodies.

2.6 Gaps, Challenges & Limitations in Current Multiple Myeloma Biomarker Literature

Despite the rapid advancement in molecular profiling technologies and the increasing availability of patient data, biomarker discovery in Multiple Myeloma (MM) still faces several critical scientific, technical, and translational challenges that limit its clinical utility. While numerous candidate biomarkers have been proposed based on genomic and transcriptomic studies, only a few have progressed to clinical adoption. This slow translation is due to both methodological limitations and biological complexities intrinsic to MM.

The major gap lies in the underutilization of phylogenetic and evolutionary analyses in biomarker research. Traditional biomarker frameworks often rely on static genomic or transcriptomic snapshots, which do not adequately account for the temporal and spatial clonal evolution characteristic of MM. Phylogenetic methods, which model how cancer clones evolve and diversify over time, remain largely absent in most biomarker studies. This omission is critical, as evolutionary-informed biomarkers may better capture subclonal heterogeneity and identify emerging resistant clones particularly in cases involving branching evolution, where minor subclones later drive relapse or treatment resistance [59]. Without such a dynamic context, prognostic biomarkers may lose relevance across disease stages or therapeutic interventions.

In parallel, the integration of multi-omics data remains a significant challenge. Although genomic, transcriptomic, and proteomic datasets are increasingly available. Although, most MM studies continue to rely on single-omics approaches, often overlooking epigenetic, metabolomic, and post-translational modifications that play critical roles in regulating gene expression and protein function [22], [58]. This lack of integration limits the depth of biomarker discovery and hinders our ability to identify functionally relevant targets that may arise only at the interface of multiple molecular layers. Furthermore, the computational and statistical frameworks needed to harmonize and analyze such complex, high-dimensional data are still underdeveloped in MM research.

Sample diversity and cohort design pose are additional limitations. Many biomarker studies are conducted in relatively small, homogeneous patient populations, often lacking representation from diverse ethnic backgrounds, age groups, or disease stages [17]. This restricts the generalizability of findings and risks producing biomarkers that may not perform consistently across broader, real-world populations. Moreover, studies tend to focus on baseline samples collected at diagnosis or initial treatment, which do not reflect the evolving nature of MM. The

disease is marked by both spatial heterogeneity with distinct subclonal populations residing in different bone marrow niches and temporal heterogeneity, where tumor composition changes in response to therapy. For example, the emergence of BCMA-negative subclones following CAR T-cell therapy has been documented, undermining the utility of BCMA as a reliable long-term biomarker [51].

Another gap lies in the limited functional annotation of non-coding mutations, structural variants, and copy number alterations. While these genomic events are increasingly detected in MM through whole-genome and exome sequencing, their biological impact is often unclear, particularly for alterations located in enhancer regions, long non-coding RNAs (lncRNAs), or structural rearrangements [60]. Without mechanistic insights, it is difficult to prioritize these findings for translational development, leaving a large portion of the mutational landscape unexplored.

Finally, translational and implementation barriers significantly hinder the clinical adoption of proposed biomarkers. Issues such as lack of assay standardization, cost constraints, regulatory hurdles, and the absence of prospective validation in large-scale clinical trials slow down the pathway from discovery to practice. Even well-studied biomarkers like TP53 deletion, which is associated with high-risk MM, or BCMA, a target for immunotherapy, face challenges related to assay reproducibility and real-time monitoring in diverse clinical settings [21]. To address these multifaceted limitations, future research must embrace longitudinal cohort designs, deploy multi-omics integration pipelines, and incorporate evolutionary and spatial modeling to account for MM's dynamic nature. In addition, there is a critical need for functional validation, using CRISPR-based gene editing or patient-derived xenografts, to confirm the biological relevance of candidate biomarkers. By expanding cohort diversity and investing in translational frameworks, the field can move toward more adaptive, personalized, and clinically actionable biomarker strategies in Multiple Myeloma.

2.7 Critical Analysis of the Literature Review

The reviewed literature highlights significant progress in biomarker discovery for MM particularly through transcriptomic, genomic and bioinformatics driven approaches. However, it also reveals critical gaps directly relevant to this study's focus on phylogenetic methods. While numerous studies have identified hub genes and DEGs with prognostic relevance using techniques such as

WGCNA, DEG analysis and enrichment mapping but these methods often lack evolutionary context. Traditional biomarkers remain clinically valuable but are limited by their inability to capture subclonal heterogeneity and dynamic tumor evolution. Importantly, current literature underscores a lack of integration between phylogenetic insights and conventional biomarker panels despite evidence that clonal evolution patters significantly impact prognosis and therapy resistance. Moreover, the scarcity of multi-omics integration, limited validation across diverse cohorts and inadequate longitudinal tracking of biomarkers further constrain clinical applicability. This analysis supports the need for a phylogenetically informed biomarker discovery framework in MM that accounts for clonal dynamics, enhance risk stratification, and aligns with emerging precision medicine paradigms.

In conclusion, the investigation of prognostic biomarker in MM using bioinformatics approaches has revealed important new information about the molecular underpinning of the disease. As summarized in Table 2.1, the discovery of important biomarkers like SSBP1, FCER1G, FGFR3, MMP9, FH, TSTA3, POLR3G, ITGA9, LAMB1 and those linked to cell adhesion underscores the potential of these molecular markers to improve patient stratification and guide individualized treatment plans. Further, the integration of several analytical methods like DEG, WGCNA and pathway enrichment analyses highlights the importance of a multi-pronged strategy in the development of biomarkers. As the field advances, the findings of these studies pave the way for additional research aimed at improving treatment outcomes and prognosis accuracy for MM patients, ultimately enhancing precision medicine in oncology. Further study is still needed to validate the identified biomarkers, examine their functional significance and assess their therapeutic utility in larger and carefully organized trails. Our proposed study intends to bridge these gaps and provide a comprehensive assessment of the prognostic significance and potential therapeutic implications of the identified biomarkers in MM by combining data from multiple databases, using bioinformatics techniques like phylogenetic analysis and carrying out computational validation.

Table 2.1: Critical analysis of previous research on biomarker identification for MM.

Ref	Objectives	Datasets	Methodology	Results	Limitations
[4]	Identification of indicative biomarkers of MM or significant co-expressed gene clusters	GSE6477 GSE24080 GSE99636	•WGCNA •GO & KEGG •PPI (cytoscape) •GSEA •Statistical Analysis •Animal Studies •In Vitro & Vivo Studies	SSBP1 appeared promising candidate biomarker that shows involvement in MAPK pathway.	Lack of phylogenetic analysis.
[3]	Identifying and characterizing key genes associated and molecular pathways with MM	GSE39754 GSE13591 GSE2658	•DEGs •WGCNA •PPI (cytoscape) •ROC Analysis •GO & KEGG •GSEA	FCER1G; a promising biomarker for MM patients	Experimental validation needed. Lack of phylogenetic analysis.
[56]	To explore aetiological explanations and identify biological targets, specifically hub genes in MM	GSE125364 GSE39754 GSE13591	•DEGs •GO & KEGG •PPI •WGCNA •ROC Analysis •GSEA	RPN1, SEC61A1, SPCS1, SRPR, SRPRB, SSR1 & TRAM1; were highly expressed and representative biomarkers for MM diagnosis.	Experimental validation needed. Lack of phylogenetic analysis.
[57]	Exploring the newly tested biomarkers in multiple myeloma using (WGCNA)	GSE47552	•DEGs •WGCNA •Significant Modules Identification •Hub Genes Identification •GO & KEGG •ROC & AUC •qRT-PCR exp	SNORA71A; a novel biomarker for early detection and a possible therapeutic target.	Lack of phylogenetic analysis.

			•Statistical Analysis		
[54]	To identify and characterize the mechanisms and hub genes associated with MM	GSE80140 GSE80545	•DEGs •GO & KEGG •PPI •Survival Analysis	FH, TSTA3, and POLR3G as hub genes & new biomarkers for diagnosis and prognosis of MM.	Experimental validation needed. Lack of phylogenetic analysis. WGCNA analysis missing.
[53]	To investigate novel treatment targets and biomarkers for diagnosis and prognosis in MM.	GSE6477 GSE13591	•DEGs •GO & KEGG •PPI •Module Analysis •GSEA •Cell Viability Assay •Western Blot •Cell Cycle Analysis •Statistical Analysis	RRM2; a diagnostic biomarker for MM	Lack of phylogenetic analysis. WGCNA analysis missing.
[61]	To determine significant genes involved in cell adhesion in MM.	GSE6477 GSE2658 GSE136324	•DEGs •KO & KEGG •PPI •ROC Analysis •Survival Analysis •GSEA	ITGAM, ITGB2, ITGA5, ITGB5, CDH1, IL4, ITGA9, LAMB1; potential diagnostic and prognostic markers in MM.	Experimental validation needed. Lack of phylogenetic analysis. WGCNA analysis missing.

Chapter 3
Materials and Methods

3 Materials and Methods

3.1 Study Context & Design

This chapter outlines the methodology followed in this study to identify prognostic biomarkers for MM. It provides description about the platforms used for datasets collection, software and tools used for carrying out the methodology to investigate and identify the prognostic biomarker for MM. The methodology is designed to provide a comprehensive assessment of gene expression patterns and their clinical relevance, ultimately aiming to enhance patient stratification and inform personalized treatment strategies. Figure 3.1 shows a detailed methodology flowchart that present the sequential steps of data collection, analysis and biomarker identification.

3.1.1 Data Description

The literature reviewed includes the identification of biomarkers for MM and bone marrow cancer. The microarray datasets are used containing gene expression data of healthy and diseased individuals. The gene expression datasets were retrieved from NCBI GEO and ArrayExpress database based on usage of human samples and relevant disease stages and due to their high-quality, publically available data relevant to MM.

3.1.2 Significance of the Study

This study has the potential to fundamentally alter the management of MM by identifying novel prognostic biomarkers that can aid in the early detection of relapses and directs the efficient treatment regimens for advanced stages. This may further improve the understanding of molecular mechanisms underlying MM by discovering significant regulatory networks and utilizing a phylogenetic approach to confirm hub genes connected to disease progression. This would enable clinicians to tailor treatments according to each patient's molecular profile. The development of personalized therapies that target specific genetic alteration which could improve patient's outcomes and quality of life with the help of reliable biomarkers.

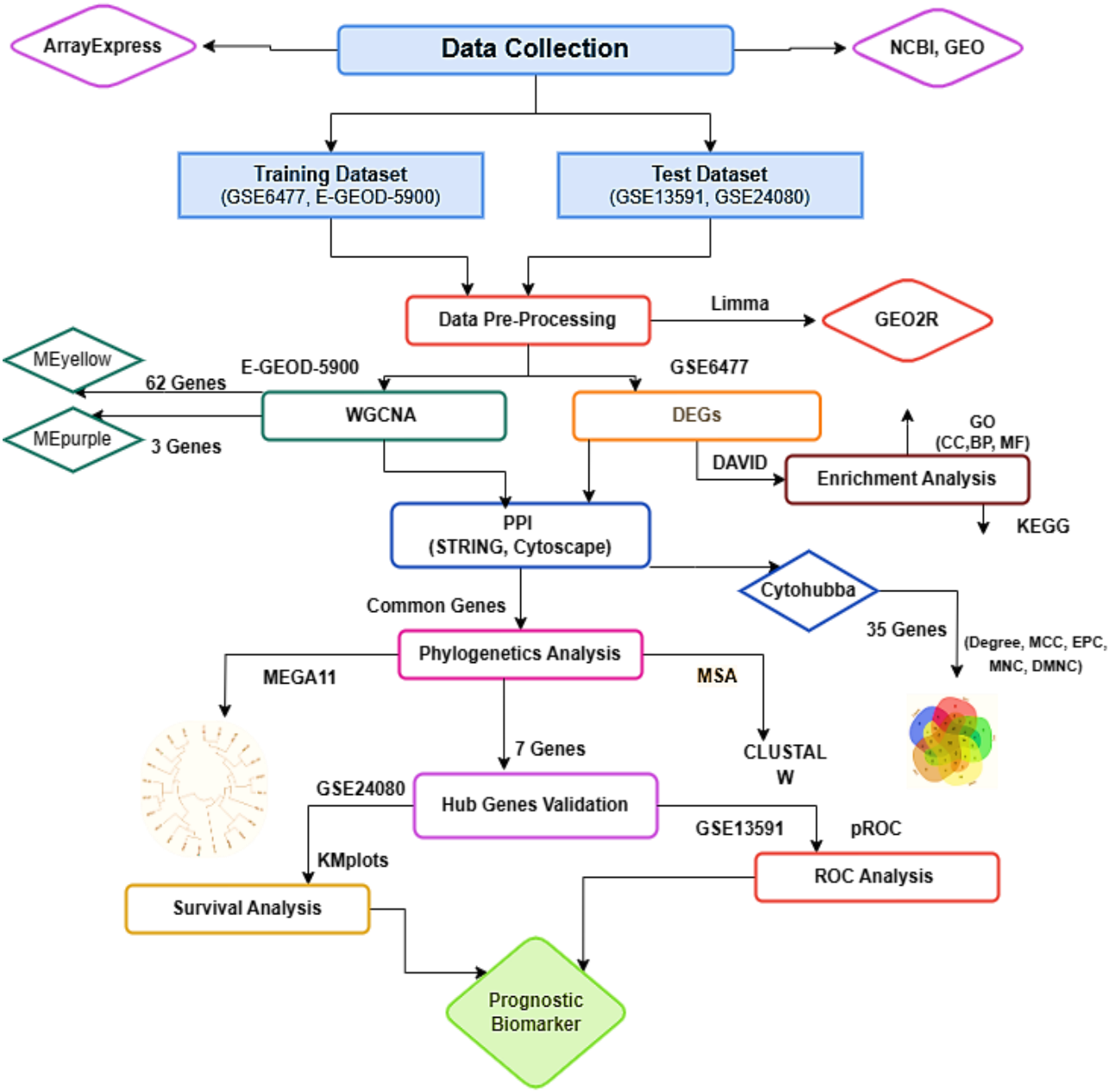


Figure 3.1: Illustration of sequential steps involved in the research methodology of prognostic biomarker identification of MM.

3.1.3 Conceptualization of Work

Conceptualization includes the specific dimensions and operational aspects of sample study. Present study provides the biomarker discovery design by using bioinformatics tools and soft

wares against multiple myeloma. This concept will provide the route map and will act as a blue print in identification of potential biomarker for manufacturing personalized drugs and therapeutics for multiple myeloma.

3.2 Methodology

3.2.1 Data Collection

The first step of methodology involved collection of relevant gene expression microarray datasets from publically available databases. The datasets of Multiple Myeloma were downloaded from GEO(https://www.ncbi.nlm.nih.gov/geo/)andArrayExpress(https://www.ebi.ac.uk/biostudies/arrayexpress) database by using the GEOquery package of R version(4.4.0) mentioned in Table 3.1. The selected dataset contains both diseased and healthy controls data, with the platform type limited to "microarray". In this study four datasets are used; two training (GSE6477 and E-GEOD-5900) datasets and two test (GSE13591 and GSE24080) datasets. The total number of samples in GSE6477 dataset are 162, 15 control and 147 diseased. While in E-GEOD-5900 dataset, the total number of samples are 78, 22 control and 56 diseased. This targeted approach ensured the inclusion of appropriate comparisons for robust downstream analyses.

Table 3.1: Summary of Gene Expression Microarray Datasets for Multiple Myeloma

Dataset	Organism	Experiment Type	Disease	Platform	Sample
GSE6477	Homo Sapiens	Expression profiling by array	MM	Affymetrix Human Genome U133A Array	162
E-GEOD-5900	Homo Sapiens	Expression profiling by array	MM	Affymetrix Human Genome U133 Plus 2.0 Array	78
GSE13591	Homo Sapiens	Expression profiling by array	MM	Affymetrix Human Genome U133A Array	158
GSE24080	Homo Sapiens	Expression profiling by array	MM	Affymetrix Human Genome U133 Plus 2.0 Array	559

3.2.2 Data Preprocessing

Data preprocessing and normalization of the microarray data was achieved with the help of the 'limma' and 'affy' method in GEO2R analyzer to remove any low-quality samples and normalize the expression data for accurate downstream analyses. The Robust Multiarray Average (RMA)

method was used to preprocess and analyze the data by executing background correction, quantile normalization and log transition. The selection of these packages was done on the basis of specific requirement of the research, including the nature of the disease, dataset characteristics, and analytical goals. Other tools were considered but did not align as effectively with the study's needs.

3.2.3 DEGs Identification

A differential gene expression analysis was performed through GEO2R online statistical tool. This analysis aimed to identify genes that express differentially at expression level between MM and healthy samples. The Benjamin and Hochberg adjustmet was utilized to modify the P-value. The cutoff criteria for investigating and selecting the differentially expressed genes (DEGs) was P value < 0.05 and log FC >1 [61]. Upregulated DEGs were considered if the logFC >=1 and Downregulated were considered if logFC <=1. LogFC values reveals the percentage of gene expression profiles. The R packages "ggplot2 (version 3.3.2)" and "pheatmap (version 0.7.7)" were used to visualize DEGs and display top 100 significant DEGs using volcano plots and heatmap in order to highlight the most significant DEGs.

3.2.4 Function and Pathway Enrichment Analysis

Function and pathway enrichment analysis was done using web tools such as Gene ontology GO (http://www.geneontology.org/) and Kyoto encyclopedia of genes and genomes KEGG (http://www.genome.jp/kegg/pathway.html) to examine the biological processes and pathways correlated with the identified DEGs. The Database for Annotation, Visualization and Integrated Discovery online tool (DAVID; https://david-d.ncifcrf.gov/)) was utilized to conduct enrichment studies for both GO and KEGG analysis. The most significant pathways were summarized using bubble plots generated by "ggplot2 (version 3.3.2)" in order to display the analysis results. P value < 0.05 and counts ≥ 2 were considered significant thresholds [37].

3.2.5 Weighted Gene Co- Expression Network Analysis

WGCNA was performed in order to identify hub genes and investigate the relationships between DEGs. The R package "WGCNA (version 1.6.9)" was utilized to perform weighted gene co-expression network analysis, which creates gene co-expression networks by clustering the genes that has similar expression patterns. These co-expression networks may demonstrate the relationship between gene modules and clinical phenotypes. Once the samples were clustered,

outliers were eliminated. For the scale-free co-expression network relationship, the soft threshold power $\beta = 17$ and $R^2 > 0.85$. In order to obtain large modules for the purpose of identifying hierarchical clustering genes, the least module scale was assigned 30. Sixteen Module Eigen vectors (MEs) were obtained by setting the default height cut to 0.25 ensuring an optimal balance between merging and preservation of distinct modules. Co-expressed gene modules were identified, and their relationship with clinical traits were analyzed. Furthermore, a heatmap displayed the relation between modules and clinical features in order to identify the most relevant modules to MM. Hub genes within these modules were chosen for their connectivity and significance, providing insights into potential key regulators of MM. Module screening cut-offs were correlation coefficient $|cor| > 0.5$ and P-value < 0.05 [4]. For further analysis, key modules with Module Membership $|MM| > 0.8$ and Gene Significance $|GS| > 0.2$ were selected [3].

3.2.6 Protein-Protein Interaction

To further investigate the interactions among the identified hub genes and DEGs, PPI analysis was performed. STRING (http://string-db.org) database was utilized to draw the PPI interactions of DEGs and hub gene modules. While, Cytoscape visualized the PPI networks (https://cytoscape.org/). This analysis aimed to detect key interactions and potential regulatory pathways involved in MM. In the candidate gene clustering, the interaction degree was identified using Cyto-Hubba plug-ins of Cytoscape based on five different methods [Degree, Density of Maximum Neighborhood Component (DMNC), Edge Percolated Component (EPC), Maximal Clique Centrality (MCC), and Maximum Neighborhood Component (MNC)]. The overlapped genes in DEGs and WGCNA modules were displayed using a Venn diagram created using web based application; Bioinformatics & Evolutionary Genomics (http://bioinformatics.psb.ugent.be/webtools/Venn/). The centrality and significance of key nodes in the PPI network were analyzed, providing insights into potential therapeutic targets. The key nodes in co-expression network and PPI network were considered to be the candidate genes.

3.2.7 Hub Gene Identification

This study used a multifaceted strategy to identify hub genes, utilizing the advantages of both evolutionary and computational analysis to identify important molecular players in the pathophysiology of multiple myeloma.

3.2.8 Crossover Candidate Genes

In order to obtain the crossover genes, the candidate genes from the PPI and WGCNA were intersected. The overlapping genes were displayed using Venn diagram.

3.2.9 Phylogenetic Analysis

Phylogenetic analysis was incorporated in order to assess the evolutionary relationships of identified hub genes. To construct the Phylogenetic tree, the genetic sequences of candidate genes were retrieved from NCBI, which enabled the examination of their evolutionary conservation and functional significance in MM. Subsequently, CLUSTAL W and MEGA 11 were used to align the genetic sequences to construct and visualize maximum likelihood phylogenetic tree. On the basis of sequence similarity and node distances, the most closed distant genes were selected from phylogenetic tree to facilitate further validation because these genes are likely to share functional relationships and may play critical roles in the biological pathways associated with MM. To increase the relevance of the validation efforts, literature survey for each gene is done by using Online Mendelian Inheritance in Man (OMIM) (https://www.omim.org/), GeneCards (https://www.genecards.org/), DisGeNET (https://www.disgenet.com/), PubMed (https://pubmed.ncbi.nlm.nih.gov/), The Human Protein Atlas (https://www.proteinatlas.org/), Open Targets (https://www.opentargets.org/), COmmon Software Measurement International Consortium (COSMIC) (https://cancer.sanger.ac.uk/census), and IntOGen (https://www.intogen.org/search) database. The selection of these databases was done on the basis of comprehensive, reliable and widely accepted information provided for gene-disease association, which is essential for validation the involvement of genes in MM progression.

3.2.10 Hub Gene Evaluation and Validation

To validate the prognostic significance of the identified hub genes, we used survival analysis and Receiver Operating Characteristic (ROC) analysis, two crucial analytical methods that together provide a comprehensive evaluation of the predictive capacity and clinical relevance of the identified hub genes, to confirm their prognostic value.

3.2.11 Receiver Operating Characteristic Analysis

We used survival analysis and Receiver Operating Characteristic (ROC) analysis that provides comprehensive evaluation of the predictive capacity, clinical relevance of the identified hub genes and confirm their prognostic value. In order to measure the accuracy of the hub genes serving as prognostic biomarkers, ROC analysis was implemented. The "pROC" package was used to determine sensitivity and specificity of hub genes in training and test datasets. Predictive capacity of the hub genes was evaluated using typical threshold of area under curve [AUC > 0.7]; indicating acceptable discrimination. An AUC near to 1 demonstrates good sensitivity and specificity of genes that distinguish between MM and normal samples [42], [62]. ROC plots were displayed using "ggplot2" package.

3.2.12 Survival Analysis

Survival analysis was carried out to establish the predictive value of hub genes using Kaplan-Meier Plotter (Kmplot, https://kmplot.com/analysis/) to plot the KM curves. Survival curves were compared based on high vs low levels of gene expression. In the analysis of overall survival (OS) and event-free survival (EFS) of MM patients Cox P value <0.05 was employed as statistically significant threshold [63]. It provides a holistic view of the efficacy of treatments and progression of disease that is crucial in refining treatments and patient outcomes. By evaluating OS and EFS, clinicians can identify high-risk patients requiring aggressive or alternative treatment combinations and understand the effectiveness of treatment over time.

Chapter 4

Results

4 Results

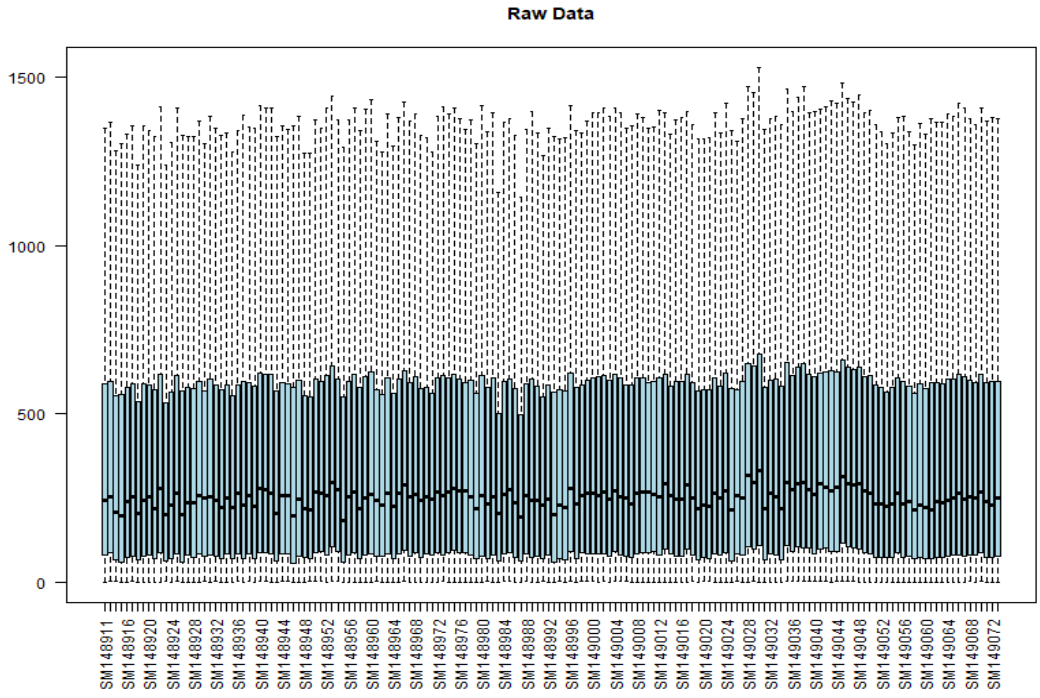
In this section, the results of the analyses performed has been presented. It shows and describes each and every step that have been performed for the identification of prognostic biomarker.

4.1 Data Collection

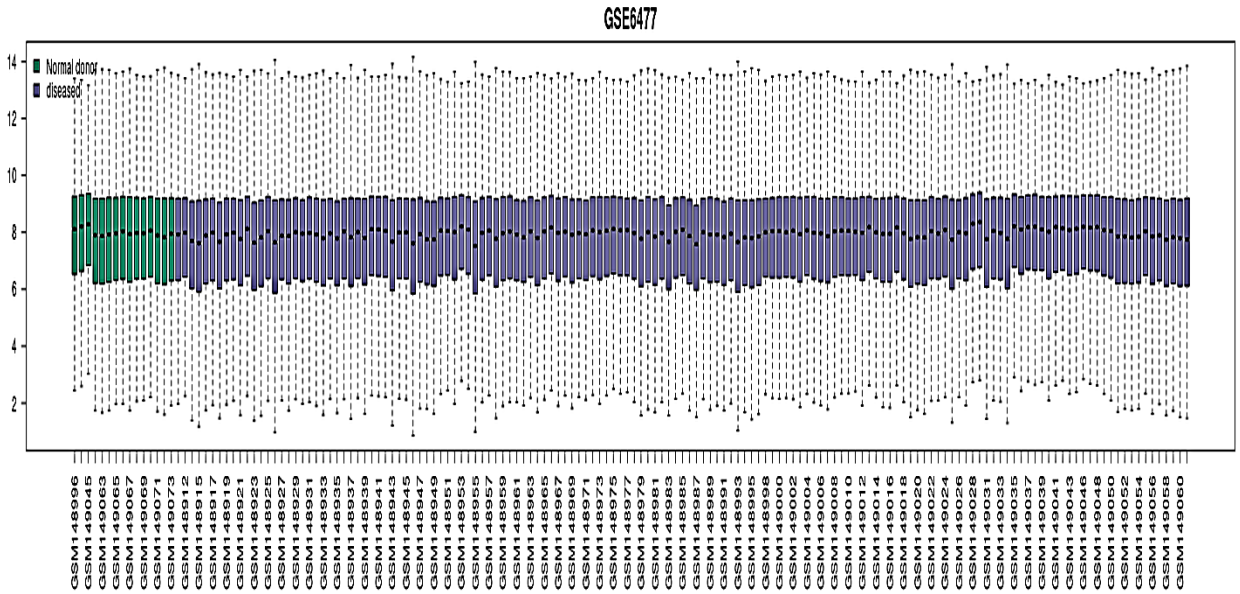
A particular keyword “multiple myeloma” was used to filter datasets according to the parameters such as the study type (expression profiling by array) and organism (Homo sapiens). Datasets were filtered using the specific keyword “multiple myeloma” and other parameters such as organism (Homo sapiens) and study type (gene expression profiling by array). This filtering resulted in a selection of datasets that were relevant to our research objectives. The datasets were downloaded in .CEL format. The GEOquery package was utilized to facilitate the retrieval of the raw data into RStudio. The files were thoroughly curated to ensure the compatibility of data with the study objectives.

4.2 Data Preprocessing

Following the data collection, preprocessing was carried out to prepare the datasets for analysis. The raw data was initially normalized using Robust Multi-array Average (RMA) method which is particularly effective for Affymetrix microarray data. The RMA approach consists of three primary steps: background correction, quantile normalization and summarization of probe-level data to gene-level expression values. GEO2R analyzer was used to process and normalize the raw data files. This revealed the overall distribution, central tendency, and variability of the expression levels of the relevant biomarkers We were able to verify that normalization successfully decreased skewness and enhanced data consistency across samples by looking at boxplots in Figure 4.1.



(A) Raw expression data



(B) Normalized expression data

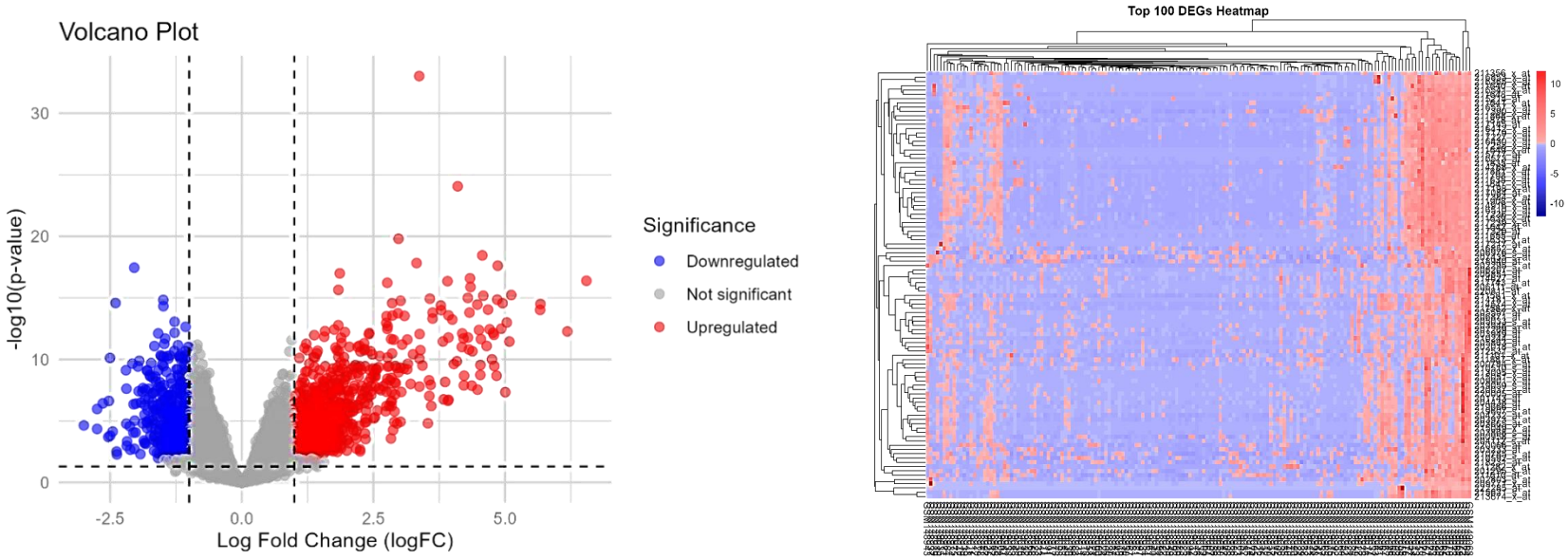
Figure 4.1. Comparison of (A) raw and (B) normalized expression data box plots showing the effectiveness of normalization in reducing variability across samples.

4.3 DEGs Identification

We used expression data of 22,283 genes from the GEO: GSE6477 datasets, which included 15 healthy adult donors and 147 MM patients. First, we performed quantile normalization, background correction and log transition of microarray data using the robust multiarray average (RMA) approach using GEO2R analyzer. The dispersal of DEGs among MM and healthy samples derived from GEO: GSE6477 was displayed using volcano plots, as seen in Figure 4.2A. Blue dots indicate genes that are downregulated ($p < 0.05$, $|log2FC| < -1$) and red dots indicate genes that are considerably upregulated ($p < 0.05$, $|log2FC| > 1$). The top 100 differentially expressed genes in the GEO: GSE6477 datasets were displayed in heatmap (Figure 4.2B). A significant expressed genes were represented by each row, and an MM or donor samples were indicated by each column. From the GEO: GSE6477 dataset, we were able to identify 1505 significant DEGs, 556 downregulated genes, and 949 upregulated genes.

4.4 Function and Pathway Enrichment Analysis

Using DAVID bioinformatics tool, GO annotation and KEGG pathway mapping was done to assess the functions of DEGs. In GO enrichment analysis, the DEGs were analyzed using three sub-ontologies: BP, MF, and CC. A total of 85 GO MF, 104 GO CC, 330 GO BP and 66 KEGG pathways were obtained. Based on the rank of enrichment count, the top 10 GO annotations and KEGG pathways were illustrated through bubble diagram in (Figure 4.3). For the BP annotation, DEGs were found largely engaged in T cell activation, inflammatory response, defense response and immune response (Figure 4.3A). For the MF annotation, DEGs were significantly correlated with RAGE receptor binding, protein binding, peptide antigen binding, lipopolysaccharide binding and transmembrane transporter activity (Figure 4.3C). For the CC annotation, the results showed that the majority of the proteins expressed by DEGs were situated inside transport vesicle membrane, endoplasmic reticulum (ER), ER membrane, lysosomal membrane, extracellular exosome, and cytosol (Figure 4.3B). The KEGG pathway analysis results demonstrated that DEGs had a substantial enrichment in the pathways correlated with ER protein processing, ribosome, hematopoietic cell lineage, cell adhesion molecules, Th17 cell differentiation, and osteoclast differentiation (Figure 4.3D).

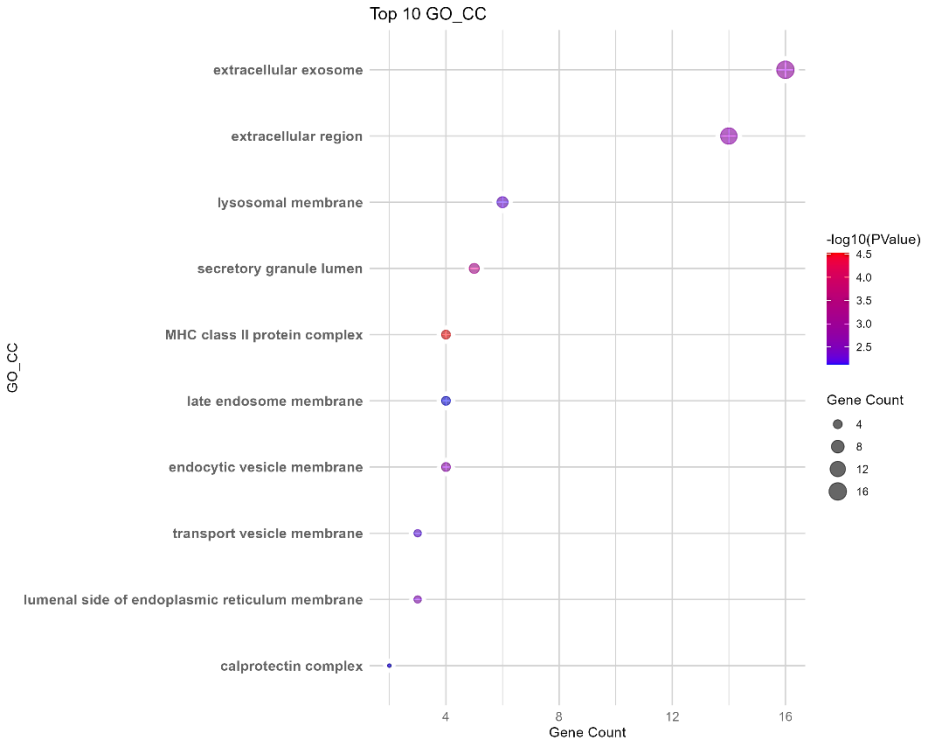


(A) (B)

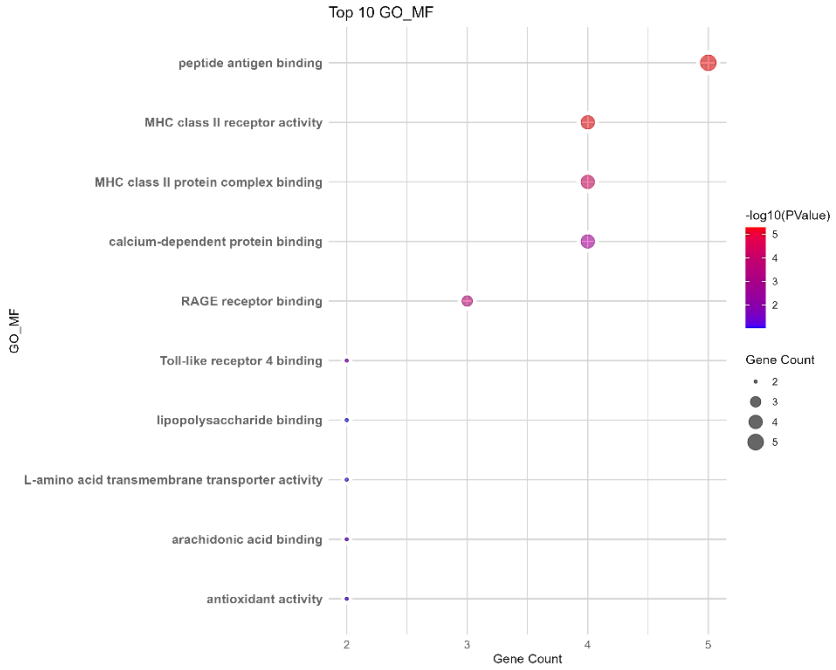
Figure 4.2. Identification of differentially expressed genes in GEO: GSE6477. (A) Volcano plot of DEGs. Red dots shows significantly upregulated genes, Blue dots shows significantly downregulated genes and gray dots represents no significance. (B) heatmap of top 100 significantly altered genes of differentially expressed genes in GEO: GSE647


Top 10 GO_BP
immune response
defense response to bacterium
positive regulation of immune response
positive regulation of T cell activation
positive regulation of interleukin-8 production
peptide antigen assembly with MHC class II protein complex
antigen processing and presentation of peptide or polysaccharide antigen via MHC class II
antigen processing and presentation of exogenous peptide antigen via MHC class II
antigen processing and presentation
leukocyte migration involved in inflammatory response
GO_BP
3 4 5 6 7 8
Gene Count
Gene Count
3
4
5
6
7
8
-log10(PValue)
5.0
4.5
4.0
3.5
3.0

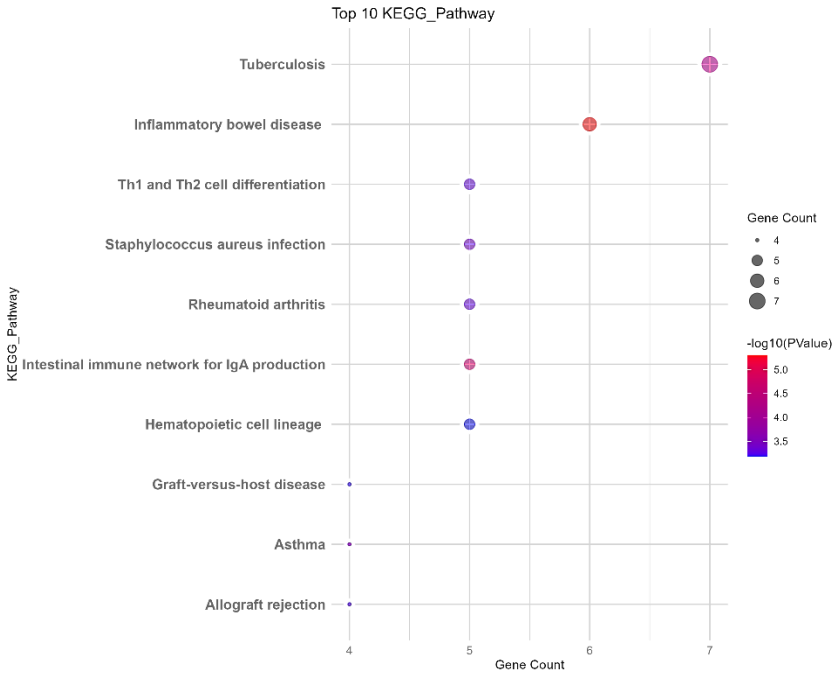
(A) GO BP


Top 10 GO_CC
extracellular exosome
extracellular region
lysosomal membrane
secretory granule lumen
MHC class II protein complex
late endosome membrane
endocytic vesicle membrane
transport vesicle membrane
lumenal side of endoplasmic reticulum membrane
calprotectin complex
GO_CC
4 8 12 16
Gene Count
-log10(PValue)
4.5
4.0
3.5
3.0
2.5
Gene Count
4
8
12
16

(B) GO CC



(C) GO MF

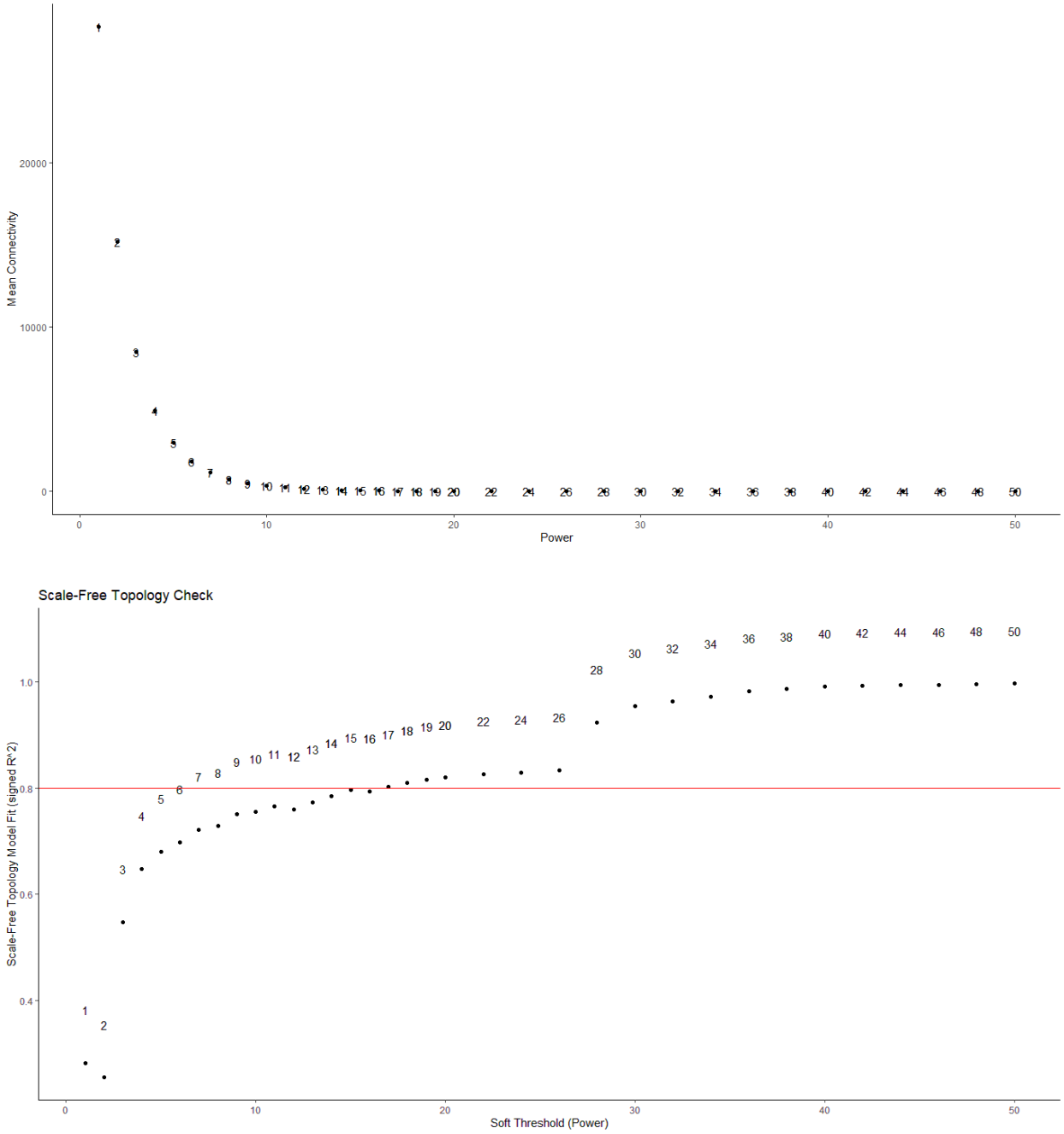


(D) KEGG Pathways

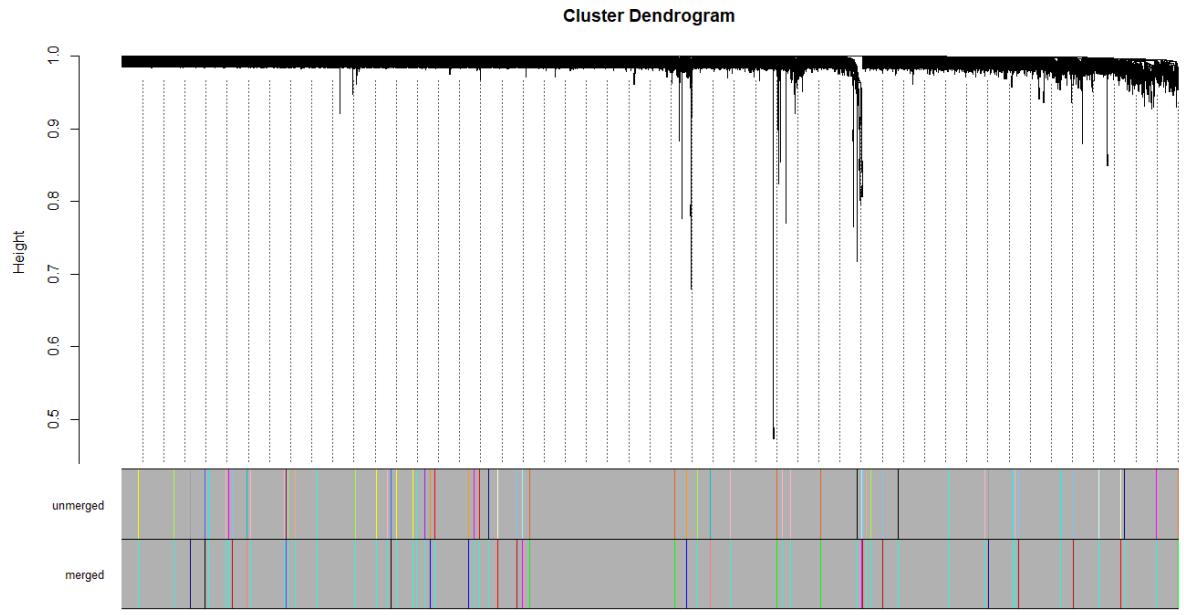
Figure 4.3. GO annotation and KEGG pathway mapping of differential expression genes. (A-D) The bubble diagrams shows top 10 annotation results. (A) Biological Processes BP. (B) Molecular Functions MF. (C) Cellular Components CC. (D) KEGG pathways.

4.5 Weighted Gene Co- Expression Network Analysis

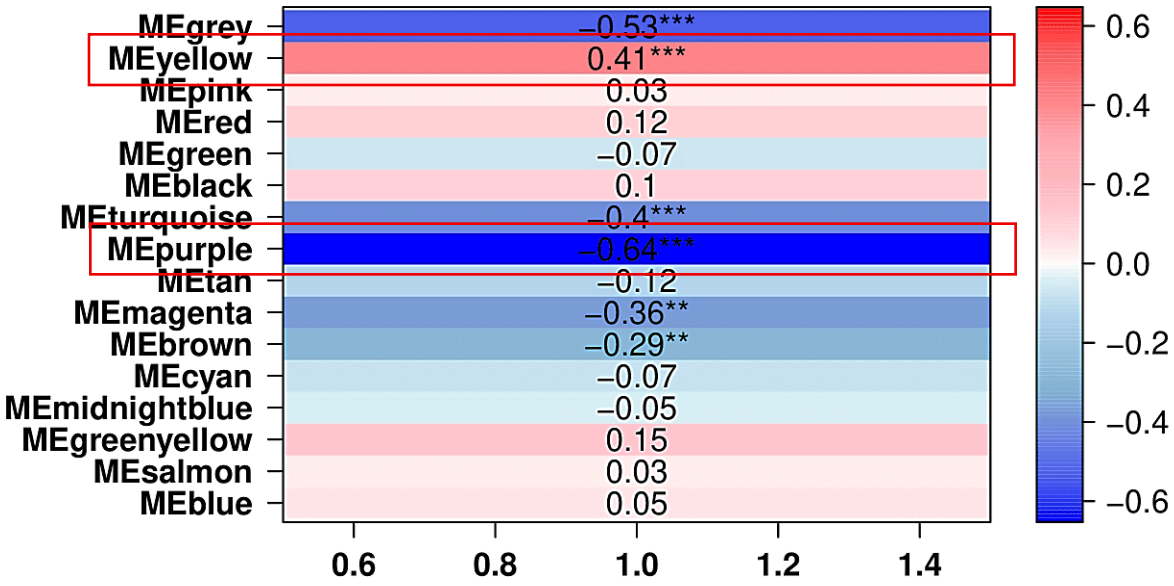
Co-expression analysis was performed on 78 samples, including diseased and controlled with complete information in the ArrayExpress: E-GEOD-5900 dataset. The "WGCNA" package in R was used to group genes exhibiting similar expression profiles into similar module. To construct a scale-free network, we selected $\beta = 17$ (scale free $R^2 = 0.85$) as the soft-threshold (Figure 4.4 (A)). 16 modules were eventually obtained at a height cut of 0.25 (Figure 4.4(B)). Using the dynamic tree cut method, the least module scale was assigned as 30. The hierarchical gene clustering yields 16 modules. The modules were selected based on the heatmap in Figure 4.4(C) that shows the association between module eigengene and clinical characteristics. The results showed significant correlation between 16 modules and clinical characteristics of multiple myeloma. 7 clusters (Yellow, Pink, Red, Black, Greenyellow, Salmon, and Blue) were positively correlated and contains overexpressed genes. Alternatively, 9 clusters (Grey, Green, Turquoise, Purple, Tan, Magenta, Brown, Cyan, and Midnigthtblue) were observed to be negatively linked with MM traits. Gene Significance GS > 0.2 and Module Membership MM > 0.8 were assigned to assess significant genes in Purple, and Yellow module. eventually, 65 significant genes were added for further screening. Furthermore, eigengene correlation is typically used to assess module similarity in the eigengene dendrogram (Figure 4.4B) and heatmap (Figure 4.4C). The trend associated with different stages of the disease is shown by heat maps that show the expression levels of all genes and the eigengene values for the purple and yellow modules. According to the results, the Purple and Yellow modules had the strongest correlations with the tumor evolution (Figure 4.4(D)). Consequently, the Purple and Yellow modules were regarded as clinically significant and tumor evolution modules.



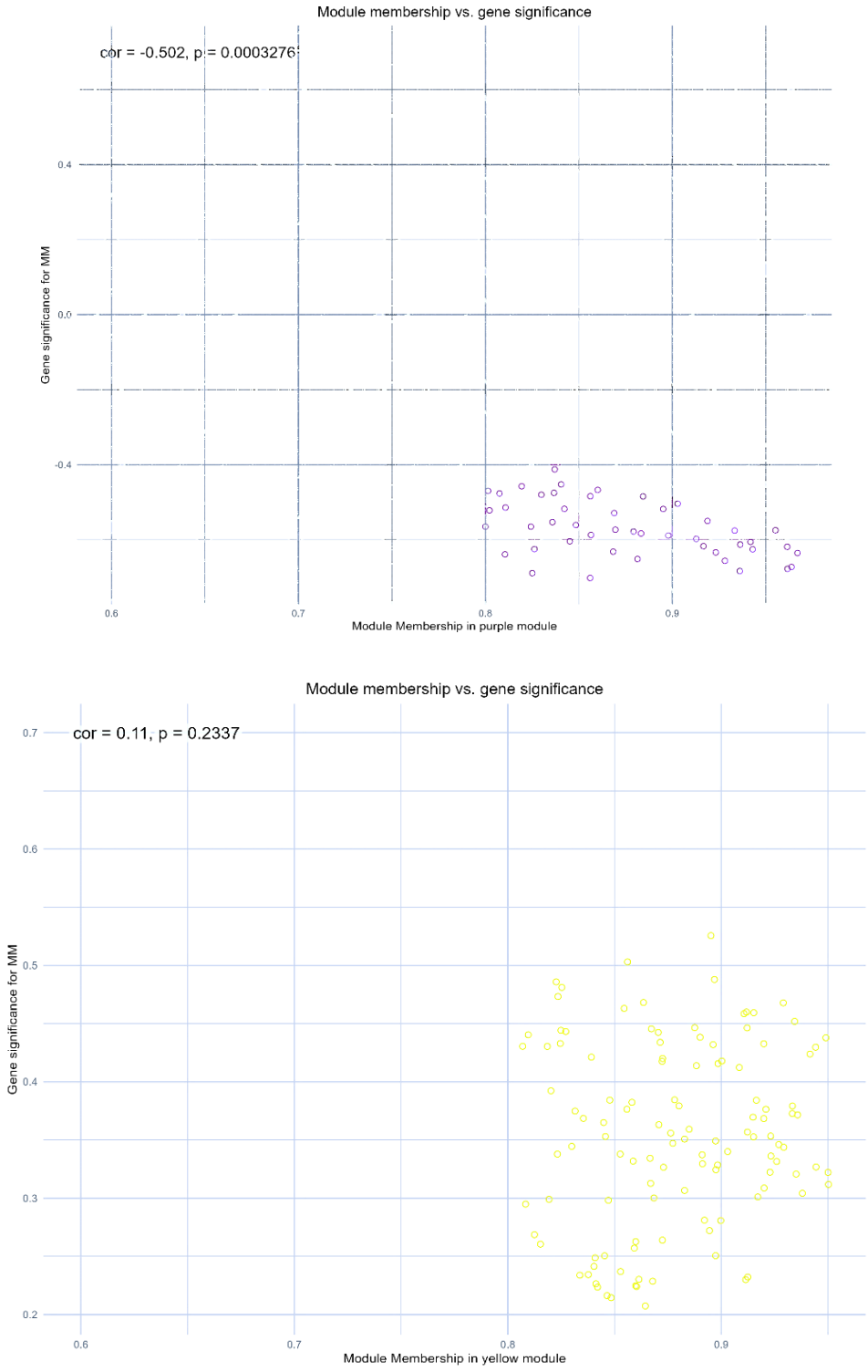
(A) Mean Connectivity & Scale Free Topology



(B) Clustering Dendrogram



(C) Module-Trait Relationship Heatmap



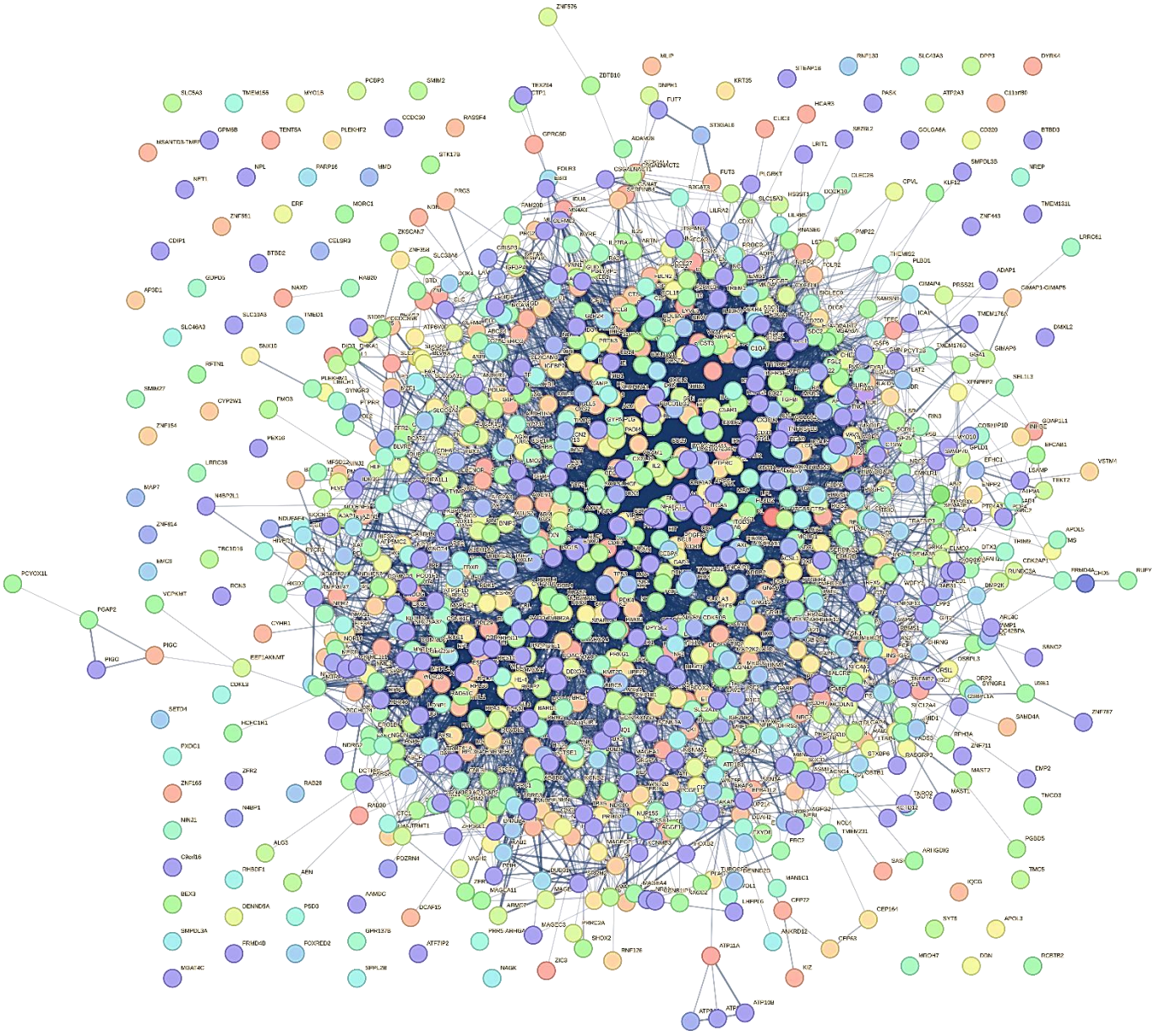
(D) Module Membership vs Gene Significance

Figure 4.4. WGCNA analysis of ArrayExpress: E-GEOD-5900. (A) The network topology analysis of several soft thresholding powers; (Left panel) the scale-free fit index (on the y-axis) is

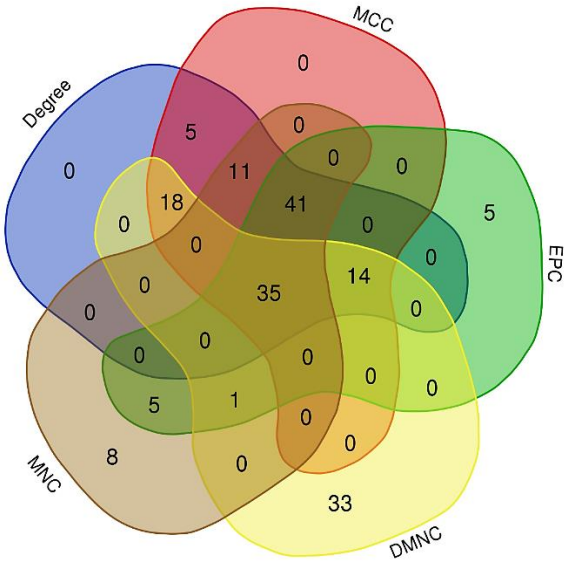
plotted as a function of the soft-thresholding power (on the x-axis); (right panel) the mean connectivity (degree, on the y-axis) plotted as a function of soft-thresholding power (on the x-axis); $\beta = 17$ was set as the soft-thresholding power for further analysis. (B) Using the assigned module colors and topological overlap, clustering dendrogram for the 54,675 genes with dissimilarity was created; 31 co-expression modules were created using different colors; up and down of the image shows the association between gene modules and gene dendrogram. (C) Module-trait relationships; each cell contains P value and the corresponding correlation; each row contains a module eigengene, while each column contains a trait. Among them, the purple and yellow module are the significant modules. (D) Scatterplot of Gene Significance and Module Membership in purple and yellow module; and mentioned correlation (cor) & P value.

4.6 Protein-Protein Interaction

The PPI network was constructed in order to identify the hub genes and understand the correlation of candidate DEGs. To construct the PPI network, 1505 DEGs in total were imported into the STRING. After removing the outlier proteins, an association network of 1075 genes was acquired. The protein network displayed in Figure 4.5A, was analyzed using Cytoscape (version 3.8.0) to screen all of the DEGs based on node rank. Using five Cytohubba methods, 1075 genes were ranked in order to assess the gene interactions. The top 100 genes acquired using each algorithm are displayed in Venn diagram, and 35 candidate genes were derived from the intersecting part (Figure 4.5B). After that, the PPI network was constructed for identified modules from WGCNA analysis. To identify hub genes with high confidence, each gene's hub score was examined using a CytoHubba plug-in in Cytoscape. Our investigation found that 62 genes in the yellow module and 3 highly connected genes in the purple module were hub genes. To identify potential candidate hub genes for downstream validation, genes identified from both the PPI network and WGCNA were prioritized for further analysis.



(A) PPI Network



(B) Cytohubba Algorithms Venn Diagram

Figure 4.5. PPI network of differentially expressed genes. (A) PPI network of 1075 protein coding genes. (B) Venn diagram demonstrates common genes based on 5 Cytohubba Algorithms [Degree, DMNC, EPC, MCC, and MNC].

4.7 Hub Gene Identification

Hub genes were identified through an integrative approach utilizing Venn diagram analysis to determine the overlapping genes across different datasets. This method allowed us to pinpoint common genes that may play significant roles in MM. Following this identification, these shared genes were subjected to phylogenetic analysis to explore their evolutionary relationships and further explain their functional significance in the context of MM.

4.7.1 Crossover Candidate Gene

By intersecting the aforementioned 65 WGCNA module genes and 35 PPI protein coding genes shown in Figure 4.6, A total of 23 genes (RPL10, RPL27, RPL24, RPS25, RPS17, RPL18, RPL32, RPL12, RPL14, RPL29, RPL10A, RPL17, RPL35, RPS10, RPS19, RPL28, RPL35A, RPL18A, EEF1B2, RPL36, RPS21, RPL13, RPL30) were identified as hub genes. The resulting hub genes were subjected to phylogenetic analysis to facilitate further validation.

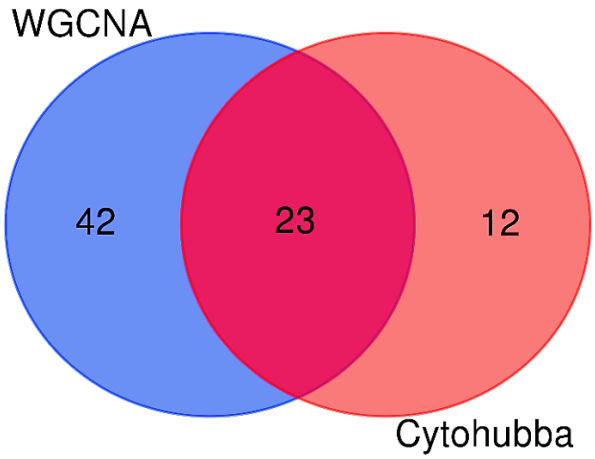
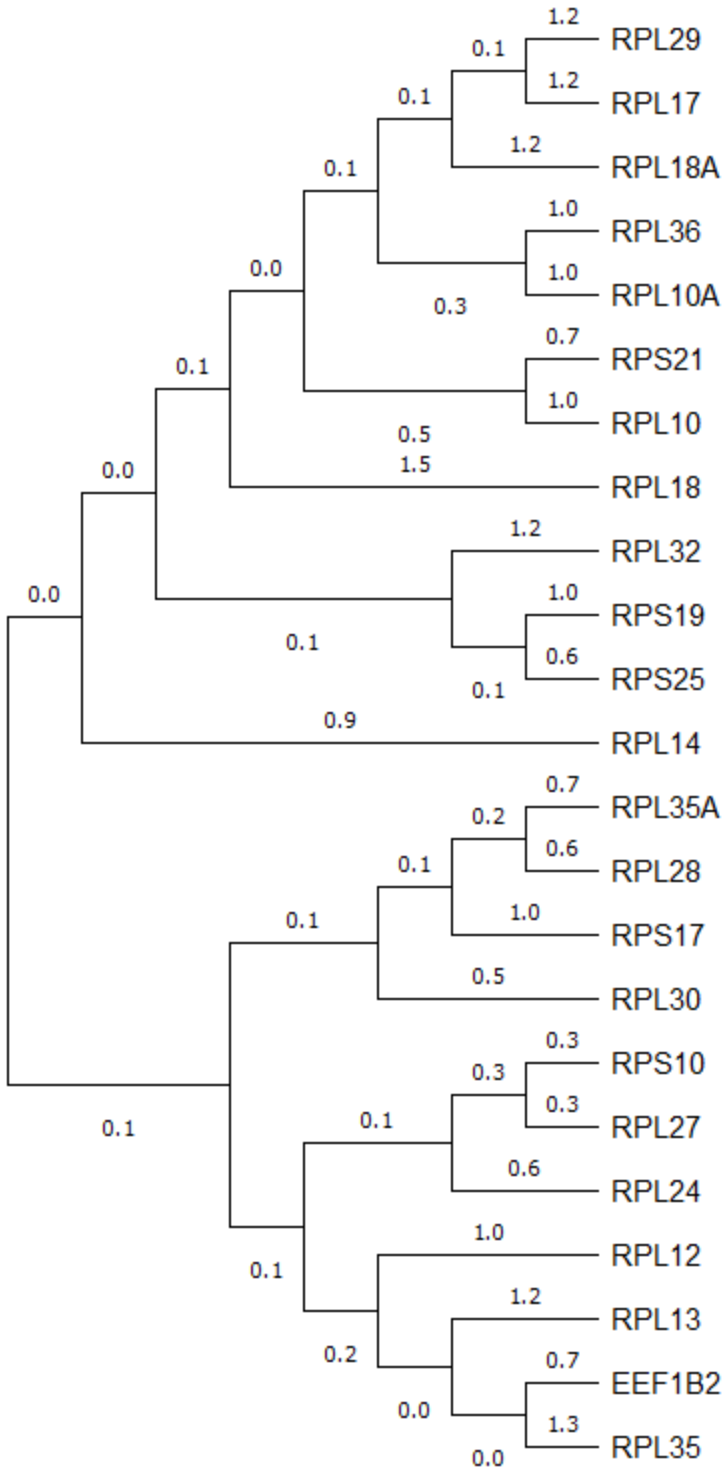


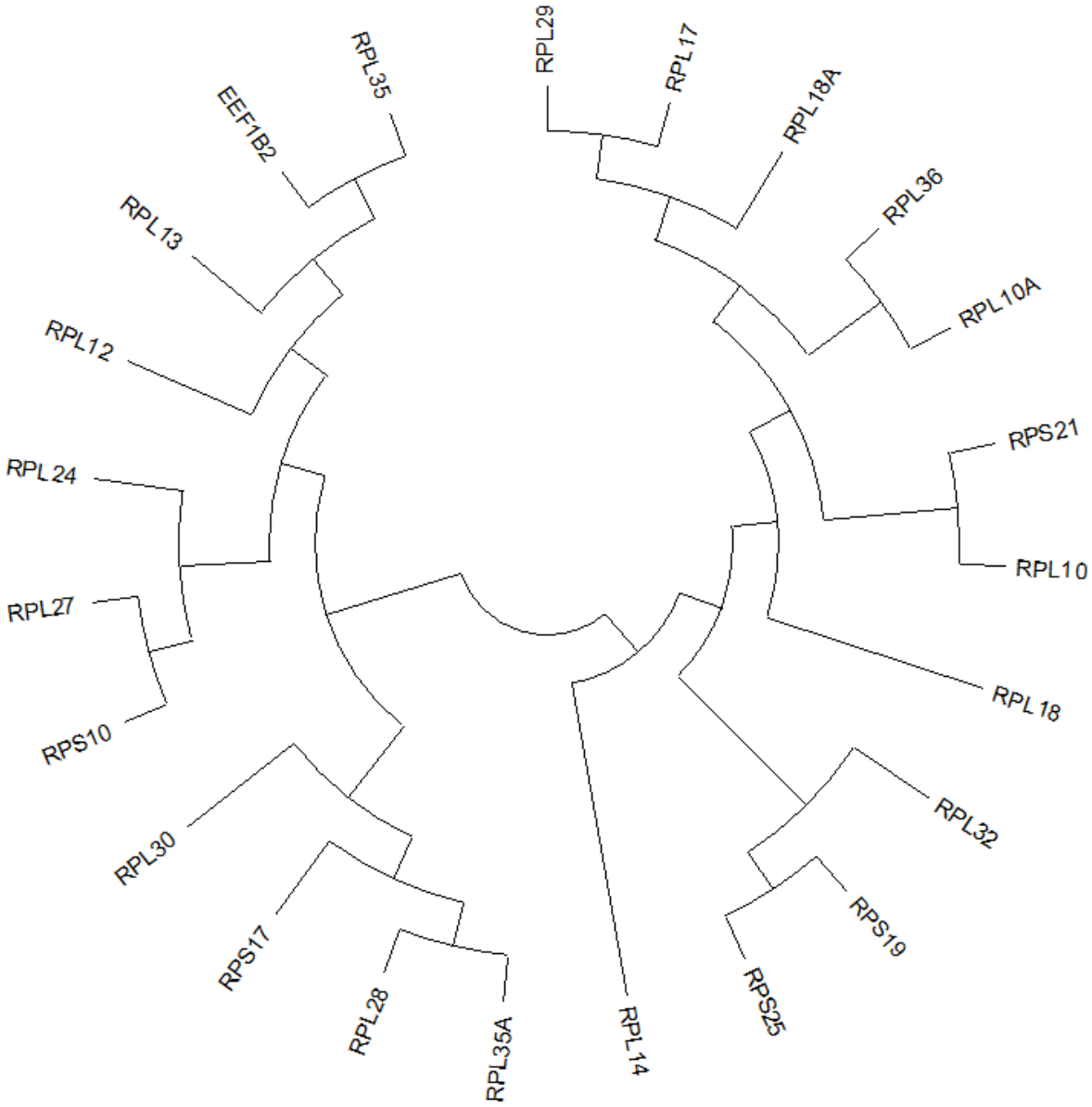
Figure 4.6. Venn diagram shows the common hub genes identified through WGCNA and PPI analysis utilizing Cytohubba.

4.7.2 Phylogenetic Analysis

CLUSTAL W, a multiple sequence alignment tool is used to construct a phylogenetic tree using genetic sequences and displayed using MEGA 11 software. A Phylogenetic relationship between the gene clusters along with distances is shown in the Figure 4.7 A. The pairwise distances based on maximum composite likelihood can be seen from Figure 4.7 B. Lastly, eight databases were consulted to identify and analyze the functions of each gene in humans, which includes OMIM, GeneCards, DisGeNET, PubMed, The Human Protein Atlas, Open Targets, COSMIC, and IntOGen database. Out of 23 genes analyzed, 19 genes exhibited close relationships based on shorter distances and similar expression rates, indicates that those genes are more responsible for MM progression. The identified genes are RPS17: Ribosomal Proteins S17, RPS25: Ribosomal Protein S25, RPS19: Ribosomal Protein S19, RPL10: Ribosomal Protein L10, RPL29: Ribosomal Protein L29, RPL30: Ribosomal Protein L30, RPL35A: Ribosomal Protein L35A. Those 7 genes were considered as hub genes for further evaluation and validation through statistical analysis such as ROC analysis and survival analysis.


RPL29
RPL17
RPL18A
RPL36
RPL10A
RPS21
RPL10
RPL18
RPL32
RPS19
RPS25
RPL14
RPL35A
RPL28
RPS17
RPL30
RPS10
RPL27
RPL24
RPL12
RPL13
EEF1B2
RPL35

(A)


RPL29
RPL17
RPL18A
RPL36
RPL10A
RPS21
RPL10
RPL18
RPL32
RPS19
RPS25
RPL14
RPL35A
RPL28
RPS17
RPL30
RPS10
RPL27
RPL24
RPL12
RPL13
EEF1B2
RPL35

(B)

		EEF1E	RPL14	RPL30	RPL13	RPS21	RPL36	RPL18	RPL35	RPL28	RPS19	RPS10	RPL35	RPL10	RPL29	RPL12	RPL32	RPL18	RPS17	RPS25	RPL24	RPL27	RPL17	RPL10
2	EEF1B2																							
3	RPL14	1.927																						
4	RPL30	1.333	1.212																					
5	RPL13	1.903	2.027	1.859																				
6	RPS21	2.041	2.196	1.787	2.853																			
7	RPL36	2.401	2.095	2.001	2.276	2.458																		
8	RPL18	2.376	2.230	1.907	2.549	2.062	2.593																	
9	RPL35	1.734	1.590	1.186	2.144	1.921	2.674	2.315																
10	RPL28	1.486	1.823	1.066	2.185	2.067	2.409	2.187	1.155															
11	RPS19	2.068	2.165	1.839	2.187	2.114	2.395	2.013	1.881	2.226														
12	RPS10	1.445	1.567	1.147	2.066	2.060	2.155	2.172	1.601	1.537	1.927													
13	RPL35	1.908	2.068	1.899	2.328	2.336	2.553	2.354	2.234	2.224	2.368	1.886												
14	RPL10	2.022	1.997	1.898	2.510	1.894	1.967	2.433	2.132	2.157	2.193	2.169	2.507											
15	RPL29	2.081	2.263	1.827	2.865	3.165	2.479	2.543	1.964	2.103	2.047	2.014	2.401	2.446										
16	RPL12	1.662	2.043	1.544	2.106	2.126	2.380	2.635	2.041	1.717	2.195	1.656	2.234	2.296	2.398									
17	RPL32	2.214	1.987	1.883	2.854	2.938	2.724	2.628	2.027	1.989	2.135	1.902	2.737	2.624	3.242	2.512								
18	RPL18	2.227	2.159	2.101	2.903	2.250	2.278	2.791	2.243	2.052	2.370	2.135	2.765	2.633	2.970	2.537	3.141							
19	RPS17	2.070	1.595	1.425	2.286	1.748	2.172	2.420	1.634	1.515	2.100	1.874	2.476	2.095	2.310	2.276	2.308	2.512						
20	RPS25	1.913	1.309	1.304	2.212	2.142	2.343	2.363	1.310	1.367	1.443	1.331	2.182	2.211	2.207	2.375	1.808	1.960	1.492					
21	RPL24	1.415	1.697	1.076	2.181	2.132	2.166	2.389	1.666	1.437	1.952	1.235	1.975	2.233	2.022	1.657	2.054	2.130	1.653	1.493				
22	RPL27	1.464	1.543	1.148	1.846	2.207	2.473	2.000	1.536	1.333	1.828	0.645	1.782	1.989	2.046	1.599	1.863	2.056	1.767	1.429	1.143			
23	RPL17	2.150	2.058	2.054	2.319	3.133	2.630	2.386	2.363	2.090	2.358	2.147	2.813	2.568	2.309	2.457	2.733	2.538	2.443	2.223	2.044	2.073		

(C)

Figure 4.7. (A) and (B) presents Phylogenetic tree of 23 genes in Multiple Myeloma. (C) Pairwise distances based on maximum likelihood.

4.8 Hub Gene Evaluation and Validation

The evaluation of hub genes was conducted using Receiver Operating Characteristic (ROC) analysis and survival analysis. ROC analysis was employed to assess the predictive power of these genes to distinguish different patient outcomes, while survival analysis provided insights into their associations with overall survival in MM. These analyses confirmed the relevance of the identified hug genes in clinical settings.

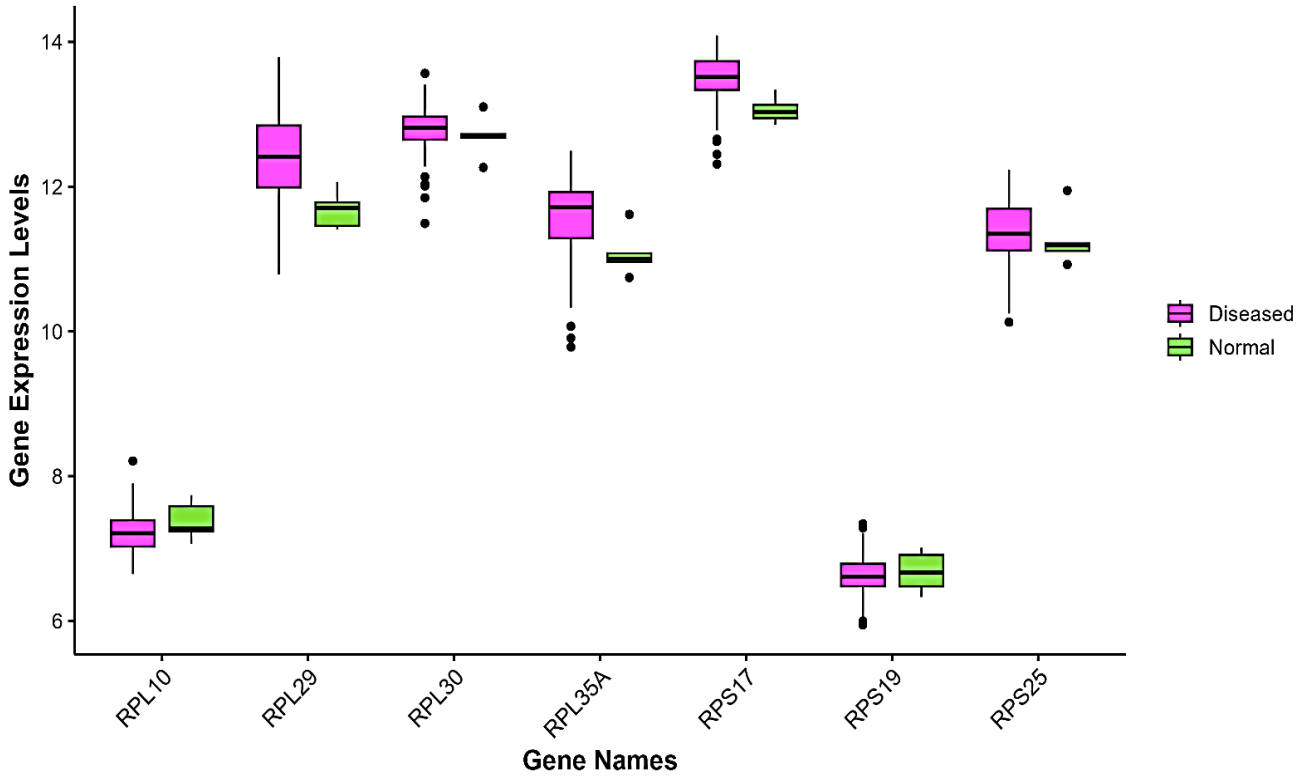
4.8.1 Receiver Operating Characteristic Analysis

In order to assess the capacity and predictive value of the candidate hub genes to detect MM, ROC analysis was performed using GEO: GSE6477 as training dataset and GEO: GSE13591 as test dataset. The expression levels of genes were visualized using boxplot (Figure 4.8A). In comparison to control samples, MM samples had higher expression levels of RPS25, RPS17, RPL10, RPL29, RPL35A, and RPL30 with P value < 0.05. Subsequently, in the validation set, the genes (RPS25, RPS19, RPL10, RPL29, RPL35A, and RPL30) had area under curve (AUC) less than 0.7, except RPS17 shown in (Figure 4.8B-H). This indicates that the RPS17 has high sensitivity (59.44%) and

specificity (76%) for MM in both training and test datasets. It is discovered that RPS17 is highly effective at differentiating between MM and healthy samples, indicating it as a biomarker for MM progression.

4.8.2 Survival Analysis

Furthermore, the influence of all aforementioned 7 hub genes was assessed to determine the expression levels of the genes that are associated with survival outcomes of MM patients using Kaplan-Meier plotter on the GEO: GSE24080 through K-M curves. Therefore, survival analysis was carried out and analyzed the predictive significance of the genes. Results reveals that higher expression level of RPS17 and RPL35A possesses significantly shorter survival times (Cox p = 0.021 and Cox p= 0.016 respectively, Figure 4.9), while higher level of expression of RPS19 (Cox p = 0.028) and RPS25 (Cox p = 0.05) had significantly better survival times. The results showed significant relationship between the hub genes RPS25, RPS19, RPL35A and RPS17, as well as with event-free survival (EFS) and the overall survival (OS) of MM patients with P value <0.05.



(A) Expression level of hub genes

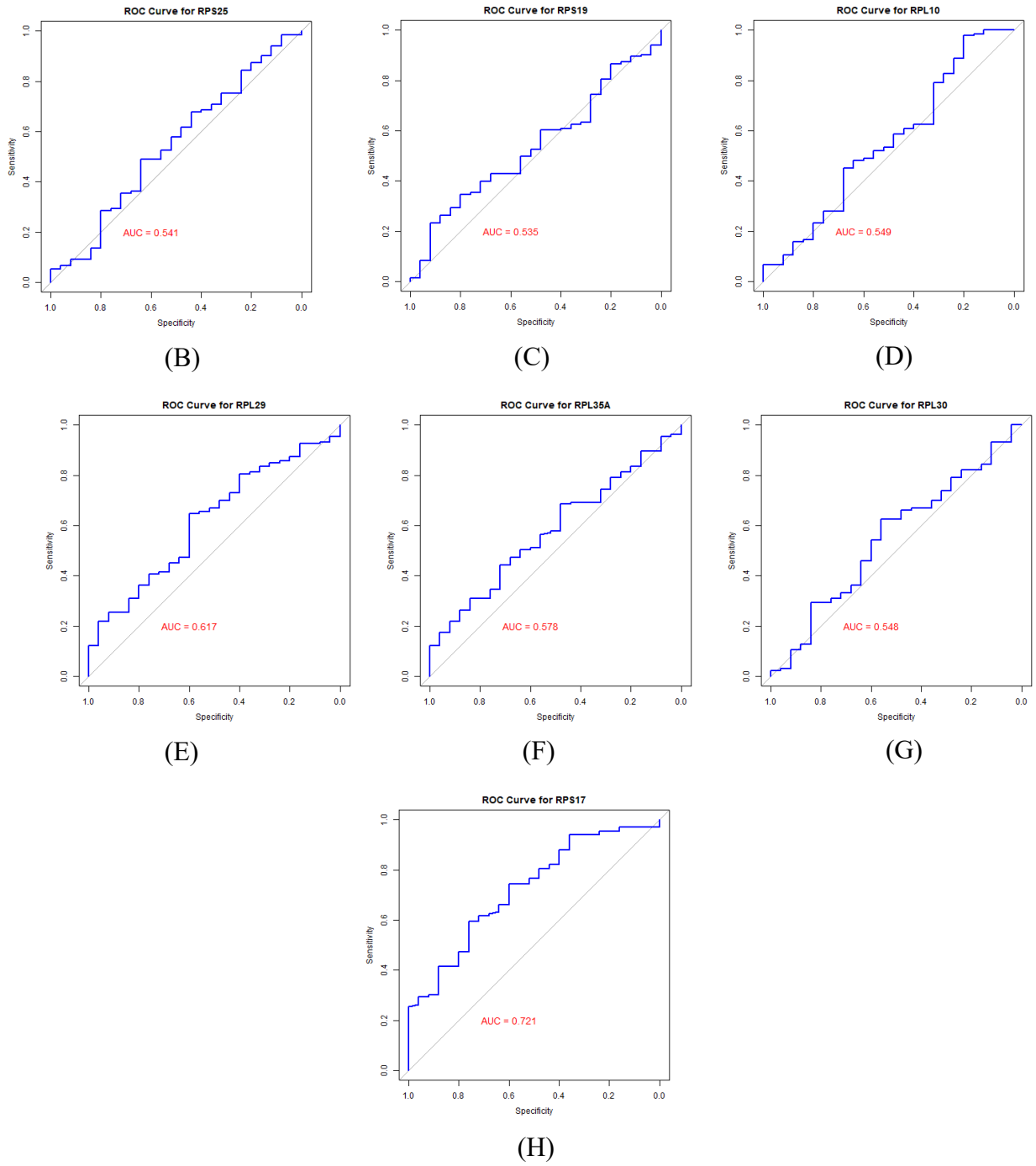
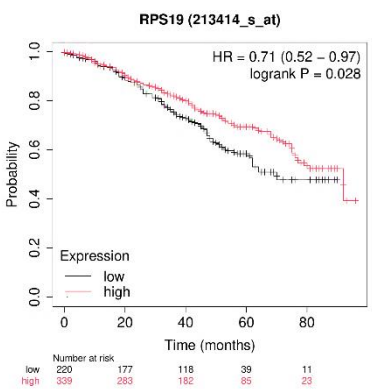
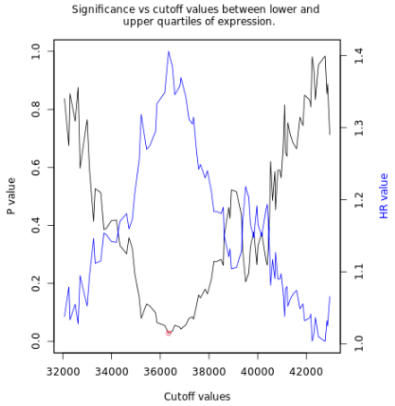
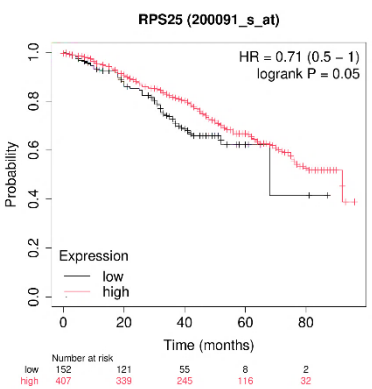
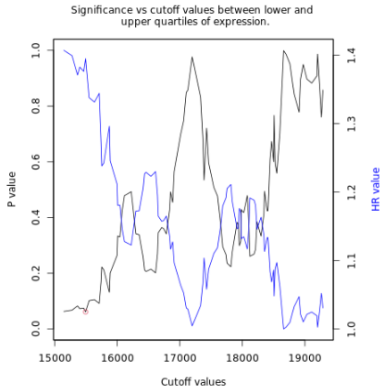


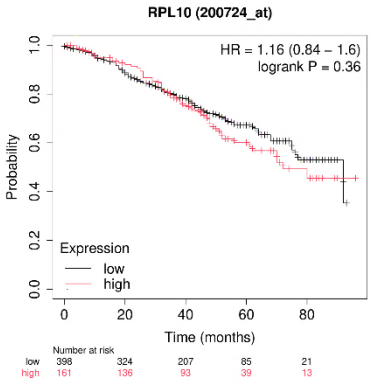
Figure 4.8. Evaluation and validation of hub genes predictive value along with their expression levels. (A) Expression level of seven hub genes in GEO: GSE13591. (B-H) ROC curves of seven hub genes (RPS19, RPS25, RPL10, RPL29, RPL35A, RPL30 and RPS17) in the GEO: GSE13591 for diagnosing MM. Data is displayed by cut-off values (sensitivity, specificity).

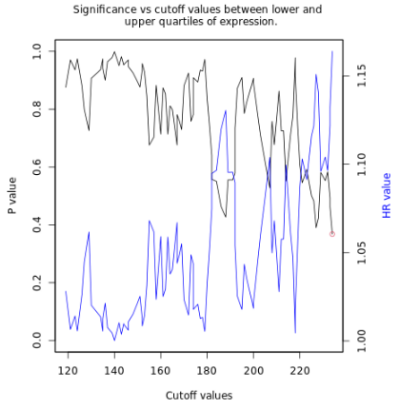

RPS19 (213414_s_at)
HR = 0.71 (0.52 – 0.97)
logrank P = 0.028
Probability
Expression
low
high
Time (months)
Number at risk
low 220 177 118 39 11
high 339 283 182 85 23

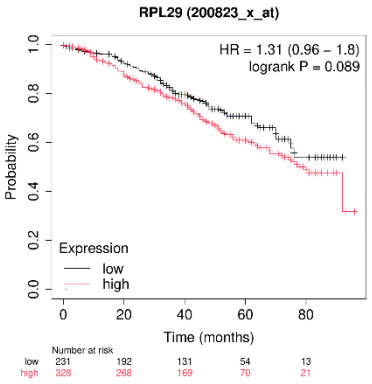

Significance vs cutoff values between lower and upper quartiles of expression.
P value
HR value
Cutoff values

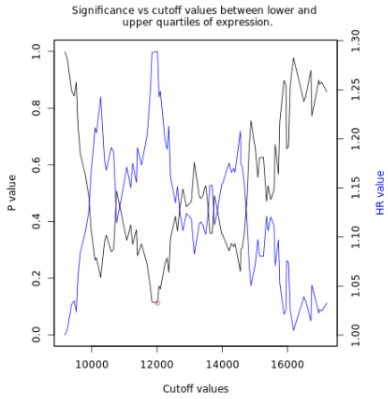

RPS25 (200091_s_at)
HR = 0.71 (0.5 – 1)
logrank P = 0.05
Probability
Expression
low
high
Time (months)
Number at risk
low 152 121 55 8 2
high 407 339 245 116 32


Significance vs cutoff values between lower and upper quartiles of expression.
P value
HR value
Cutoff values


RPL10 (200724_at)
HR = 1.16 (0.84 – 1.6)
logrank P = 0.36
Probability
Expression
low
high
Time (months)
Number at risk
low 398 324 207 85 21
high 161 136 93 39 13


Significance vs cutoff values between lower and upper quartiles of expression.
P value
HR value
Cutoff values


RPL29 (200823_x_at)
HR = 1.31 (0.96 – 1.8)
logrank P = 0.089
Probability
Expression
low
high
Time (months)
Number at risk
low 231 192 131 54 13
high 328 268 169 70 21


Significance vs cutoff values between lower and upper quartiles of expression.
P value
HR value
Cutoff values

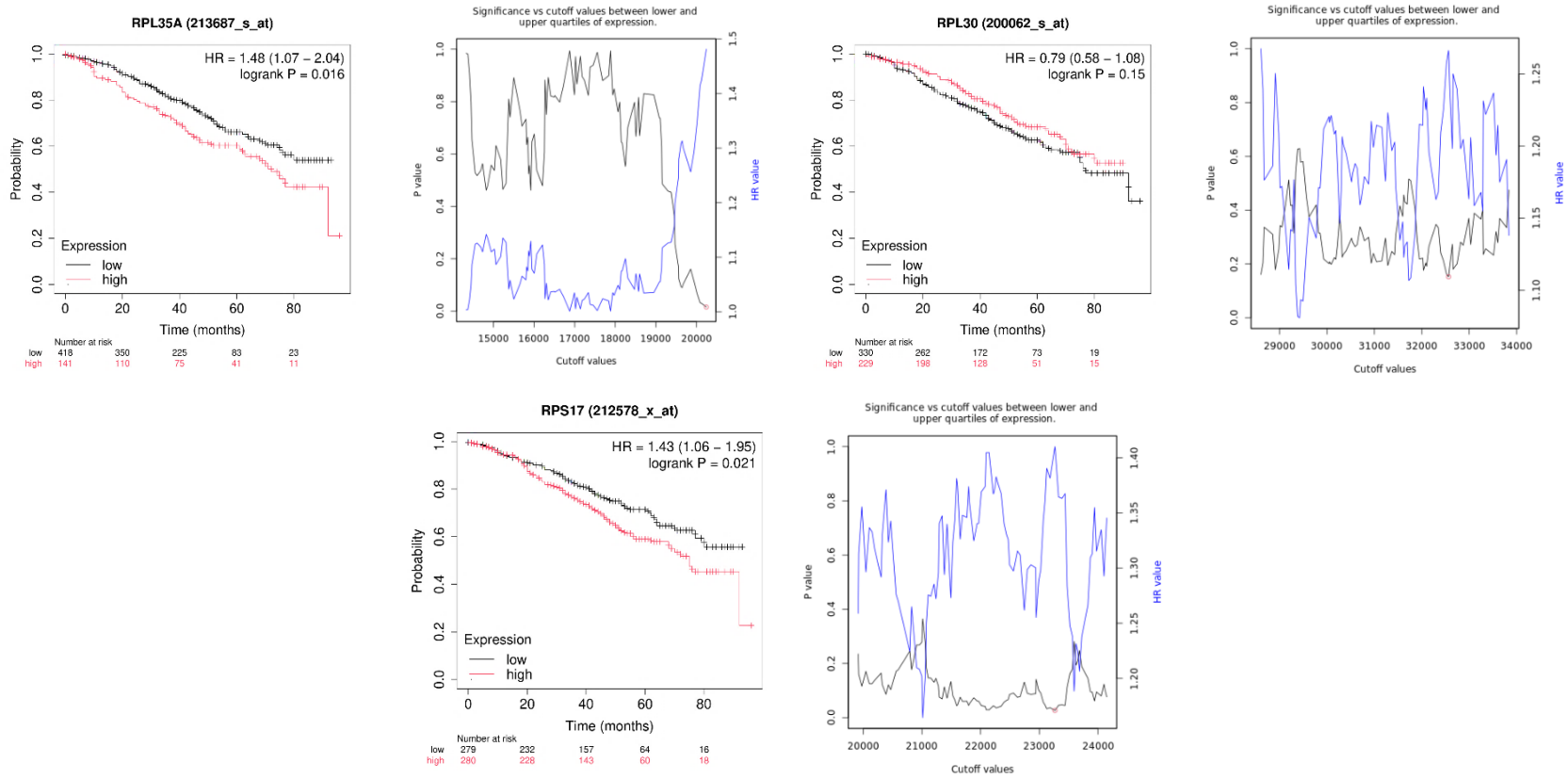


Figure 4.9. Validation of hub genes prognostic value. Kaplan Meier analysis and expression values of hub genes (RPS19, RPS25, RPL10, RPL29, RPL35A, RPL30 and RPS17) in MM in the GSE24080. MM; multiple myeloma.

Chapter 5

Discussion

5 Discussion

Multiple Myeloma exhibits a broad clinical spectrum, progressing from MGUS to SMM and ultimately to MM and relapsed MM. The incidence has risen over recent decades, establishing it as a disease of significant clinical and economic relevance [4], [53]. It's well-established fact that Immunomodulatory agents and proteasome inhibitors remain fundamental components of contemporary treatment. However, CD38-targeting antibodies are increasingly recognized as significant elements in both relapse and first-line therapies. Over recent times, many new medications, such as carfilzomib, selinexor, and belantamab mafodotin have been discovered to treat relapsed MM [3]. Despite the fact that the existing treatments are highly efficacious in preventing disease progression, but still numerous patients continue to experience relapses and disease refractory. Accordingly, identifying novel targets associated with the progression and prognosis of MM are therefore of great significance. The WGCNA algorithm can identify highly relevant gene modules to their clinical phenotypes. It has continued to be broadly used to filter significant genes in multiple cancer types, including gastric, lymphoma, ovarian, cholangiocarcinoma and hepatocellular, as well as in an increasing number of non-neoplastic diseases [64]–[67]. However, WGCNA has rarely been applied in MM so far, but over time more researchers are using it for the identification of new diagnostic and prognostic biomarkers in cancers [3], [68]. To the extent of our insight, this is the first study to utilize the GEO and ArrayExpress databases for potential multiple myeloma biomarkers using a combination analysis of WGCNA, DEGs, and a phylogenetic approach. Therefore, gene expression profiling and co-expression network analysis can reveal genes that might facilitate the onset and progression of disease. Herein, an integrated bioinformatics and phylogenetic analyses was done to detect biomarkers related to MM progression and prognosis.

Using the WGCNA approach, we have found potential genetic biomarkers involved in MM progression and screened for DEGs. The WGCNA approach and DEGs analysis was used to find key modules and significant DEGs correlated with MM progressing using the ArrayExpress: E-GEOD-5900 and GEO: GSE6477 data sets, respectively. The key module yellow and purple observed being significantly linked to the disease state in MM. Subsequently, 62 genes in yellow module and 3 genes in purple module were identifies as hub genes. DEGs screening revealed that there were 1505 significant DEGs between MM patients and healthy donors, with 556

downregulated genes and 949 upregulated genes with logFC > 1 and $p < 0.05$. Additionally, the analyses of KEGG pathways and GO annotation were performed to look into the biological roles of the genes that were enriched in the training datasets. KEGG analysis revealed enrichment pathways linked to protein processing in the ER, ribosome, hematopoietic cell lineage, cell adhesion molecules, osteoclast differentiation, and Th17 cell differentiation. While GO enrichment analysis in BPs identified the functional enrichment of DEGs in T cell activation, inflammatory response, defense response and immune response. In contrast, genes of the purple module were linked to the immune system, extracellular space, immune response, positive regulation of B cell activation and B cell receptor signaling. Whereas, genes of the yellow module were primarily involved in cytoplasmic translation, RNA binding, protein binding, endoplasmic reticulum, ribosome biogenesis, structural constituent of ribosome and COVID-19. We also used Degree, MNC, EPC, MCC, and DMNC methods in Cytohubba to identify the top 100 protein-coding genes and perform protein-protein interactions of DEGs. Furthermore, we screened overlapped genes from Cytohubba in DEGs and genes in WGCNA modules. The 23 hub genes were listed: RPL10, RPL27, RPL24, RPS25, RPS17, RPL18, RPL32, RPL12, RPL14, RPL29, RPL10A, RPL17, RPL35, RPS10, RPS19, RPL28, RPL35A, RPL18A, EEF1B2, RPL36, RPS21, RPL13, RPL30. CLUSTAL W, a multiple sequence alignment tool, is used to construct a phylogenetic tree using the sequences of 23 genes and displayed using MEGA 11 software. Lastly, eight databases were consulted to identify and analyze the functions of each gene in humans, which includes OMIM, GeneCards, DisGeNET, PubMed, The Human Protein Atlas, Open Targets, COSMIC, and IntOGen database. Out of 23 genes analyzed, 19 genes exhibited close relationships based on shorter distances and similar expression rates, indicating that those genes are more responsible for MM progression. The identified genes are RPS17: Ribosomal Protein S17, RPS25: Ribosomal Protein S25, RPS19: Ribosomal Protein S19, RPL10: Ribosomal Protein L10, RPL29: Ribosomal Protein L29, RPL30: Ribosomal Protein L30, RPL35A: Ribosomal Protein L35A. Upon reviewing the relevant literature, the resultant 7 genes were subjected to statistical validation analysis. Four genes (RPS17, RPS19, RPS25 and RPL35A) were eventually taken as candidate hub genes. Among the four hub genes, the expression levels of RPS19 and RPS25 were observed as notably lower while the expression levels of RPS17 and RPL35A were notably higher in MM samples in the validation dataset. Furthermore, RPS17 demonstrated outstanding diagnostic efficacy for both MM patients and healthy individuals in both the training and test datasets, as

indicated by the AUCs curves of ROC analysis. Afterwards, survival analysis was performed in order to further validate the hub genes. Using the Kaplan-Meier analysis, we discovered that high levels of expression of RPS17 and RPL35A and low levels of expression of RPS19 and RPS25 were significantly related to poor overall survival (OS) and event-free survival (EFS) of MM patients ($p < 0.05$). Regarding the role of RPS17, RPS19, RPS25 and RPL35A further investigations are required.

RPS17, a ribosomal protein, critically implicated in ribosome biogenesis and protein synthesis. The increased expression of RPS17 has been demonstrated to stimulate the growth and survival of MM cells, potentially contributing to disease progression. This would mean RPS17 could be an important factor in the aggressive behavior of MM cells either by promoting their proliferation or by helping maintain the malignant phenotype. Correlation of RPS17 with other apoptotic and cell cycle control pathways may indicate that it contributes in the progression processes of MM [53]. Further, due to its role in the modulation of cellular stress responses, the efforts of achieving therapeutic success might become much more challenging with increased resistance to therapy. The association of RPS17 with poor survival identifies the potential for using it as a target for therapy, such that methodologies aimed to down-regulating RPS17 will likely improve patient's outcomes [69]. Therefore, it is suggested that RPS17 may become a promising biomarker for MM prognosis. However, there is need of further in-depth research about its function as well as its therapeutic potential as a target of interest.

RPS19, a ribosomal protein, critical for proper ribosome assembly and function. Its downregulation may impair cellular homeostasis, increasing myeloma cell apoptosis and several other affects [70]. The RPS19 play the role as a potential preventive factor against the progression of disease. Whereas, Daptomycin (DAP), an antibiotic possessing anti-tumor activities has recently been shown in a report to block the progression of MM through downregulating the RPS19 in the cell line RPMI 8226**.** More importantly, it has been shown by research evidence that upregulation of RPS19 would enhance protein synthesis and treatment responses [53]. Recent studies suggest that the role of RPS19 in MM has been implicate and the expression of RPS19 is important to ensure a balance between cell division and apoptosis. As we know that RPS19 is crucial for cell survival and a reduced ribosomal functionality. It is suggested that the low RPS19 levels could lead to lessened synthesis of proteins and protein synthesis deficiency. Therefore,

myeloma cells might have a higher chance of stress-induced apoptosis which could lower the survival rate of patients [71]. Targeting the expression of RPS19 may hold the potential to maximize cell death in tumor cells and reverse the function of ribosomes, thereby providing a dual benefit toward the treatment of MM. These findings hold immense importance for treatment approaches. Moreover, therapeutic approaches focusing on RPS19 may lead to dual benefits by either restoring normal levels of RPS19 or blocking its downregulation during the treatment with drugs like DAP. This could result in improved protein synthesis and induction of apoptotic pathways in myeloma cells. Thus, resulting in overall improved treatment response and longer survival for the MM patients. Finally, these studies uncover the complex interaction between ribosomal proteins while RPS19 stands out to be an important gene for further exploration as therapeutic target and a biomarker in MM.

RPS25, a part of the ribosomal subunit, is necessary for proper protein synthesis as well as regulation of cellular development. Dysregulation of RPS25 may negatively impact ribosome function, contributing to the malignant phenotype seen in MM. Moreover, research suggests that inhibiting the protein RPS25 would restore normal ribosome function, making myeloma cells sensitive to apoptosis and making currently used therapies more effective [53], [72] According to new studies, RPS25 associates with IRES under ER stress, aside from being required for ribosome biogenesis, RPS25 acts as a trans-acting factor regulating translation of c-myc. RPS25 interacts with the hnRNP A1 RNA binding protein, that is required for IRES-dependent translation of c-myc during ER stress that is caused by bortezomib. Although the phosphorylation of eIF-2α inhibits global protein synthesis and mTORC1 is depressed in activity. This interaction is crucial for maintaining the levels of c-myc protein [69]. Indeed, it seems that IRES activity is induced in MM cells by stress, which leads to the survival of myeloma cells independent of apoptosis. The interference of the RPS25-hnRNP A1 complex could modulate c-myc translation, further sensitizing myeloma cells to drugs such as bortezomib. Furthermore, Knowledge of the dual role of RPS25 in ribosome biogenesis and as a c-myc translation regulator could thus lead to new treatment modalities targeted at improving the outcome of patients suffering from MM.

RPL35A is significant in regulating cell growth and protein synthesis. Overexpression of RPL35A might cause enhanced malignant plasma cell proliferation, thus aggravating the disease [53]. Consequently, a link between increased RPL35A levels and low survival reveals some

compelling basis for further functional investigation of this protein within the tumor microenvironment. The potential therapeutic strategy appears to be directed at RPL35A, which may interfere with the pertinent ribosomal function for myeloma cell survival and proliferation. Understanding the role of RPL35A in MM biology will help researchers better understand its potential as a therapeutic target and prognostic marker.

Our study thus provides practical and computational basis for future investigations on the basis of these particular targets and confirms the regulation of significant genes in the pathogenic progression of MM. Notwithstanding these significant results, our study may have a number of drawbacks. First, there could be potential biases in dataset selection, despite the fact that samples from the various datasets were rigorously quality controlled. However, it is indisputable that the impacts of inherent variability might not have been totally eradicated. Second, it was not possible to investigate the dynamic nature of these gene expressions during disease progression due to the limitations of the available data and dependency on computational models, which may not fully represent biological complexity. Therefore, more in vitro and in vivo experimental data is required to confirm the clinical value of discovered genes. Finally, based on computational predictions, exploring the drug-gene interactions for the biomarkers RPS17, RPS19, RPS25, and RPL35A might shed light on their functions in MM resistance and therapeutic responses.

Chapter 6

Conclusion and Future Directions

6 Conclusion

As a conclusion, several biologically and clinically significant biomarkers for multiple myeloma were discovered as a result of this investigation. Key identified differentially expressed genes were significantly enriched in ribosomal and translational pathways. Ribosomal proteins including RPS17, RPS19, RPS25 and RPL35A were among the most frequently mentioned. These proteins were shown to be possible hub genes with improved survival correlations. This research outlines the key roles of RPS17, RPS19, RPS25 and RPL35A in the progression of MM. The survival outcomes showed substantial correlation between the expression levels of hub genes and the overall prognosis of the patients. Higher levels of RPS17 and RPL35A are linked to shorter survival, while RPS19 and RPS25 expression at lower levels is associated with a shorter survival rates. This indicates that these genes can be used as prognostic biomarkers. Moreover, the involvement of these genes in MM pathobiology was further highlighted by their involvement in essential cellular functions such as cell proliferation, apoptosis, stress response and translation. The phylogenetic conservation of these genes adds further weight to their fundamental biological roles. Collectively, the results points to the clinical significant of these genes and their potential use as targets for treatment. This study points out how these biomarkers could be useful in the treatment and prediction of MM. Studying MM with both bioinformatics analysis and experimental validation will reveal the important molecular mechanism involved in its progression. This work advances the growing field of computational oncology by providing a set of candidate genes for future functional validation and translational studies. The experimental validation of this research prepares the ground for new studies on how to improve patient management strategies.

Future Directions: Looking ahead, this study opens up several significant avenues for further research directions. First, long-term studies that examine how the expression of these biomarkers changes over the progression of disease would provide crucial insight into their dynamic roles in MM. Understanding the evolution process of these biomarkers could help us predict patient outcomes and tailor treatment strategies. Second, investigating treatment-biomarker interactions related to RPS17, RPS19, RPS25 and RPL35A offer crucial information on how these genes influence treatment outcomes and resistance mechanisms. These analyses may aid in developing individualized treatment regimens that optimize therapeutic efficacy based on the distinct

biomarker profiles of each patient. Lastly, further research is required to identify the functional role of these ribosomal proteins in the pathophysiology of MM. Analyzing their role in biological processes like apoptosis, proliferation and stress responses may lead to novel treatment strategies in the discovery of new therapeutic targets.

In conclusion, this study provides a foundation for further research into ribosomal protein biomarkers in MM. Our understanding of their roles in disease progression and response to treatment will guide us to create more personalized and effective treatment strategies for MM patients.

Chapter 7

References

[1] O. I. Awe, N. En najih, M. N. Nyamari, and L. B. Mukanga, "Comparative study between molecular and genetic evolutionary analysis tools using African SARS-CoV-2 variants," *Informatics Med. Unlocked*, 2023, doi: 10.1016/j.imu.2022.101143.

[2] J. Daugelaite, A. O' Driscoll, and R. D. Sleator, "An Overview of Multiple Sequence Alignments and Cloud Computing in Bioinformatics," *ISRN Biomath.*, 2013, doi: 10.1155/2013/615630.

[3] X. Qiu *et al.*, "Identification of FCER1G as a key gene in multiple myeloma based on weighted gene co-expression network analysis," *Hematol. (United Kingdom)*, 2023, doi: 10.1080/16078454.2023.2210904.

[4] P. Xiao *et al.*, "SSBP1 is a novel prognostic marker and promotes disease progression via p38MAPK signaling pathway in multiple myeloma," *Mol. Carcinog.*, 2024, doi: 10.1002/mc.23684.

[5] S. K. R. Mukkamalla and D. Malipeddi, "Myeloma bone disease: A comprehensive review," *International Journal of Molecular Sciences*. 2021, doi: 10.3390/ijms22126208.

[6] J. P. Abeykoon, R. K. Tawfiq, S. Kumar, and S. M. Ansell, "Monoclonal gammopathy of undetermined significance: evaluation, risk assessment, management, and beyond," *Fac. Rev.*, 2022, doi: 10.12703/r/11-34.

[7] M. Ho, A. Patel, C. Y. Goh, M. Moscvin, L. Zhang, and G. Bianchi, "Changing paradigms in diagnosis and treatment of monoclonal gammopathy of undetermined significance (MGUS) and smoldering multiple myeloma (SMM)," *Leukemia*. 2020, doi: 10.1038/s41375-020-01051-x.

[8] G. Gkoliou *et al.*, "Differences in the immunoglobulin gene repertoires of IgG versus IgA multiple myeloma allude to distinct immunopathogenetic trajectories," *Front. Oncol.*, 2023, doi: 10.3389/fonc.2023.1123029.

[9] J. Nunnelee *et al.*, "Improvement in Post-Autologous Stem Cell Transplant Survival of Multiple Myeloma Patients: A Long-Term Institutional Experience," *Cancers (Basel).*, 2022, doi: 10.3390/cancers14092277.

[10] S. V. Rajkumar, "Multiple myeloma: 2020 update on diagnosis, risk-stratification and

management," *Am. J. Hematol.*, 2020, doi: 10.1002/ajh.25791.

[11] R. Kotb, C. Hart, and H. Goubran, "Multiple Myeloma," in *Paraproteinemia and Related Disorders*, 2022.

[12] N. Bolli, G. Martinelli, and C. Cerchione, "The molecular pathogenesis of multiple myeloma," *Hematol. Rep.*, 2020, doi: 10.4081/HR.2020.9054.

[13] M. A. Dimopoulos *et al.*, "Multiple myeloma: EHA-ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up†," *Ann. Oncol.*, 2021, doi: 10.1016/j.annonc.2020.11.014.

[14] S. V. Rajkumar and S. Kumar, "Multiple myeloma current treatment algorithms," *Blood Cancer J.*, 2020, doi: 10.1038/s41408-020-00359-2.

[15] G. Kozalak, İ. Bütün, E. Toyran, and A. Koşar, "Review on Bortezomib Resistance in Multiple Myeloma and Potential Role of Emerging Technologies," *Pharmaceuticals*. 2023, doi: 10.3390/ph16010111.

[16] N. Giuliani *et al.*, "Novel targets for the treatment of relapsing multiple myeloma," *Expert Review of Hematology*. 2019, doi: 10.1080/17474086.2019.1624158.

[17] S. V. Rajkumar and S. Kumar, "Multiple Myeloma: Diagnosis and Treatment," *Mayo Clinic Proceedings*. 2016, doi: 10.1016/j.mayocp.2015.11.007.

[18] S. A. Padala *et al.*, "Epidemiology, Staging, and Management of Multiple Myeloma," *Medical sciences (Basel, Switzerland)*. 2021, doi: 10.3390/medsci9010003.

[19] S. E. Mousavi, M. Ilaghi, A. Aslani, Z. Yekta, and S. A. Nejadghaderi, "A population-based study on incidence trends of myeloma in the United States over 2000–2020," *Sci. Rep.*, 2023, doi: 10.1038/s41598-023-47906-y.

[20] H. Ludwig, S. Novis Durie, A. Meckl, A. Hinke, and B. Durie, "Multiple Myeloma Incidence and Mortality Around the Globe; Interrelations Between Health Access and Quality, Economic Resources, and Patient Empowerment," *Oncologist*, 2020, doi: 10.1634/theoncologist.2020-0141.

[21] A. Palumbo and K. Anderson, "Medical Progress Multiple Myeloma," *N Engl J Med*,

2011.

[22] S. K. Kumar *et al.*, "Continued improvement in survival in multiple myeloma: Changes in early mortality and outcomes in older patients," *Leukemia*, 2014, doi: 10.1038/leu.2013.313.

[23] J. Bladé *et al.*, "Extramedullary disease in multiple myeloma: a systematic literature review," *Blood Cancer Journal*. 2022, doi: 10.1038/s41408-022-00643-3.

[24] J. Geng *et al.*, "Global, regional, and national burden and quality of care of multiple myeloma, 1990-2019," *J. Glob. Health*, 2024, doi: 10.7189/JOGH.14.04033.

[25] W. C. Yang and S. F. Lin, "Mechanisms of Drug Resistance in Relapse and Refractory Multiple Myeloma," *BioMed Research International*. 2015, doi: 10.1155/2015/341430.

[26] P. Sonneveld *et al.*, "Treatment of multiple myeloma with high-risk cytogenetics: A consensus of the International Myeloma Working Group," *Blood*. 2016, doi: 10.1182/blood-2016-01-631200.

[27] A. Ahmed *et al.*, "Outcomes of Salvage VDT-PACE-like Regimens in Relapsed-Refractory Multiple Myeloma: 10-Year Experience of a Large Academic Institution," *Blood*, 2023, doi: 10.1182/blood-2023-182972.

[28] M. Ortiz-Estévez *et al.*, "Integrative multi-omics identifies high risk multiple myeloma subgroup associated with significant DNA loss and dysregulated DNA repair and cell cycle pathways," *BMC Med. Genomics*, 2021, doi: 10.1186/s12920-021-01140-5.

[29] J. Chen and L. Zhang, "A survey and systematic assessment of computational methods for drug response prediction," *Briefings in Bioinformatics*. 2021, doi: 10.1093/bib/bbz164.

[30] S. Ovejero and J. Moreaux, "Multi-omics tumor profiling technologies to develop precision medicine in multiple myeloma," *Explor. Target. Anti-tumor Ther.*, 2021, doi: 10.37349/etat.2020.00034.

[31] J. G. Lohr *et al.*, "Widespread genetic heterogeneity in multiple myeloma: Implications for targeted therapy," *Cancer Cell*, 2014, doi: 10.1016/j.ccr.2013.12.015.

[32] A. Broyl *et al.*, "Gene expression profiling for molecular classification of multiple

myeloma in newly diagnosed patients," *Blood*, 2010, doi: 10.1182/blood-2009-12-261032.

[33] N. Amodio, P. D'Aquila, G. Passarino, P. Tassone, and D. Bellizzi, "Epigenetic modifications in multiple myeloma: recent advances on the role of DNA and histone methylation," *Expert Opinion on Therapeutic Targets*. 2017, doi: 10.1080/14728222.2016.1266339.

[34] A. M. Poos *et al.*, "Resolving therapy resistance mechanisms in multiple myeloma by multiomics subclone analysis," *Blood*, 2023, doi: 10.1182/blood.2023019758.

[35] T. Barrett *et al.*, "NCBI GEO: Archive for functional genomics data sets - Update," *Nucleic Acids Res.*, 2013, doi: 10.1093/nar/gks1193.

[36] D. Sean and P. S. Meltzer, "GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor," *Bioinformatics*, 2007, doi: 10.1093/bioinformatics/btm254.

[37] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nat. Protoc.*, 2009, doi: 10.1038/nprot.2008.211.

[38] P. Langfelder and S. Horvath, "Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinform 9: 559," *BMC Bioinformatics*, 2009.

[39] P. Shannon *et al.*, "Cytoscape: A software Environment for integrated models of biomolecular interaction networks," *Genome Res.*, 2003, doi: 10.1101/gr.1239303.

[40] G. P. Rédei, "CLUSTAL W (improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice)," in *Encyclopedia of Genetics, Genomics, Proteomics and Informatics*, 2008.

[41] K. Tamura, G. Stecher, and S. Kumar, "MEGA11: Molecular Evolutionary Genetics Analysis Version 11," *Mol. Biol. Evol.*, 2021, doi: 10.1093/molbev/msab120.

[42] X. Robin *et al.*, "pROC: An open-source package for R and S+ to analyze and compare ROC curves," *BMC Bioinformatics*, 2011, doi: 10.1186/1471-2105-12-77.

[43] G. B., L. A., E. A.C., D. C., B. J., and L. Q., "An online survival analysis tool to rapidly

assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients," *Breast Cancer Res. Treat.*, 2010.

[44] W. C. Cheng *et al.*, "Microarray meta-analysis database (M2DB): A uniformly pre-processed, quality controlled, and manually curated human clinical microarray database," *BMC Bioinformatics*, 2010, doi: 10.1186/1471-2105-11-421.

[45] K. Rathnam, S. V. Saju, and S. R. Honey, "Management of Relapsed and Refractory Multiple Myeloma: Recent advances," *Indian Journal of Medical and Paediatric Oncology*. 2022, doi: 10.1055/s-0042-1758537.

[46] P. Bhatt, C. Kloock, and R. Comenzo, "Relapsed/Refractory Multiple Myeloma: A Review of Available Therapies and Clinical Scenarios Encountered in Myeloma Relapse," *Current Oncology*. 2023, doi: 10.3390/curroncol30020179.

[47] D. Ribatti, "A historical perspective on milestones in multiple myeloma research," *European Journal of Haematology*. 2018, doi: 10.1111/ejh.13003.

[48] F. Maura *et al.*, "Genomic landscape and chronological reconstruction of driver events in multiple myeloma," *Nat. Commun.*, 2019, doi: 10.1038/s41467-019-11680-1.

[49] M. Heider, K. Nickel, M. Högner, and F. Bassermann, "Multiple Myeloma: Molecular Pathogenesis and Disease Evolution," *Oncology Research and Treatment*. 2021, doi: 10.1159/000520312.

[50] A. Seesaghur *et al.*, "Clinical features and diagnosis of multiple myeloma: A population-based cohort study in primary care," *BMJ Open*, 2021, doi: 10.1136/bmjopen-2021-052759.

[51] N. C. Munshi *et al.*, "Idecabtagene Vicleucel in Relapsed and Refractory Multiple Myeloma," *N. Engl. J. Med.*, 2021, doi: 10.1056/nejmoa2024850.

[52] I. J. Cardona-benavides, C. de Ramón, and N. C. Gutiérrez, "Genetic abnormalities in multiple myeloma: Prognostic and therapeutic implications," *Cells*. 2021, doi: 10.3390/cells10020336.

[53] J. Zhou, M. Zhang, Y. Zhang, X. Shi, L. Liu, and R. Yao, "Identification of Potential

Prognostic Biomarker for Predicting Survival in Multiple Myeloma Using Bioinformatics Analysis and Experiments," *Front. Genet.*, 2021, doi: 10.3389/fgene.2021.722132.

[54] J. Yang, F. Wang, S. Zhong, and B. Chen, "Identification of hub genes with prognostic values in multiple myeloma by bioinformatics analysis," *Hematol. (United Kingdom)*, 2021, doi: 10.1080/16078454.2021.1943617.

[55] X. Chen, X. Y. He, Q. Dan, and Y. Li, "FAM201A, a long noncoding RNA potentially associated with atrial fibrillation identified by ceRNA network analyses and WGCNA," *BMC Med. Genomics*, 2022, doi: 10.1186/s12920-022-01232-w.

[56] S. Zhao *et al.*, "Comprehensive bioinformatics analysis reveals the hub genes and pathways associated with multiple myeloma," *Hematol. (United Kingdom)*, 2022, doi: 10.1080/16078454.2022.2040123.

[57] M. Xu, Y. Meng, Q. Li, A. Charwudzi, H. Qin, and S. Xiong, "Identification of biomarkers for early diagnosis of multiple myeloma by weighted gene co-expression network analysis and their clinical relevance," *Hematol. (United Kingdom)*, 2022, doi: 10.1080/16078454.2022.2046326.

[58] L. Gong, L. Qiu, and M. Hao, "Novel Insights into the Initiation, Evolution, and Progression of Multiple Myeloma by Multi-Omics Investigation," *Cancers*. 2024, doi: 10.3390/cancers16030498.

[59] E. Schwenger and U. Steidl, "An Evolutionary Approach to Clonally Complex Hematologic Disorders," *Blood Cancer Discovery*. 2021, doi: 10.1158/2643-3230.BCD-20-0219.

[60] E. Taiana *et al.*, "Genomic instability in multiple myeloma: A 'non-coding rna' perspective," *Cancers*. 2021, doi: 10.3390/cancers13092127.

[61] Y. Peng, D. Wu, F. Li, P. Zhang, Y. Feng, and A. He, "Identification of key biomarkers associated with cell adhesion in multiple myeloma by integrated bioinformatics analysis," *Cancer Cell Int.*, 2020, doi: 10.1186/s12935-020-01355-z.

[62] G. S. Goh, W. M. Yue, C. M. Guo, S. B. Tan, and J. L. Chen, "Defining threshold values on the neck disability index corresponding to a patient acceptable symptom state in

patients undergoing elective surgery for degenerative disorders of the cervical spine," *Spine J.*, 2020, doi: 10.1016/j.spinee.2020.05.004.

[63] Z. A. Wainberg *et al.*, "Event-Free Survival as a Surrogate for Overall Survival in Gastric and Gastroesophageal Junction Adenocarcinoma: A Meta-analysis in the Neoadjuvant ∓ Adjuvant Setting," *Clin. Cancer Res.*, 2023, doi: 10.1158/1078-0432.CCR-22-2920.

[64] J. Long *et al.*, "Transcriptional landscape of cholangiocarcinoma revealed by weighted gene coexpression network analysis," *Brief. Bioinform.*, 2021, doi: 10.1093/bib/bbaa224.

[65] R. Su *et al.*, "Construction of a ceRNA network of hub genes affecting immune infiltration in ovarian cancer identified by WGCNA," *BMC Cancer*, 2021, doi: 10.1186/s12885-021-08711-w.

[66] J. A. You, Y. Gong, Y. Wu, L. Jin, Q. Chi, and D. Sun, "WGCNA, LASSO and SVM Algorithm Revealed RAC1 Correlated M0 Macrophage and the Risk Score to Predict the Survival of Hepatocellular Carcinoma Patients," *Front. Genet.*, 2022, doi: 10.3389/fgene.2021.730920.

[67] Z. Rezaei *et al.*, "Identification of early diagnostic biomarkers via WGCNA in gastric cancer," *Biomed. Pharmacother.*, 2022, doi: 10.1016/j.biopha.2021.112477.

[68] L. Liu *et al.*, "An interactive nomogram based on clinical and molecular signatures to predict prognosis in multiple myeloma patients," *Aging (Albany. NY).*, 2021, doi: 10.18632/aging.203294.

[69] Y. Shi *et al.*, "Retraction Note: Therapeutic potential of targeting IRES-dependent c-myc translation in multiple myeloma cells during ER stress (Oncogene, (2016), 35, 8, (1015-1024), 10.1038/onc.2015.156)," *Oncogene*. 2023, doi: 10.1038/s41388-023-02820-5.

[70] Y. Zhuang, Y. Zhang, C. Chen, J. Chen, Q. Xu, and P. Wang, "Daptomycin Inhibits Multiple Myeloma Progression through Downregulating the Expression of RPS19," *Comb. Chem. High Throughput Screen.*, 2024, doi: 10.2174/0113862073283460240129104114.

[71] H. Sadaf *et al.*, "Multiple myeloma etiology and treatment," *Journal of Translational Genetics and Genomics*. 2022, doi: 10.20517/jtgg.2021.36.

[72] M. F. O'Donohue, V. Choesmel, M. Faubladier, G. Fichant, and P. E. Gleizes, "Functional dichotomy of ribosomal proteins during the synthesis of mammalian 40S ribosomal subunits," *J. Cell Biol.*, 2010, doi: 10.1083/jcb.201005117.

[73] Y. Zhu, J. Liu, and B. Wang, "Identification of biomarkers in multiple myeloma: A comprehensive study combining microarray analysis and Mendelian randomization," J. Cell. Mol. Med., vol. 28, no. 12, p. e18504, 2024. [Online]. Available: https://doi.org/10.1111/jcmm.18504

[74] L. Zhuge, X. Lin, Z. Fan, M. Jia, C. Lin, M. Zhu, et al., "Global, regional and national epidemiological trends of multiple myeloma from 1990 to 2021: a systematic analysis of the Global Burden of Disease study 2021," Front. Public Health, vol. 13, p. 1527198, 2025.

[75] D. Szklarczyk, K. Nastou, M. Koutrouli, R. Kirsch, F. Mehryary, R. Hachilif, et al., "The STRING database in 2025: protein networks with directionality of regulation," Nucleic Acids Res., vol. 53, no. D1, pp. D730–D737, 2025.

[76] A. J. Madu, C. Nonyelu, H. C. Okoye, and S. Ocheni, "Multiple Myeloma: An In-depth Review of the Historical and Pathogenetic Processes," Asian Hematol. Res. J., vol. 4, no. 4, pp. 21–33, 2021.