Design and Development of AI-based Computational Tools for Identifying Predictive Biomarkers and Signaling Pathways for Blood Cancer

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

VIVEK RUHELA
PhD18202

ADVISORS
PROF. ANUBHA GUPTA, IIIT DELHI
PROF. (DR.) RITU GUPTA, AIIMS, DELHI

DEPARTMENT OF COMPUTATIONAL BIOLOGY
INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI
NEW DELHI - 110020

May 4, 2024

Certificate

This is to certify that the thesis titled **Design and Development of AI-based Computational Tools for Identifying Predictive Biomarkers and Signaling Pathways for Blood Cancer**, submitted by **Vivek Ruhela**, to the Indraprastha Institute of Information Technology, Delhi, for the award of the degree of **Doctor of Philosophy**, is a bonafide record of the research work done by him under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Prof. Anubha Gupta
SBILab, Deptt. of ECE
Indraprastha Institute of Information
Technology Delhi-110020, India

Prof. (Dr.) Ritu Gupta
Laboratory Oncology, Dr. BRAIRCH
All India Institute of Medical Sciences
New Delhi-110029, India

Place: New Delhi

Date: May 03, 2024

Acknowledgements

With deep gratitude and immense joy, I am delighted to announce the successful completion of my thesis. My Ph.D. journey has been a remarkable period of growth and learning, and this achievement would not have been possible without the invaluable support and assistance of numerous individuals. Expressing my gratitude to them feels like an understatement.

I am profoundly thankful to Prof. Anubha Gupta for her belief in my research abilities and for providing me with continuous inspiration and encouragement throughout this academic journey. Her insightful feedback and dedication to research have transformed me into an independent and confident researcher. I am eternally thankful for her kindness, invaluable guidance, and unwavering assistance. I am also indebted to Prof. Ritu Gupta for her innovative insights and recommendations, which significantly enhanced the quality of my research. Her guidance and the resources she made available to me were invaluable in developing my thesis.

My warmest gratitude extends to Dr. Gurvinder Kaur and Dr. Lingaraja Jena for their insightful discussions, consistent support, and invaluable contributions throughout my Ph.D. journey. I would like to acknowledge Dr. Sriram K. for generously sharing his time and providing constructive feedback, which enriched my work. I have gained valuable insights through our discussions. I sincerely thank my Ph.D. monitoring committee members, Prof. G. P. S. Raghava and Dr. Ganesh Bagler, for their invaluable observations and insightful suggestions on my research.

I wish to extend my heartfelt appreciation to my parents, Sh. Mahesh Kumar Ruhela and Smt. Maya Devi Ruhela, my elder brother, Gaurang Ruhela, and my elder sister, Shweta Ruhela. Their unwavering support, love, and encouragement have been the cornerstone of my journey. The unconditional love, motivation, and support from my wife, Naina Naiwar, and my kid, Vivansh Ruhela, have been my pillars of strength throughout this challenging endeavour. Their sacrifices, understanding, and patience have been my driving force, and I am profoundly grateful for their presence in my life.

I also wish to acknowledge the enduring friendships and the collaborative spirit of my fellow researchers and colleagues (Neetesh Pandey, Ashutosh Vaish, Sachin Motwani, Priyadarshini Rai, Sarita Poonia, and Omkar Chandra), who have shared this journey with me. Your camaraderie and shared experiences have enriched this phase of my life.

Last but not the least, I offer my sincere gratitude to all those who have supported me in ways both seen and unseen. This accomplishment reflects the collective effort of many, and I am humbled by the support and encouragement I have received.

RuhVik

Abstract

Blood cancer has emerged as a growing concern over the past decade, necessitating early detection for timely and effective treatment. Traditional methods of diagnosing blood cancers involve a series of pathological tests and consultations with medical experts, a process that is not only time-consuming but also financially burdensome. The advent of genomic data analysis offers a promising avenue for understanding the pathogenesis of blood cancers, providing valuable insights into crucial biomarkers that could serve as potential therapeutic targets, ultimately impeding the progression of the disease. In the scope of this study, we have delved into the genomic intricacies of two prominent blood cancer types: Chronic Lymphocytic Leukemia (CLL) and Multiple Myeloma (MM). The treatment decisions for CLL and MM rely heavily on patient symptoms and are underpinned by the genetic anomalies in the patient's genome. Here, we have undertaken a comprehensive omics data analysis, employing novel pipelines and methodologies developed in-house. Our objective has been to unearth the genetic aberrations that underlie these diseases' development and identify pivotal biomarkers that hold promise as therapeutic targets for each category of haematological malignancy.

Our first objective was to identify clinically relevant small non-coding RNAs (sncRNAs) in CLL through a comprehensive genome-wide study of RNA-Seq data. This analysis revealed a distinct pattern of dysregulated miRNAs in the CLL cohort. Among these, three miRNAs were up-regulated (hsa-mir-1295a, hsa-mir-155, and hsa-mir-4524a), while five miRNAs were down-regulated (hsa-mir-30a, hsa-mir-423, hsa-mir-486*, hsa-let-7e, and hsa-mir-744). Moreover, our investigation identified seven novel miRNA sequences with elevated expression in CLL, including tRNAs, piRNAs (piRNA-30799, piRNA-36225), and snoRNAs (SNORD43). Notably, we observed a significant correlation between the increased expression of hsa-mir-4524a and a shorter time to first treatment (TTFT) (HR: 1.916, 95% CI: 1.080–3.4, p-value: 0.026) and higher expression of hsa-mir-744 with a longer TTFT (HR: 0.415, 95% CI: 0.224–0.769, p-value: 0.005) in CLL patients. These findings suggest that further research may establish the potential integration of these differentially expressed miRNA (DEM) markers into risk stratification models and prognostic approaches for CLL.

We proceeded by developing an integrated and reproducible workflow for RNA-Seq data analysis, known as miRPipe. This pipeline was designed to identify dysregulated sncRNAs, including miRNAs and piRNAs, and functionally similar miRNAs, often

called miRNA paralogues. To evaluate the performance and benchmark miRPipe, we introduced an in-house synthetic sequence simulator called miRSim. miRSim utilizes seed and xseed information from sncRNA sequences to generate synthetic sequences. Additionally, it provides ground-truth data in a user-friendly comma-separated file format, offering comprehensive information on known miRNAs, piRNAs, novel miRNAs, their sequences, chromosome locations, expression counts, and CIGAR strings for all sequences. We rigorously benchmarked miRPipe against seven existing state-of-the-art pipelines using synthetic and publicly available real RNA-Seq expression datasets (lung cancer, breast cancer, and CLL). In synthetic datasets, miRPipe demonstrated superior performance to existing pipelines, achieving an accuracy of 95.23% and an F1-score of 94.17%. Furthermore, our analysis of all three cancer datasets indicated that miRPipe excelled in extracting a more significant number of known dysregulated miRNAs and piRNAs than existing pipelines.

Then, we designed an innovative AI-driven bio-inspired deep learning architecture to identify altered signaling pathways (BDL-SP) and determine the pivotal genomic biomarkers that can distinguish MM and its precursor stage, named Monoclonal gammopathy of undetermined significance (MGUS). The proposed BDL-SP model comprehends gene-gene interactions using the protein-protein interaction (PPI) network and analyzes genomic features using deep learning (DL) architecture to identify significantly altered genes and signaling pathways in MM and MGUS. The exome sequencing data of 1174 MM and 61 MGUS patients were analyzed for this. In the quantitative benchmarking with the other popular machine learning models, BDL-SP performed almost similarly to the best-performing predictive machine learning (ML) models of Random Forest and CatBoost. However, an extensive post-hoc explainability analysis, capturing the application-specific nuances, clearly established the significance of the BDL-SP model. This analysis revealed that BDL-SP identified a maximum number of previously reported oncogenes (OG), tumour-suppressor genes (TSG), both oncogene and driver gene (ODGs) and actionable genes (AGs) of high relevance in MM as the top significantly altered genes.

Further, the post-hoc analysis revealed a significant contribution of single nucleotide variants (SNVs) and genomic features associated with synonymous SNVs in disease stage classification. Finally, the pathway enrichment analysis of the top significantly altered genes showed that many cancer pathways are selectively and significantly dysregulated in MM compared to its precursor stage of MGUS. At the same time, a few that lost their significance with disease progression from MGUS to MM were related to the other disease types. These observations may pave the way for appropriate therapeutic interventions to halt the progression to overt MM in the future.

Lastly, we designed a curated, comprehensive, targeted sequencing panel focusing on

282 MM-relevant genes and employing clinically oriented NGS-targeted sequencing approaches. To identify these 282 MM-relevant genes, we designed an innovative AI-based Biological Network for Directed Gene-Gene Interaction Learning (BIO-DGI) model for detecting biomarkers and gene interactions that can potentially differentiate MM from MGUS. The BIO-DGI model leverages gene interactions from nine PPI networks and analyzes the genomic features from 1154 MM and 61 MGUS samples. The proposed model outperformed baseline ML and DL models, demonstrating quantitative and qualitative superiority by identifying the largest number of MM-relevant genes in the post-hoc analysis. The pathway analysis underscored the importance of top-ranked genes by highlighting the MM-relevant pathways as the top-significantly altered pathways. The 282-gene panel encompasses 9272 coding regions and has a length of 2.577 Mb.

Additionally, the 282-gene panel showcased superior performance compared to previously published panels, excelling in detecting genomic and transformative events. Notably, the proposed gene panel also highlighted highly influential genes and their interactions within gene communities in MM. The clinical relevance is confirmed through a two-fold univariate survival analysis. The study's findings shed light on essential gene biomarkers and their interactions, providing valuable insights into disease progression.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AD	Allele Depth
AI	Artificial Intelligence
AIIMS	All India Institute of Medical Sciences
ALL	Acute Lymphocytic Leukemia
ATL	Adult T-cell Leukemia/Lymphoma
AUC	Area Under the Curve
AUPRC	Area Under the Precision-Recall Curve
BM	Bone Marrow
BAM	Binary Alignment Map
BDL-SP	Bio-Inspired Deep Learning-based architecture for identification of altered signaling pathways
BIO-DGI	Bio-Inspired Graph Network Learning-based Gene-Gene Interaction
CNN	Convolutional Neural Network
CNV	Copy-Number Variation
CLL	Chronic Lymphocytic Leukemia
CS-Cat	Cost-Sensitive CatBoost
CSDT	Cost-Sensitive Decision Tree
CSLR	Cost-Sensitive Logistic Regression
CS-SVC	Cost-Sensitive Support Vector Machine
CS-XGB	Cost-Sensitive XGBoost
CTL	Cytotoxic T lymphocytes
CSRF	Cost-Sensitive Random Forest

DEM	Differentially Expressed miRNAs
DL	Deep Learning
EGA	European Genome-phenome Archive
EHR	Electronic Health Records
FAMD	Factor Analysis of mixed data
FC	Fold-Change
FN	False Negative
FP	False Positive
GATK	Genomic Analysis ToolKit
GCN	Graph Convolutional Networks
GDC	Genomic Data Commons
GDL	Geometric Deep Learning
GEO	Gene Expression Omnibus
GHIS	Genome-wide Haploinsufficiency Score
GVHD	Graft-Versus-Host Disease
HTLV-1	Human T-cell leukemia virus 1
Ig	Immunoglobulin
IGH	Immunoglobulin Heavy Chain
IGHV	Immunoglobulin Heavy Chain Variable region
Indel	Insertions and Deletions
IPI	International Prognostic Index
KEGG	Kyoto Encyclopedia of Genes and Genomes
KSHV	Kaposi sarcoma-associated herpesvirus
LA	Leiden Algorithm
LNA	Locked Nucleic Acid

LIME	Local Interpretable Model-agnostic Explanations
LOF	Loss-of-Function
lncRNA	Long Non-Coding RNA
MGUS	Monoclonal Gammopathy of Undetermined Significance
ML	Machine Learning
MM	Multiple Myeloma
MNCs	Mononuclear Cells
moRNAs	MicroRNA-offset RNAs
ncRNA	Non-coding RNAs
NDMM	Newly Diagnosed Multiple Myeloma
NGS	Next Generation Sequencing
NK	Nature Killer
NLL	Negative Log Loss
ODG	Both Oncogenes and Driver Genes
OG	Oncogenes
OS	Overall Survival
PCA	Principal Component Analysis
PCL	Plasma Cell Leukemia
PCR	Polymerase Chain Reaction
PFS	Progression Free Survival
piRNA	Piwi-interacting RNAs
PolyPhen-2	Polymorphism Phenotyping v2
PPI	Protein-Protein Interaction
PROVEAN	Protein Variation Effect Analyzer
RIN	RNA Integrity Number

RT-qPCR	Real-time quantitative Polymerase Chain Reaction
SAM	Sequence Alignment Map
ShAP	SHapley Additive exPlanations
SIFT	Sorting Intolerant from Tolerant
smRNA	Small RNA
SMM	Smoldering Multiple Myeloma
sncRNA	Small Non-Coding RNA
SNV	Single Nucleotide Variant
SNP	Single Nucleotide Polymorphisms
snoRNA	small nucleolar RNA
SPMs	Second Primary Malignancies
SV	Structural Variation
TGen	Translational Genomics Research Institute
TN	True Negative
TP	True Positive
tRNA	Transfer RNA
TSG	Tumor-Suppressor Genes
TTFT	Time To First Treatment
VAF	Variant Allele Frequency
VCF	Variant Call Format
VEP	Variant Effect Predictor
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

Chapter 1

Introduction

1.1 Background

Cancer is an intricate and multifaceted spectrum of diseases characterized by the unbridled proliferation and division of cells. It is a formidable global health challenge, with ongoing research continually enriching our understanding of its underlying biology. The increasing global burden of cancer is rapidly outpacing our current capacities for control and intervention. In 2020 alone, there were over 19 million new cancer cases diagnosed worldwide, and cancer claimed the lives of 10 million individuals [6]. Projections from the Global Cancer Observatory indicate that by 2040, the annual total of new cancer cases will increase to approximately 30 million, with 16 million people succumbing to the disease [7]. This relentless progression of cancer is rooted in the accumulation of genetic mutations and epigenetic modifications that disrupt the intricately orchestrated regulatory mechanisms governing cell division and growth. On a molecular level, cancer stems from genetic abnormalities that subvert the finely tuned processes governing cell proliferation. These genetic alterations may be hereditary, induced by environmental factors, or occur spontaneously. Among the common genetic aberrations observed in cancer are mutations in OGs, which drive unchecked cell growth, and mutations in tumor-suppressor genes (TSGs), which act as crucial brakes on this growth.

CLL is the most frequently diagnosed hematologic malignancy worldwide, distinguished by the accumulation of mature B lymphocytes. Its age-standardized incidence rate has exhibited a range from 3.8 to 5.0 per 100,000 person-years since the 2000s [8, 9, 10, 11, 12]. The clinical trajectory of CLL is remarkably heterogeneous, encompassing an indolent form that often remains stable for prolonged periods without the need for treatment, as well as more aggressive variants that exhibit rapid relapse after initial therapeutic interventions. As the population of long-term CLL survivors continues to expand, it becomes imperative to gain insight into their long-term health outcomes. The emergence of second primary malignancies (SPMs), which refer to cancers diagnosed after the CLL diagnosis, represents a significant concern. These SPMs may contribute to increased morbidity and potentially counteract the extended longevity achieved by CLL patients. Thus, comprehending the nature and extent of SPMs in the context of CLL is indispensable for health-related planning and ongoing surveillance efforts [13, 14, 15, 16, 17].

MM is one of the most prevalent hematologic malignancies on a different hematologic front, marked by the clonal expansion of malignant plasma cells within the bone marrow. In 2018, the number of new MM cases recorded worldwide amounted to 159,985, accounting for 0.9% of all newly diagnosed tumors. Notably, this figure nearly surpassed the corresponding mortality rate, with MM-related deaths accounting for 1.1% of all cancer-related fatalities [18]. The incidence rate of MM exhibits regional disparities, with the highest rates observed in North America, Australia, New Zealand, and Europe. At the same time, Asia, excluding Western Asia, registers the lowest incidence rates [19]. The median age at which MM is diagnosed is approximately 70 years, with 37% of the patients falling below the age of 65, 26% in the age range of 65-74, and 37% at or above the age of 75 [20]. Intriguingly, MM is an exceedingly rare diagnosis in patients under the age of 30, accounting for only 0.02-0.3% of cases, with a slightly higher incidence among males compared to females [21].

Recent advances in sequencing methods have not only illuminated the genomic intricacies and diversity within cancer but also offered promising avenues to impede progression to full-blown malignancy in the future. Traditional sequencing methods, such as Sanger sequencing, were costly and time-consuming. The advent of Next-Generation Sequencing (NGS) technologies, including whole exome sequencing (WES), whole genome sequencing (WGS), RNA-Seq, and more, has revolutionized the field. NGS methods are high-throughput, allowing for the parallel sequencing of multiple genomic regions across numerous samples in a single run. In WGS, the entire genome is sequenced from a large DNA sample. Achieving high sequencing coverage in WGS is essential for detecting clinically relevant mutations, but this can be prohibitively expensive and time-intensive.

On the other hand, WES focuses exclusively on the coding regions (exons) of the genome, which is roughly 2% of the entire human genome. As a result, WES offers a more cost-effective and time-efficient alternative to WGS and is widely utilized in cancer genomics to identify rare and common variants. RNA-Seq plays a pivotal role in deciphering changes in gene expression profiles, identifying alternative gene-spliced transcripts, and detecting gene fusions. NGS technologies are also indispensable for investigating epigenetic modifications, broadening their utility in cancer research.

Various NGS platforms are available today, with Illumina/Solexa, SOLiD (Sequencing by Oligonucleotide Ligation and Detection), and Ion Torrent being well-known options. Additionally, there are third and fourth-generation sequencers, such as PacBio sequencing and nanopore sequencing, which, while slightly less accurate than NGS, boast longer read lengths, making them suitable for specific applications. The 10x genomics technology in 2016 enabled the genome and transcriptome analysis at a single-cell level, utilizing the Chromium system. These sequencing platforms continue to evolve, becoming faster,

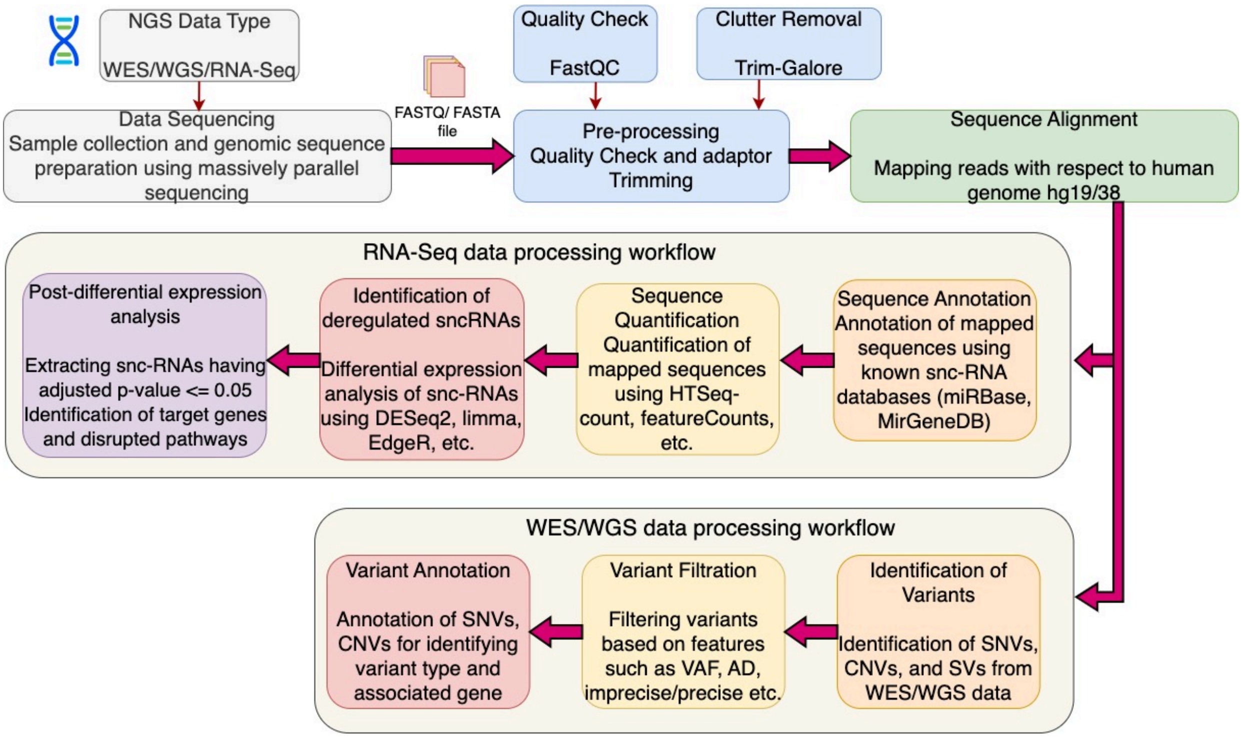

NGS Data Type
WES/WGS/RNA-Seq
Data Sequencing
Sample collection and genomic sequence preparation using massively parallel sequencing
FASTQ/ FASTA file
Quality Check
FastQC
Clutter Removal
Trim-Galore
Pre-processing
Quality Check and adaptor Trimming
Sequence Alignment
Mapping reads with respect to human genome hg19/38
RNA-Seq data processing workflow
Sequence Annotation
Annotation of mapped sequences using known snc-RNA databases (miRBase, MirGeneDB)
Sequence Quantification
Quantification of mapped sequences using HTSeq-count, featureCounts, etc.
Identification of deregulated sncRNAs
Differential expression analysis of snc-RNAs using DESeq2, limma, EdgeR, etc.
Post-differential expression analysis
Extracting snc-RNAs having adjusted p-value <= 0.05
Identification of target genes and disrupted pathways
WES/WGS data processing workflow
Identification of Variants
Identification of SNVs, CNVs, and SVs from WES/WGS data
Variant Filtration
Filtering variants based on features such as VAF, AD, imprecise/precise etc.
Variant Annotation
Annotation of SNVs, CNVs for identifying variant type and associated gene

Figure 1.1: An overview of the NGS pipeline for processing RNA-Seq and DNA-Seq data. The workflow for both RNA-Seq and DNA-Seq begins with data generation and pre-processing of raw sequencing data, provided in FASTQ/FASTA file format. During data generation, nucleotide base calls are extracted from the sample using massively parallel sequencing, and the resulting genomic information is stored in FASTQ/FASTA files. In the pre-processing phase, FASTQ/FASTA files undergo quality checks using FASTQC [3] to retain high-quality reads. Subsequently, high-quality reads are denoised by trimming adaptor sequences with Trim-Galore. The trimmed reads are then aligned with respect to the known human genome (hg19/38) using sequence alignment tools, yielding BAM files. For RNA-Seq data processing, BAM files are further processed to annotate sequences using known sncRNA databases. Annotated reads are counted to determine the expression levels of known sncRNAs, which are compared with sncRNA expression counts from healthy samples to identify deregulated sncRNAs. Once deregulated sncRNAs are determined, post-differential expression analysis is performed to identify associated signaling pathways and target genes. In the case of DNA-Seq data processing, mapped reads in the BAM files are further processed to identify SNVs, copy number variants (CNVs), and structural variants (SVs). Identified variants are filtered based on features such as allele depth (AD), variant allele frequency (VAF), etc. The filtered variants (SNVs, CNVs, and SVs) are then annotated to identify associated genes.

more efficient, and cost-effective. Consequently, they generate vast amounts of NGS data, necessitating proficient computational and bioinformatics skills for data analysis. Subsequently, substantial advancements have been made in data storage and management capabilities and the development of computational methods to process and derive insights from this massive genomic dataset. This evolving synergy between NGS technologies and computational advancements paves the way for deeper insights into cancer genomics and other complex biological processes.

The comprehensive NGS workflow for RNA and DNA-Seq data is depicted in Figure-1.1, encompassing multiple critical steps. The journey commences with sample collection and the generation of sequencing data, followed by the subsequent stages. Initially, the collected sample is meticulously preserved at -80°C to maintain its integrity. Subsequently, the sample undergoes processing to extract either RNA or DNA, which is then transformed into a library of fragments suitable for sequencing. High-throughput sequencing systems, such as Illumina, are employed to sequence this library, yielding raw sequencing data as FASTQ/FASTA files. The next phase involves the essential pre-processing of the raw sequencing data, focusing on quality assurance and adaptor trimming. This pivotal step encompasses various operations, including filtering and adaptor trimming, all preceded by a thorough quality assessment of the sequenced reads. Tools like FastQC [3] are commonly employed for this purpose, generating comprehensive reports that furnish extensive insights into the quality and characteristics of the FASTQ files.

The information obtained from the FastQC report guides the subsequent filtering of reads. This filtering process is based on parameters such as base call quality (Phred score) and read length, ensuring that only high-quality and relevant reads are retained for further analysis. Moreover, the final aspect of this pre-processing stage involves the removal of library adaptor sequences from the ends of the reads. This critical step is imperative to prevent interference with subsequent processes, including read mapping and assembly. Trim-Galore [22] stands out as a widely employed tool for proficiently executing this adaptor trimming process, thereby streamlining data preparation for downstream analyses.

After pre-processing the raw sequencing data, the reads are aligned to the known human reference genome (hg19/38) using alignment tools like BWA [23], Bowtie1 [24], Bowtie2 [25] or similar tools. This alignment process generates SAM (Sequence Alignment Map) or BAM (Binary Alignment Map) files. Subsequent post-alignment data processing involves extracting alignment information and removing poorly mapped reads from the SAM/BAM files, typically using tools like SAMTools [26]. For RNA-Seq data, the BAM files are further processed to estimate the expression counts of sncRNAs. In the case of DNA-Seq data, the BAM files are processed to identify variants, including SNVs, CNVs, or SVs. In RNA-Seq data analysis, the mapped reads are annotated using well-established sncRNA databases such as miRBase [4], piRNAdb database (version 1.7.6) (https://www.pirnadb.org/), DASHR [27], snoDB 2.0 [28], GtRNAdb [29], and others. Subsequently, the annotated reads are counted to determine the expression levels of sncRNAs. This counting process is typically performed using tools like HTSeq-count [30]. To identify deregulated sncRNAs (those that are up-regulated or down-regulated), the sncRNA counts obtained in the previous step are compared with their corresponding counts in healthy samples. Differential expression analysis tools such as DESeq2 [31], edgeR [32], limma [33], and others are commonly employed for this purpose. Once deregulated sncRNAs are identified, their target genes and associated signaling pathways can be determined using specialized databases like miRNet 2.0 [34]. This comprehensive analytical pipeline allows for extracting valuable insights from NGS data, shedding light on the regulatory roles of sncRNAs in various biological processes.

In DNA-Seq data analysis, several intermediate steps are crucial before the variant calling step after the reads have been aligned to the reference genome. These steps include duplicate removal, local read alignment, base quality recalibration, etc. During library preparation, the Polymerase Chain Reaction (PCR) technique can generate duplicate reads, potentially introducing false positives. To mitigate this, duplicate reads are identified and removed from the analysis using Picard tools. The presence of insertions and deletions (InDels) can lead to read mismatches, and thus, local read alignment is employed to reduce such discrepancies. Additionally, base quality scores, which are generated by the sequencing platform, can be affected by various factors, such as sequencing errors, library preparation artefacts, etc. As a result, these scores may not accurately represent the true base-calling error rate. Therefore, recalibrating the base quality scores is essential to enhance the accuracy of variant calling. This recalibration process is typically performed using tools like the Genome Analysis Toolkit (GATK) [35]. Following these pre-processing steps, variant identification begins in the post-processed BAM file. Various variant callers, relying on Bayesian approaches, likelihood calculations, or machine learning algorithms, have evolved significantly in recent years. Most variant callers produce variant calling format (VCF) files as their output. The identified variants can encompass a wide range, including SNVs, InDels, and complex chromosomal alterations like translocations, inversions, and CNVs. To identify SNVs and INDELs from WES or WGS data, tools such as MuSE [36], Mutect2 [37], SomaticSniper [38], and VarScan2 [39] are commonly employed. On the other hand, to detect CNVs and SVs from WGS data, tools like Delly [40], BreakDancer [41], and Pindel [42] are used. In the subsequent variant annotation step, SNVs identified in the previous stage are subjected to annotation using tools such as ANNOVAR [43] and Variant Effect Predictor (VEP) [44]. These annotation tools take the VCF files obtained from the variant callers and provide valuable information, including the precise chromosomal location of the variants and their biological impact. The biological impact is crucial for determining whether the variant is missense, nonsense, or synonymous or if it results in a stop-gain or stop-loss mutation, among other possibilities. By filtering variants based on their impact, the analysis becomes more focused, increasing the likelihood of identifying actionable or driver variants. When the objective is to uncover rare disease-causing variants within the dataset, it is common practice to exclude variants commonly found in the general population, such as Single Nucleotide Polymorphisms (SNPs). SNPs are single nucleotide variants present in at least 1% of the population. Databases like dbSNP [45] are typically used for this purpose, allowing the removal of these common variants from the analysis. To further refine the selection of variants, multiple scoring systems, such as SIFT [46], Polyphen [47], FATHMM-XF [48], and CADD [49], PROVEAN [50] are employed to filter out benign variants. These scoring systems assess the potential functional impact of the variants based on factors like protein structure, sequence conservation, and predicted pathogenicity. In addition to these filters, population databases like COSMIC [51], ClinVar [52], and OncoKB [53] play a vital role in determining the clinical relevance and association of variants. These databases provide information on known associations of variants with diseases, especially cancer-related ones. These tools and methods have become indispensable in identifying and characterizing genetic variants, providing valuable insights into the genomic basis of various diseases, including cancer.

The sequencing data furnishes a wealth of valuable information, enabling the identification of up-regulated and down-regulated genes, detection of genomic variants, copy number variations, and chromosomal alterations. As depicted in Figure-1.1, a wide array of computational methods is continually being developed to analyze genomic data with increasing precision and accuracy to identify genomic biomarkers. In the context of CLL, many studies have delved into the transcriptional profiling of miRNAs, unveiling numerous DEMs. Intriguingly, our focus gravitated toward comprehending the collective influence of DEMs, alongside small nucleolar RNAs (snoRNAs), PIWI-interacting RNAs (piRNAs), and transfer RNAs (tRNAs), on CLL survival outcomes. As we ventured further into sncRNA data analysis, we encountered a critical challenge concerning the reliability and reproducibility of workflows used to identify deregulated sncRNAs. To address this challenge comprehensively, we delved deeper into sncRNA analysis, culminating in designing a unified, reproducible, and highly accurate workflow for identifying known and novel miRNAs and piRNAs.

Additionally, our explorations extended to other subtypes of blood cancer, mainly focusing on MM and MGUS, incorporating WES data analysis. While analyzing the WES data from MM and MGUS samples, we developed a bio-inspired model for pinpointing

Table 1.1: List of challenging problems and their computational workflow designed to address these problems.

Problems	Studied data types	Outcomes
RNA-Seq profiling of deregulated miRNAs in CLL and their impact on clinical outcome	NGS RNA-Seq data and CLL patients survival data	1. We proposed a unique sncRNA signature containing deregulated known and novel miRNAs, tRNAs, piRNAs and snoRNAs to characterize their impact on the clinical outcomes in CLL. 2. Multivariate analysis showed that miR-4524a (HR: 1.916, 95% CI: 1.080–3.4, p-value: 0.026) and miR-744 (HR: 0.415, 95% CI: 0.224–0.769, p-value: 0.005) were significantly associated with risk and time to first treatment.
Reproducible workflow for accurate identification of novel miRNAs	NGS RNA-Seq data	1. We developed an in-house seed-based small non-coding synthetic RNA-Seq simulator, called miRSim, to generate synthetic RNA-Seq data with the help of seed and xseed region information from miRNA sequence. 2. We developed a unified workflow, miRPipe, for the accurate and reliable identification of miRNAs and piRNAs from Next-Generation Sequencing RNA-Seq data.
Identification of pivotal biomarkers that can distinguish MM and MGUS	NGS whole exome sequencing data	1. We designed an innovative AI-based model, the Bio-inspired Deep Learning architecture, to identify altered Signaling Pathways (BDL-SP) and discover pivotal genomic biomarkers that can potentially distinguish MM from MGUS. 2. Using the application-aware interpretability analysis of the trained AI model, we demonstrated a way to identify the best AI model from among the multiple machine learning or deep learning models that may have performed similarly on the quantitative metrics on the available data.
A Directed Gene-Gene Interactions in Multiple Myeloma	NGS whole exome and whole genome sequencing data and MM patients survival data	1. We designed a clinically oriented targeted panel of 282 genes that potentially cause MGUS-to-MM transition and influence survival outcomes in MM. 2. We designed an attention-based graph neural network, namely BIO-DGI, that extracts gene-gene interactions utilizing a-priori information from nine PPI databases. 3. The proposed BIO-DGI model outperformed baseline machine learning (ML) and deep learning (DL) models, demonstrating quantitative and qualitative superiority by identifying the highest number of MM-relevant genes in the post-hoc analysis. 4. We identified five gene communities using five distinct learned PPI adjacency matrices from five trained BIO-DGI classifiers.

pivotal biomarkers capable of distinguishing MM from MGUS, a critical endeavour given the evolving landscape of these hematologic malignancies. Further, we employed the agnostic models for bio-inspired model post-hoc explainability to pinpoint the pivotal biomarkers (genes and genomic features) to distinguish MM from MGUS. The agnostic models represent interpretability approaches aimed at understanding the predictive response of a model rather than its original response. These models are flexible and independent of the original ML/DL model, allowing for broad applicability across different models [54]. Intriguingly, we observed the presence of several vital distinguishing genes that appeared to serve as epicentres in the disruption of disease-initiating and disease-transformative genes associated with MM. Our in-depth analysis of MM WES and WGS profiles revealed genes responsible for driving disease progression, further advancing our understanding of the molecular mechanisms at play. The multifaceted challenges addressed in this thesis, spanning diverse types of genomic data, are summarized in Table-1.1.

1.2 Literature Review

CLL is a clinically heterogeneous malignancy where a large molecular inter-individual heterogeneity is observed, which is fundamentally governed by differences in the underlying genetic vulnerabilities of individual cases [55]. Congruent molecular and pathological studies have identified several potential genomic biomarkers for CLL prognosis or response to therapy. The most persistently observed somatic CNVs with prognostic significance in CLL include del(13q14), del(11q22.3), del(17p), trisomy 12, amp (8q24.21), amp(3q26.32) and del(8p) [56]. Recurrent mutations among genes believed to act as putative drivers of CLL, such as *TP53, SF3B1, NOTCH1, MYD88, ATM, SAMHD1, NRAS*, and *BIRC3* have also been shown to exhibit a significant prognostic association [57]. Recent genomic studies using parallel high throughput technologies like NGS and microarrays have revealed that the molecular heterogeneity of CLL is further complicated by alterations in gene expression patterns and epigenetic regulatory events and abundance of sncRNAs such as miRNAs,tRNA, piRNA, snoRNA, etc and long noncoding RNAs (lncRNA) [58, 59]. A plethora of studies on transcriptional profiling of miRNAs have identified a variety of DEMs in CLL [60, 61, 62, 63, 64].

In the landmark study, a 13 miRNA signature was reported in CLL patients with high Zeta-chain-associated protein kinase 70 (ZAP70) expression and unmutated immunoglobulin heavy chain variable region gene (IGHV) status [60]. Differential expression of various miRNAs, including hsa-mir-15a, hsa-mir-16, hsa-mir-29a/b/c, hsa-mir-223 and hsa-mir-150 have been consistently reported to be associated with well-established prognostic factors such as IGHV status, ZAP70/CD38 expression, β2 microglobulin levels and disease progression in CLL [65]. Several studies have delineated miRNA signatures specific to karyotype variations in CLL, enabling discrimination among patients with del(17p), del(11q), del(13q), trisomy 12, and a normal karyotype [66, 67]. In cases of the commonly encountered del(13q14), the concurrent deletion of the tumor suppressive hsa-mir-15a and hsa-mir-16-1 occurs, resulting in increased BCL-2 expression and the initiation of CLL [68]. Del(11q) has been linked to the co-deletion of hsa-mir-34b/c clusters and elevated levels of hsa-mir-769-5p and miR-338-3p. Meanwhile, trisomy 12 is associated with the up-regulation of miR-181a and downregulation of miR-155, miR-148a, and miR-483-5p in CLL [66, 69].

In poor prognostic subgroup with del(17p), differential regulation of various miRNAs such as hsa-mir-34a, hsa-mir-29b/c, hsa-mir-17-5p, hsa-mir-223, hsa-mir-150, hsa-mir-181, hsa-mir-33b, hsa-mir-96, and hsa-mir-21 has been observed [66, 70]. Owing to the noteworthy prognostic potential of miRNAs, cumulative prognostic scores in combination with other prognostic factors have also been proposed in CLL [71, 72]. Considering the growing diverse miRNA repertoire, their immense translational potential and advances in technology for their detection, we have studied whole sncRNA transcriptome for identifying unique patterns of DEMs, targets and deregulated piRNAs and snoRNA-related molecules in CLL.

While a substantial array of computational methods has been devised to analyse exceedingly complex sequencing datasets systematically, the quest for a singular approach that robustly attains the necessary precision in detection and estimation remains unfulfilled. In this relentless pursuit, numerous methods have emerged, undergone rigorous testing, and been deployed. However, it is noteworthy that, to date, these methods face limitations in accessibility as open-source tools and often fall short of the anticipated accuracy in their functionality. The underlying molecular mechanisms by which miRNAs mature and silence their target transcripts have been extensively studied. However, due to their centralized position in regulating key cellular processes, a thorough understanding of their identity and, hence, their function, both under the homeostatic and pathological state, is an ever-daunting task due to the limited availability of computational methods for their reliable detection. Likewise, in cancer, microRNAs have been centrally classified in the networks of OGs and TSGs [73] and, therefore, reported to influence diverse transcripts with distinct functions. Loss of function-related experiments in cancer cells pinpointed the underlying mechanistic pathways by which miRNAs regulate cancer initiation and progression.

Moreover, due to the specific expression of miRNAs in cancer, many of them have been proposed as potential biomarkers for cancer detection. Despite their immense importance, reliable computational methods are required to identify novel miRNAs and estimate their expression levels systematically. Although several methods have been proposed in the past decade for detecting known and novel miRNAs from the sequencing data, differences in the data processing pipelines of RNA-Seq data lead to varying results on the same dataset. Some of the state-of-the-art pipelines are miRDeep2 [74], miRDeep* [75], mirPRo [76], mirnovo [77], miRge2.0 [78], sRNAtoolbox [79], and MiR&moRe2 [80]. These pipelines for the analysis of smRNAs (small RNAs) yield multiple false positives, do not identify paralogues of existing miRNAs, and often fail to identify the reverse complement sequences of known miRNAs. Although the above methods can detect several deregulated miRNAs, it is important to detect statistically significantly dysregulated miRNAs. These results generally vary across methods because of the algorithm for extracting the miRNAs and their count values. Hence, there is a need to develop robust methods to detect accurate and statistically significant dysregulated miRNAs and their count values.

To overcome the aforementioned limitations, we designed a robust computational workflow for the reliable identification and expression estimation of known and novel miRNAs from RNA-seq data, namely miRPipe. We have performed a comparative analysis of miRPipe with other well-known methods and found that miRPipe outperformed all these methods when benchmarked with synthetic (known ground truth) and CLL RNA-Seq expression datasets. No synthetic data simulators are available to generate ground truth to benchmark miRNA pipelines. Therefore, we have also developed a highly flexible, innovative, and faster synthetic sequence simulator tool, miRSim, for the comparative analysis of various existing pipelines with miRPipe. Our analysis of CLL datasets identified 31 known and eight novel dysregulated miRNAs, which we have experimentally validated using real-time PCR on clinical samples. An open-source, friendly synthetic data simulator, miRSim and a free dockerized version of miRPipe are available for deployment in a clinical setup. By providing the dockerized version of miRPipe pipeline, our goal is to make our miRPipe pipeline accessible to bioinformaticians of all skill levels, enabling effortless utilization and ensuring consistent reproducibility across various computing environments. They can subsequently share the analyzed results with the clinicians for further inference.

MM is a neoplasm of malignant plasma cells in the bone marrow, preceded by the precancerous stage, or MGUS. Presently, the distinction between MM and its precursor states (MGUS and smouldering multiple myeloma (SMM)) is based on the clinical symptoms and disease load, including the percentage of aberrant plasma cells in the bone marrow, levels of monoclonal protein secreted by the aberrant plasma cells, and the extent of dysregulation of normal homeostasis. However, the distinction between different stages is ambiguous sometimes in clinical practice. The role of early treatment and the type of such treatment to prevent progression to MM or to reduce the associated morbidity is also not clear. Although survival in MM has improved notably over the last few years, myeloma remains an incurable disease with an overall median survival of 2 to 10 years, depending on the response to the treatment. Thus, it would be interesting to decipher genes, genomic biomarkers and crucial pathogenic prognostic factors representative of MGUS and MM to develop appropriate therapeutic interventions to halt the progression to overt MM.

Multiple studies involving exome data have been performed to understand the genomic abnormalities driving tumor progression in MM. Exome data analysis of MM patients has revealed that the primary events in MM are either hyperdiploidy, i.e., trisomy of chromosomes 3, 5, 7, 9, 11, 15, 17 and/or 21, or non-hyperdiploidy involving translocations affecting the genes encoding immunoglobulin (Ig) heavy chains (IGH)-mainly t(4;14), t(6;14), t(11;14), t(14;16), and t(14;20) [81]. Primary events are then followed by multiple secondary events that are secondary translocations: t(8;14) linked with MYC overexpression, loss of heterozygosity, CNVs, acquired mutations, and epigenetic modifications [81], contributing to tumorigenesis. Initial deep sequencing studies on 38 WGS and 23 WES MM patients revealed frequent mutations in the NF-kB signaling pathway and activating mutations in the oncogene *BRAF* [82]. In another study based on the WES data of 84 MM patients, *SP140, LTB, ROBO1*, and *EGR1* genes were identified as the novel drivers of MM [83]. Similarly, the analysis of 463 WES data of MM patients revealed 15 recurrently mutated genes: *IRF4, KRAS, NRAS, MAX, HIST1H1E, RB1, EGR1, TP53, TRAF3, FAM46C, DIS3, BRAF, LTB, CYLD*, and *FGFR3* [84]. Further, the analysis of the same 463 MM samples reported RAS and NF-Kappa-B pathways as the most altered signaling pathways. Furthermore, the same study reported that the mutations in *CCND1* and DNA repair pathway genes-*TP53, ATM*, and *ATR* adversely impacted the overall survival (OS), while the alterations in *IRF4* and *EGR1* was associated with favourable OS.

Another study on the exome data analysis of 203 MM patients demonstrated tumor heterogeneity with a subclonal pattern of mutations and multiple mutations within the same pathway in the same patient [85]. A recent study on 62 newly diagnosed MM (NDMM) patients reported the association of changes in the cellular prevalence of mutations with disease progression [86]. Another study explored oncogenic dependencies between mutations in driver genes, hyperdiploidy events, primary translocations, and copy number alterations in MM patients [87]. Associations were established between t(4;14) and mutations in *FGFR3, DIS3*, and *PRKD2*; t(11;14) and mutations in *CCND1* and *IRF4*; t(14;16) and mutations in *MAF, BRAF, DIS3*, and *ATM*; and hyperdiploidy with gain 11q and mutations in *FAM46C*, and *MYC* rearrangements [87]. A recent study demonstrated the co-occurrence of mutations within the same or a different clone and the clonal shifts in the co-occurring and mutually exclusive mutations with progression in MM [88]. Similar phenomena may occur from the stage of MGUS to overt MM and require to be evaluated. Analysis of WES data of unpaired samples of MGUS and MM has been carried out by several groups [89, 90, 91, 92]. These studies have demonstrated a less complex genomic architecture in MGUS compared to MM, which has fewer mutations and lower TMB in MGUS. In a landmark study, the analysis of MGUS and MM paired samples reaffirmed the clonal heterogeneity and presence of most genomic changes at MGUS stage [93]. The existence of the majority of genomic abnormalities seen in MM at the MGUS stage poses a challenge in distinguishing MM from MGUS based on the genomic signatures and in defining critical genomic events responsible for the progression of MGUS to MM [89, 90, 91, 92, 93].

The early diagnosis of MM and identifying relevant differentiating genomic biomarkers between MGUS and MM present several challenges at the genomic level and the subject level. The unavailability of paired sequencing data (that is, sequencing data of MGUS and MM from the same sample) because all the MGUS subjects do not progress to MM, and the unavailability of reliable workflows for analyzing a pool of large mutational information to decipher accurate and reliable genomic information, biomarkers, and significantly altered pathways pose the key challenges at the genomic-level. Moreover, at the subject level, limited information in the studies about the time intervals of a subject's treatment and death times poses key challenges in pursuing disease progression and reliable identification of critical genes, genomic features, and signaling pathways for targeted therapeutics. With advancements in bioinformatics and increasing inclination toward ML or DL, newer methods are being developed for deducing salient information from the genomic data. For example, ML models have been developed to predict the survival outcome and treatment sensitivity in MM [94, 95]. Similarly, AI-assisted risk stratification models for predicting survival and deciding the treatment regimen have been developed for newly diagnosed MM patients [96, 97]. Pathway enrichment analysis and classification have been shown to improve with the imputation of missing values in the microarray data of blood cancers via ML methods [98, 99]. ML/DL methods have also been proposed to detect somatic mutations from WES data [100, 101], prediction of CNVs from whole exome data [102, 103, 104], driver genes in cancer [105, 106, 107, 108, 109] and, prediction of the survival-outcome and treatment-sensitivity in MM [94, 95].

However, the multi-dimensional analysis of exonic mutational profiles from exome sequencing data with gene-gene interaction has not yet been explored. This can be a promising direction for detecting key biomarkers in any cancer type. In recent years, geometric deep learning (GDL) has emerged to incorporate graph structures into a deep learning framework. Graph Convolutional Networks (GCNs) [110, 111], a type of GDL, can learn gene regulatory networks and do disease classification based on the network topology and disease-associated features, enabling an integration of graph-based data with genomic profiles [112]. The PPI network captures the physical interactions between proteins in an organism. Since the level of proteins and their interplay govern the molecular, cellular, and signaling controls, which are the key to gene-level functionality and help capture disease-specific information, PPI networks can be constructive if utilized alongside genomic information. A study on the exploration of the PPI network reported that the disease-related components in the PPI network are likely to be found in the network-based vicinity of disease components [113]. Similarly, another study on the PPI network revealed that the genes that contribute to a common disorder show an increased tendency of their protein-protein interactions [114]. These observations indicate that due to the interconnected nature of a PPI network, genes belonging to similar diseases have a high predilection for interacting with other genes, forming a disease module. Therefore, identifying such genes or disease modules with the help of the PPI network can divulge the disease-related signaling pathways or other disease genes. These observations motivated us to incorporate the biological interactions between genes as a key attribute of the bio-inspired BDL-SP model. Thus, we have incorporated the PPI information from the STRING database [115], the most comprehensive and global PPI network.

We addressed the problem of identifying significant biomarkers that differentiate MGUS from MM by incorporating a multidimensional analysis of exome profiles and their PPI network in a BDL-SP model. One of the challenges with this task is the ability to analyze a large amount of mutational information, a significant amount of which overlaps in MGUS and MM samples. Since this mutational information is difficult to decipher when extracting differentiable patterns among MGUS and MM, the current literature shows this gap. To address the above gap, we have designed and implemented a GCN-based model, BDL-SP, to extract important genomic information to discern MGUS and MM. The BDL-SP model uses single nucleotide variant (SNV) profiles of the significantly altered genes, the genes exhibiting statistically significant alterations as compared to other genes, from the exome sequencing data and the topological features of the PPI network to identify pivotal biomarkers that can distinguish MGUS from MM. An in-depth analysis has been carried out to identify significantly altered genes and pathways that are specifically associated with MM and may be beneficial for the early identification of MGUS patients at a high risk of progression to the malignant MM stage. This work can further lead to the identification of novel therapeutic targets, thereby preventing or delaying the malignant transformation of MGUS to MM.

For post-hoc model explainability, several state-of-the-art agnostic models are available, including SHapley Additive exPlanations (ShAP) [116], Local Interpretable Model-agnostic Explanations (LIME) [117], GNNExplainer [118], etc. The LIME algorithm implements the local surrogate models that are interpretable and used to explain individual predictions of black box machine learning or deep learning models. Instead of global explainability, LIME incorporate the local surrogate model to explain the individual predictions. To explain the model predictive response, LIME generates a new dataset consisting of perturbed samples and the corresponding predictions of the black box model. On this new dataset, LIME trains an interpretable model (Ridge Regression or Lasso Regression), which is weighted by the proximity of the sampled instances to the instance of interest. The strength of LIME lies in its local view, providing interpretable explanations for individual predictions.

The ShAP algorithm is considered one of the emerging and preferred approaches for decoding a DL model as well as for estimating feature importance based on their contribution to the model's predictions. The ShAP algorithm is a game theoretic approach that explains the output of any machine learning or deep learning model. The goal of ShAP is to explain the prediction of an instance by computing the contribution of each feature to the prediction. For this, the ShAP explanation method computes Shapley values, highlighting the contribution of that feature in making a prediction. To get the global importance of the feature, the average absolute shapley value per feature across the data is estimated. Lastly, The GNNExplainer is a local explainer with a model-agnostic approach that provides interpretable explanations for any graph neural network (GNN)–based model. To explain the predictive behavior of the GNN-based model, the GNNExplainer tries to find an explanatory sub-graph structure and its key node features that define the reasons behind the GNN's prediction of a specific label for that node. By maximizing mutual information between the explanatory sub-graph and the original computational graph, GNNExplainer provides local explanations for GNN predictions. In our work, we choose the ShAP algorithm for post-hoc explainability due to the following reasons:

1. ShAP builds upon the idea of Shapley values and extends it to provide a unified framework for explaining the output of any machine learning model. It connects the local explanations LIME provides to the global explanations offered by Shapley values. ShAP values capture the average marginal contribution of each feature across all possible coalitions of features, providing both local and global insights into model predictions.

2. With ShAP, global interpretations are consistent with the local explanations since the Shapley values are the "atomic unit" of the global interpretations.

3. The explanations provided by LIME and GNNExplainer are limited to a single instance, making it difficult for these explainers to apply in the inductive setting because the explanations are hard to generalize to other unexplained nodes.

We ranked the significantly altered genes based on their contribution to disease classification using the ShAP score. Among all the ML models trained in this study, the BDL-SP model reported the largest numbers of previously reported driver genes, OGs, TSGs, and AGs in the top-ranked significantly altered genes compared to the other models. The BDL-SP model also shows novel genes in the top-ranking genes not reported in MM but found to significantly alter and contribute significantly to disease prediction. We performed pathway enrichment analysis for the top 500 significantly mutated genes. We analyzed whether an altered signaling pathway becomes more or less significant with disease progression from MGUS to MM. We observed that several signaling pathways either become significant (from being insignificant at MGUS) or more significant with disease progression from MGUS to MM. We benchmarked the BDL-SP with several baseline ML models both quantitatively (that is, in terms of performance matrices, such as balanced accuracy and AUPRC) and qualitatively (that is, the model ability to prioritize the MM-relevant genes for sample classification) and observed that BDL-SP outperformed the other models in both aspects. With the help of the BDL-SP model, we identified the genes and their corresponding enriched signaling pathways that significantly contributed to MM disease development. The BDL-SP model's findings helped us to improve the understanding of cell transformation from premalignant to malignant state and strategic diagnosis to support the early detection of transformation to MM.

For comprehensive genomic profiling of MM and MGUS, several targeted sequencing panels have been devised to decipher the genomic complexity of MM [119, 120, 121, 122, 123]. These panels encompass critical genomic aberrations related to MM (such as SNVs, CNVs, and SVs). For instance, a 26-gene panel focused on prevalent mutations in previously published MM-relevant genes [123] but lacked validation for SVs. Similarly, another panel of 182 genes validated for SNVs, CNVs, and specific translocations (related to IGH only) in previously published MM-relevant genes [121]. A more extensive 228-gene panel covered various alterations, including SNVs, CNVs, and translocations involving *IGH* and *MYC* genes [120]. In a similar quest for comprehensive genomic profiling of MM, a 47-gene panel was crafted, encompassing dysregulated and frequently mutated genes in MM and those targeted by common therapies, validated for SNVs only [119]. Lastly, the largest gene panel of 465 genes was designed and validated for MM-related SNVs, CNVs, and translocations related to the *IGH* gene only [122]. However, these panels were designed using only MM samples and lacked markers and interactions distinguishing MM from MGUS that can give insights into MM pathogenesis.

We addressed the challenge of designing an efficient gene panel for comprehensive genomic profiling of MM by analyzing the unique characteristics of MM and MGUS in terms of genomic profile and interactions. We meticulously analyzed SNV, CNV, and SV profiles of key distinguishing genes between MM and MGUS to envision a targeted sequencing panel. This effort resulted in the identification of 282 genes for inclusion in the panel. To design the 282-genes targeted sequencing panel, we introduced a novel

AI-based BIO-DGI model, employing graph network learning to discern differentiating biomarkers and gene-gene interactions in MM and MGUS. In this proposed model, we integrated bio-inspired learning, utilizing the topological information gathered from nine PPI networks and exonic mutational profiles. This empowered the BIO-DGI model to rank genes and genomic features based on their role in disease progression more efficiently, with fewer GCN layers and multi-head attention modules than traditional ML or DL models relying solely on exonic mutational profiles. Moreover, our proposed BIO-DGI model outperformed exhaustive benchmarking against several baseline ML and DL models, including quantitative and qualitative evaluations. We further identified top-ranked genes and genomic features utilizing the ShAP algorithm.

To delve deeper, we identified five distinct gene communities using the Leiden algorithm (LA) [124] by utilizing the adjacency matrices derived from five trained BIO-DGI classifiers. This analysis sheds light on the influential genes within these communities, quantified through Katz centrality scores [125]. We have also highlighted the key gene interactions involving highly haploinsufficient genes and the genes participating in disease-initiating and disease-transformative genomic events within the gene communities to emphasize the genes mediating the progression of MM from MGUS. Interestingly, we observed several corner genes (a group of genes interacting with one central gene) were found in previously reported MM-related genes and were highly haploinsufficient with high node influence.

We meticulously analysed various variant profiles, including SNVs, CNVs, SVs, and Loss of Function (LOF) mutations. This detailed investigation drove the design of a clinically tailored 282-gene panel, aiming for a clinically and biologically relevant comprehensive genomic profiling of MM. Including genes from our proposed panel in MM-related pathways strongly underlines their pivotal roles in the disease progression from MGUS to MM and their potential impact on treatment outcomes in MM. This discovery underscores the clinical relevance and potential of the targeted sequencing panel designed for comprehensive genomic profiling in MM. Our study firmly establishes this panel as a promising novel strategy, particularly in identifying MGUS samples likely to progress to MM and pinpointing high-risk MM samples.

Lastly, it is essential to highlight that the biomarkers can be categorised as either prognostic biomarkers or diagnostic biomarkers. The prognostic biomarkers are the biomarkers that help in measuring the risk of disease progression or potential response to therapy. On the other hand, diagnostic biomarkers are the biomarkers that help to identify the early onset of disease [126, 127]. A gene and sncRNA can be used as prognostic and diagnostic biomarkers as they are subjective to the disease and study hypothesis. Some of the examples of genes and sncRNAs reported as prognostic and diagnostic biomarkers in the literature are shown below:

1. Genes as a prognostic and diagnostic biomarker: Genes represent a broader set of biomarkers, and their mutational spectrum and expression profile help to understand the underlying biology of the disease pathogenesis. For instance, mutations in mutations in *KRAS, NRAS, TP53*, and cytogenetic abnormalities such as del(17p) del(14q) are well-established prognostic biomarkers in MM [128, 129]. While *IGH* translocations (t(4;14), t(14;16)) are a well-established diagnostic biomarker in MM [128, 130, 91]. Similar examples can also be found for other diseases.

2. SncRNA as a prognostic and diagnostic biomarker: The association of sncRNAs with post-transcriptional regulation and other biological processes has been extensively studied in several diseases, including Cancer. The specificity of sncRNAs is the disease-specific expression pattern that makes them a unique prognostic and diagnostic biomarker. For instance, hsa-mir-21, hsa-mir-221, and hsa-mir-155 are well-established prognostic biomarkers of MM [131]. Meanwhile, hsa-mir-15a and hsa-mir-16-1 are well-established diagnostic biomarkers of MM [132, 133].

1.3 Thesis Contributions

The major contributions of the thesis are summarized below:

1. We comprehensively analysed the whole small RNA transcriptome in CLL and identified a unique pattern of differential regulation of eight miRNAs. Among these, three were up-regulated (hsa-mir-1295a, hsa-mir-155, hsa-mir-4524a) and five were down-regulated (hsa-mir-30a, hsa-mir-423, hsa-mir-486*, hsa-let-7e, and hsa-mir-744) in CLL. RT-qPCR validated the altered expression of all these eight DEMs. Besides, seven novel sequences identified to have elevated expression levels in CLL turned out to be tRNA, piRNAs (piRNA- 30799, piRNA-36225), and snoRNA (SNORD43) related. Multivariate analysis showed that hsa-mir-4524a (HR: 1.916, 95% CI: 1.080–3.4, p-value: 0.026) and hsa-mir-744 (HR: 0.415, 95% CI: 0.224–0.769, p-value: 0.005) were significantly associated with risk and time to first treatment.

2. We designed a robust, reproducible workflow for accurately identifying the novel miRNAs, namely miRPipe. The miRpipe workflow detects unique novel miRNAs by incorporating the sequence information of seed and non-seed regions, concomitant with clustering analysis. miRPipe can jointly identify miRNAs and piRNAs and carry out parallel batch processing to efficiently utilise the computational resources. We validated the performance of miRPipe with the available state-of-the-art pipelines using both synthetic datasets generated using the newly developed miRSim tool and three cancer datasets (CLL, Lung cancer, and breast cancer). In the experiment over the synthetic dataset, miRPipe is observed to outperform the existing state-of-the-art pipelines (accuracy: 95.23% and F1-score: 94.17%). Analysis of all three cancer datasets shows that miRPipe can extract

more known dysregulated miRNAs or piRNAs from the datasets than the existing pipelines.

3. We designed an innovative, bio-inspired AI-based workflow BDL-SP to identify pivotal genomic biomarkers to distinguish MGUS from MM. The proposed GCN-based BDL-SP model can discover discriminative genomic biomarkers that can distinguish MM from MGUS. BDL-SP outperformed all the baseline ML-based models. Further, using the application-aware interpretability analysis of the trained AI model, we have demonstrated a way to identify the best AI model from among the multiple ML or DL models that may have performed similarly on the quantitative metrics on the available data. In the post-hoc interpretability benchmarking, BDL-SP outperformed all the baseline models by identifying the largest number of previously reported genes such as *KRAS, BRAF, LTB, NRAS, FGFR3, NF1, NFKBIA, ARID2, RB1, HLA- A, TP53, SP140, TRAF3, EGR1, IRF1, SAMHD1, DIS3, CYLD, KMT2B/C, MAX, ZFHX3* and *NCOR1*, that are of high relevance in MM. Further, some of the genes that acted as differentiable biomarkers included TSGs (*HLA- B/C, NOTCH1, SDHA, MITF, ARID1B, FANCD2, KMT2D, APC, CMTR2*, and *AMER1*) and OGs (*CARD11, NOTCH1, VAV1, IRS1, MGAM, ABL2, TCL1A, PGR, MITF, RPTOR, TERT, BRD4, MECOM*, and *TAL1*) that have not yet been identified as MM drivers. These require validation by future studies before being declared as MM drivers. We further validated our findings by performing pathway analysis on the top mutated genes. It was inferred from the pathway analysis that several signaling pathways, such as the Calcium signaling pathway, B-Cell receptor signaling pathway, PI3K-Akt signaling pathway, MAPK signaling pathway, etc., are selectively and more significantly dysregulated with disease progression. Additional mutations in driver genes, critical OGs, TSGs, and AGs fostered the transformation of benign MGUS to MM. Similarly, the genomic mutation associated with the Synonymous SNV group (synonymous SNVs, UTR3, and UTR5) was found to be the most significantly contributing biomarker differentiating MM from MGUS. These observations may hold great significance from a therapeutic point of view. We observed that the number of OGs, driver genes, and AGs in the MGUS samples of European and Indian populations was statistically different. However, no population-specific differences were observed in our analysis of the MM data, which consists of the American and Indian populations. The results of the MGUS data indicate that ethnicity's impact on MM's disease biology should be further explored.

4. We proposed a clinically oriented targeted sequencing panel of 282 genes harbouring key genomic biomarkers for early detection of MM. For 282 genes panel crafting, we introduced the novel AI-based BIO-DGI model, encompassing gene interactions from nine PPI databases and exonic mutational profiles from three global MM and MGUS repositories (AIIMS, EGA, and MMRF). The BIO-DGI model demonstrated quantitative and qualitative superior performance, ensuring application-aware interpretability. Notably, the model identified the most previously reported genes, including OGs, TSGs, ODGs, and AGs, which are known for their high relevance in MM. Further exploration of these genes is recommended

to unveil novel driver genes. The validation of the top 500 genes set against MM-related datasets using Geo2R confirmed disruption in 488 out of 500 genes, underscoring their pertinence to MM. Similarly, pathway analysis of top-ranked genes further corroborated the relevance of top-ranked genes, revealing a shift in pathway deregulation from MGUS to MM. Key pathways like PI3K-AKT, NFKBIA, and MAPK were prominently altered, emphasizing their role in MM progression. Moreover, in the post-hoc analysis, the functional significance of nonsynonymous mutations, allele depth of synonymous SNVs and total number of other SNVs were found to be the most contributing genomic biomarkers in distinguishing MM from MGUS. Through meticulous analysis of variant profiles and validation using Geo2R, we curated a targeted sequencing panel comprising 282 MM-relevant genes. These observations hold immense potential for informed therapeutic interventions and may facilitate early detection and interception of disease progression in MM.

1.4 Thesis Organization

The rest of the thesis is organized into different chapters. In Chapter 2, we studied the RNA-Seq profiling of CLL patients to identify the deregulated sncRNAs and their impact on the clinical outcomes of CLL patients survival outcomes. We designed a novel workflow for the RNA-Seq data analysis. We also demonstrated the joint impact of multiple sncRNAs (such as miRNAs, piRNAs, snoRNAs, and tRNAs). Lastly, we identified the target genes of disrupted sncRNAs. We also performed the pathway analysis to gain deeper insights into deregulated miRNAs on altered pathways. Finally, the survival analysis was performed to study the correlation between deregulated miRNAs with several clinical parameters.

In Chapter 3, we continued our analysis on RNA-Seq data analysis as the reproducibility and reliability of the workflow for identifying novel sncRNAs were not appropriately addressed. We designed a unified, robust workflow, namely miRPipe, to identify statistically deregulated miRNAs, miRNA paralogues, functionally similar miRNAs and miRNAs from the reverse complement sequence of known miRNAs from RNA-Seq data. For the performance assessment of the miRPipe pipeline, we developed an in-house synthetic sequence simulator, named miRSim, to generate the synthetic RNA-Seq data with known ground truth. Using miRSim, we generated the synthetic RNA-Seq data in three read-depth categories (50K, 0.1 million, and 1 million) with known ground truth and assessed the miRPipe performance using the anticipated accuracy and f1-score. We benchmarked miRPipe with three publicly available real RNA-Seq expression datasets (CLL, lung and breast cancer). Further, the miRPipe identification accuracy was checked with the RT-qPCR results obtained from CLL patients. Finally, once the miRPipe pipeline is benchmarked, the source codes of the miRPipe pipeline and miRSim tool were uploaded to the public repository and the dockerized version of the miRPipe was developed to ensure the reproducibility of the RNA-Seq data analysis.

In Chapter 4, we designed an innovative AI-based model, the Bio-inspired Deep Learning architecture, for identifying altered Signaling Pathways (BDL-SP) to discover pivotal genomic biomarkers that can potentially distinguish MM from MGUS. The proposed BDL-SP model comprehends gene-gene interactions using the PPI network and analyzes genomic features using DL architecture to identify significantly altered genes and signaling pathways in MM and MGUS. For this, whole exome sequencing data of 1174 MM and 61 MGUS patients obtained from three global repositories (MMRF [134], European Genome-phenome Archive (EGA) [135] and All India Institute of Medical Sciences (AIIMS) [136]) were analyzed. In the quantitative benchmarking with the other popular ML models, BDL-SP performed almost similarly to the two best-performing predictive ML models of Random Forest and CatBoost. However, an extensive post-hoc explainability analysis, capturing the application-specific nuances, clearly established the significance of the BDL-SP model. This analysis revealed that BDL-SP identified a maximum number of previously reported OGs, TSGs, and AGs of high relevance in MM as the top significantly altered genes. Further, the post-hoc analysis revealed a significant contribution of the total number of SNVs and genomic features associated with synonymous SNVs in disease stage classification. Finally, the pathway enrichment analysis of the top significantly altered genes showed that many cancer pathways are selectively and significantly dysregulated in MM compared to its precursor stage of MGUS. At the same time, a few that lost their significance with disease progression from MGUS to MM were related to the other disease types. These observations may pave the way for appropriate therapeutic interventions to halt the progression to overt MM in the future.

Chapter 5 presents a curated, comprehensive, targeted sequencing panel focusing on 282 MM-relevant genes and employing clinically oriented NGS-targeted sequencing approaches. To identify these 282 MM-relevant genes, we designed an innovative AI-based Biological Network for Directed Gene-Gene Interaction Learning (BIO-DGI) model for detecting biomarkers and gene interactions that can potentially differentiate MM from MGUS. The BIO-DGI model leverages gene interactions from nine PPI networks and analyzes the genomic features from 1154 MM and 61 MGUS samples. The proposed model outperformed baseline ML and DL models, demonstrating quantitative and qualitative superiority by identifying the largest number of MM-relevant genes in the post-hoc analysis. The pathway analysis underscored the importance of top-ranked genes by highlighting the MM-relevant pathways as the top-significantly altered pathways. The 282-gene panel encompasses 9272 coding regions and has a length of 2.577 Mb. Additionally, the 282-gene panel showcased superior performance compared to previously published panels, excelling in detecting genomic and transformative events. Notably, the proposed gene panel also highlighted highly influential genes and their interactions within gene communities in MM. The clinical relevance is confirmed through a two-fold univariate survival analysis. The study's findings shed light on essential gene biomarkers and their interactions, providing valuable insights into disease progression.

Chapter 6 summarizes the thesis work and provides suggestions for future work.

1.5 Publications and Preprints from the Thesis

1.5.1 Journals and preprints

1. Gurvinder Kaur, **Vivek Ruhela**, Lata Rani, Anubha Gupta, Sriram K., Ajay Gogia, Atul Sharma, Lalit Kumar, and Ritu Gupta. "RNA-Seq profiling of deregulated miRNAs in CLL and their impact on clinical outcome", *Blood cancer journal*, 10(1), 1–9, January 13, 2020.

2. **Vivek Ruhela**, Anubha Gupta, Sriram K., Gaurav Ahuja, Gurvinder Kaur and Ritu Gupta, "miRPipe: A Unified Computational Framework for a Robust, Reliable, and Reproducible Identification of Novel miRNAs from the RNA Sequencing Data". *Frontiers in Bioinformatics*, 2, p.842051, July 8, 2022.

3. **Vivek Ruhela**, Lingraja Jena, Gurvinder Kaur, Ritu Gupta and Anubha Gupta. BDL-SP: A Bio-inspired DL model for the identification of altered Signaling Pathways in Multiple Myeloma using WES data. *American Journal of Cancer Research*, 13(4), p.1155, April 15, 2023.

4. **Vivek Ruhela**, Rupin Oberoi, Anubha Gupta, and Ritu Gupta. A comprehensive targeted panel of 282 genes: Unveiling key biomarkers in multiple myeloma. bioRxiv, 2023. doi: https://doi.org/10.1101/2023.10.28.564536.

1.5.2 Software tool developed

1. **Vivek Ruhela**, Ritu Gupta, Sriram K., Gaurav Ahuja, and Anubha Gupta. "miR-Sim: Seed-based Synthetic Small Non-coding RNA Sequence Simulator". *Zenodo*, June 14, 2021 https://doi.org/10.5281/zenodo.6546356

1.5.3 Posters

1. Akanksha Farswan, Anubha Gupta, **Vivek Ruhela**, Lingaraja Jena, Gurvinder Kaur, Sriram K, Ritu Gupta, Prognostic value of TMB and its association with overall survival in newly diagnosed multiple myeloma patients, Presented (online) at Center for Molecular Medicine Cologne (CMMC) Symposium, Sept 26-28, 2021, Cologne, Germany.

2. **Vivek Ruhela**, Akanksha Farswan, Anubha Gupta, Sriram K., Gurvinder Kaur and Ritu Gupta. "P-035: AI-based models for the identification of critical genetic biomarkers to distinguish MM from MGUS using the WES data". *Clinical Lymphoma Myeloma and Leukemia*, October 1, 2021, 21, p.S57.

Chapter 2

RNA-Seq profiling of dysregulated miRNAs in CLL and their impact on clinical outcome

2.1 Introduction

CLL stands out as a highly diverse malignancy characterized by substantial molecular and clinical heterogeneity that primarily stems from distinct genetic susceptibilities among individuals [55]. Cutting-edge genomic investigations utilizing high-throughput technologies like NGS and microarrays have unveiled the intricate molecular complexity of CLL. This heterogeneity is further compounded by variations in gene expression patterns, epigenetic regulatory mechanisms, and the abundance of noncoding RNAs, including lncRNAs and sncRNAs such as miRNAs, tRNA, piRNA, and snoRNA [58, 59]. Recognizing the substantial prognostic potential of miRNAs, researchers have proposed cumulative prognostic scores in conjunction with other factors to enhance prognostication in CLL [71, 72]. In this study, we analyze the whole small RNA transcriptome in CLL to identify the unique patterns of DEMs. Further, We have also co-analyzed genome-wide gene expression profiles to gain a deeper bimodal insight into the CLL miRNome circuitry and its mechanistic functional pathways. Moreover, this study has demonstrated the first time deregulated piRNAs and snoRNAs-related molecules in CLL. Further analysis has revealed a significant impact of specific DEMs on clinical outcomes in CLL.

2.2 Materials and Methods

2.2.1 Subjects

CLL patients diagnosed as per the diagnostic criteria of the International Workshop on CLL-sponsored Working Group [137] and 10 age-matched healthy controls (5 males and 5 females) were enrolled. The demographic, clinical and laboratory-based details of the cases evaluated for different sets of experiments are provided in Table-2.1. The study was conducted in accordance with the Declaration of Helsinki guidelines. Ethical clearance for the study was obtained from the institute's ethics committee, and written informed consent was obtained from all the participants.

2.2.2 Genome-wide miR sequencing by NGS

Total RNA was extracted from the Mononuclear cells (MNCs) of CLL patients and MACS-sorted CD19+ B cells (cat no. 130050301, Milteneyi Biotech, Germany) of healthy controls using the miRVana kit (Thermofisher Scientific, MA, USA). The samples having RNA Integrity Number (RIN)≥7 were processed further. Small RNA libraries were generated for 28 CLL cases and 2 pooled healthy controls (pooled from 5 males and 5 females) using the “TruSeq small RNA sample preparation kit" (Illumina, San Diego, CA, USA). The libraries with 76 nucleotide inserts were subsequently sequenced on NextSeq 500 (Illumina).

Table 2.1: Baseline demographic, laboratory, and clinical characteristics of CLL patients as per different experimental cohorts.

Parameter	NGS (n=28) Numbers	Gene Expression Array (n=21) Numbers	RT-qPCR (n=89) Numbers (%)
Gender			
Male	21 (75%)	16 (76%)	68 (76.5%)
Female	07 (25%)	05 (24%)	21 (23.5%)
Median age	60	60	60
≤65 years	20 (71.4%)	16 (76.2%)	69 (77.5%)
>65 years	08 (28.6%)	05 (23.8%)	20 (22.5%)
Rai stage			
Stage 0/I/II	4/6/07	4/8/09	15/14/28
Stage III/IV	6-May	-/-	13/19
Beta2 Microglobulin*			
≤3.5	2 (7.1%)	6 (30%)	15 (17.2%)
>3.5	26 (92.9%)	14 (70%	72 (82.8%)
IGHV mutational status**			
Mutated	10 (35.7%)	10 (48%)	47 (56%)
Unmutated	18 (64.3%)	11 (52%)	37 (44%)
Genetic abnormality***			
No abnormality	09 (32%)	08 (42%)	34 (40.5%)
Del (13q)+	07 (25%)	04 (21%)	22 (26.2%)
Del (11q)+	07 (25%)	03 (16%)	07 (8.3%)
Del (17p)+	01 (4%)	02 (10.5%)	14 (16.7%)
Trisomy12	04 (14%)	02 (10.5%)	07 (8.3%)

*Beta2 Microglobulin data was available for 20/21 and 87/89 patients of Gene expression (GE) array and RQ-PCR cohorts, respectively. **IGHV mutational status was available for 84/89 patients of the RT-qPCR cohort. ***Genetic aberrations data was available for 19/21 and 84/89 patients of GE array and RQ-PCR cohort respectively

2.2.3 RNASeq pipeline for analysis of NGS data

The FASTQ files, as obtained from RNA-Seq experiments, were further analyzed with the bioinformatics pipeline developed in-house, which was pre-validated on publicly available published data on Acute lymphocytic leukemia (ALL) [2]. A schematic representation of the RNA-Seq analysis pipeline and related workflow is shown in Figure-2.1.

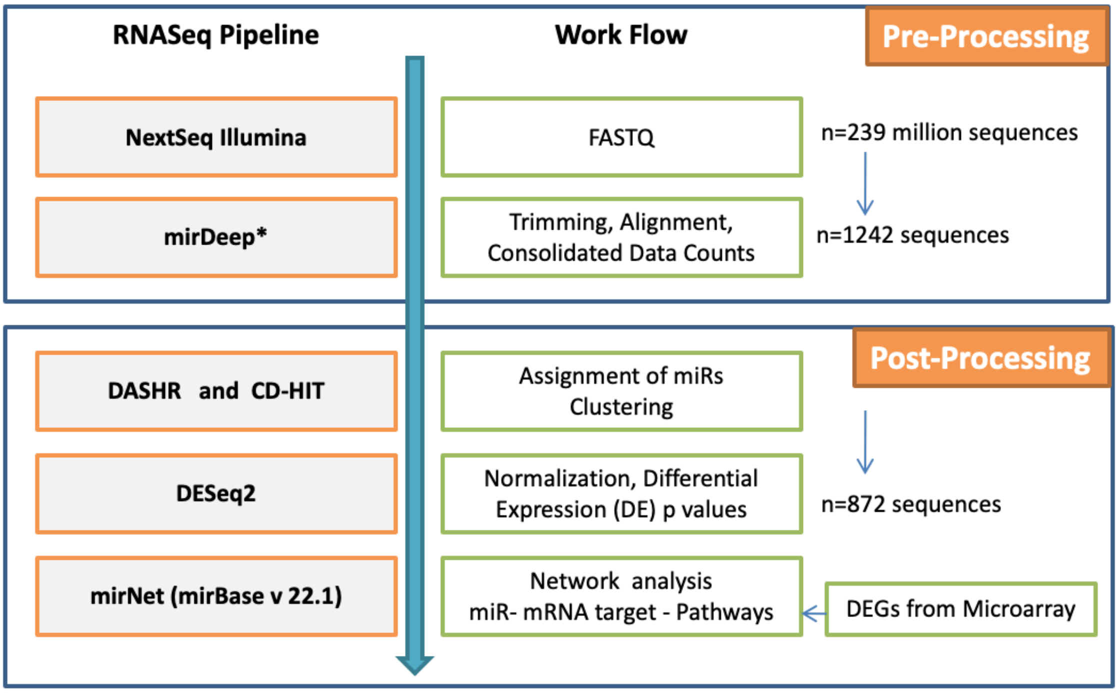


Figure 2.1: Bioinformatics workflow for the processing and analysis of RNASeq data.

The base quality of the raw reads (>Q30) was initially checked with java based FastQC developed by Babraham Bioinformatics. This was followed by adapter trimming and sequence alignment with GRCh37 human genome assembly database using miRDeep* [75] and miRBase v19 [4]. A custom Python script was used to compute the consolidated count matrix from the result files obtained from miRDeep*, and the duplicates were merged. The unaligned potential novel miRNAs were clustered with CD-HIT [138] and their unique IDs were generated. Sequence annotations of potential novel miRNAs were ascertained with DASHR [139]. The trimmed data obtained after post-processing using CD-HIT and DASHR was further processed with Bioconductor DESeq2 [31] where the consolidated count data were normalized and differentially expressed miRNAs (DEMs) were identified along with corresponding Wald statistic p and Benjamini-Hochberg adjusted p values to avoid false discovery rates. The miRNAs with adjusted p-values $\leq$0.05 and fold change (FC) $\geq$2 were considered to be significantly different.

2.2.4 Validation of DEMs by Real-time quantitative PCR (RT-qPCR)

Eight miRNAs found to be differentially expressed in miRNA deep sequencing analysis were validated on CLL (n=89) patients using Locked Nucleic Acid (LNA) – based primers specific to each miRNA and SYBR green master mix as per the manufacturer's recommendations (Exiqon, South Korea) on Quantstudio 12K Flex system (Thermofisher Scientific). The laboratory staff was kept unaware and blind to sample details. The data was normalized using the geometric mean of two endogenous controls, SNORD44 and SNORD48, and relative expression was evaluated using the comparative Ct method.

2.2.5 Prediction of gene targets of DEMs and their functional pathways

To predict the gene targets and functional pathways of all the eight significant DEMs, a hypergeometric test was applied, and the Kyoto Encyclopedia of Genes and Genomes (KEGG) database was evaluated in miRNet [140].

2.2.6 Statistical analysis

The differences in gene expression obtained from RT-qPCR between CLL and healthy controls were compared using the Mann-Whitney Rank Sum test. The chi-square test was used to correlate the expression of DEMs with prognostic parameters. The TTFT was calculated from the date of diagnosis to the date of the start of the first therapy. OS was calculated from the date of diagnosis to the date of last follow-up or date of death due to any cause. For comparing cumulative incidence curves for risk of treatment, Gray's test was used where death previous to any treatment was also considered as a competing event. The log-rank test was used to compare Kaplan Meier curves of OS. Variables with significant differences in univariate analysis were subsequently subjected to multivariate analysis using Fine-Gray and Cox regression models for TTFT and OS, respectively. Gray's test and Fine-Gray modelling were performed using the cmprsk library from the CRAN R repository [141]. At the same time, the rest of the statistical analysis was carried out with SigmaPlot V13.0 (Systat Software, Inc.).

2.2.7 Data access

NGS and gene expression data generated in the study have been submitted to the NCBI Gene Expression Omnibus (GEO)) under accession number GSE123436 and GSE81935, respectively.

2.3 Results

2.3.1 Identification of DEMs in CLL

A total of 239,039,053 raw reads were analyzed through the RNA-Seq pipeline, which resulted in 872 miRNA sequences. Of these, fifteen sequences (8 known miRNAs: hsa-let-7e, hsa-mir-1295a, hsa-mir-155, hsa-mir-30a, hsa-mir-423, hsa-mir-4524a, hsa-mir-486, hsa-mir-744 and seven novel miRNAs) were found to be differentially distributed and significantly deregulated in CLL (p adj $\leq$ 0.05; Figure-2.2). Among the significant DEMs, three were upregulated: hsa-mir-1295a (log2FC= 8.28), hsa-mir-4524a (log2FC= 7.39) and hsa-mir-155 (log2FC= 2.06) and five were downregulated: hsa-mir-30a (log2FC=-4.19), hsa-let-7e (log2FC=-3.59), hsa-mir-744 (log2FC=-2.63), hsa-mir-486* (log2FC=-1.54) and hsa-mir-423 (log2FC=-1.41).

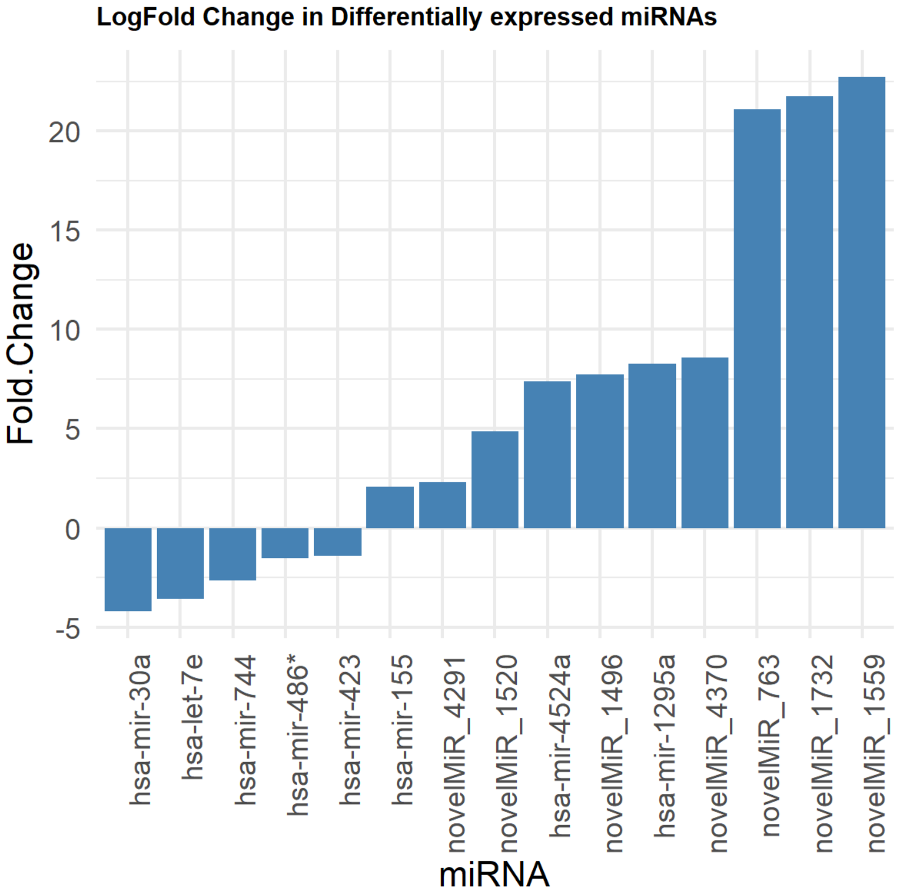


Figure 2.2: Histograms of relative fold changes of the eight differentially expressed miRNAs (DEM) as identified by RNA-seq.

2.3.2 Annotation of novel miRNAs

The seven differentially expressed novel miRNAs identified with NGS (p adj. $\leq$0.05) were analyzed with DASHR and UCSC human genome browser for sequence annotations. Five of the novel miRNAs (hsa-novelmiR_4291, hsa-novelmiR_1520, hsa-novelmiR_1559, hsa-novelmiR_1732 and hsa-novelmiR_4370) showed homology with a multitude of tRNA molecules located on chromosomes 1, 6, 7, 9, 11, 12, 14, 15, 16 and 17. The novelmiR_4370 showed homology with a piRNA-36225 (alias piRNA-28374; GenBank: DQ598159.1) as well. The novelmiR_763 got assigned on chr22 as piRNA-30799 (GenBank: DQ600599) and snoRNA U43 (SNORD43) (GenBank: X96642). Further characteristics of snoRNA were obtained from snoRNA Base and Rfam (version-14.1); and of piRNAs from piRNABase and IsopiRBank. The novel miR_1496 with mature miR sequence did not yield any result in the DASHR database, but when aligned using BLAST showed homology with a predicted uncharacterized LOC107984496. When BLAST analyzed the precursor sequence of novel miR_1496, it showed full alignment with 3' end of tRNA-Ile (TAT)1-1 at Chr-19.

2.3.3 Prediction of DEM targets and networking of functional pathways

Eight significant DEMs identified by NGS were analyzed for putative gene targets in miRNet. A list of gene targets that were predicted for each of the 8 DEMs by miRNet is shown in Supplementary Table 2. These DEM targets consisted of several crucial genes, including driver genes of CLL such as ATM (targeted by hsa-mir-30a-5p), TP53 (hsa-mir-30a-5p), SF3B1 (hsa-mir-423-3p), NOTCH1 (hsa-mir-30a-5p) and MYD88 (hsa-mir-155-5p). MirNet-based network analysis of inter-miR interactions of 8 DEMs and with their targets suggested significant enrichment of various KEGG-derived pathways (Supplementary Table 3) such as RNA transport (p<0.001), Pathways in cancer (p<0.001), cell cycle (p<0.001), mTOR signaling pathway (p<0.001) and p53 signaling pathway (p<0.001).

2.3.4 Validation of DEMs by RT-qPCR

Eight differentially expressed miRNAs identified by NGS were validated using RT-qPCR. As compared to healthy B-cells, hsa-mir-30a (FC=0.06; p=0.05), hsa-mir-423 (FC=0.21; p=0.034), hsa-mir-744 (FC=0.03; p=0.024), let-7e (FC=0.14; p=0.038) and mir 486 (FC=0.166; p=0.096) were down-regulated while hsa-mir-155 (FC=6.39; p=0.019), hsa-mir-1295 (FC=74.2; p=0.017) and hsa-mir-4524a (FC=52.2; p=0.026) were significantly up-regulated in CLL. These results are congruent with the pattern obtained from miRNA deep sequencing.

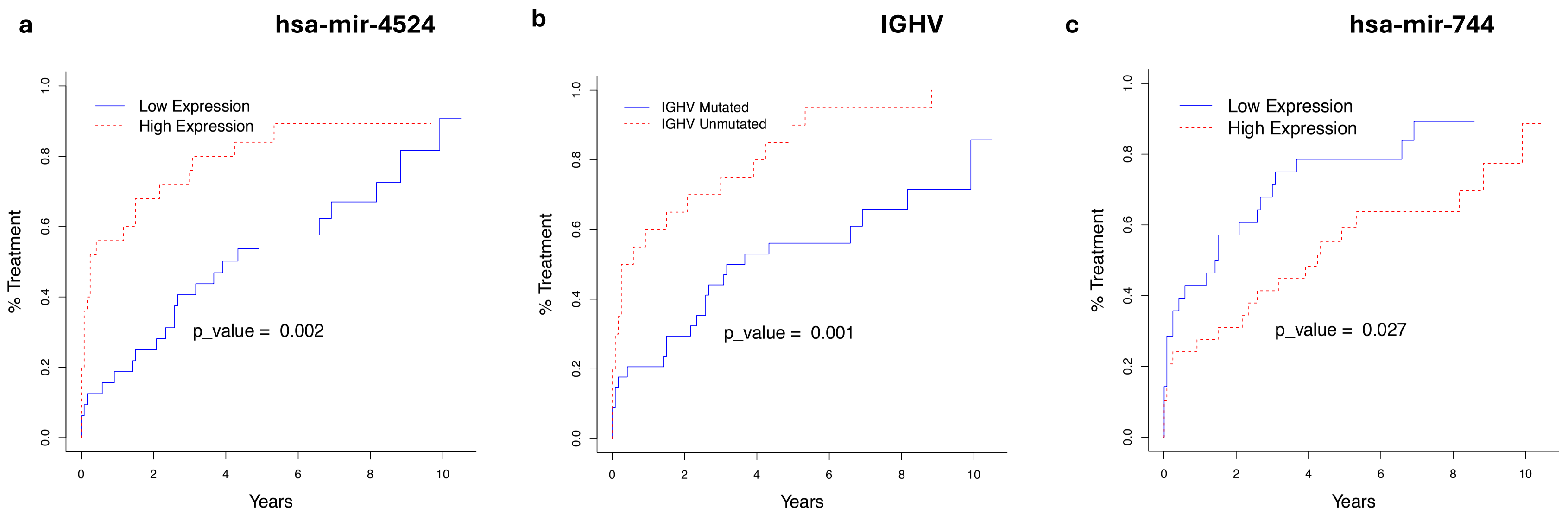
2.3.5 Association of miRNA expression with prognostic factors and clinical outcome

In the validation cohort of 89 patients, IGHV mutation status, beta 2 microglobulin levels and genetic aberration data were available for 84, 87, and 84 patients, respectively. The CLL-International prognostic index (IPI) [142] could be calculated for 78 patients, of which 13 were assigned as low risk, 31 as intermediate risk, 28 as high risk and 6 as very high-risk patients. No significant association was observed for any of the DEM with IGHV status, beta 2 microglobulin levels and IPI score. The number of patients in each subgroup based on the genetic aberrations was too small to draw any statistical conclusion.

The expression level of DEMs was further investigated in terms of their association with TTFT and OS. The median fold change values of individual miRNAs measured by RT-qPCR were used as cut-offs to group the samples into high and low-expression groups. TTFT was compared in early-stage CLL patients (Rai stage 0-II; n = 57) between the groups. The univariate analysis using Gray's test identified hsa-mir-4524 (p = 0.002), hsa-mir-744 (p = 0.027) and IGHV mutation status (p = 0.001) as significant parameters for the evaluation of TTFT. Multivariate analysis with the significant parameters suggested that high expression of hsa-mir-4524a (HR: 1.916, 95% CI: 1.08–3.40, p = 0.026; Figure-2.3a) and IGHV unmutated status (HR: 2.84; 95% CI = 1.58–5.120; p = 0.0005; Figure-2.3b) were significantly associated with shorter time to first treatment while higher expression of hsa-mir-744 was found to be associated with longer TTFT (HR: 0.415; 95% CI= 0.224–0.769; p = 0.005; Figure-2.3c). The OS was calculated for all the patients of the validation cohort (n = 89). IGHV-mutated patients displayed longer OS as compared to unmutated patients (p = 0.011). No association was observed between the expression of any of the 8 DEMs with OS.

2.4 Discussions

The miRNA profiling in this study has identified differential regulation of eight important known miRNAs and seven novel sncRNA species related to tRNA, piRNA, and snoRNA that might contribute to the development/progression of CLL by targeting various crucial genes. Of the eight known miRNAs, differential regulation of four miRNAs, i.e., hsa-mir-155 [61, 62, 63], hsa-mir-486-5p [143], hsa-mir-423 [144] and hsa-let-7e [144] has been earlier reported in CLL. The hsa-mir-30a miRNA, which has been shown to have a tumor suppressor role in lung cancer [145] and breast cancer [146], was found to be down-regulated in the present study. Of the various gene targets of hsa-mir-30a, GAB1, which is a key molecule in the pathogenesis and progression of CLL, was found to be up-regulated in the present study [70]. Higher mRNA levels of GAB1 have been shown



Years	2	4	6	8	10
No. at risk in Low expression group (FC ≤52.2; n=32)	25	50	57	67	90
No. at risk in High expression group (FC > 52.2; n=25)	68	80	89	89	NA

Years	2	4	6	8	10
No. at risk in IGHV mutated group (n=34)	29	53	56	56	86
No. at risk in IGHV unmutated group (n=20)	65	80	95	95	NA

Years	2	4	6	8	10
No. at risk in Low expression group (FC ≤0.03; n=28)	57	78	79	89	NA
No. at risk in High expression group (FC > 0.03; n=20)	31	48	63	64	89

Figure 2.3: Cumulative incidence plots demonstrating the risk of treatment in CLL patients stratified on the basis of the level of expression of a hsa-mir-4524a, b IGHV mutation status, and c hsa-mir-744. The cut-offs for defining low and high expression of miRNA and the number of cases in each subgroup are shown below the curves. The p-values and hazard ratios obtained in the Fine-Gray multivariate analysis model are shown inside the curve.

to be associated with strong B-Cell receptor responsiveness and disease outcome in CLL [147]. Higher mRNA levels of GAB1 with low expression of hsa-mir-30 in the present study point towards a regulatory connection between these two, which might be important for the malignant behaviour of CLL cells. Like previous studies, the present study also reports up-regulation of oncogenic hsa-mir-155 in CLL [61, 62, 63]. Similar to previous reports, expression of hsa-mir-155 was not found to be associated with IGHV mutation status and survival outcome [61, 148]. Unlike a previous report in CLL, down-regulation of hsa-mir-486 was observed in CLL patients in the present study [143]. Downregulation of hsa-mir-486 has been reported to contribute to the progression and metastasis of lung cancer due to increased expression of its target Rho GTPase activating protein 5 (ARHGAP5) [149]. In the present study, a concomitant downregulation of hsa-mir-486 and an upregulation of mRNA expression of ARHGAP5 suggests a pathogenic role of hsa-mir-486 in CLL. The perturbed levels of the hsa-mir-1295 cluster located on chromosome 1 have been implicated in tumorigenesis in colorectal cancer and follicular lymphoma [150, 151]. As per a very recent report in CLL, hsa-mir-1295 was amongst one of the five most up-regulated miRNAs in CLL [152]. In the present study, hsa-mir-1295 was also the most abundantly expressed miRNA in CLL. Consistent up-regulation of hsa-mir-1295 suggests that it could be an important molecule in the pathobiology of CLL. However, a detailed investigation is required to functionally characterize its role in this process. The hsa-let-7e is yet another important miRNA that was downregulated in CLL in this study and has been implicated in several cancers. Similar studies on CLL have also reported lower expression of hsa-let-7e in CLL as well as poor prognostic CD38+ CLL subgroup, which further supports observations of this study [144].

Incidentally, three of the eight DEMs observed in this study, namely hsa-mir-423, hsa-mir-4524a and hsa-mir-744, are located on chromosome 17. The levels of expression of hsa-mir-423, located at 17q11.2, have been shown to be influenced by an SNP rs6505162 C > A, which has been shown to correlate with risk in a, wide range of cancers although the mechanistic processes remain elusive [153]. In this study, expression levels of hsa-mir-423 were found to be reduced in CLL. An RT-qPCR-based study has also reported reduced expression of hsa-mir-423 in CLL patients, particularly in the context of higher lactate dehydrogenase (LDH) activity [144]. Higher expression of hsa-miR-744 among older female patients with ovarian carcinoma has been reported to correlate with prolonged disease-free survival, suggesting its protective influence. This is comparable to findings in this study, where higher expression of hsa-mir-744 correlates with extended time to first treatment in CLL patients compared to those with reduced expression. The hsa-mir-4524a present at 17q24.2 is located within intron 22 of the host gene ABCA6. ABCA6 is an ATP binding cassette superfamily A, member 6 transporter, which plays a role in macrophage lipid homeostasis. As per the present study, the expression of hsa-mir-4524a was found to be upregulated in CLL, and a significant association was observed between its high expression levels and shorter TTFT. Furthermore, a study has shown upregulation of the expression of ABCA6 in CLL [143]. It has been further shown that hsa-mir-4524a/b targets LDH A that promotes aerobic glycolysis in colorectal cancer and that it could become an important therapeutic target of cancer energy metabolism [154].

In our study, we observed the sequence homology of seven novel DEMs to known ncRNAs (Supplementary Table 1). Through sequence homology analysis, we observed that the novel DEMs exhibit similarities with various classes of ncRNAs, including tRNAs, piRNAs, and snoRNAs. Further, the matched ncRNAs were found to be associated with several biological processes, such as cell proliferation, apoptosis, cell cycle stability, gene silencing, etc. The dysregulation in these ncRNAs may directly affect these biological processes and support cancer progression. The significance of these findings is multifaceted:

1. Functional Conservation: The observed sequence homology suggests that the novel DEMs may share functional attributes with the known ncRNAs, which implies potential involvement in similar biological processes or regulatory pathways. For instance, piRNAs primarily contribute to maintaining genome stability in germline cells. Therefore, if a novel DEM displays sequence homology with piRNAs, it may possess similar functional characteristics.

2. Diagnostic and Therapeutic Implications: The homology between novel DEMs and known ncRNAs could have diagnostic and therapeutic implications. If the known ncRNAs are linked to specific diseases or conditions, the novel DEMs might serve as biomarkers for diagnosis or as targets for therapeutic intervention.

Therefore, exploring the homology between novel DEMs and known ncRNAs offers valuable insights into their potential functions, regulatory roles, and relevance in disease contexts. Two novel miRNA sequences (novel hsa-mir-4370 and novel hsa-mir-763) having differential regulation in the present study showed homology with two piRNA sequences (hsa-piR-36225 and hsa-piR-30799) and a C/D box snoRNA (U43 or SNORD43). Recent molecular studies have rediscovered the structural and functional diversity of snoRNAs [155] and piRNAs [156] and their aberrant expression in cancer. Induction of C/D box snoRNAs has been reported to favor leukemogenesis [157]. A number of snoRNAs have been reported to predict the clinical outcome in the early stage as well as IGHV mutated CLL [158, 159]. Various tRNA fragments have been reported to induce transient translational arrest, and tRNA-derived small RNAs can function similarly to miRNAs [155]. Identifying tRNA molecules in the present study suggests these might also be involved in developing CLL. The observed sequence homology of novel DEMs with known ncRNAs paves the way for further investigation into their biological significance and clinical applications.

2.5 Limitations of the study

In this study, our primary focus was to delve into the transcriptional patterns of well-established sncRNAs, whose functional characteristics are extensively documented in the literature. The identification of novel DEMs was an additional aspect of our research, and we found their sequence homology with known piRNAs, tRNAs, and snoRNAs particularly intriguing, shedding light on their potential functions. Although we discovered the novel DEMs, we didn't validate the novel DEMs using RT-qPCR as the RT-qPCR validation entails significant resources and time commitments. Therefore, before embarking on validating novel DEMs, it is imperative to establish robust literature support confirming that similar novel sncRNA sequences have been previously identified as DEMs by other studies. This approach ensures the judicious allocation of resources and enhances the validity of our findings.

2.6 Conclusion

Extensive studies aimed at a better elucidation of the global transcriptional landscape of sncRNAs and their effects on clinical outcomes could help refine the patient stratification schemes, and sncRNAs could surface as additional molecular biomarkers for improved prognosis and exploration of therapeutic targets in future. These investigations are crucial for a comprehensive understanding of the intricate molecular mechanisms that underlie this malignancy. By unravelling the roles of sncRNAs and their impacts on clinical outcomes, we aim to enhance patient stratification schemes, providing a more precise and tailored approach to CLL treatment. Moreover, the identification of sncRNAs, including miRNAs and piRNAs, holds immense promise, as these molecules could serve as an additional molecular biomarker. These biomarkers not only contribute to refining prognosis but also present novel opportunities for the identification and exploration of therapeutic targets in the future. In the following chapter, we will discuss a robust workflow for identifying the known and novel miRNAs, as well as piRNAs. The extensive benchmarking of this workflow not only ensures accuracy, reproducibility, and reliability but also paves the way for a deeper understanding of the roles of these sncRNAs in cancer pathogenesis and progression.

Chapter 3

A Unified Computational Framework for a Robust, Reliable, and Reproducible Identification of Novel miRNAs from the RNA Sequencing Data

3.1 Introduction

High-throughput sequencing techniques have ushered in a compelling need for reliable computational tools to accurately discern sequenced molecular entities. Despite numerous computational methods designed for the analysis of intricate sequencing datasets, none have achieved the desired precision for detection and estimation. In the context of cancer, miRNAs hold a central role within oncogene and tumor suppressor gene networks, influencing a wide array of transcripts with distinct functions, often serving as potential biomarkers for cancer detection. To address these challenges, we introduce miRPipe, a robust computational workflow tailored for the identification and expression estimation of known and novel miRNAs from RNA-seq data. miRPipe excelled compared to state-of-the-art methods, proven through benchmarking against synthetic data (with known ground truth) and real RNA-Seq expression data from CLL. In response to the current absence of synthetic data simulators for miRNA pipeline benchmarking, we present miRSim, a versatile and expedited synthetic sequence simulator. This tool facilitates a comprehensive comparative analysis of various existing pipelines in conjunction with miRPipe. Our exploration of CLL datasets led to the identification of 31 known and eight novel dysregulated miRNAs, validated through RT-qPCR on clinical samples. We offer the freely accessible miRSim synthetic data simulator and a dockerized version of miRPipe, catering to bioinformaticians of all levels and fostering seamless collaboration with clinicians to gain further insights.

3.2 Materials and Methods

We have used synthetic and real RNA-Seq expression datasets for benchmarking against the available state-of-the-art miRNA pipelines. Of note, we have developed an in-house tool miRSim [160] to assess the pipeline performance by comparing pipeline outcomes with matched ground truth. For miRPipe validation, we have considered three publicly available GEO datasets, that is, the CLL dataset (GSE123436)[1], breast cancer dataset (GSE171282) [161] for miRNA identification and Lung Cancer dataset (GSE37764)

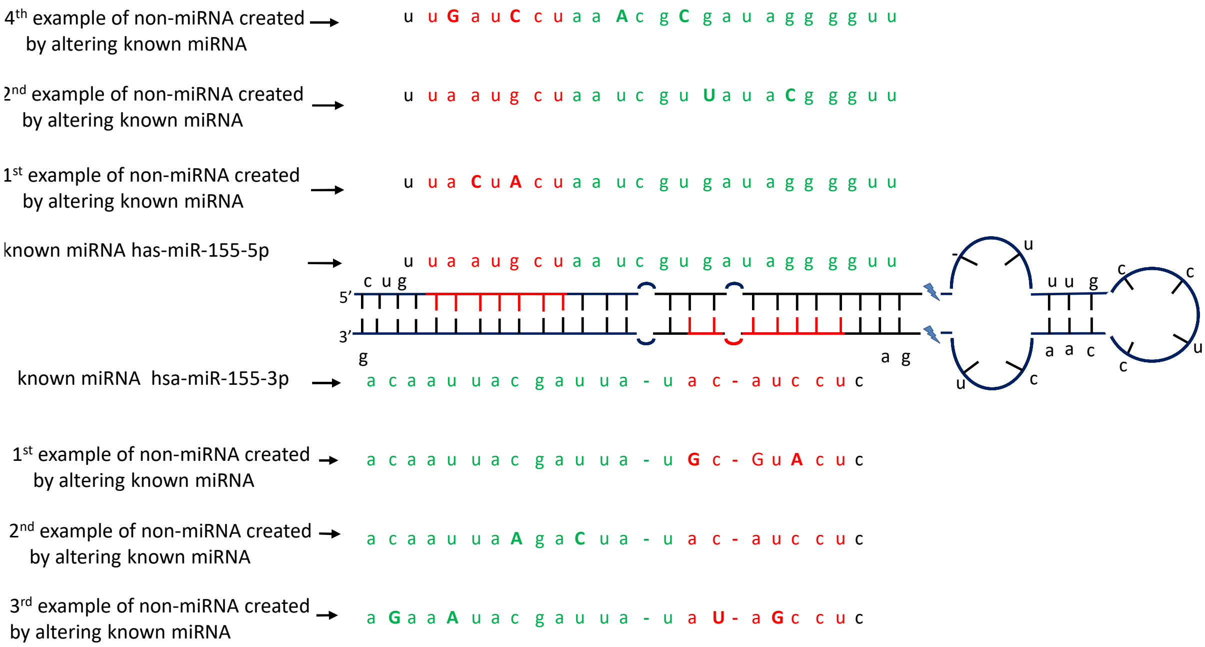
[162] for piRNA identification.

3.2.1 Proposed miRSim tool: Synthetic data simulator

Assessing the performance of existing bioinformatics tools or developing new ones, such as sequence aligners or quantification tools, relies heavily on the availability of ground truth data. Presently, several synthetic sequence simulators available for the generation of synthetic sequencing data, such as ART [163], pIRS [164], Flux [165], polyester [166], RSEM [167], CAMPAREE [168], BEERS2 [169], NEAT [170], DWGSIM [171], WGSIM [172], SimNGS [173], SimSeq [174], ISS [175], Mason [176], and RNA-Seq simulator [177] to generate synthetic RNA-Seq sequencing data. These synthetic RNA-Seq data generator tools are generic in nature, and data generation depends on platform-specific parameters. However, out of these simulators, only five simulators (ART, NEAT, BEERS2, CAMPAREE, and RSEM) provide the ground truth for the generated synthetic data. Additionally, it's worth highlighting that most of these simulators focus on generating full-length synthetic mRNA transcriptome sequencing data; our study aims at sncRNA sequencing, encompassing miRNAs and piRNAs. To bridge this gap, we developed a small sncRNA sequencing simulator, miRSim.

The design and workflow outline of the miRSim tool is illustrated in Fig.3.1. Mechanistically, the standard miRNA sequences and their genomic location can be stored in FASTA and GFF file formats (gff3) as the reference input to the miRSim tool. The miRNA, piRNA, and novel miRNA sequences were collected from the miRBase [4], piRNAdb database (version 1.7.6) (https://www.pirnadb.org/), and the recent literature [1, 2], respectively. For the robustness of any RNA-Seq pipeline, it is essential to detect known miRNAs, novel miRNAs, and their paralogues robustly. Hence, miRSim provides the option to add a selected percentage of altered sequences of miRNAs and piRNAs as the 'Other' category, which acts as a true negative to assess the efficiency of the pipeline. In the 'Other' category, the new miRNA sequence is generated by altering either the seed region's nucleotides (red-colored nucleotides in Figure-3.1A) or by altering the xseed region's nucleotides (green-colored nucleotides in Figure-3.1A) or both by at least 2nt. The altered nucleotides are shown in capital letters in each of the seed and xseed regions. The seed and xseed regions (regions other than the seed region) of a miRNA govern the functionality of miRNA in biological processes [178]. The 2nt cut-off was based on the fact that the recommended tolerance used in the standard RNA-seq aligners such as STAR [179], TopHat2 [180], miRDeep2, and miRDeep*. The resulting sequence will not be a miRNA or piRNA. The fraction of sequences for each error type is provided in the form of an error profile as input to the miRSim tool by the user.

(A) Example of synthetic reads based on hsa-miR-155 miRNA hairpin structure


4th example of non-miRNA created by altering known miRNA → u u G a u C c u a a A c g C g a u a g g g u u

2nd example of non-miRNA created by altering known miRNA → u u a a u g c u a a u c g u U a u a C g g g u u

1st example of non-miRNA created by altering known miRNA → u u a C u A c u a a u c g u g a u a g g g u u

known miRNA has-miR-155-5p → u u a a u g c u a a u c g u g a u a g g g u u

known miRNA hsa-miR-155-3p → a c a a u u a c g a u u a - u a c - a u c c u c

1st example of non-miRNA created by altering known miRNA → a c a a u u a c g a u u a - u G c - G u A c u c

2nd example of non-miRNA created by altering known miRNA → a c a a u u a A g a C u a - u a c - a u c c u c

3rd example of non-miRNA created by altering known miRNA → a G a a A u a c g a u u a - u a U - a G c c u c

(B) One example calculation of synthetic data generation from miRSim pipeline

- ❖Total no. of reads for all chromosomes = 500
- ❖No. of miRNA reads generated for chr1 = $\frac{No\ of\ miRNA\ in\ \ chr1\ (=156)}{Total\ no\ of\ miRNA\ present\ in\ 23\ chromosome\ pair\ (=1918)} * 500 \approx 41$

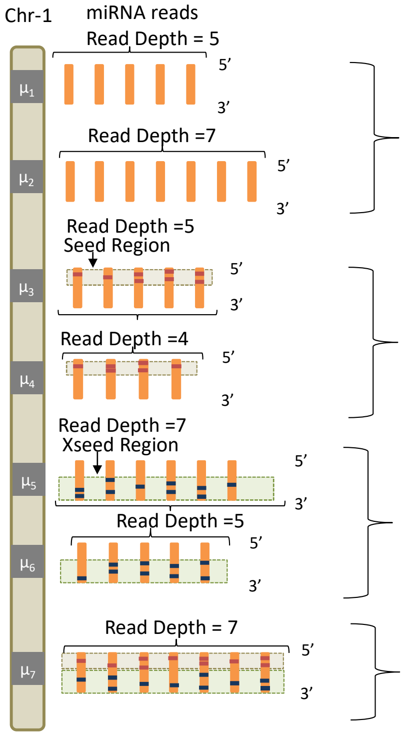
Legend

▭ Chromosome

μ_1 miRNA/PiRNA/novel-miRNA region

🟥 Alteration in seed region

🟦 Alteration in xseed region



standard sequences shown at μ_1 & μ_2 $30\% = \frac{30}{100} * 41 = 12$

sequences shown at μ_3 & μ_4 with alteration in seed region $22\% = \frac{22}{100} * 41 = 9$

sequences shown at μ_5 & μ_6 with alteration in xseed region $30\% = \frac{30}{100} * 41 = 12$

sequence shown at μ_7 with alteration in both seed and xseed region $18\% = \frac{18}{100} * 41 = 7$

$\text{Total } (100\%) = 41$

Figure 3.1 *(previous page)*: (A) Example of synthetic reads based on hsa-mir-155 miRNA hairpin structure. The red color shows the seed region, the green color shows the xseed region and capital letters denote altered nucleotide. (B) One example of data from the miRSim pipeline. Here, the miRNA/piRNA region is represented by μ_1, μ_2, ... Here μ_1 represents original miRNA and μ_2-μ_7 are derived from μ_1 by alterations in the seed and xseed sequence of μ_1 and may or may not constitute a valid miRNA. The number of miRNAs present in chromosome-1 and the total number of miRNAs in all chromosomes are taken from miRBase (version22) [4]

One example is shown in Fig.3.1. Here, Fig.3.1(A) shows the hairpin structure of hsa-mir-155 with highlighted seed (red-colored nucleotides in Fig.1A) and xseed (green-colored nucleotides in Figure-3.1(A)) regions. Similarly, Figure-3.1(B) shows one example calculation of synthetic sequence generation by miRSim by doing alterations in the miRNA sequences.

Workflow of miRSim Tool: The miRSim tool accepts reference sequences of miRNAs and their genomic location from the input fasta and gff3 files provided by the user. In addition, the user provides other input parameters such as the total number of sequences to be generated, % of known miRNAs, % of novel miRNAs, % of known piRNAs, quality score encoding, minimum depth and expression profile distribution for generating the synthetic data. The miRSim tool utilizes this information first to infer the distribution of reads (that is, the number of reads per chromosome). The read distribution is directly proportional to the number of miRNAs present in each chromosome. Using the read distribution, the number of miRNAs per chromosome is computed such that each miRNA gets a read depth greater than or equal to the minimum required read depth. Using this miRNA distribution, each chromosome's expression profile is generated based on either the Poisson or the gamma distribution. Finally, the synthetic sequences are generated by adding the adaptor sequence and primer sequence so that the overall sequence length becomes 75, which is written in the output FASTQ/FASTA file. The miRSim tool supports parallel thread operation to write the synthetic sequences in multiple threads simultaneously in order to generate data expeditiously.

miRSim also provides the ground truth in a readable comma-separated file format that contains information about known miRNAs, piRNAs, and novel miRNAs along with their sequences, chromosome location, expression counts, and the CIGAR string for all the sequences. The 'Other' category sequences also specify the known miRNAs/piRNAs (with chromosome location) from which these altered sequence reads are generated besides the above information. Any robust pipeline should discard these noisy reads. miRSim delivers output in the form of a compressed FASTQ or FASTA file format. As of now, the miRSim tool has been developed and tested for the human genome only. For other genomes, miRSim parameters can be readjusted accordingly. For other non-human genomes, a user has to adjust 1). RNA-sequence length for that genome, and 2). seed region and xseed region location. The core algorithm will remain the same. The source code of miRSim is available in GitHub and zenodo in both the command line version and the jupyter notebook version.

3.2.2 Synthetic RNA-seq expression dataset used in this study

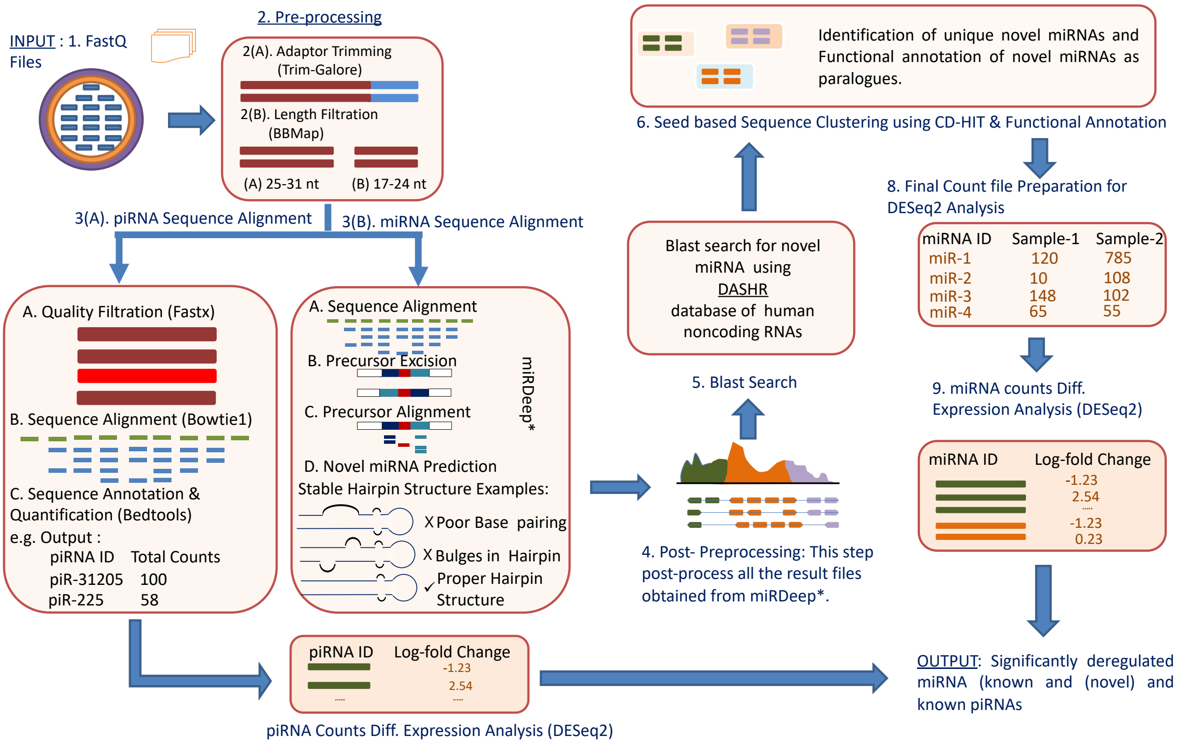
In this study, we have generated the miRSim simulated synthetic dataset for the pipeline benchmarking on the identification of known/novel miRNAs and known miRNAs. Using miRSim, we generated thirty synthetic FASTQ files with a varying read depth of 50K reads, 0.1 million reads, and 1 million reads with known proportions of both altered and unaltered known/novel miRNAs and known piRNAs (Table-3.2). The reason behind considering the multiple read-depth categories is to assess the pipeline for low-depth as well as high-depth data. However, the synthetic data experiments can extend to a higher depth (more than 1 million reads). For known miRNA identification experiments, reads were generated using high-confidence miRNAs taken from miRBase (version 22) to ensure the least false positives or false negatives in the experimental design. Similarly, miRpipe includes the genomic and structural features for novel miRNA identification. Novel miRNAs detected recently in [1, 2] were added to the synthetic data experiments as the ground truth sequences. The complete list of novel miRNAs used in synthetic data experiments is provided at Supplementary Material S9. Moreover, for known piRNA identification, the reads were generated from the piRNAdb database (version 1.7.6). We have also generated the synthetic data file for benchmarking pipelines on the identification of reverse complement sequences known as miRNAs. For this purpose, we have generated a synthetic FASTQ file with the reverse complement reads of 887 high-confidence miRNAs (added from miRBase database version 22) with a read depth of 10 each. Thus, the synthetic FASTQ file contained 8870 reads with a reverse complement of 887 high-confidence miRNAs.

3.2.3 Real RNA-seq expression datasets used in this study

In our study, we have incorporated three publicly available datasets for miRPipe validation: the CLL dataset (GSE123436), the breast cancer dataset (GSE171282), and the lung cancer dataset (GSE37764). In the CLL dataset, the RNA-Seq profile of 28 CLL cases and 10 age-matched healthy controls were studied to identify the unique pattern of eight dysregulated miRNAs in CLL. This study also validated the altered expression levels of eight dysregulated miRNAs by RT-qPCR. The breast cancer dataset (GSE171282) consists of 3 normal and 6 tumor RNA-Seq samples. The breast cancer dataset was studied to understand the effects of two commonly used local anaesthetics, lidocaine and bupivacaine, on the malignancy of MCF-7 breast cancer cells. The original publication of the breast cancer dataset (GSE171282) has reported 11 RT-qPCR-validated dysregulated known miRNAs. We have used CLL and breast cancer datasets for the miRPipe validation in miRNA identification. Similarly, in the lung cancer dataset, the primary non-small cell lung adenocarcinoma tissues of 6 never-smoker Korean female patients were studied to identify dysregulated piRNAs to identify the altered expression patterns of non-coding RNAs in the non-smoker females. The original publication of the lung cancer (GSE37764) dataset has not reported any dysregulated piRNAs. We have used this dataset for miRPipe validation in piRNA identification.

3.2.4 Description of the proposed miRPipe

A complete block diagram of miRPipe is provided in Fig. 3.2.



INPUT : 1. FastQ Files
2. Pre-processing
2(A). Adaptor Trimming (Trim-Galore)
2(B). Length Filtration (BBMap)
(A) 25-31 nt
(B) 17-24 nt
3(A). piRNA Sequence Alignment
3(B). miRNA Sequence Alignment
A. Quality Filtration (Fastx)
B. Sequence Alignment (Bowtie1)
C. Sequence Annotation & Quantification (Bedtools)
e.g. Output :
piRNA ID Total Counts
piR-31205 100
piR-225 58
A. Sequence Alignment
B. Precursor Excision
C. Precursor Alignment
D. Novel miRNA Prediction
Stable Hairpin Structure Examples:
X Poor Base pairing
X Bulges in Hairpin
✓ Proper Hairpin Structure
miRDeep*
4. Post- Preprocessing: This step post-process all the result files obtained from miRDeep*.
5. Blast Search
Blast search for novel miRNA using DASHR database of human noncoding RNAs
6. Seed based Sequence Clustering using CD-HIT & Functional Annotation
Identification of unique novel miRNAs and Functional annotation of novel miRNAs as paralogues.
8. Final Count file Preparation for DESeq2 Analysis
miRNA ID Sample-1 Sample-2
miR-1 120 785
miR-2 10 108
miR-3 148 102
miR-4 65 55
9. miRNA counts Diff. Expression Analysis (DESeq2)
miRNA ID Log-fold Change
-1.23
2.54
-1.23
0.23
piRNA ID Log-fold Change
-1.23
2.54
piRNA Counts Diff. Expression Analysis (DESeq2)
OUTPUT: Significantly deregulated miRNA (known and (novel) and known piRNAs

Figure 3.2: Infographic representation of miRPipe workflow: This pipeline identifies differentially expressed novel miRNAs, known miRNAs, and known piRNAs.The RNA-Seq data in standard FASTQ format is cleaned for adaptor contamination. Reads of specific lengths are aligned to the human reference genome for miRNA and piRNA identification. Further, the aligned reads are processed using novel seed-based clustering for their functional annotation. Lastly, their differential expression analysis is computed using DESeq2 to find the significantly dysregulated miRNAs and piRNAs.

Input Data:

The miRPipe allows both single or multiple sample processing with input files in FASTQ format (either .fastq or .fastq.gz). For computing the differentially expressed miRNAs, miRPipe utilizes the widely accepted DESeq2 method. Importantly, the information on the adaptor sequence, human reference genome, and miRBase version can be either provided by the user, or the default options of miRPipe can be chosen.

Hardware & Software Dependencies

Pipelines: miRPipe, miRDeep*, miRDeep2, mirPRo, sRNAToolbox, miRge2.0, mirnovo, MiR&moRe2 were installed and run on a workstation with the hardware configuration of Single Intel(R) Core(TM) i7-8700 CPU 6Cores,12Threads,@Base frequency of 3.20GHz, 32GB DDR4 RAM. The docker is fully functional on the Linux platform and requires the following system configuration: Ubuntu 18.04 operating system with at least 8 GB RAM. The miRSim tool was developed on the hardware configuration of Single Intel(R) Core(TM) i5-8400 CPU 2Cores, 4Threads, @Base frequency of 2.80GHz, 8GB DDR4 RAM.

3.2.5 miRPipe Workflow

miRPipe is an integrated pipeline for the identification of statistically significant differentially expressed known/novel miRNAs and known piRNAs simultaneously by parallel threaded operations.

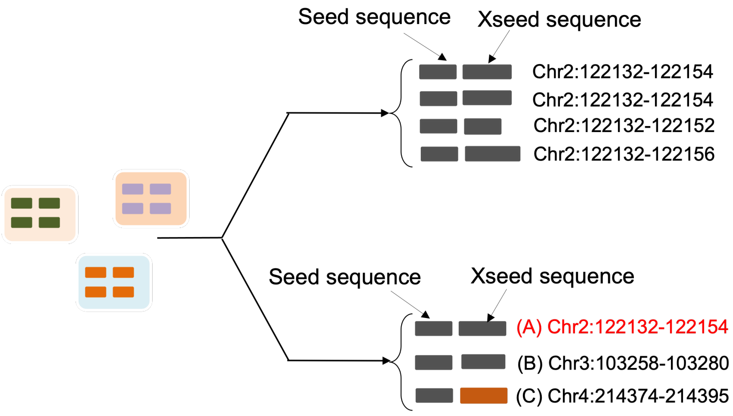
The following steps are sequentially carried out in the miRPipe (Fig. 3.2):

1. *Input:* miRPipe accepts sequencing files in the FASTQ format, along with the sample information file in the CSV format that contains a sample (or subject) IDs and sample group (whether treated and control or the data collected at two different time points).

2. *Pre-processing:* miRPipe performs adapter removal in the raw FASTQ files using the Trim-Galore tool. Post-trimming, miRPipe segregates reads based on their sequence lengths. The first category contains read sequences of 17-24nt lengths that are processed further for miRNA identification. The second category contains read sequences of 25-31nt lengths that are processed for piRNA identification. The remaining read sequences of lengths ≥32nt are rejected by miRPipe.

3. *Sequence alignment:* In Step 3, miRPipe initializes parallel threads for (a) the identification of known and novel miRNAs and (b) the identification of piRNAs. While one CPU thread is allocated for piRNA identification, the remaining CPU threads are allocated for miRNA identification.

   3 (A) *piRNA identification pipeline:* For piRNA identification, reads of length 25-31nt are screened based on their quality scores. Reads having more than



This branch shows four novel miRNAs that have identical seed and <=2nt change in the xseed sequence. All these miRNAs also shares similar genomic locations. Hence, these miRNAs will be merged and assigned one new miRNA ID in Step-6 of miRPipe.

This branch shows three miRNAs. The first miRNA (miRNA (A) with chromosome location in red color) can be either known miRNA or novel miRNA, while the remaining two miRNAs (miRNAs (B) and (C) with chromosome location in black color) are declared as novel miRNAs in Step-3 of miRPipe. The miRNAs (B) and (C) have identical seeds as that of miRNA (A). The miRNA (C) has more than 2nt change in the xseed sequence with that of miRNA (A) (as shown in orange color). Due to identical seed and different genomic locations, miRNAs (B) and (C) are called as paralogues (i.e. functionally same) of miRNA (A) and hence, will be reannotated by appending "_n" (where n = 1,2,3,..) after the name of miRNA (A) in Step-6(B) of miRPipe.

Figure 3.3: Functional annotation of novel miRNAs using seed-based clustering. The above figure shows an example of all possible scenarios for cluster formation of novel miRNAs functional annotation with known and novel miRNAs, along with their genomic location of sequence alignment.

10% bases with a phred score of less than 20 are filtered out. The remaining reads of better quality are aligned to the human genome using the Bowtie2 with the following fixed parameters: (a) minimum length of sequence l = 25,; (b) zero mismatch in the seed region ; (c) with no reverse complement alignment allowed to obtain the alignment results. All the alignment results are then post-processed using the bedtools [181] intersect. miRPipe utilizes piRNA annotations from piRNAdb. Subsequent analysis results in a final count matrix of all the annotated piRNAs across all samples that are used as input for the DESeq2 for the differential gene expression analysis.

3(B) *miRNA alignment*: The first step in miRNA identification is the sequence alignment using miRDeep*. miRDeep* sequence aligner is developed on the top of the bowtie and allows the sequence mapping with zero mismatches in both strands of the human genome reference.

4. *Post-processing*: All known miRNA and novel miRNA reads are collected from all the samples (multiple subjects) to prepare a list of reads for the DASHR blast search processing.

5. *Blast search using DASHR*: All miRNAs that are not annotated as known miRNA are blast-searched with the DASHR database to check if they match with any known miRNA sequences. Moreover, there can be some sequences that are annotated as novel miRNAs, whose annotation is missed due to it being present as

a reverse complement sequence in the fastq file. Although bowtie can map a reverse complement sequence of a known miRNA to its correct genomic location, due to differences in the mapping strand and precursor sequence from the respective mapping strand and precursor sequence of that known miRNA, miRDeep* fails to annotate the reverse complement sequence to its true known miRNA. Such cases are referred to as novel by miRDeep*, although they should have been assigned as known miRNA. Thus, in such cases, we blast search these sequences in the DASHR database and check if they match with any of the known miRNAs. The DASHR database tries to find the best possible match with known miRNAs (according to the reference genome hg19 or hg38 as specified by the user). If the DASHR database does not find any hit with any known miRNA, then we take the reverse complement of these sequences and blast search again in the DASHR database. Now, if they match any known miRNA at the same genomic location as that of the novel miRNA, the novel miRNA will be re-annotated as known miRNA, and the count of the novel miRNA will be merged with that of the known miRNA. For other reads, the miRNA nomenclature system used in miRBase [182] is followed for their renaming, as explained in the next step.

6. *Novel seed-based clustering and functional annotation of novel miRNAs:* In Step-6, miRPipe performs the seed-based clustering on both known and novel miRNA sequences. The methodology of seed-based sequence clustering and functional annotation is as follows:

   6 (A) *Novel seed-based clustering:* In this step, seed-based clustering is employed by miRPipe on known and novel miRNAs to identify unique novel miRNAs and known miRNA paralogues. Different scenarios of seed-based clustering are shown in Fig.3.3. First, we perform CD-hit [138] clustering on the seed sequences of all novel and known miRNAs. novel miRNAs whose seed sequences are identical (that is, 0nt mismatch) are subsequently checked, again via CD-hit clustering, but now on xseed region sequences. All novel miRNA reads having identical seed sequences, maximum alterations of 2nt in the xseed sequence, and similar genomic location (that is, maximum 2nt variation in the genomic position) was identified as a single novel miRNA. Their counts were merged.

   6 (B) *Functional annotation of novel miRNAs:* According to [183], if a given sequence has an identical seed sequence with a different genomic location, such a sequence is called the paralogous. Using these criteria, all novel miRNAs that share an identical seed with different genomic locations are called paralogues. If the novel miRNA has the identical seed as that of a known miRNA (say hsa-mir-x) and a different genomic location, then the novel miRNA will be called a paralogue of known miRNA and will be labeled as "hsa-mir-x_n" where n is 1,2,3,... as more paralogues are discovered. Similarly, if the novel miRNA has an identical seed as that of a novel miRNA (say novel-mir-x) and a different genomic location, then the novel miRNA will be called a paralogue of novel miRNA and will be labeled as "novel-mir-x_n" where n is 1,2,3,... as more paralogues are discovered.

Functionally, the paralogues behave identically [184] due to the same seed in their mature miRNA sequence.

7. *Final count file preparation*: Once the functional annotation of novel miRNAs is completed, miRPipe returns the final count matrix containing expression counts of all novel miRNAs, known miRNAs, and known piRNAs across all the samples. Since there are many miRNAs in miRBase whose mature sequences are identical and located at multiple genomic locations in the human genome. Such miRNAs represent the miRNA paralogues. The sequence aligner in Step 3 of the miRPipe workflow lists all these miRNA paralogues with the same expression counts. Thus, in real RNA-Seq expression data, miRPipe deduplicates the final count file to remove the multiple entries of the same mature sequence present in the count file.

8. *Differential expression analysis*: miRPipe carries out DEMs using the DESeq2 method. Any miRNA or piRNA is considered to be statistically and differentially expressed if its p-adj value after Benjamini-Hochberg (BH) correction is ≤ 0.05.

9. *Renaming of novel miRNAs*: The novel miRNAs identified in CLL datasets are renamed using the miRNA nomenclature system used in miRBase [4]. The rules for miRNA nomenclature are as follows:

   (a) *Previously annotated miRNAs:* If the novel miRNA sequence has already been annotated in another organism, then the same identifier will be used in other organisms also. For this, we have blast-searched all the novel miRNA sequences in the Rfam database with E-values less than or equal to 0.01 and then renamed them using the same identifier.

   (b) *Sequential annotation:* If the above conditions are not met for any novel miRNAs, then the renaming was done sequentially. In the end, we have added "*" in the suffix of all novel miRNA new names to represent that these names are putative names only.

10. *Output*: The final output file in the corresponding user data directory contains the significantly differentially expressed miRNAs and piRNAs.

3.3 Results

Benchmarking of miRPipe with existing standard pipelines

miRPipe is benchmarked with seven standard pipelines introduced in the recent past for the novel miRNA detection. These are mirdeep2, mirdeep*, mirPRo, mirnovo, miRge2.0, sRNAToolbox, and MiR&moRe2. We have proposed an innovative strategy for miRNA pipeline validation and benchmarking, where we have used synthetic RNA-Seq data with known ground truth and real RNA-Seq expression data. The synthetic data experiments allow us to evaluate the accuracy, sensitivity, and specificity for extensive comparison with other pipelines in identifying and annotating correct miRNAs. Hence, results are presented: (1) by comparing the workflow and architecture of all the pipelines, (2) by using the pipelines on the synthetic data generated by the miRSim tool, where the ground truth is known, and (3) by comparing the results obtained on the real RNA-Seq expression CLL, lung cancer and breast cancer dataset.

3.3.1 Summary comparison of different pipelines

We have compared various attributes of all eight bioinformatics pipelines, such as sequence quality control criteria (minimum phred quality score), miRNA sequence length criteria, sequence aligner used for sequence alignment, type of miRNA annotation database (miRBase/miRGeneDB), RNA profiles (types of other sncRNA), model types (whether based on machine learning or not), platform dependencies (Windows/Linux/MAC OSX), etc. and shown them in Table-3.1. The pre-processing of sequencing data in any bioinformatics pipeline includes 1. adaptor-trimming, 2. sequence quality control, and 3. sequence length control. The adaptor trimming and sequence quality control steps are mandatory steps to prepare reads for downstream analysis, such as sequence alignment with respect to the reference genome, miRNA annotation, etc. In the sequence length filtration step, all eight pipelines have different criteria for miRNA sequence length, e.g., miRDeep2 and miRDeep* consider sequences having a length range of 18nt-23nt as miRNA. In contrast, miRPro considers all the sequences having a length greater than 17nt as miRNA, etc.

We have observed that most of the human miRNAs lie in the range of 17nt-24nt. Also, both miRNAs and piRNAs sequences have slightly varying lengths across different miRNAs and piRNAs instead of strictly defined fixed lengths. Thus, it is challenging to find the exact cutoff of sequence length that can help infer a maximum number of true positive miRNAs or piRNAs. We have observed that there are 1.3% miRNAs that are present in the miRBase database, having sequence lengths of 25nt to 28nt. At the same time, there are 21.3% piRNAs that are present in the piRNAdb database, having sequence lengths of 25nt to 28 nt. Considering sequences of length 25nt to 28nt for miRNAs may lead to false negatives in the piRNAs pipeline, and we may miss many important piRNAs. Similarly, rejecting a sequence of length less than 18nt may lead to ignoring the actual miRNAs. Hence, we consider the sequence length range of 17nt-24nt for miRNAs and 25nt-31nt for piRNAs. Thus, all the steps in sequencing data pre-processing, especially length filtration, play a crucial role in pipeline performance evaluation.

In addition, selecting sequence aligners also plays an important role in sncRNA identification. Most of the pipelines have used bowtie1 except mirPRo, mirnovo, and miRPipe because bowtie1 is very sensitive to small and medium-length sequences. Currently, Bowtie1 is deprecated, and the development team does not provide technical support.

For miRNA annotation, miRBase is considered a standard database used in all eight pipelines except miRge2.0, which lets the user choose the annotation database (either miRBase or miRGeneDB) as per user requirement. In addition to identifying known and novel miRNAs, some pipelines like mirnovo, miRge2.0, MiR&moRe2, and sRNA-Toolbox provide information about other RNA types such as tRNA, rRNA, snoRNA, microRNA-offset RNAs (moRNA), loop-RNA, etc. Only three out of the seven existing pipelines (mirnovo and miRge2.0, sRNAToolbox) use the machine learning-based model (random forest, SVM, and Weka, respectively) for novel miRNA prediction. Also, each pipeline has been developed using a different programming language and has different platform dependencies. The difference in the miRNA selection criteria, sequence alignment strategy, annotation database, the model used for sncRNA identification, etc., makes these pipelines methodologically unique and causes different outputs for the same input data.

Table 3.1: Comparison of recently published bioinformatics pipelines on all the intermediate steps such as sequence pre-processing, de-duplication, sequence alignment, and sequence annotation. In all the intermediate steps, each pipeline uses different tools (with different versions) or their own module written in languages such as Python, Perl, R or C++. Further, each pipeline has its own miRNA consideration criteria. For example, miRge2.0 pipeline considers 16-25nt length of sequence for miRNA identification, while sRNAToolbox considers all the sequences of length less than 25nt for miRNA identification. The variation in length criteria significantly impacts the accuracy of the miRNA sequence alignment step, which is the most crucial step in the bioinformatics pipeline for miRNA identification. The majority of the above-mentioned pipelines use Bowtie1 for sequence alignment, while the mirnovo pipeline uses Bowtie2, and the mirPRo pipeline uses a Novoalign sequence aligner. In our proposed pipeline miRPipe, we have used the miRDeep* (in Step 3 of the workflow) with DASHR blast search (in Step 5 of the workflow) and seed-based clustering of the novel miRNAs (in Step 6 of the workflow). Most of the pipelines report other categories of RNAs present in the sequencing data, such as rRNA, moRNA, piRNA, etc. We have integrated our pipeline piRNA identification pipeline in miRPipe with parallel thread execution for optimum use of computational resources to facilitate less overall time to deliver the output results.

Pipeline Steps	Modules	Pipelines							
		miRDeep2 [74]	miRDeep* [75]	mirPRo [76]	Mirnovo [77]	miRge2.0 [78]	sRNAToolbox [79]	MiR&moRe2 [80]	miRPipe
Year		2012	2013	2015	2017	2018	2019	2020	
Pre-processing	Adaptor Trimming	✓	✓	✓	Reaper	Cutadapt (v1.11)	✓	Cutadapt (v2.5)	Trim-Galore
	Quality Control (Min Phred Score)	20	20	✓	×	20	20	20	20
	Length Control	18-23 nt	18-23 nt	>17 nt	×	16-25 nt	<25 nt	15-30 nt	17-24 nt (miRNA) 25-31 nt (piRNA)
Sequence Deduplication	Collapsed to Unique reads	✓	✓	✓	Tally + vsearch + CD-HIT	✓	✓	✓	✓
Sequence Alignment w.r.t. Reference Genome	Sequence Alignment	Bowtie (v1.1)	Bowtie (v0.1)	Novoalign + HTSeq	Bowtie2 + mirnovo_ analysis.pl module	Bowtie (v1.1.1) + Samtools	Bowtie (v0.12)	Bowtie (v1.1.2) + Samtools + bedtools (v2.27)	Bowtie2
Sequence Annotation	miRNA Database	miRBase	miRBase	miRBase	miRBase	miRBase/ miRGeneDB	miRBase	miRBase	miRBase
	Other RNA Database	×	×	×	Rfam	Ensembl	Provided by user	✓ (Experimental)	piRNADb
	miRNA Identification	✓	✓	✓	✓	✓	✓	✓	✓
	Isomirs	✓	✓	✓	✓	✓	✓	✓	✓
	novel miRNAs	✓	✓	✓	✓	✓	✓	✓	✓
	Other sncRNA	×	×	×	rRNA, tRNA	Primary tRNA, rRNA, snoRNA, known RNA spike-in sequence	tRNA, snoRNA, snRNA, rRNA, piRNA	moRNA, loop RNA	piRNA
Features	Genomic features	×	×	×	9	×	×	×	×
	Coverage Profile features	×	×	×	12	×	×	×	×
	Sequence Features	✓	✓	✓	12	21	✓	✓	✓
Model Type	Machine Learning based	No	No	No	Random Forest	Support Vector Machine	Weka	No	No
Application Genome	Human	✓	✓	✓	✓	✓	✓	✓	✓
	Non-human	✓	✓	Mouse, Chicken	8 animal species and 7 plant species	novel miRNA prediction only for Mouse genome	✓	×	×
Language	Programming Language	Perl, Bash	Java	C++	Perl, R	Python 2.7	Web Server	Pytyon3, R Bash	Python3, Bash
	Packages				Random Forest R Package	Biopython, Numpy, Scipy, pandas, sklearn, reportlab, forgi python packages		data.table R Packages	
OS Supported		Linux	Linux, MACOSX Windwos	Linux	Linux MACOSX	Linux	Not required	Linux	Linux, MACOSX & Windows

3.3.2 Benchmarking of pipelines on synthetic data

Synthetic data is generated using the miRSim tool for pipeline validation with the known percentage of reads of known/novel miRNAs and known piRNAs (Refer to Table-3.2 for an example). Since the ground truth is known, pipelines are assessed on the metrics of accuracy and F_1-score. The following notations are used for four class label detection: Class-1: known miRNA, class-2: novel miRNA, class-3: known piRNA, and class-4: Not belonging to other 3 classes.

Table 3.2: An example of synthetic data generation by *miRSim* tool, where the user specifies the fraction of sequences of a particular RNA type and a particular sub-category of sequence in the synthetic data required to be generated. Say synthetic data is generated having a total of 10K reads. Of these 10K reads, 20% belong to known miRNAs, 10% belong to novel miRNAs (sequences collected from the [1, 2]), and 10% reads are of the known piRNAs. The reads of the 'Other' class are generated by making random alterations to the sequences of known miRNA, novel miRNAs, and to those of the known piRNAs. In this example, the user has specified to generate 10% reads (of overall total read count) by making alterations to the seed regions of known miRNAs,10% by making alterations to the xseed regions, and 5% by making alterations to both seed and xseed regions of known miRNAs to simulate the reads of the 'Other' class. The % fraction of each sub-category of other classes for other sncRNAs (novel miRNAs and known piRNAs) are also mentioned in this example.

RNA-Type	% fraction of actual miRNA	% fraction of other class		
		Generated by altering seed regions of miRNAs/piRNAs	Generated by altering xseed regions of miRNAs/piRNAs	Generated by altering both seed and xseed regions of miRNAs/piRNAs
Known miRNA	20	10	10	5
Novel miRNA	10	7	7	3.5
Known piRNA	10	7	7	3.5

- A read is counted as true positive (TP) if the pipeline correctly identifies it.
- A read is considered as false positive (FP) to class-'x' when it belongs to one of the other classes but is identified as a class-'x' read.
- A read is considered as false negative (FN) when it belongs to class-'x' but is rejected by the pipeline to class-4 (Others).
- A read is considered as true negative (TN) when it is an altered miRNA/piRNA and is also labeled the same by the pipeline, that is, all the reads with sequences not belonging to valid miRNAs or piRNAs or novel miRNAs (in other words, not belonging to any of the above three classes) are called as true negatives.

The performance metrics are computed as follows.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$

$$\text{F1 Score} = 2\Big(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\Big),$$

where

$$\text{Precision} = \frac{TP}{TP + FP}, \text{ and Recall} = \frac{TP}{TP + FN}.$$

Benchmarking of pipelines on the identification of known miRNAs

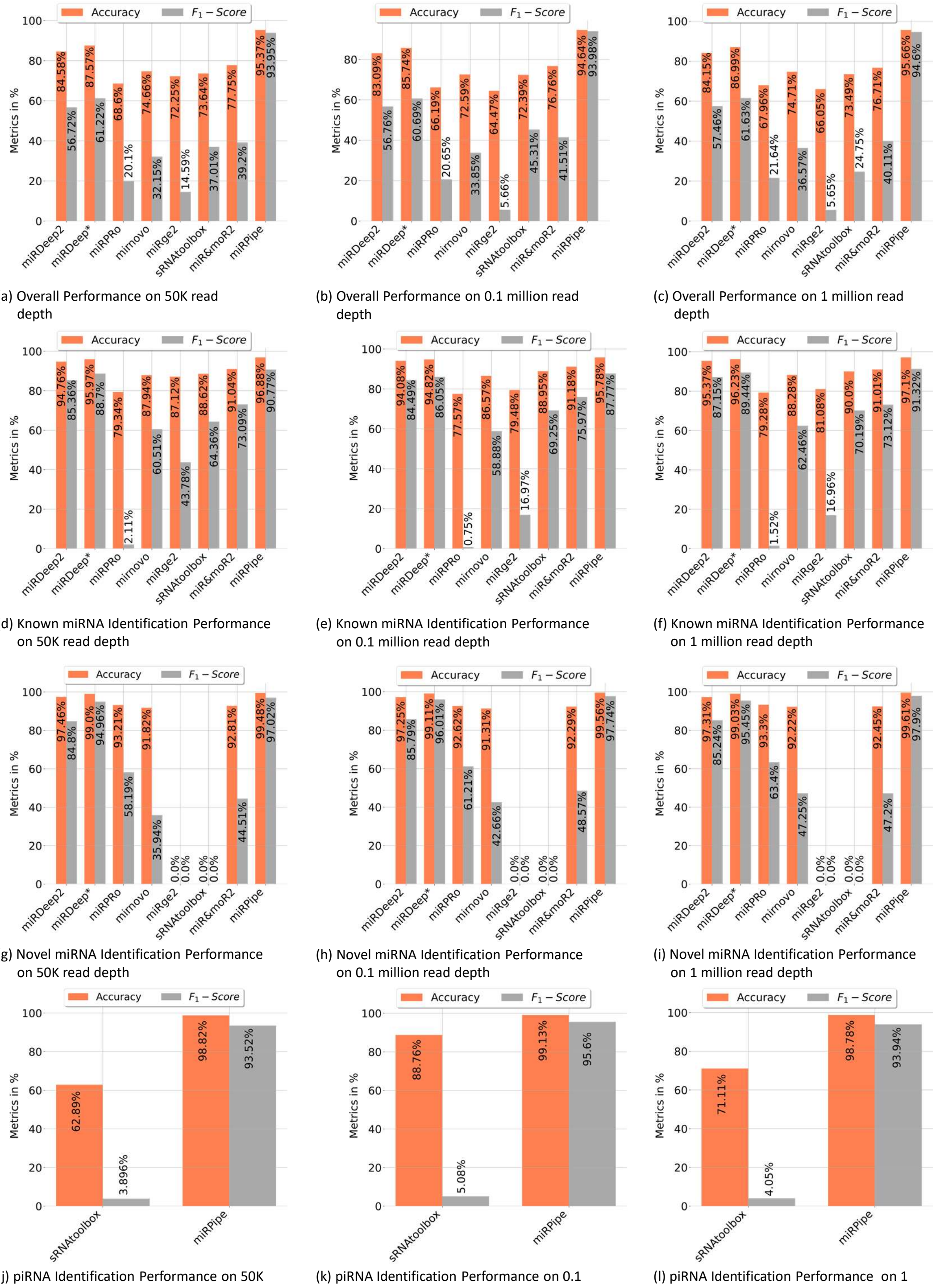
In the synthetic data experiment for identifying known miRNAs, we have observed that many mature miRNA sequences can match at multiple genomic locations on the human genome, wherein the miRNAs at these different genomic locations correspond to different precursor sequences. For such cases, we have compared the miRPipe outcome with the miRSim-generated ground truth for miRPipe pipeline assessment. The comparative analysis with miRPipe revealed that miRPipe outperforms existing pipelines with an average accuracy (across all depths) of 96.58% and an average F_1-score of 89.95% on the identification of known miRNAs (Table-3.4, Fig. 3.4, and Table-3.3a, Supplementary Material S1-S3, synthetic data experiment results for 50K, 0.1M and 1M read depth).

Benchmarking of pipelines on the identification of novel miRNAs

Comparison of inter-computational pipelines revealed that miRPipe outperformed all the other computational methods, with an average accuracy of 99.55% and average F1-score of 97.55% across all depths in synthetic data experiments on novel miRNA identification (Table-3.4, Fig. 3.4, and Table-3.3b, Supplementary Material S1-S3.

Benchmarking of pipelines on the identification of Known piRNAs

We performed a comparative analysis with sRNAToolbox, which uses RNAcentral for piRNA and other sncRNA annotations for piRNA identification. Notably, it is the only dedicated computational workflow that allows simultaneous identification of miRNAs and piRNAs. While miRPipe yielded an average accuracy of 98.91% and an average F1-score of 94.35%, sRNAToolbox yielded an average accuracy of 74.25% and an average F1-score of 4.34% across all depths (Table-3.4, Fig. 3.4, and Table-3.3c, Supplementary Material S1-S3.



(a) Overall Performance on 50K read depth

(b) Overall Performance on 0.1 million read depth

(c) Overall Performance on 1 million read depth

(d) Known miRNA Identification Performance on 50K read depth

(e) Known miRNA Identification Performance on 0.1 million read depth

(f) Known miRNA Identification Performance on 1 million read depth

(g) Novel miRNA Identification Performance on 50K read depth

(h) Novel miRNA Identification Performance on 0.1 million read depth

(i) Novel miRNA Identification Performance on 1 million read depth

(j) piRNA Identification Performance on 50K read depth

(k) piRNA Identification Performance on 0.1 million read depth

(l) piRNA Identification Performance on 1 million read depth

Figure 3.4: Benchmarking of miRPipe with the existing pipelines on synthetic data. Averaged results are reported over 10 FASTQ files generated for each read depth of 50k, 0.1 million, and 1 million reads. The overall performance of all existing pipelines is shown in (a) to (c). All the pipelines are benchmarked against miRPipe for known miRNAs in (d) to (f) and for novel miRNAs in (g) to (I). Among the existing pipelines, only sRNAToolbox identifies piRNA, and hence, comparison results of miRPipe for piRNA are compared with only sRNAToolbox in (j) to (l).

Table 3.3: Average performance of pipelines for the identification of (A) known miRNAs, (B) novel miRNAs and (C) known miRNAs for three categories of read depth (50K, 0.1M, and 1M, where M represent a million reads). The cells with '-' indicate that the pipeline does not identify that particular type of RNA.

(a) Average performance of pipelines for identification of Known miRNA

Pipelines	50K read depth		0.1M read depth		1M read depth	
	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
miRDeep2	94.76	85.36	94.08	84.49	95.37	87.15
miRDeep*	95.97	88.70	94.82	86.05	96.23	89.44
mirPRo	79.34	2.11	77.57	0.75	79.28	1.52
mirnovo	87.94	60.51	86.57	58.88	88.28	62.46
miRge2.0	87.12	43.78	79.48	16.97	81.08	16.96
sRNAToolbox	88.62	64.36	88.95	69.25	90.00	70.19
MiR&moRe2	91.04	73.09	91.18	75.97	91.01	73.12
miRPipe	**96.88**	**90.77**	**95.78**	**87.77**	**97.10**	**91.32**

(b) Average performance of pipelines for identification of Novel miRNA

Pipelines	50K read depth		0.1M read depth		1M read depth	
	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
miRDeep2	97.46	84.79	97.25	85.79	97.31	85.24
miRDeep*	98.99	94.96	99.11	96.01	99.03	95.45
mirPRo	93.20	58.19	92.62	61.21	93.30	63.40
mirnovo	91.82	35.94	91.31	44.66	92.22	47.25
miRge2.0	-	-	-	-	-	-
sRNAToolbox	-	-	-	-	-	-
MiR&moRe2	92.81	44.51	92.30	48.57	92.45	47.21
miRPipe	**99.48**	**97.01**	**99.56**	**97.74**	**99.61**	**97.90**

(c) Average performance of pipelines for identification of Known piRNA

Pipelines	50K read depth		0.1M read depth		1M read depth	
	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
miRDeep2	-	-	-	-	-	-
miRDeep*	-	-	-	-	-	-
mirPRo	-	-	-	-	-	-
mirnovo	-	-	-	-	-	-
miRge2.0	-	-	-	-	-	-
sRNAToolbox	62.89	3.98	88.76	5.08	71.11	4.06
MiR&moRe2	-	-	-	-	-	-
miRPipe	**98.82**	**93.52**	**99.13**	**95.61**	**98.79**	**93.95**

(d) Overall Average performance of pipelines for identification of sncRNA

Pipelines	50K read depth		0.1M read depth		1M read depth	
	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
miRDeep2	84.58	56.72	83.09	56.76	84.15	57.46
miRDeep*	87.57	61.22	85.74	60.69	86.99	61.63
mirPRo	68.60	20.10	66.19	20.65	67.96	21.64
mirnovo	74.66	32.15	72.59	33.85	74.71	36.57
miRge2.0	72.25	14.59	64.47	5.66	66.05	5.65
sRNAToolbox	73.64	37.01	72.39	45.31	73.49	24.75
MiR&moRe2	77.75	39.20	76.76	41.51	76.71	40.11
miRPipe	**95.37**	**93.96**	**94.64**	**93.99**	**95.66**	**94.60**

Table 3.4: Average performance of pipelines for known miRNA, novel miRNA and known piRNAs. The cells with '-' indicate that the pipeline does not identify that particular type of RNA.

Pipelines	Average Accuracy across all depths (in %)			Average F1 score across all depths (in %)		
	known miRNA	novel miRNA	known piRNA	known miRNA	novel miRNA	known piRNA
miRDeep2	94.74	97.33	-	85.66	85.27	-
miRDeep*	95.67	99.04	-	88.06	95.47	-
mirPRo	78.73	93.04	-	1.45	60.93	-
mirnovo	87.59	91.78	-	60.61	41.95	-
miRge2.0	82.56	0.0	-	25.90	0.0	-
sRNAToolbox	89.18	0.0	74.25	67.93	0.0	4.34
MiR&moRe2	91.07	92.52	-	74.05	46.76	-
miRPipe	**96.58**	**99.55**	**98.91**	**89.95**	**97.55**	**94.35**

Table 3.5: Comparison of pipeline performance in CLL real RNA-Seq expression dataset

S. No.	Pipeline	No. of dysregulated miRNA identified by pipeline	Number of miRNAs validated with RT-qPCR results	% False Positive	% False Negative
1	miRDeep2	29	9	68.96	70.96
2	miRDeep*	22	10	54.54	67.74
3	miRPro	34	12	64.70	61.29
4	mirnovo	32	6	81.25	80.64
5	miRge2.0	25	4	84	87.09
6	sRNAToolbox	5	1	80	96.77
7	MiR&moRe2	5	1	80	96.77
8	**miRPipe**	**31**	**17**	**45.16**	**45.16**

Overall benchmarking of all the pipelines

The overall cumulative performance of all the pipelines is done by considering known/novel miRNAs and piRNAs in the synthetic data experiments, and they are reported in Fig. 3.4. In consistency with the previous results, cumulative performance of miRPipe revealed an average accuracy of 95.22% and an average F_1-score of 94.17% across all depths, a way higher than all tested alternative computational methods (Table-3.4, Fig. 3.4, and Table-3.3d, Supplementary Material S1-S3.

Benchmarking of pipelines on the identification of reverse complement miRNA sequence as known miRNAs

We have also benchmarked miRPipe with seven standard pipelines introduced in the recent past for the annotation of reverse complement sequences as known miRNAs. As mentioned in Section 3.2.2 (Synthetic RNA-seq expression dataset used in this study), we have generated the synthetic data for pipeline benchmarking on the annotation of reverse complement sequence using 887 high-confidence miRNAs in the miRBase database (version 22). The comparative analysis with miRPipe revealed that miRPipe outperformed existing pipelines with an accuracy of 42.16% and an F1-score of 59.31%. We have observed that miRDeep2, miRDeep*, miRPro, mirnovo, miRge2.0, sRNAToolbox, MiR&moRe2, and miRPipe has identified 4, 35, 7, 0, 6, 56, 0, and 374 miRNAs respectively out of 887 high confidence miRNAs in miRBase database (version 22). We have shown the pipeline performance comparison on the identification of the reverse complement miRNA sequence in Fig.3.5. Although miRPipe has also missed annotating many reverse complement miRNAs, miRPipe has still successfully identified the most number of reverse complement sequences as known miRNAs among all eight pipelines.

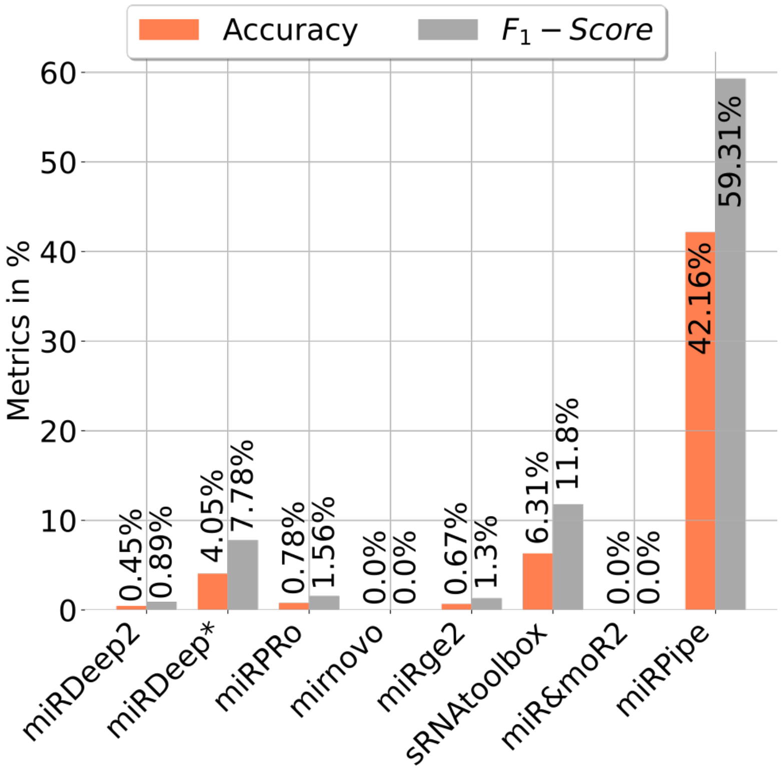


Figure 3.5: Benchmarking of miRPipe with the seven standard pipelines on the identification of reverse complement miRNA sequence as known miRNA.

3.3.3 miRPipe validation on publicly available CLL dataset (GSE123436)

We have validated miRPipe with the publicly available CLL real RNA-Seq expression dataset (GSE123436) for miRNA identification. In the CLL dataset, miRNA profiling of 28 CLL cases and ten age-matched healthy controls were studied to understand the involvement of dysregulated miRNAs in CLL and their impact on clinical outcomes.

Results of all pipelines on CLL dataset (GSE123436)

A total of 31 known miRNAs were found to be dysregulated by the miRPipe pipeline on the CLL real RNA-Seq expression dataset (GSE123436). Out of 31 dysregulated known miRNAs, 24 miRNAs were found to be upregulated, and 7 miRNAs were downregulated. On the other hand, we have observed that miRDeep2, miRDeep*, miRPro, mirnovo, miRge2.0, sRNAToolbox, and MiR&moRe2 have identified 29, 22, 34, 32, 25, 5, and 5 dysregulated known miRNAs respectively. Further, miRPipe has identified 28 dysregulated piRNAs in CLL real RNA-Seq expression data. Out of 28 dysregulated piRNAs, one piRNA was found to be up-regulated, and the remaining 27 were down-regulated (shown in Table-3.6). The average percentage of known piRNAs across CLL samples was observed as 5.94%, calculated as,

$$\% \text{ piRNAs across CLL samples} = \left(\frac{n_{pi}}{n_{pi} + n_{mi}}\right) * 100,$$

where

n_{pi} = Total no. of sequences annotated as piRNAs, and
n_{mi} = Total no. of sequences annotated as miRNAs.

Literature validation of all pipelines on CLL dataset (GSE123436)

According to the original publication of this dataset [1], eight miRNAs were found as dysregulated in the CLL real RNA-Seq expression dataset. Out of 8 dysregulated miRNAs reported in the original publication, there were five common miRNAs reported by miRPipe. In comparison of dysregulated known miRNAs identified by miRDeep2, miRDeep*, miRPro, mirnovo, miRge2.0, sRNAToolbox, MiR&moRe2 and miRPipe with the literature, 17 out of 29 (58.62%), 18 out of 22 (81.81%), 21 out of 36 (58.33%), 14 out of 25 (56%), 8 out of 25 (32%), 2 out of 5 (40%), 2 out of 5 (40%) and 27 out of 31 (87.09%) miRNAs, respectively, were found to be reported as dysregulated in the literature of CLL. Here, the dysregulated miRNAs identified by miRPipe are found to be reported in multiple CLL-related literature [65, 185, 186, 187, 188, 189, 190, 143, 191, 192, 193]. To the best of our knowledge, we have also reported 28 dysregulated piRNAs in CLL, which no one has reported to date.

Comparison of all pipeline results with RT-qPCR on CLL dataset (GSE123436)

Next, we compared the results of DEMs obtained from all the pipelines on the CLL real RNA-Seq expression dataset. Further, we validated these findings using semi-quantitative real-time PCR. The miRNA profiling was carried out on treatment naïve 28 CLL cases using the TaqMan Array Human MicroRNA Card A+B v2.0 (Applied Biosystems, CA, USA), each of which profiles 380 TaqMan MicroRNA Assays enabling the simultaneous quantitation of 754 (377+377) human miRNAs plus 4 endogenous controls. The data was normalized using three endogenous controls: U6 snRNA, RNU48 and RNU44. The results obtained were also validated in additional cohorts of de novo CLL patients using the miRCURY LNA™ Universal RT microRNA PCR System (Exiqon). Our group also generated additional data on 89 CLL patients in the same CLL cohort

Table 3.6: Differentially expressed piRNAs in CLL dataset

S. No.	piRNA	up/down Regulation	Fold Change
1	hsa-piR-23019	down	-3.15
2	hsa-piR-23020	down	-3.27
3	hsa-piR-32157	down	-3.21
4	hsa-piR-32158	down	-3.22
5	hsa-piR-32159	down	-3.22
6	hsa-piR-32160	down	-3.22
7	hsa-piR-32161	down	-3.22
8	hsa-piR-32166	down	-3.21
9	hsa-piR-32178	down	-3.21
10	hsa-piR-32181	down	-3.15
11	hsa-piR-32185	down	-3.27
12	hsa-piR-32186	down	-3.22
13	hsa-piR-32194	down	-3.27
14	hsa-piR-32234	down	-3.27
15	hsa-piR-32237	down	-3.27
16	hsa-piR-32838	down	-3.22
17	hsa-piR-32839	down	-3.22
18	hsa-piR-32845	down	-3.18
19	hsa-piR-32852	down	-3.19
20	hsa-piR-32978	down	-3.72
21	hsa-piR-32995	down	-3.72
22	hsa-piR-33013	down	-3.75
23	hsa-piR-32963	up	3.46
24	hsa-piR-32990	down	-1.54
25	hsa-piR-33010	down	-1.54
26	hsa-piR-33053	down	-1.59
27	hsa-piR-32847	down	-1.99
28	hsa-piR-32920	down	-1.16

enrolled in the study, as mentioned in Kaur et al. [1]. A unique set of 68 DEMs was validated out of 754 miRNAs tested with TaqMan Array Human MicroRNA Card A+B v2.0 (Applied Biosystems, CA, USA). The complete list of RT-qPCR validated 68 DEMs in CLL is provided at Supplementary Material S5.

We have compared the number of DEMs identified by the existing pipelines to the results of RT-qPCR to assess the pipeline performance. In the comparison of pipeline performance for miRNA identification, we have observed that each pipeline (including miRPipe) has detected many dysregulated miRNAs (DEMs). However, after comparing them with the results of RT-qPCR, the true DEM count decreased considerably. This is because it is practically difficult to test and validate all the predicted DEMs in the laboratory for at least three reasons: (a) The assays used for DEM validation may not contain all the predicted DEMs, (b) the limitation of sample material available, and (c) it adds a huge cost and extra working hours. Hence, only the topmost or prioritized DEMs are preferably tested and validated. The false-positive miRNAs are the miRNAs that are identified as dysregulated by the pipeline but not validated in RT-qPCR experiments. Similarly, the false-negative miRNAs are the miRNAs that are not identified as dysregulated by the pipeline but are RT-qPCR validated. The $\%$ false positives and $\%$ false negatives of the pipeline are computed as,

$$\% \text{ False Positives} = \left(1 - \frac{N_{RTmiR}}{N_{TotalmiR}}\right) * 100,$$

where

N_{RTmiR} = Number of RT-qPCR validated miRNAs identified by the pipeline and
$N_{TotalmiR}$ = Total number of miRNA identified by the pipeline, and

$$\% \text{ False Negatives} = \left(1 - \frac{N_{RTmiR}}{N_{TotalRTmiR}}\right) * 100,$$

where

N_{RTmiR} = Number of RT-qPCR validated miRNAs identified by the pipeline and
$N_{TotalRTmiR}$ = Total number of RT-qPCR validated miRNA.

We combined the RT-qPCR-validated miRNAs identified by all eight pipelines to get the total number of RT-qPCR-validated miRNAs, which gives a total of 134 miRNAs. Out of 134 miRNAs, 31 miRNAs were found to be RT-qPCR validated. The miRPipe has outperformed all other pipelines with the least false positives and false negatives. In miRPipe, out of 31, 17 miRNAs are found as RT-qPCR validated, giving the least false positives, that is, (1-17/31*100=45.16%) and least false negatives (1-17/31*100=45.16%) among all eight pipelines. The $\%$ false positives and $\%$ false negatives for the remaining seven pipelines are shown in Table-3.5 and Supplementary Material S4.

3.3.4 miRPipe validation on publicly available Lung Cancer dataset (GSE37764)

We have validated miRPipe with the publicly available lung cancer data set (GSE37764) for the identification of piRNAs. In the lung cancer dataset, the role of dysregulated miRNAs and piRNAs in nonsmoking female lung cancer patients was studied. Among eight pipelines used for benchmarking, only miRPipe and sRNAToolbox identify piRNAs. According to the original publication of this dataset [162], no piRNAs were found to be dysregulated in RNA-Seq samples of non-smoker females. However, a total of 18 and 20 dysregulated piRNAs were identified by the miRPipe and sRNAToolbox, respectively. The complete list of dysregulated piRNAs in lung cancer dataset obtained from miRPipe and sRNAToolbox pipeline is provided at Supplementary Material S7. The two pipelines detected no common piRNA. Out of the 18 piRNAs (identified by miRPipe), 6 piRNAs (33.33%) were found to be reported as dysregulated in lung adenocarcinoma in the literature [194]. On the contrary, none of the piRNAs identified by sRNAToolbox were found to be reported in the literature.

3.3.5 miRPipe validation on publicly available Breast Cancer dataset (GSE171282)

We have also validated the miRNA identification pipeline in miRPipe with a publicly available breast cancer dataset (GSE171282). In [161], 11 dysregulated miRNAs were identified to understand their involvement in the effects of anaesthetics on breast cancer cells. We have observed that miRDeep2, miRDeep*, miRPro, mirnovo, miRge2.0, sRNAToolbox, MiR&moRe2, and miRPipe have identified 22, 8, 31, 29, 34, 14, 42, and 21 known dysregulated miRNAs respectively. The complete list of dysregulatd miRNAs in the breast cancer dataset obtained from all eight pipelines is provided at Supplementary Material S8. In comparison with the literature reported miRNA 9 out of 22 (40.90%), 7 out of 8 (87.5%), 10 out of 31 (32.25%), 8 out of 29 (27.58%), 19 out of 34 (55.88%), 12 out of 14 (85.71%), 23 out of 42 (54.76%), and 19 out of 21 (90.47%) miRNAs were found to be reported as dysregulated in the literature of breast cancer. Here, the dysregulated miRNAs identified by miRPipe are found to be reported in multiple breast cancer-related research papers [195, 196, 197, 198, 199, 200, 201, 202, 203, 204, 153, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215]. Only 6 out of 11 (54.54%) miRNAs reported in the original publication of this dataset [161] were found to be reported as dysregulated in the literature. Of the pipelines compared, miRPipe and MiR&moRe2 reported miRNAs matched most with the literature (19 and 23, respectively). However, miRPipe has the least number of FPs because of the 21 reported by miRPipe, 19 matched with the literature.

3.4 Discussions

In this work, we have benchmarked our pipeline, miRPipe, with seven recent pipelines (miRDeep2, miRDeep*, mirPRo, mirnovo, miRge2.0, sRNAToolbox, and MiR&moRe2) using a newly developed synthetic RNA-sequence simulator, miRSim tool that generates FASTQ file with known fraction of altered/unaltered known/novel miRNAs and piRNAs, and help evaluate pipelines on identifying true positives and rejecting false miRNA/piRNA reads looking similar to known miRNAs/piRNAs.

3.4.1 Specificities of miRSim Synthetic Sequence Simulator

Numerous read simulators for generating synthetic RNA-Seq data are documented in the literature. The list of 15 such simulators and their respective characteristics is summarized in Table- 3.7. While most of these simulators adhere to Illumina-based error models, a subset—such as RNA-Seq Simulator, RSEM, and Polyester offers the flexibility to customize error profiles by adjusting parameters like substitution rate and fraction of indels. Typically, users are required to input read abundances alongside their desired error profiles. However, simulators like RSEM and CAMPAREE accept real RNA-Seq samples in FASTQ format, and BEERS2 utilizes output from CAMPAREE to generate synthetic RNA-Seq data.

Notably, only five simulators — ART, NEAT, BEERS2, CAMPAREE, and RSEM provide ground truth for the generated synthetic data. Additionally, it's worth highlighting that most of these simulators focus on generating full-length synthetic mRNA transcriptome sequencing data; our study aims at sncRNA sequencing, encompassing miRNAs and piRNAs. To bridge this gap, we developed a sncRNA sequencing simulator named miRSim. To the best of our knowledge, the specificities of the miRSim synthetic sequence simulator are as follows:

1. While numerous synthetic sequence simulators exist for generating RNA-Seq data, the existing simulators are primarily designed for generating full-length mRNA transcriptome data, neglecting the need for synthetic sncRNA Seq data (Table- 3.7). To address this gap, we developed miRSim, which is designed explicitly for generating synthetic sncRNA Seq data, which is crucial for assessing the performance of sncRNA identification pipelines.
2. We performed synthetic data experiments by generating the synthetic RNA Seq data using the available simulators and observed that none of them were producing valid miRNAs. To the best of our knowledge, these simulators overlook the essential criteria for valid miRNA sequences, zero alterations in seed and xseed region and the presence of a valid precursor sequence with a stable hairpin structure. To bridge this gap, we developed miRSim to generate the synthetic sncRNA Seq data by leveraging the sequence information (seed and xseed region) and accepting the customized reference from the miRbase database to generate a synthetic sequence with valid precursor and stable hairpin structure.

Table 3.7: Characteristics of 15 previously published synthetic DNA/RNA sequence simulators.

Synthetic sequence simulator	Simulator Characteristics							
	Read Length	Read Type	Input	Error Model	Quality Score Profile	Multi-threading	Ground Truth	Sequence Type
RNA-Seq Simulator	Variable	Single-end, paired-end	Number of reads and positional error profile	Custom	No	No	No	mRNA
RSEM	Variable	Single-end, paired-end	Fastq Format Sequence	Custom	Yes	Yes	Yes	mRNA
Polyester	Variable	Single-end, paired-end	Number of reads	Custom	No	No	No	mRNA
CAMPAREE	Variable	Single-end, paired-end	Fastq format sequence	Illumina	Yes	No	Yes	mRNA
BEERS2	Variable	Single-end, paired-end	Accept output from CAMPAREE	Illumina	Yes	No	Yes	mRNA
ART	36bp-250bp	Single-end, paired-end	Number of reads, custom error profile	Illumina	Yes	No	Yes	DNA/mRNA
NEAT	Variable	Single-end, paired-end	Coverage	Illumina	No	No	Yes	DNA/mRNA
DWGSIM	Variable	Single-end, paired-end	Number of reads	Illumina	Fixed Score	No	No	DNA/mRNA
ISS	125bp-301bp	Paired-end	Number of reads	Basic	No	Yes	No	DNA/mRNA
Mason	Variable	Single-end, paired-end	Number of reads	Illumina	Yes	Yes	No	DNA/mRNA
wgsim	Variable	Single-end	Number of reads	Illumina	No	No	No	DNA/mRNA
Flux	50bp-150bp	Single-end, paired-end	Number of reads	Illumina	Yes	No	No	mRNA
SimNGS	Variable	Single-end, paired-end	Covariance of noise between the base	Illumina	Yes	No	No	DNA/mRNA
pIRS	Variable	Paired-end	Number of reads, error profile	Illumina	Yes	Yes	No	DNA/mRNA
SimSeq	Variable	Paired-end, mate-pair	Number of reads, error profile	Illumina	Yes	No	No	mRNA
miRSim	**17nt-31nt**	**Single-end**	**Number of reads, error profile**	**Custom**	**Yes**	**Yes**	**Yes**	**sncRNA (miRNA and piRNA)**

3. Only a few available RNA-Seq simulators provide ground truth data with synthetic sequences, yet the format is often cumbersome and not user-friendly for performance assessment. For instance, the formats such as BAM/SAM provided by simulators like ART, BEERS, CAMPAREE, MASON, NEAT and RSEM require additional parsing to extract meaningful information, such as expression of altered sequence and location of alterations in the sequence, etc. In contrast, miRSim generates ground truth data in a user-friendly format, presenting altered sequences alongside their original counterparts, corresponding CIGAR strings, and expression counts extracted from associated FASTQ files in a comma-separated file.
4. miRSim offers both command-line and user-friendly Jupyter Notebook versions. The Jupyter Notebook version allows users to effortlessly modify parameters such as seed and xseed region information, error profiles, and output file types (Fastq or Fasta). Moreover, miRSim's adaptability extends to non-human genomes, ensuring its utility across diverse research domains. This user-friendly approach sets miRSim apart from other simulators, enhancing accessibility and ease of use for researchers.
5. While most existing simulators adhere to Illumina-based error models, miRSim offers the flexibility to customize error profiles by adjusting parameters like the number of alterations in seed and xseed region, the number of synthetic reads to be generated, etc.

3.4.2 Difference in miRDeep2, miRDeep* and miRPipe

We have added the following methods in miRPipe to make it better than miRDeep2 and miRDeep*:

1. *Novel Seed-based clustering*: Both miRDeep* and miRDeep2 do not report the known miRNA paralogues and yield many false positives and false negatives, which reduce their accuracy and F1 score. In Step 3 of the miRPipe workflow (that is, by miRDeep*), there can be many novel miRNA sequences that are not assigned to their correct miRNA family; in other words, they are not detected properly. For example, some known miRNAs paralogues are declared novel miRNAs by the sequence aligner in Step 3 of miRPipe (miRDeep*). However, they should have been assigned to their respective known miRNA families. miRPipe clusters such as miRNAs are declared as novel in Step 3 of miRPipe using novel seed-based clustering (Step-6). In Step-6, miRPipe identifies novel miRNAs and known miRNA paralogues by comparing the seed, xseed sequence (other than the seed sequence), and their genomic locations. Similarly, Step-6 of the miRPipe workflow also combines novel miRNAs sharing the same seed sequence as that of a known miRNA (or another novel miRNA), a maximum of two alterations in xseed sequence and similar genomic location through seed-based clustering. After Step-6, miRPipe eventually yields uniquely identified novel miRNAs and their paralogues. This step helps miRPipe to yield the least false positives and false negatives. For example, let us consider a sequence "tccctgtcctccaggagctc" identified as a novel miRNA (say novelMir-1) in Step 3 of the miRPipe workflow.

The novelMir-1 has an identical seed as that of hsa-mir-339, has more than 2nt alteration in the xseed region, and is mapped at a genomic location other than that of hsa-mir-339. Therefore, novelMir-1 should be called a paralogue of hsa-mir-339 and should be labeled as hsa-mir-339_1. Thus, the naming of novelMir-1 leads to a false positive for the novel miRNA class and a false negative for the known miRNA class. In the miRPipe pipeline, the false positive and false negative would be reduced by assigning the correct class to novelMir-1 as hsa-mir-339_1.

2. *Identification of reverse complement miRNAs as known miRNA using DASHR blast search*: miRPipe checks whether the miRNAs identified as a novel miRNA in Step 3 of the miRPipe pipeline are indeed novel. In Step 3 of the miRPipe workflow, some sequences can be annotated as novel miRNAs, whose annotation is missed due to it being present as a reverse complement sequence in the fastq file. miRDeep2 fails to identify the reverse complement sequence known as miRNA. Out of 887 high-confidence known miRNAs, miRDeep2 has correctly annotated only four reverse complement sequences as known miRNAs. Moreover, miRDeep* can annotate only those reverse complement sequences, known as miRNA, already annotated in the miRBase database, regardless of the mapping strand of the reverse complement sequence with the human genome. For example, the miRNAs hsa-mir-3529-5p (aggtagactgggatttgttgtt) and hsa-mir-7-2 (aacaacaaatcccagtctacct) are reverse complementary pairs. The reverse complement of hsa-mir-3529-5p (or hsa-mir-7-2) will be mapped to hsa-mir-7-2 (or hsa-mir-3529-5p) in the opposite strand. Similarly, for the reverse complimentary pair hsa-mir-103a-3p (agcagcattgtacagggctatga) and hsa-mir-103b-1 (tcatagccctgtacaatgctgct), the reverse complement of hsa-mir-103a-3p (or hsa-mir-103b-1) will be mapped to hsa-mir-103b-1 (or hsa-mir-103a-3p) in the same strand. Out of 887 high-confidence known miRNAs, miRDeep* has correctly annotated only 35 reverse complement sequences as known miRNAs. However, in many cases, due to different mapping strands and precursor sequence of the reverse complement sequence with the respective mapping strand and precursor sequence of that known miRNA, miRDeep* failed to annotate the reverse complement sequence to its true known miRNA and annotated them as novel miRNAs. Due to the incorrect annotation of reverse complement sequences as known miRNAs, miRDeep2 and miRDeep* yield many false positives and false negatives. On the other hand, miRPipe correctly annotates the reverse complement sequence to its true known miRNA in Step 5 of the workflow (DASHR blast search). For example, a sequence "ctacagaggcgacatgggggtca" (say mir-1), which is the reverse complement of hsa-mir-6859-3p (tgacccccatgtcgcctctgtag). The sequence of mir-1 is mapped at the genomic location chr1:17369-17391 (chromosome_number:chromosome_start, chromosome_end), which is the same as the genomic location of hsa-mir-6859-3p reported in the miRBase database. The mapping strand of mir-1 is opposite to the respective strand of hsa-mir-6859-3p. The precursor sequence generated by miRDeep* for mir-1 is the reverse complement of the respective precursor sequence of hsa-mir-6859 in miRBase. Hence, miRDeep* will annotate mir-1 as a novel miRNA, while miRPipe will correctly annotate mir-1 to hsa-mir-6859 in Step-5 of the workflow (DASHR blast search).

3. *Identification of piRNA*: Unlike most bioinformatics pipelines that either identify miRNAs or piRNAs, miRPipe also identifies piRNAs along with the miRNAs from the RNA-Seq data.
4. *Customized reference genome*: miRPipe allows users to choose the reference genome hg19/hg38) and miRBase version (version 19/20/21/22) as per the requirement. The sequence aligner used in miRPipe uses the miRBase database for sequence annotation. If required, a user can add another database for miRNA annotation. For example, MirGeneDB can be used instead of miRBase, and the sequences can be annotated according to this database. If a user replaces the miRBase annotation files with that of mirGeneDB, then miRPipe will annotate the miRNA according to the MirGeneDB database.
5. *Batch-mode operation*: Since miRDeep* is a single-threaded memory-intensive sequence aligner, the sequential operation increases the time the pipeline takes when data of multiple subjects is required to be processed. On the other hand, miRPipe allows the execution of sequence alignment in batch mode for multiple subjects' data analysis and, therefore, significantly reduces execution time in downstream analysis. In order to provide operational flexibility in miRPipe, a user can control whether to run a job in the sequential mode (one subject's file or one sample file at a time) or in the batch mode (multiple subjects' files or multiple samples' files). In sequential mode, miRPipe will align one file at a time. Similarly, in batch mode, the entire dataset consisting of multiple files is divided into several small batches. All these batches are processed parallelly on dedicated (individual) CPU threads. Further, the user can also control the number of threads and memory allocation per thread (as per the system hardware RAM limits). This operation is faster and less time-consuming than the sequential operation for a big dataset.
6. *Cohort analysis and identification of dysregulated miRNAs*: miRPipe can perform cohort analysis (dataset containing multiple samples) and report the dysregulated known miRNAs, novel miRNAs, and known piRNAs via the statistical test of DESeq2. For cohort analysis, miRPipe can split the cohort into multiple batches, process each batch on a dedicated thread parallel, and then use DESeq2 to report dysregulated miRNAs or piRNAs. On the contrary, since miRDeep* can process only one sample at a time, it does not report the dysregulated miRNAs or dysregulated piRNAs. Still, it can only detect miRNAs present in the fastq file of a subject.
7. *Synthetic Data Generator (miRSim)*: We have also developed the miRSim tool to generate synthetic data for the extensive benchmarking of different pipelines.
8. Both miRSim and miRPipe are open-source and available publicly in an interactive jupyter notebook at the GitHub repositories.
9. *Selective pipeline execution*: We have developed miRPipe in an interactive jupyter notebook. The miRPipe pipeline is developed so that both piRNA and miRNA pipelines can run together. If a user wants to run only one pipeline at a time, that can be done easily in the jupyter notebook.

3.4.3 Comparison of all pipelines on known miRNA identification

Of the existing pipelines, miRDeep2 identifies miRNAs by hierarchical sequence alignment followed by RNA secondary structure prediction of potential precursors and estimation of the performance statistics of all potential precursors to filter false positives. However, it allows mismatches of 1 to 2nt in the reads while matching the corresponding sequence to those of known miRNAs, introducing false positives. In addition, if a known miRNA sequence has a reverse complement, it either rejects it or annotates it as novel miRNA. On the other hand, miRdeep* follows the same methodology as miRDeep2, except that it incorporates an improved strategy for miRNA precursor sequence identification and additional isomiR detection capacity. Further, it does not allow any mismatch with known miRNAs, unlike miRdeep2, reducing the false positives. mirPRo also follows the same methodology as miRDeep2 except that it imposes a stringent condition, wherein only perfectly mapped reads are allowed for known miRNA prediction. mirPRo also includes isomiR detection. mirPRo pipeline does not report the paralogues of known or novel miRNAs. mirPRo uses a Novoalign sequence aligner for the identification of known miRNAs, allowing a maximum 2nt mismatch or three indels in one opening gap. This could be the reason for more false positives with mirPRo.

We observed that six out of eight pipelines performed well on known miRNA. Of these, miRPipe, miRDeep2, and miRDeep* performed best, while mirnovo, miRge2.0, and sRNAToolbox yielded average performance, while mirPRo comparatively underperformed. The performance of miRDeep2 was close to miRDeep* except for a few miRNAs, whose precursors were inconsistent for dicer processing. On the other hand, the miRDeep* tool has an improved precursor excision strategy over miRDeep2, leading to better performance on known miRNA identification. We have observed that the average accuracy and average F1-score of miRDeep2 and miRDeep* across all depths for known miRNA identification was 94.74%, 85.66% and 95.67%, 88.06%, respectively. At the same time, miRPipe has an average accuracy and average F1-score of 96.58% and 89.95% (Table-3.4). The improvement in the miRPipe performance on the identification of known miRNAs was due to the DASHR blast search and seed-based clustering method.

3.4.4 Comparison of all pipelines on novel miRNA identification

For novel miRNA identification, miRDeep2, miRDeep*, mirPRo, MiR&moRe2, and miRPipe use the hybrid approach that includes both genomic features and hairpin structural features. A sequence has to pass through 6 conditions to be annotated as novel miRNAs, such as 1. Position of potential mature sequence to potential hairpin sequence, 2. Potential star sequence, 3. Potential loop sequence, 4. Number of base pairs between mature and star sequence, 5. Percentages of reads aligned to the location of mature miRNA for proper dicer processing (at least 90% read should be aligned) and,

6. Log-odds probability score for potential mature miRNA. These six conditions are used to rigorously scan the precursor sequence to identify a read as a novel miRNA. miRDeep* additionally employs the improved precursor excision strategy compared to miRDeep2, which leads to better performance. mirPRo has improved performance on novel miRNA compared to known miRNA detection. It also performs better on novel miRNA identification than miRge2.0, sRNAToolbox, and MiR&moRe2 because mirPRo follows the same six conditions and allows a maximum mismatch of 1nt. It considers mapped read lengths between 18 to 25nt and the fold-change criterion (that is, keep only mapped reads with the highest read stack with at least two-fold change compared to the second-highest read stack) to reduce the false positives. Since miRPipe is an improvisation for reducing false positives and false negatives by incorporating DASHR blast search and seed-based clustering on novel miRNA sequences, it yields better results than these tools and other pipelines. Of note is that miRPipe has the lowest false positives and false negatives compared to other pipelines.

sRNAToolbox imposes stringent conditions for novel miRNA prediction, such as within-cluster ratio, 5' fluctuation, minimum number of hairpin bindings, minimum number of mature bindings, length intervals, and minimum reads. The threshold for each feature is derived from the same machine-learning model training dataset used in miRAnalyzer [216]. We have observed that no novel miRNA was identified in miRSim simulated synthetic data due to the sRNAToolbox stringent conditions. The sRNAToolbox has also not identified any novel miRNAs in the synthetic data experiment on the identification of reverse complement sequences as known miRNAs. Moreover, sRNAToolbox has reported only three, one, and three novel miRNAs in the CLL dataset (GSE123436), lung cancer dataset (GSE37764), and breast cancer dataset (GSE171282) dataset, respectively. None of the novel miRNAs were found as dysregulated in differential expression analysis in any of the datasets. Similarly, miRge2.0 utilizes an SVM machine-learning model that uses 22 structural and compositional features for novel miRNA predictions. The SVM model has been trained on 17 tissues of the human and mouse datasets. Due to these stringent conditions, miRge2.0 did not report any novel miRNAs in synthetic data benchmarking experiments. Moreover, the miRge2.0 pipeline identified 18, zero, and zero novel miRNAs in the CLL dataset (GSE123436), lung cancer dataset (GSE37764), and breast cancer dataset (GSE171282) dataset, respectively. None of the identified novel miRNAs were found as dysregulated in differential expression analysis in any of the datasets. This could be due to the lack of generalizability of the SVM model trained by miRge2.0, which has led to such high false negatives.

Similarly, the mirnovo pipeline uses machine learning (random forest model) with 12 coverage profile features, 12 sequence complexity, and nine genomic features hairpin structural features for novel miRNA identification. It provides not only novel miRNAs but also other non-coding RNAs such as tRNA or rRNAs. It is also observed to have high false negatives. All three above (sRNAToolbox, miRge2.0, and mirnovo) are simple methods that do not impose many stringent conditions for detecting novel miRNAs and, hence, lead to many false positives.

MiR&moRe2 identifies loop-RNAs, moRNAs, and novel miRNAs with the precursor excision methodology similar to miRDeep2, except that the candidate precursor sequences are extended to 30nt on both upstream and downstream for the identification of the possible moRNAs. It also checks for the sequences that are aligned in the offset region or loop region of the miRNAs hairpin and can be annotated as moRNAs and loop-RNAs. The miRNA sequences that are neither moRNAs nor loop-RNAs and located in close proximity to the mature sequence of the hairpin precursor are considered novel miRNAs. MiR&moRe2 lacks the identification paralogues and has many false negatives due to an inefficient precursor excision strategy. miRPipe addresses the issues of identification of paralogues functional annotation of novel miRNAs, utilizing both the genomic and precursor features and, hence, outperforming all the other pipelines.

3.4.5 Comparison of all pipelines on known piRNA identification

Among all these pipelines, only miRPipe and sRNAToolbox identify piRNAs and, hence, reported these in the synthetic data experiments. We observed average accuracy and a low F_1-score for piRNA identification in the sRNAToolbox due to high false negatives. In miRPipe, the stringent condition of zero nucleotide mismatch in the seed region and no reverse complement alignment helped in reducing the false positives during piRNA identification. Compared to other pipelines, sRNAToolbox also reports other non-coding RNAs (long non-coding RNAs, piRNAs etc.) using blast search for all unmapped/unassigned reads to several remote databases hosted at NCBI (such as GenBank, EMBL etc.) with the help of several helper tools in sRNAToolbox such as Ensembl Parser, NCBI Parser, RNA central parser, and Genomic tRNA database parser.

3.4.6 Assessment of bioinformatics workflow using sequence homology

In this study, we assessed the performance of the bioinformatics workflow by introducing the notation of TP read, defined as "A read is counted as TP if it is correctly identified by the pipeline". In the context of sequence analysis, the following two key parameters are considered for a sequence to be “correctly identified”:

1. Sequence alignment with Reference Genome: The first step for a sequence to be a TP sequence is to precisely be mapped with zero nucleotide mismatch with respect to the reference genome. This criterion ensures the accuracy and robustness of sequence alignment.
2. Valid miRNA Characteristics: In the second step, a valid miRNA sequence must

exhibit specific characteristics, including a valid precursor sequence within the reference genome and forming a stable hairpin structure.

Only the sequences meeting both criteria are deemed true positives, ensuring their conformity to established miRNA attributes. In this scenario, the sequence homology is accounted for in identifying true positive sequences and calculating accuracy as follows:

1. Sequence Alignment: Within the miRPipe workflow, we considered the sequence of length 17 – 24nt for identifying miRNA sequence. We employed miRDeep* for miRNA sequence alignment. For mapping the query sequence to the human reference genome, the miRDeep* aligner does not allow any nucleotide mismatch. It maps the query sequence to the reference genome only when all the nucleotides of the query sequence match precisely with the reference sequence. Here, the sequence homology is considered by assessing the similarity between the query sequence and reference genome sequence.
2. piRNA Sequence Alignment: Similarly, for the piRNA sequence, we incorporated the stringent criteria of zero nucleotide mismatch while mapping with the reference genome, hence ensuring maximum similarity between the query sequence and reference genome.

3.4.7 General remarks and limitations of the study

It is possible that the combination of different methods can improve the results. The combination of multiple methods can be either the consensus of results of all methods or the union of results of all methods. If we consider the consensus results, it is possible to reduce false positives. However, it may lead to high false negatives because of the methodological differences of pipelines that impact miRNA identification. Similarly, considering the union results, it may lead to high false positives, which is also not good. We believe that miRPipe addresses this issue because miRPipe is an end-to-end unified workflow that can report all important miRNAs/piRNAs in one go with the least false positives or false negatives, as shown in the benchmarking results.

We have validated miRPipe using miRSim simulated synthetic data with ground truth and three publically available real RNA-Seq expression datasets (GSE123436, GSE37764, GSE171282). The bioinformatics pipeline can also be validated using some publicly available sequencing data with added synthetic microRNAs, usually using an equimolar mixture of 962 synthetic microRNAs miRXplore Universal Reference from Miltenyi [217]. Further, miRPipe or any other bioinformatics pipeline can also be tested on the comprehensive atlas of the human transcriptome from "The RNA Atlas expands the catalogue of human non-coding RNAs." [218], which includes small polyA RNA as well as total RNA from 300 human tissues and cell lines. Since miRPipe is an open-source bioinformatics pipeline, any future researcher can test the pipeline on these datasets.

miRPipe is a generic workflow and can be used for both human and non-human datasets. Currently, the miRPipe pipeline has been tested for human datasets only and has default parameters, such as miRNA sequence length, piRNA sequence length, etc., adjusted according to the human genome. However, the miRPipe pipeline can also be used for the non-human genome. For this, the user needs to replace the human genome and its reference index with the corresponding non-human reference genome, its reference index and the corresponding sequence annotation database in the sequence aligner step (Step 3) of the miRPipe pipeline. After replacing the reference and annotation files, miRPipe can be used for the non-human genome, as the core algorithm will remain the same. Similarly, the applicability of the miRSim simulator to the non-human genome can be extended by providing the non-human genome reference sequences and adjusting the seed and xseed region location within the miRSim tool. To make this adjustment, we have provided miRSim with command-line and Jupyter Notebook versions, enabling users to modify parameters according to their genome requirements easily.

Notably, miRPipe was developed between 2019 and 2021. The developer team was actively supporting the Bowtie 1 sequence aligner during this period. The most recent update for Bowtie 1 was posted on September 13, 2021. The source codes of the miRSim synthetic simulator are available at Zenodo open-source repository, published on June 14, 2021. Further, we have provided the dockerized version of the miRPipe pipeline to ensure reproducibility. This approach ensures the smooth deployment of the miRPipe pipeline irrespective of system configurations and prevents package dependency or conflict issues. Furthermore, the dockerization also ensures the operational consistency of miRPipe even when Bowtie1 is deprecated. However, we acknowledge the current status of Bowtie 1 being deprecated, with Bowtie 2 now available as an alternative. In the future, we will focus on releasing the next version of the miRPipe docker, incorporating the recent versions of the tools used within the miRPipe workflow and the capability to analyze non-human genome data.
All the source codes necessary to reproduce the results given in figure-3.4 (a-l) and the Table-3.3 are available in the GitHub for the synthetic data. The synthetic fastq data files are also available in the same repository. The open-source synthetic data simulator tool miRSim is available at GitHub. The RNA-Seq CLL real RNA-Seq expression data data can be accessed from the repository GSE123436. Similarly, lung cancer and breast cancer real RNA-Seq expression data can be accessed from the repository GSE37764 and GSE171282 respectively.

3.5 Conclusion

The synthetic data experiment validation and benchmarking strategy, along with the validation on real RNA-Seq expression data, establishes miRPipe as a robust, reliable, and reproducible pipeline for the detection of known/novel miRNAs, paralogues, and piRNAs from the RNA-Seq data. miRPipe outperforms recent state-of-the-art pipelines. miRPipe can jointly identify miRNAs and piRNAs and carries out parallel batch processing for the efficient utilization of the computational resources. The Jupyter Notebook for bioinformatics pipeline and containerization of tools makes its configuration and deployment easy with minimum effort. As we delved deeper into CLL, we learned about the critical importance of MM, a vital subtype of blood cancer that exhibits a unique, benign precursor stage. MM is characterized by the presence of abnormal plasma cells in the blood, which presents a distinctive challenge in distinguishing between this precursor stage, known as MGUS, and the disease stage itself. Additionally, there is a pressing need to identify reliable biomarkers for tracking disease progression in MM. Our concentrated efforts to pinpoint differentiating biomarkers between MM and MGUS serve as a foundational step toward gaining a more profound understanding of MM pathogenesis and identifying the genomic events that drive disease progression. In the forthcoming chapter, we delve into the application of artificial intelligence (AI) in genomics, focusing on developing application-aware models aimed at discerning biomarkers capable of distinguishing MM from MGUS. We introduce an innovative GCN-based bio-inspired model meticulously engineered to identify pivotal genomic biomarkers, including genes and genomic features, and to dissect disrupted signaling pathways, providing insights into their roles in disease progression. The intricacies of this model and post-hoc interpretability analysis are presented in-depth in the following chapter.

Chapter 4

Bio-inspired DL model for the identification of altered Signaling Pathways in Multiple Myeloma using WES data

4.1 Introduction

Multiple Myeloma is a neoplasm of malignant plasma cells in the bone marrow, preceded by the precancerous stage of MGUS. However, in clinical practice, the distinction between different stages is, at times, ambiguous. The role of early treatment and the type of such treatment to prevent progression to MM or to reduce the associated morbidity is also not clear. Thus, it would be interesting to decipher genes, genomic biomarkers and crucial pathogenic prognostic factors that are representative of MGUS and MM in order to develop appropriate therapeutic interventions to halt the progression to overt MM. In this study, we address the challenge of identifying significant biomarkers that can effectively distinguish MGUS from MM by employing a multidimensional analysis of exome profiles and their PPI network within a bio-inspired deep learning-based architecture named the "BDL-SP" model. Additionally, we underscore the importance of selecting models based on their interpretability within the context of the application domain. We rank the genes based on their distinguishing ability in MM and MGUS. The pathway analysis of these top-ranked genes sheds light on the disruptive role of pathways in MM pathogenesis. These novel findings hold the potential to pave the way for tailored therapeutic interventions aimed at halting the progression to overt MM in the future.

4.2 Material and Methods

4.2.1 Whole-exome sequencing datasets of MM and MGUS patients

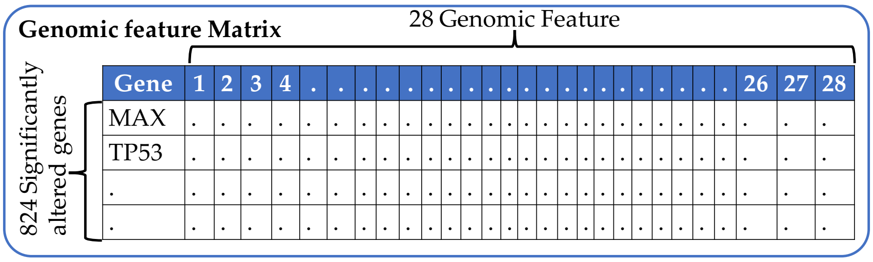
In this work, we utilized two external WES datasets available with controlled access and one in-house WES dataset of MM and MGUS patients. These datasets are: 1) Multiple Myeloma Research Foundation (MMRF) CoMMpass data (of the American population), 2) EGA dataset (of the European population), and 3) AIIMS WES dataset (of the Indian population). The MMRF CoMMpass (https://research.themmrf.org) is an open-source, extensive clinical and molecular database of MM. The majority of MM samples in the MMRF CoMMpass dataset (>75%) were collected from people of American ethnicity. The MMRF CoMMpass dataset is aimed to provide molecular characterization and to correlate clinical datasets of MM patients for finding new, actionable targets to facilitate future clinical trial designs [134]. In our study, we have included 1092 bone marrow (BM) samples of MM collected from the Genomic Data Commons (GDC) portal via dbGaP authorized access (phs000748; phs000348). This is to note that the MMRF dataset also contained 20 peripheral blood samples that were not included in this study for the uniformity of the data. Similarly, the EGA contains over 700 studies of multiple diseases (including cancer and non-cancer) worldwide. EGA (http://www.ebi.ac.uk/ega/) was launched in 2008 by the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) to provide secure storage of biological data and distribution only to authorized users [135]. The whole exome sequencing data of 33 MGUS European patients were obtained from the EGA repository EGAD00001001901. Besides the above two external datasets, we also included the WES data collected in-house from patients of Indian origin registered at AIIMS, New Delhi, India. This dataset included 82 MM and 28 MGUS samples. We have used the tumor-normal matched pairs of all BM samples obtained from MMRF, EGA and AIIMS WES data repository. Thus, we have included MGUS and MM WES datasets from three different databases.

4.2.2 Methods

Data pre-processing

Four variant callers, namely, MuSE, Mutect2, VarScan2, and SomaticSniper were used for finding the variants in MM patients from the MMRF CoMMpass study. Therefore, for each patient, four VCF files corresponding to each variant caller were downloaded from the GDC portal via dbGaP authorized access (phs000748; phs000348). Exome data obtained from EGA and AIIMS were processed with an exome sequencing pipeline [219] using BWA and GATK, which is also considered a standard pipeline and mostly adopted to process the exome sequencing data. Similar to the MMRF data, the SNVs in EGA and AIIMS exome sequencing data were extracted using MuSE, Mutect2, VarScan2, and SomaticSniper variant callers. SNVs were annotated using ANNOVAR tool that provides information about mutated genes, mutation type, the property of being deleterious or not, and clinical validation. In our study, we considered 23 types of functionally significant SNVs clustered into three groups based on their functional impact as follows: 1) Non-Synonymous (NS) SNV Group: This group consists of non-synonymous SNVs, exonic, ncRNA_exonic, stop gain, stop loss, start loss, exonic; splicing, splicing, frameshift insertion, and frameshift deletion type SNVs; 2) Synonymous SNV Group: This group consists of synonymous SNVs, UTR3 and UTR5 SNVs; and 3) Other SNV Group: This group consists of non-frameshift insertion/deletion/substitution, intronic, intergenic, ncRNA_intronic, upstream, downstream, unknown, and ncRNA_splicing SNVs.

The benign SNVs were filtered out using the FATHMM-XF method. Genomic annota-



Description of genomic features used in the feature matrix

Feature Number	Feature Name
1	Total number of the SNVs
2	Total number of the SNVs in synonymous group
3	Maximum VAF of the SNVs in synonymous group
4	Median VAF of the SNVs in synonymous group
5	Mean VAF of the SNVs in synonymous group
6	VAF's standard deviation of the SNVs in synonymous group
7	Maximum AD of the SNVs in synonymous group
8	Median AD of SNVs in synonymous group
9	Mean AD of SNVs in synonymous group
10	AD's standard deviation of the SNVs in synonymous group
11	Total number of SNVs in non-synonymous group
12	Maximum VAF of SNVs in non-synonymous group
13	Median VAF of SNVs in non-synonymous group
14	Mean VAF of SNVs in non-synonymous group
15	VAF's standard deviation of the SNVs in non-synonymous group
16	Maximum AD of SNVs in non-synonymous group
17	Median AD of SNVs in non-synonymous group
18	Mean AD of SNVs in non-synonymous group
19	AD's standard deviation of the SNVs in non-synonymous group
20	Total number of SNVs in other group
21	Maximum VAF of SNVs in other group
22	Median VAF of SNVs in other group
23	Mean VAF of SNVs in other group
24	VAF's standard deviation of the SNVs in other group
25	Maximum AD of SNVs in other group
26	Median AD of SNVs in other group
27	Mean AD of SNVs in other group
28	AD's standard deviation of the SNVs in other group

Figure 4.1: Schematic layout of genomic feature matrix used for the training of proposed BDL-SP model. The dimension of the genomic feature matrix is 824×28 with 824 significantly altered genes (See Table S1 of supplementary material) and 28 genomic features obtained from MMRF, EGA and AIIMS WES datasets using the AI-based workflow shown in Figure 2. The genomic features were extracted from three groups of SNVs, namely: 1. Nonsynonymous SNV group, 2. Synonymous SNV group, and 3. Other SNV groups. A total of nine features were extracted for each SNV group to learn the distributive statistics (maximum, mean, median, and standard deviation).

tions of SNVs (i.e., SNV type, mutated gene name, etc.) obtained from ANNOVAR were pooled and analyzed to identify the top significantly mutated genes using the 'dndscv' tool [220] based on the q-value (≤ 0.05) in both MM and MGUS individually. Union of significantly mutated genes from all four variant callers for MM (1174 patients) and MGUS (61 patients) groups led to 617 and 362 genes, respectively, and further union of the genes mentioned above yielded a total of 824 genes (Table S1 of supplementary material). For each gene, a total of 28 genomic features were created that include the total variant count and the distributive statistics (maximum, mean, median, and standard deviation) of VAF and AD of each of the three groups of SNVs (nonsynonymous SNV group, synonymous SNV group, and Other SNV group). A detailed description of the 28 genomic features is presented in Figure-4.1. The complete AI workflow is presented in Figure-4.2. For gene-gene interaction network information, we used the STRING database to get the PPI of 824 significantly altered genes. The STRING database contains all the known and predicted associations of protein-protein interactions, including physical and functional associations for more than 14000 organisms.

Proposed shallow bio-inspired deep learning architecture from signaling pathways (BDL-SP):

The conventional convolutional neural network (CNN) often fails to learn data from non-Euclidean space because non-Euclidean data cannot be modelled into n-dimensional linear space. The PPI network used in our model has a similar underlying non-Euclidean structure. Thus, a GCN could help us learn the PPI data of non-Euclidean space. The proposed BDL-SP model performs disease classification using a graph convolutional network that learns significant features from the exonic mutational profiles of genes interacting among each other according to the PPI network interactions. The mathematical description of the GCN model is as follows:

For a given undirected graph, g = (υ, ϵ) where υ is a collection of a finite set of nodes and ϵ is a collection of the finite set of edges, a graph convolution network learn the node representation by applying the graph laplacian with the input feature matrix $X \in \mathbb{R}^{N \times p}$ (where N denotes the number of nodes and p the number of features) and follows the propagation rule for each layer shown below:

$$H^{(l+1)} = \sigma(LH^{(l)}W^{(l)}) \tag{4.1}$$

Where L denoted the normalized graph laplacian defined below.

$$L = I - D^{-\frac{1}{2}}\tilde{A}D^{-\frac{1}{2}} = U\Lambda U^T \tag{4.2}$$

Where $D_{i,j} = \Sigma_{i=1}^{n} A(i,j)$, degree matrix of the graph and $\tilde{A} = A + I$, where A is the

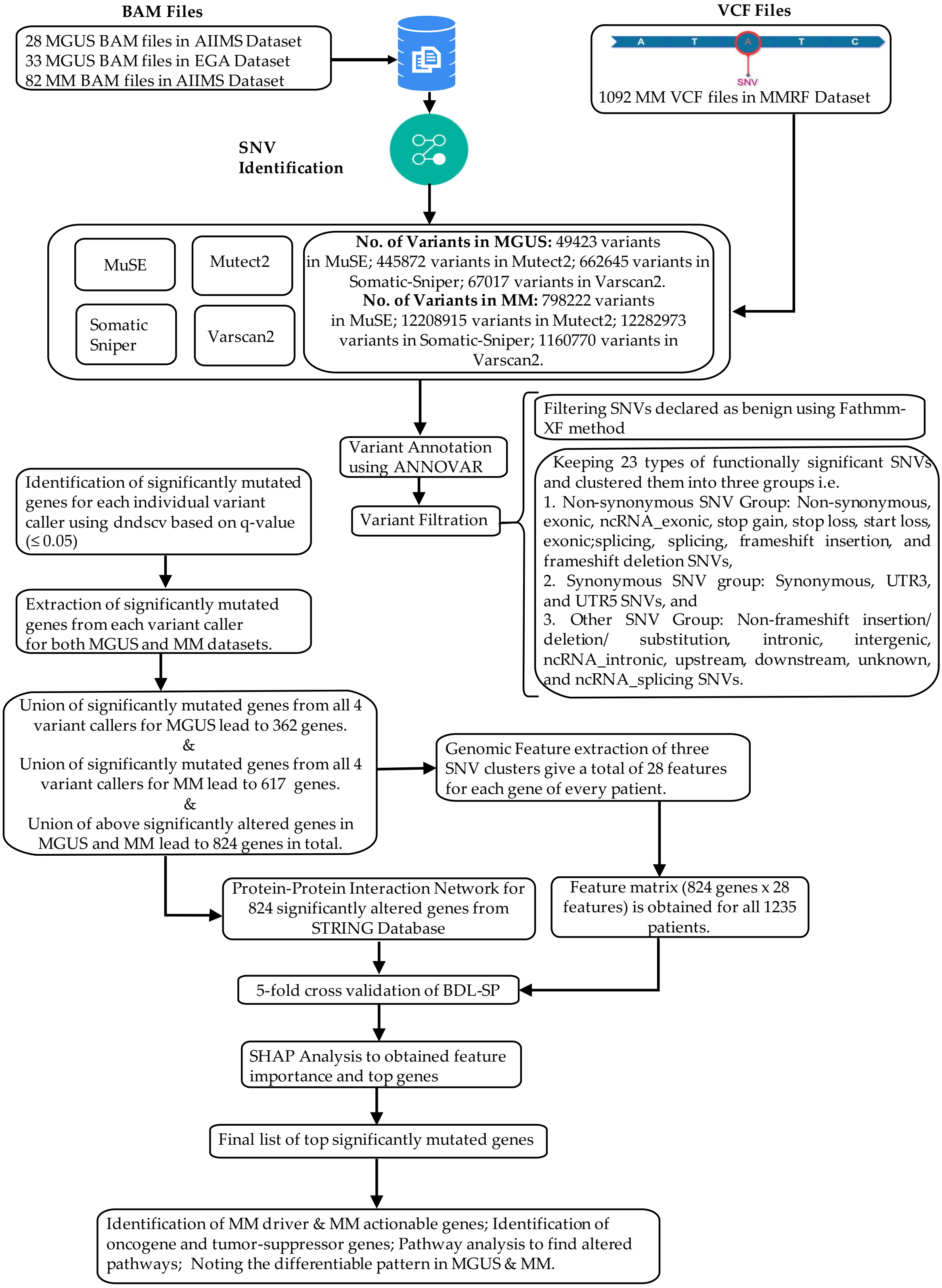

BAM Files
28 MGUS BAM files in AIIMS Dataset
33 MGUS BAM files in EGA Dataset
82 MM BAM files in AIIMS Dataset
VCF Files
A T A T C
SNV
1092 MM VCF files in MMRF Dataset
SNV Identification
MuSE
Mutect2
Somatic Sniper
Varscan2
No. of Variants in MGUS: 49423 variants in MuSE; 445872 variants in Mutect2; 662645 variants in Somatic-Sniper; 67017 variants in Varscan2.
No. of Variants in MM: 798222 variants in MuSE; 12208915 variants in Mutect2; 12282973 variants in Somatic-Sniper; 1160770 variants in Varscan2.
Variant Annotation using ANNOVAR
Variant Filtration
Filtering SNVs declared as benign using Fathmm-XF method
Keeping 23 types of functionally significant SNVs and clustered them into three groups i.e.
1. Non-synonymous SNV Group: Non-synonymous, exonic, ncRNA_exonic, stop gain, stop loss, start loss, exonic;splicing, splicing, frameshift insertion, and frameshift deletion SNVs,
2. Synonymous SNV group: Synonymous, UTR3, and UTR5 SNVs, and
3. Other SNV Group: Non-frameshift insertion/ deletion/ substitution, intronic, intergenic, ncRNA_intronic, upstream, downstream, unknown, and ncRNA_splicing SNVs.
Identification of significantly mutated genes for each individual variant caller using dndscv based on q-value (≤ 0.05)
Extraction of significantly mutated genes from each variant caller for both MGUS and MM datasets.
Union of significantly mutated genes from all 4 variant callers for MGUS lead to 362 genes. & Union of significantly mutated genes from all 4 variant callers for MM lead to 617 genes. & Union of above significantly altered genes in MGUS and MM lead to 824 genes in total.
Genomic Feature extraction of three SNV clusters give a total of 28 features for each gene of every patient.
Protein-Protein Interaction Network for 824 significantly altered genes from STRING Database
Feature matrix (824 genes x 28 features) is obtained for all 1235 patients.
5-fold cross validation of BDL-SP
SHAP Analysis to obtained feature importance and top genes
Final list of top significantly mutated genes
Identification of MM driver & MM actionable genes; Identification of oncogene and tumor-suppressor genes; Pathway analysis to find altered pathways; Noting the differentiable pattern in MGUS & MM.

Figure 4.2: AI-based workflow to infer differentiable genomic biomarkers to identify MGUS and MM using the WES data.

adjacency matrix, U is the matrix of eigenvectors of the graph, Λ, denote the respective eigenvectors, and $W \in \mathbb{R}^{p \times m}$ (where m corresponds to the number of filters in the graph convolution) denotes a learnable weight matrix. A GCN model transforms a graph into the spectral domain by graph Fourier transformation defined as below:

$$x * g = UgU^{T}x \tag{4.3}$$

The above Fourier transformation can be computed by approximating Chebyshev polynomials and the renormalization trick mentioned in [30] as:

$$Z = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}XW \tag{4.4}$$

The infographic representation of the architecture of BDL-SP with an end-to-end pipeline is shown in Figure 3 and is explained here. The BAM files from the EGA and AIIMS datasets and VCF files from the MMRF dataset are processed to extract 824 significantly altered genes using the dndscv tool (as shown in the WES Data pre-processing block) in Figure-4.3. The interaction among these 824 genes is extracted using the PPI network (from the STRING database). A network of nodes and edges is set up using this information, where each node denotes one of these 824 genes, and each link implies that the two nodes/genes of that link were connected as per the PPI network. Each node is set up with its 28-length feature vector extracted earlier (as shown in Figure-4.1). Hence, the 28-length genomic feature vectors of all 824 genes are added to the network established using the PPI network. This input layer is followed by two hidden layers of GCN, which are further followed by one fully connected layer of 824 neurons to 2 neurons, giving output through a log-softmax activation function. Since there were 95% samples of MM class and 5% samples of MGUS class, which made the data highly imbalanced (class imbalance ratio = 19.22), a cost-sensitive loss function was utilized to train the BDL-SP model in order to deal with the data imbalance problem. BDL-SP is trained in a supervised fashion, where the MM/MGUS target class label, along with the feature matrix of 824×28, is provided as input to the architecture. The network is trained until the loss reduces and saturates. Five-fold cross-validation was performed, which led to the training of five BDL-SP classifiers, one for each fold of test data. Next, the ShAP algorithm is used on these five trained BDL-SP classifiers to obtain the top genomic features and significantly altered signaling pathways, as explained in the next subsection. The setting of layers of BDL-SP and the values of the hyperparameters are shown in Table-4.1.

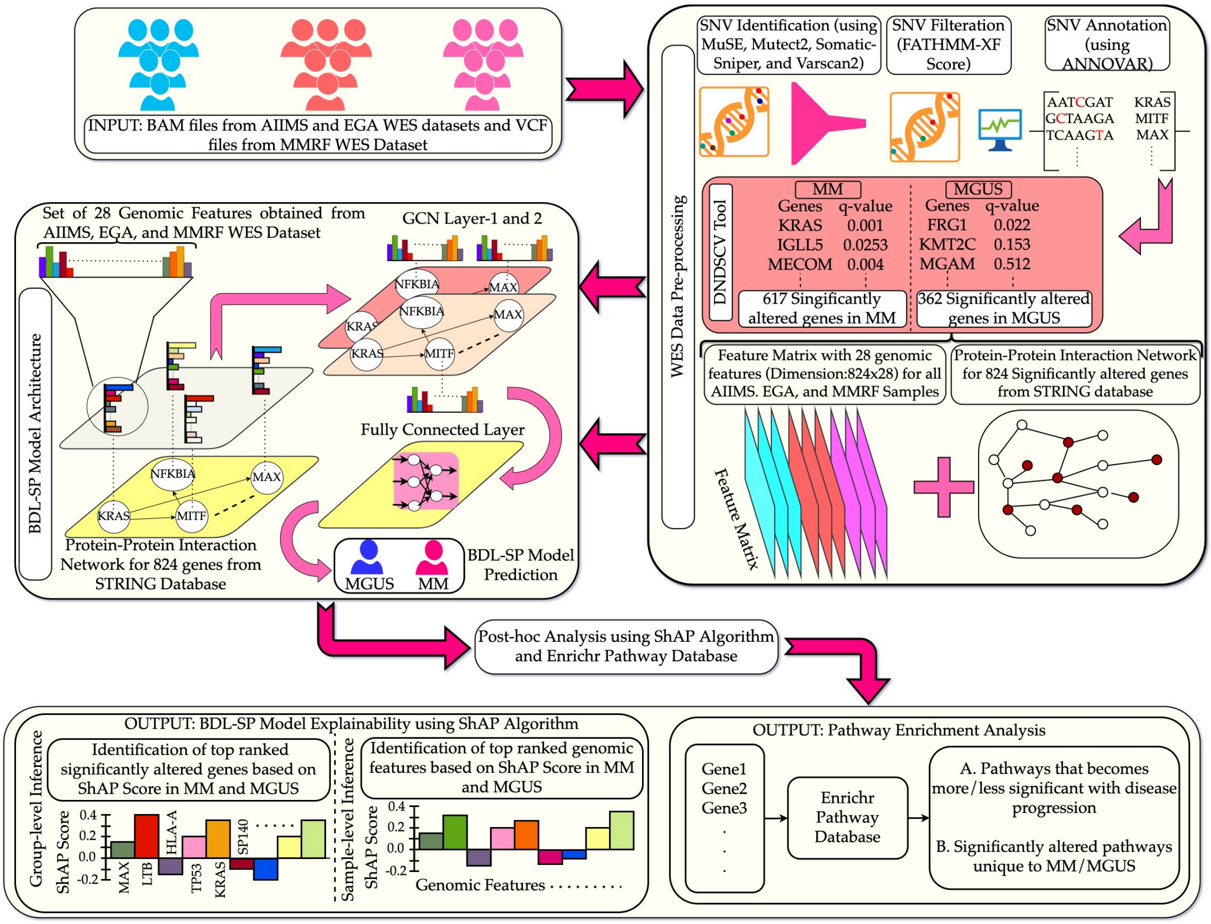

INPUT: BAM files from AIIMS and EGA WES datasets and VCF files from MMRF WES Dataset
WES Data Pre-processing
SNV Identification (using MuSE, Mutect2, Somatic-Sniper, and Varscan2)
SNV Filteration (FATHMM-XF Score)
SNV Annotation (using ANNOVAR)
AATCGAT
GCTAAGA
TCAAGTA
KRAS
MITF
MAX
DNDSCV Tool
MM
Genes q-value
KRAS 0.001
IGLL5 0.0253
MECOM 0.004
617 Singificantly altered genes in MM
MGUS
Genes q-value
FRG1 0.022
KMT2C 0.153
MGAM 0.512
362 Significantly altered genes in MGUS
Feature Matrix with 28 genomic features (Dimension:824x28) for all AIIMS. EGA, and MMRF Samples
Protein-Protein Interaction Network for 824 Significantly altered genes from STRING database
Feature Matrix
BDL-SP Model Architecture
Set of 28 Genomic Features obtained from AIIMS, EGA, and MMRF WES Dataset
GCN Layer-1 and 2
NFKBIA
MAX
KRAS
MITF
Fully Connected Layer
Protein-Protein Interaction Network for 824 genes from STRING Database
MGUS
MM
BDL-SP Model Prediction
Post-hoc Analysis using ShAP Algorithm and Enrichr Pathway Database
OUTPUT: BDL-SP Model Explainability using ShAP Algorithm
Group-level Inference
Identification of top ranked significantly altered genes based on ShAP Score in MM and MGUS
ShAP Score
0.4
0.2
0.0
-0.2
MAX
LTB
HLA-A
TP53
KRAS
SP140
Sample-level Inference
Identification of top ranked genomic features based on ShAP Score in MM and MGUS
ShAP Score
0.4
0.2
0.0
-0.2
Genomic Features
OUTPUT: Pathway Enrichment Analysis
Gene1
Gene2
Gene3
Enrichr Pathway Database
A. Pathways that becomes more/less significant with disease progression
B. Significantly altered pathways unique to MM/MGUS

Figure 4.3: Infographic representation of the proposed AI-based BDL-SP model architecture and the application-aware post-hoc analysis for the identification of pivotal genomic biomarkers that distinguish MGUS from MM. The BAM files from the EGA and AIIMS datasets and VCF files from the MMRF dataset are processed to extract 824 significantly altered genes using the dndscv tool (as shown in the WES Data pre-processing block). The interaction among these 824 genes is extracted using the PPI network (from the STRING database). A network of nodes and edges is set up using this information, where each node denotes one of these 824 genes, and each link implies that the two nodes/genes of that link were connected as per the PPI network. Each node is set up with its 28 genomic features extracted for the corresponding gene, as explained earlier. This input layer is followed by two hidden layers of GCN, one fully connected layer, and a softmax layer at the output. Thus, each subject's WES data is analyzed, and the feature vectors of all 824 genes are extracted. These are given as input along with the subject's MM/ MGUS target class label to train the GCN in a supervised mode. Once the BDL-SP model is learned to distinguish MGUS from MM, the top genomic features and significantly altered signaling pathways were obtained from the ShAP algorithm and the Enrichr Pathway Database.

Table 4.1: Hyperparameters values and layer dimensions of the BDL-SP model architecture

GCN Architecture Attribute/ Hyperparameter	Hyperparameter Value
Number of GCN layers	2
GCN layer dimensions	Input sample dimension: 824x28 1st layer dimension (for each node): 28x7 2nd layer dimension (for each node): 7x1 Output dimension: 824x1
Output linear layer dimension	824x2 (number of classes = 2)
Dropout	0.75
Cost function and adjusted cost for class imbalance	Cost function: Cross-entropy loss Cost adjusted: 20.0
GCN weight initlization	Uniform Xavier

Quantitative benchmarking of BDL-SP model with traditional machine learning classifiers

We have benchmarked the proposed BDL-SP model with six baseline ML models (random forest, decision tree, logistic regression, XGBoost, CatBoost, and SVM from scikit-sklearn [221]). The conventional cost-blind machine learning models do not account for the imbalanced classes in the data and tend to make decisions favoring the majority class resulting in misclassification. In the case of medical diagnosis, such misclassification can lead to erroneous direction of strategic treatment, causing patients to suffer. In our study, there were 95% samples of MM class and 5% samples of MGUS class, which made the data highly imbalanced (class imbalance ratio = 19.22). Therefore, we have used cost-sensitive ML models to account for the class imbalance in our data. During training, the cost-sensitive loss function penalizes the mistake in identifying each MGUS sample (minority class) more than the mistake in identifying each MM sample (majority class). This ensures that the classifier is not biased toward the majority class and learns to identify the samples of both classes. These baseline models are trained with the traditional data pre-processing pipeline using principal component analysis (PCA). Each baseline ML model was trained exhaustively with five-fold cross-validation, where the confusion matrix of the hold-out set was kept separate for each fold. The final confusion matrix was obtained by adding the confusion matrices of all five hold-out sets, and the performance metrics were calculated for each ML model.

Qualitative application-aware post-hoc benchmarking of BDL-SP model using ShAP

ShAP is an algorithm that measures the significance of an attribute in the prediction of a model, scoring each attribute proportional to its contribution. Therefore, it was utilized to explain the post-hoc explainability of the BDL-SP model. The most-contributing genomic features and significantly altered genes at the group (i.e., either MGUS or MM) as well as at the individual sample-level were identified. Since five-fold validation was carried out during training, the ShAP algorithm was applied to each trained classifier to obtain the significant genomic attributes (both genes and genomic features) for each sample. Note that the ShAP score can either be positive or negative. Here, the positive ShAP score for an attribute indicates its contribution to the model's prediction toward the MGUS class (positive class). In contrast, the negative score indicates its contribution to the model's prediction toward the MM class (negative class). Therefore, the higher the magnitude of the ShAP score, the higher its impact on the model's positive class outcome. Moreover, only those samples were considered for extracting ShAP interpretability that were correctly predicted by at least one of the five classifiers.

Next, we devised the algorithms for the estimation of the best ShAP score: 1) for all 824 significantly altered genes (Figure-4.4(A)) and 2) for all 28 genomic features (Figure-4.4(B)) at a sample-level to understand their contribution to the BDL-SP model's prediction. The pseudo-codes with mathematical descriptions for estimating the best ShAP scores for genes and genomic features are provided in Table-4.2, Algorithm-A, and Algorithm-B. The algorithms shown in Figure-4.4(A) and 4.4(B) take the sample feature matrix as input and estimate the best ShAP scores for genes and genomic features at a sample-level. For each sample feature matrix, the corresponding sample class was predicted using all five trained classifiers of the BDL-SP model and the ShAP algorithm was applied only to those classifiers that predicted the sample's class correctly. Here, the ShAP score for all the genomic attributes was collected at the classifier-level and the sample-level. For each genomic attribute, the best ShAP score was first calculated at a classifier level. Then, the final best ShAP score was estimated among all classifiers at a sample-level. For each gene, we first collected the ShAP score of all 28 genomic features at a sample-level and then grouped them based on their positive and negative signs. Next, we compared the absolute value of the sum of ShAP scores of genomic features having positive ShAP scores with the absolute sum of those having negative ShAP scores. The ShAP score having the largest absolute value was considered the best ShAP score for that gene and the classifier. This step was repeated for all those classifiers that predicted the sample's class correctly, and the best ShAP score was saved for each of them. The ShAP score, having the largest absolute value among all the classifiers, was considered the best ShAP score for a gene at a sample-level. For a better clarity of the steps employed in the estimation of the best ShAP scores of significantly altered genes

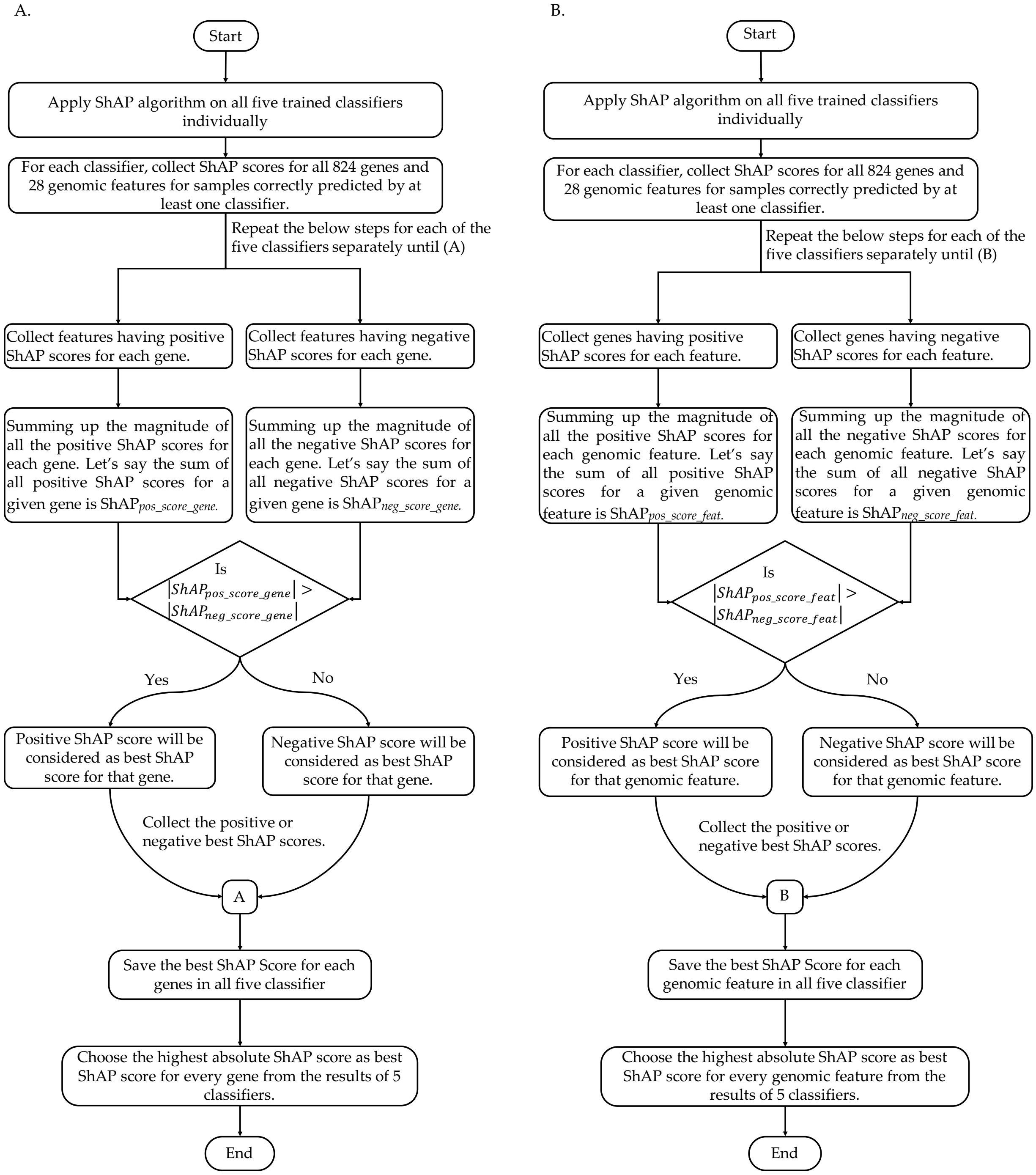

A.
Start
Apply ShAP algorithm on all five trained classifiers individually
For each classifier, collect ShAP scores for all 824 genes and 28 genomic features for samples correctly predicted by at least one classifier.
Repeat the below steps for each of the five classifiers separately until (A)
Collect features having positive ShAP scores for each gene.
Collect features having negative ShAP scores for each gene.
Summing up the magnitude of all the positive ShAP scores for each gene. Let's say the sum of all positive ShAP scores for a given gene is ShAP$_{pos\_score\_gene.}$
Summing up the magnitude of all the negative ShAP scores for each gene. Let's say the sum of all negative ShAP scores for a given gene is ShAP$_{neg\_score\_gene.}$
Is $|ShAP_{pos\_score\_gene}| > |ShAP_{neg\_score\_gene}|$
Yes
No
Positive ShAP score will be considered as best ShAP score for that gene.
Negative ShAP score will be considered as best ShAP score for that gene.
Collect the positive or negative best ShAP scores.
A
Save the best ShAP Score for each genes in all five classifier
Choose the highest absolute ShAP score as best ShAP score for every gene from the results of 5 classifiers.
End
B.
Start
Apply ShAP algorithm on all five trained classifiers individually
For each classifier, collect ShAP scores for all 824 genes and 28 genomic features for samples correctly predicted by at least one classifier.
Repeat the below steps for each of the five classifiers separately until (B)
Collect genes having positive ShAP scores for each feature.
Collect genes having negative ShAP scores for each feature.
Summing up the magnitude of all the positive ShAP scores for each genomic feature. Let's say the sum of all positive ShAP scores for a given genomic feature is ShAP$_{pos\_score\_feat.}$
Summing up the magnitude of all the negative ShAP scores for each genomic feature. Let's say the sum of all negative ShAP scores for a given genomic feature is ShAP$_{neg\_score\_feat.}$
Is $|ShAP_{pos\_score\_feat}| > |ShAP_{neg\_score\_feat}|$
Yes
No
Positive ShAP score will be considered as best ShAP score for that genomic feature.
Negative ShAP score will be considered as best ShAP score for that genomic feature.
Collect the positive or negative best ShAP scores.
B
Save the best ShAP Score for each genomic feature in all five classifier
Choose the highest absolute ShAP score as best ShAP score for every genomic feature from the results of 5 classifiers.
End

Figure 4.4: Flowchart showing steps for estimating the best ShAP score for (A) 824 significantly altered genes and (B) 28 genomic features at sample-level to reveal their contribution to the BDL-SP model prediction.

and genomic features at a sample-level, one may refer to Figure-4.4(A) and Algorithm-A of Table4.2, Figure-4.4(B) and Algorithm-B of Table4.2, respectively.

Similarly, for each genomic feature, we first collected the ShAP score of all 824 genes at a sample-level and grouped them based on their positive and negative signs. Next, we compared the absolute value of the sum of ShAP scores of genes having positive scores with the sum of ShAP scores of genes having negative ShAP scores. The ShAP score having the largest absolute value was considered the best ShAP score for a genomic feature and the classifier. We repeated the above step for all the classifiers that predicted the sample's class correctly and saved the best ShAP score for each classifier. The ShAP score, having the largest absolute value among all the classifiers, was considered the best ShAP score for that genomic feature at a sample-level. Once the best ShAP scores were obtained for all the genes and all the genomic attributes, the top-ranked genes and the top-ranked genomic attributes were identified at the group-level and at the sample-level.

Further, the top-ranked significantly altered genes revealed by BDL-SP were also compared with the MM-related studies to identify the previously reported significantly altered genes. We included information from multiple databases for model validation and post-hoc analysis at gene level analysis (OncoKB, COSMIC, IntoGen, and TargetDB databases). We downloaded a list of 1064 cancer genes from OncoKB to deduce the OGs and TSGs in our top mutated genes. Further, 318 OGs and 320 TSGs obtained from the COSMIC database were also used to deduce OGs and TSGs in our top-mutated genes. Similarly, we created a list of MM driver genes reported by [222, 87]. MM Driver genes were also extracted from the IntoGen database [223] to infer MM driver genes present in our top mutated gene list. Finally, a list of 180 AGs from the COSMIC database and 135 AGs from the TargetDB database [224] was used to infer the AGs present in our top mutated gene list. The top-ranked significantly altered genes were grouped into four categories based on their functional significance as follows: 1. OGs; 2. TSGs; 3. ODGs; 4. AGs. The top-ranked significantly altered genes in each of the above gene categories were then collected at the group-level (MM/MGUS) and the sample-level. We also checked the role of genomic features on the disease classification in post-hoc interpretability analysis of the BDL- SP model.

Statistical analysis

We performed the unpaired Wilcoxon rank-sum statistical analysis to study the impact of ethnicity in MM. In this analysis, we first extracted the top significantly altered genes from the WES data of MGUS/MM patients of American (MMRF), European (EGA), and Indian (AIIMS) populations using the top-performing BDL-SP model. Next, for each sample, we computed the total number of significantly altered genes that belonged to the reported categories of OGs, TSGs, ODGs, and AGs of MM literature. Then, we performed a statistical comparison of the number of significantly altered genes of the reported category of OGs, TSGs, ODGs, and AGs on the groups of American (MMRF), European (EGA), and Indian (AIIMS) populations to study the impact of ethnicity on individual gene category.

Gene pathway analysis

The significant genes identified by BDL-SP, which helped differentiate MM from MGUS, were mapped back to the significant gene list obtained for MM and MGUS using the dndscv tool. Some genes were found to be common in both groups, while some were found to be significantly mutated either in MGUS or in MM only. Pathway analysis was done on the top 500 genes obtained from the BDL-SP model. KEGG and Reactome pathways were found via Enrichr gene set enrichment analysis web server [225, 226, 227].

Impact of Nonsynonymous SNVs on protein structure

To evaluate the influence of nonsynonymous SNVs on protein structure and function, we initially collate those classified as pathogenic by the FATHMM-XF algorithm and identified by all four variant callers (MuSE, Mutect2, SomaticSniper, and Varscan2). These SNVs are then refined using four deleteriousness scores: SIFT, PolyPhen2-HDIV, PolyPhen2-HVAR, and PROVEAN, considering only those deemed deleterious or damaging by all four scores. Subsequently, utilizing ANNOVAR, we ascertain the affected Pfam domains to gain deeper insights into the ramifications of these SNVs on protein function. Finally, we utilized SWISS-MODEL [228] for visualizing their corresponding structural conformation.

4.3 Results

Using the dndscv tool (as shown in Figure-4.2), 362 and 617 significantly altered genes were identified in MGUS and MM, respectively. Of these, 155 genes were common in MGUS and MM. The complete list of all 824 genes is shown in Table S1 of supplementary material. We then inferred the important genes that were accountable for distinguishing MGUS from MM as obtained through our graph-based BDL-SP model.

Table 4.2: (A) Pseudo-codes of algorithm A for estimating the best ShAP score of 824 genes, (B) Pseudo-codes of algorithm A for estimating the best ShAP score of 28 features.

```
Algorithm A: Estimate the Best ShAP Score (BSS) for each gene at a sample level
$Fivefold\ classifiers$ ← [List of five classifiers trained on each fold of test dataset]
$CPC$ ← [List of correct prediction classifiers, i.e. classifiers that correctly predicted the sample's class]
$SFM$ ← Sample feature matrix
$Genes$ ← [List of 824 genes]
$GFPS_{g|c}$ ← [List of genomic features having positive ShAP score for a gene "g" and classifier "c"]
$GFNS_{g|c}$ ← [List of genomic features having negative ShAP score for a gene "g" and classifier "c"]
$CSG_{g|c}$ ← Best ShAP score of gene "g" and classifier "c"
$ACGS[c]_g$ ← [List of best ShAP scores of gene "g" for all the classifiers that correctly predicted the sample]
$BSG_g$ ← Best ShAP score of gene "g" among all classifiers
$LBSG_{genes}$ ← List of the best ShAP score of all the genes among all the classifiers
**procedure** $BSSgene(SFM)$
    **for** classifier in $Fivefold\ classifiers$ **do**
        Predict the sample's class with the help of a classifier
        **if** classifier predicts the correct sample class **then**
            $CPC$ ← [Append the classifier in CPC list]
            Apply ShAP algorithm on the classifier
            Collect the ShAP score for all 824 genes on their respective 28 genomic features for that classifier
    **for** gene in $Genes$ **do**
        **for** classifier in $CPC$ **do**
            $GFPS_{gene|classifier}$ ← Collect features having positive ShAP score
            $GFNS_{gene|classifier}$ ← Collect features having negative ShAP score
            **if** $\left|\sum GFPS_{gene|classifier}\right|$ > $\left|\sum GFNS_{gene|classifier}\right|$ **then**
                $CSG_{gene|classifier} \leftarrow GFPS_{gene|classifier}$
            **else**
                $CSG_{gene|classifier} \leftarrow GFNS_{gene|classifier}$
            $ACGS[classifier]_{gene} \leftarrow CSG_{gene|classifier}$
        $BSG_{gene} \leftarrow ACGS[argmax[|CSG|$ for $CSG$ in $ACSG]]$
        $LBSG_{genes}[gene] \leftarrow BSG_{gene}$
    Output ← $LBSG_{genes}$
```

Algorithm B: Estimate the Best ShAP Score (BSS) for each genomic feature (GF) at a sample level

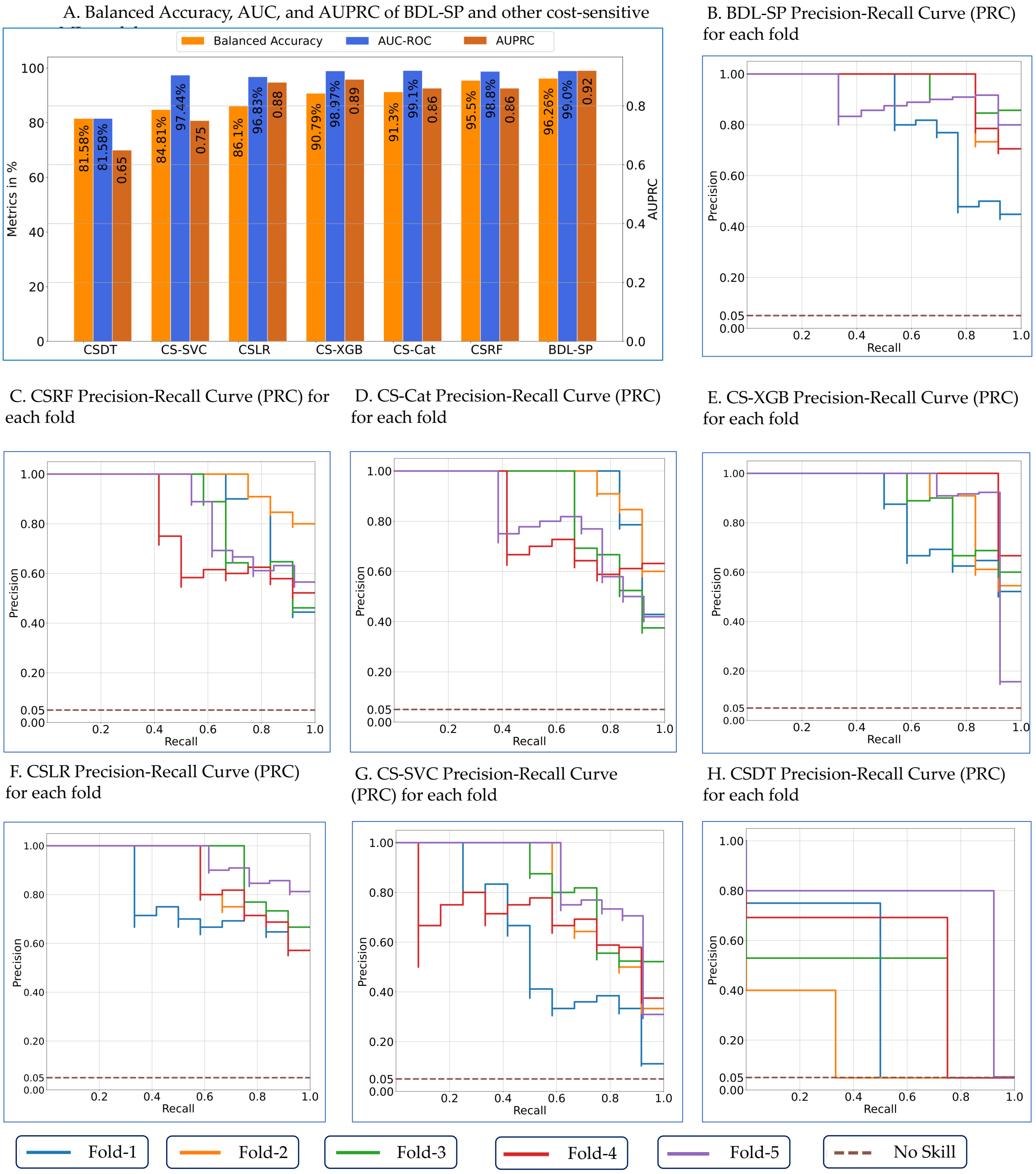
```
$Fivefold\ classifiers$ ← [List of five classifiers trained on each fold of test dataset]
$CPC$ ← [List of correct prediction classifiers, i.e. classifiers that correctly predicted the sample's class]
$SFM$ ← Sample feature matrix
$Genomic\ Features$ ← [List of 28 genomic features]
$GPS_{gf|c}$ ← [List of genes having positive ShAP score for genomic feature "gf" and classifier "c"]
$GNS_{g|c}$ ← [List of genes having negative ShAP score for genomic feature "gf" and classifier "c"]
$CSGF_{gf|c}$ ← Best ShAP score of genomic feature "gf" and classifier "c"
$ACGFS[c]_g$ ← [List of best ShAP scores of genomic feature "gf" for all the classifiers that correctly predicted the sample]
$BSGF_g$ ← Best ShAP score of genomic feature "gf" among all classifiers
$LBSGF_{genes}$ ← List of the best ShAP score of all genomic features among all the classifiers
procedure $BSSGenomicFeature(SFM)$
    for classifier in $Fivefold\ classifiers$ do
        Predict the sample's class with the help of a classifier
        if classifier predicts the correct sample class then
            $CPC$ ← [Append the classifier in CPC list]
            Apply ShAP algorithm on the classifier
            Collect the ShAP score for all 824 genes on their respective 28 genomic features for that classifier
    for features in $Genomic\ Features$ do
        for classifier in $CPC$ do
            $GPS_{feature|classifier}$ ← Collect genes having positive ShAP score
            $GNS_{feature|classifier}$ ← Collect genes having negative ShAP score
            if $\left|\sum GPS_{feature|classifier}\right|$ > $\left|\sum GNS_{feature|classifier}\right|$ then
                $CSGF_{feature|classifier} \leftarrow GPS_{feature|classifier}$
            else
                $CSGF_{feature|classifier} \leftarrow GNS_{feature|classifier}$
            $ACGFS[classifier]_{feature} \leftarrow CSGF_{feature|classifier}$
        $BSGF_{feature} \leftarrow ACGFS[argmax[|CSGF|$ for $CSGF$ in $ACGFS]]$
        $LBSGF_{GenomicFeatures}[feature] \leftarrow BSGF_{feature}$
    Output ← $LBSGF_{GenomicFeatures}$
```

4.3.1 Comparative performance of BDL-SP and standard ML models

Using our AI-based workflow of BDL-SP (Figures-4.2 and 4.3), we trained the BDL-SP model with a 5-fold cross-validation and compared its performance with six standard cost-sensitive machine learning models. Results of the BDL-SP model and all the six cost-sensitive classifiers are presented in Figure-4.5. The proposed BDL-SP model outperformed the rest of the models in terms of balanced accuracy and AUPRC (area under the precision-recall curve), while the area under the curve (AUC) was largest (and equal) for the top three models. The BDL-SP model performed best with a balanced accuracy of 96.26%. Cost-sensitive Random Forest (CSRF) performed the next best with a balanced accuracy of 95.5%, and cost-sensitive Catboost (CS-Cat) performed the third best with a balanced accuracy of 91.3% (Figure-4.5A). All three models reported an Area Under the Curve (AUC) value of 0.99. BDL-SP model also outperformed other models on AUPRC, scoring the largest AUPRC of 0.92, while the AUPRC of both CSRF and CS-Cat model was 0.86 (Figure-4.5B-D). It is worth noting that AUPRC is one of the most important quantitative metrics and is more relevant than AUC in terms of unbalanced data. BDL-SP outperformed the other models on AUPRC by a great margin. This shows that, quantitatively, BDL-SP performed best, with the CSRF model being the second-best model.

BDL-SP identified the maximum number of minority class samples, i.e., 60 out of 61 MGUS samples and 1087 MM samples out of a total of 1153 MM samples. The second-best model was CSRF, which identified 59 out of 61 MGUS samples and 1086 out of 1153 MM samples. The third best-performing model was CS-Cat, which identified 52 out of 61 MGUS samples and 1121 out of 1153 MM samples. Thus, BDL-SP outperformed other models on minority class detection, and CSRF performed next to this model. Since the performance of CSRF was close to the leading BDL-SP model on metrics other than AUPRC, we performed post-hoc interpretability benchmarking of the top three performing models (BDL-SP, CSRF, and CS-Cat). In post-hoc benchmarking, we utilized the ShAP algorithm. We tabulated the top 250 and top 500 genes identified by the top three trained models to understand the reasons for the model's predictions. Then, the top-ranked genes were further analyzed to identify previously reported OGs, TSGs, ODGs, and AGs in MM. As demonstrated later in this Section with the post-hoc interpretability analysis results, we observed that BDL-SP identified the maximum number of the previously reported genes in the top 250 and 500 genes.

Out of 824 significantly altered genes identified from the workflow shown in Figure-4.2, there were 31 OGs (e.g. *KRAS, LTB, CARD11, NOTCH1*, etc.), 43 TSGs (e.g. *HLA-A/B/C, TRAF3, TP53, SDHA*, etc.), ten genes that were ODGs (*KRAS, LTB, NRAS, FGFR3, BRAF*), and 19 AGs (e.g. *MITF, ARID1B, ARID2, RPTOR*, etc.) (Table-4.3).



A. Balanced Accuracy, AUC, and AUPRC of BDL-SP and other cost-sensitive

B. BDL-SP Precision-Recall Curve (PRC) for each fold

C. CSRF Precision-Recall Curve (PRC) for each fold

D. CS-Cat Precision-Recall Curve (PRC) for each fold

E. CS-XGB Precision-Recall Curve (PRC) for each fold

F. CSLR Precision-Recall Curve (PRC) for each fold

G. CS-SVC Precision-Recall Curve (PRC) for each fold

H. CSDT Precision-Recall Curve (PRC) for each fold

Figure 4.5: (A) The benchmarking of the performance of BDL-SP with six cost-sensitive ML models on the metrics of balanced accuracy, AUC, and AUPRC (Area under Precision-Recall Curve). Precision-Recall Curves (PRC) for all five folds of (B) BDL-SP, (C) CSRF, (D) CS-Cat, (E) CS-XGB, (F) CSLR, (G) CS-SVC, and (H) CSDT. No skill line is also shown in each of the AUPRC plots, representing the inability of the classifier to correctly classify a sample. The full form of the abbreviation used in these figures are as follows: CSDT = Cost-Sensitive Decision Tree, CS-SVC = Cost-Sensitive Support Vector Machine, CSLR = Cost-Sensitive Logistic Regression, CS-XGB = Cost-Sensitive XGBoost, CS-Cat = Cost-Sensitive CatBoost, and CSRF = Cost-Sensitive Random Forest.

Table 4.3: Types of four different gene categories (OG, TSG, ODG, and AG) and their counts in 824 significantly altered genes

Gene type based on functionality	Total number of previously reported genes present in our list of 824 significantly altered genes
Oncogenes	31
Tumor-suppressor genes	43
Both oncogene and driver gene	10
Actionable genes	19

Table 4.4: Counts of previously reported 4 categories of genes as found in the post-hoc analysis based on top 250 and top 500 genes of the top 3 models (BDL-SP, CSRF, and CS-Cat)

Top Genes	BDL-SP (Top-performing model)				CSRF (Second best model)				CS-Cat (Third best model)			
	OG	TSG	ODG	AG	OG	TSG	ODG	AG	OG	TSG	ODG	AG
top 250	**20**	**21**	**7**	**11**	7	10	1	4	6	5	1	4
top 500	**27**	**37**	**10**	**17**	7	10	1	4	6	5	1	4

The number of previously reported genes (OG/TSG/ODG/AG) obtained in each category (top 250/top 500) using the best performing model is highlighted in bold.

This full list of genes is provided in Table S1 of supplementary material. For each of the top three models, we have considered only those genes in the top 250 or top 500 gene list that have a non-zero ShAP score in the post-hoc explainability analysis. The total counts of previously reported genes as found in the top 250 and top 500 genes of the top three models is shown in Table-4.4.

From Table-4.4, we observed that the BDL-SP model identified 20 out of 31 OGs in the top 250 and 27 out of 31 OGs in the top 500 gene list, while CSRF and CS-Cat could identify only 7 and 5 OGs in top 250 and top 500 gene list, respectively. Similarly, out of 43 TSGs, the BDL-SP model identified 21 and 37 TSGs in the top 250 and top 500 gene lists, while CSRF and CS-Cat identified only 10 and 5 TSGs, respectively, in the top 250 and top 500 gene lists. Further, the BDL-SP model identified 7 and all 10 ODGs, while CSRF and CS-Cat could identify only one ODG in the top 250 and top 500 significantly altered genes. Finally, the BDL-SP model identified 11 and 17 AGs in the top 250 and top 500 genes, respectively, while CSRF and CS-Cat could identify only 4 AGs in the top 250 and top 500 significantly altered genes. The post-hoc benchmarking of the top three models is shown in Table-4.4 and the list of OGs, TSGs, ODGs, and AGs in the top 250 and top 500 significantly altered gene list of BDL-SP, CSRF, and CS-Cat models is provided in Table-4.5. Since the BDL-SP model identified the largest number of previously reported OGs, TSGs, ODGs, and AGs, this model can be inferred as the best-performing model and was used subsequently for inferring the top significantly altered genes, genomic features, and altered signaling pathways to identify the pivotal genomic biomarkers to distinguish MM and MGUS. This analysis shows that one can obtain similar quantitative results with two or more different ML models, but one should choose the model that is more interpretable with reference to the application domain.

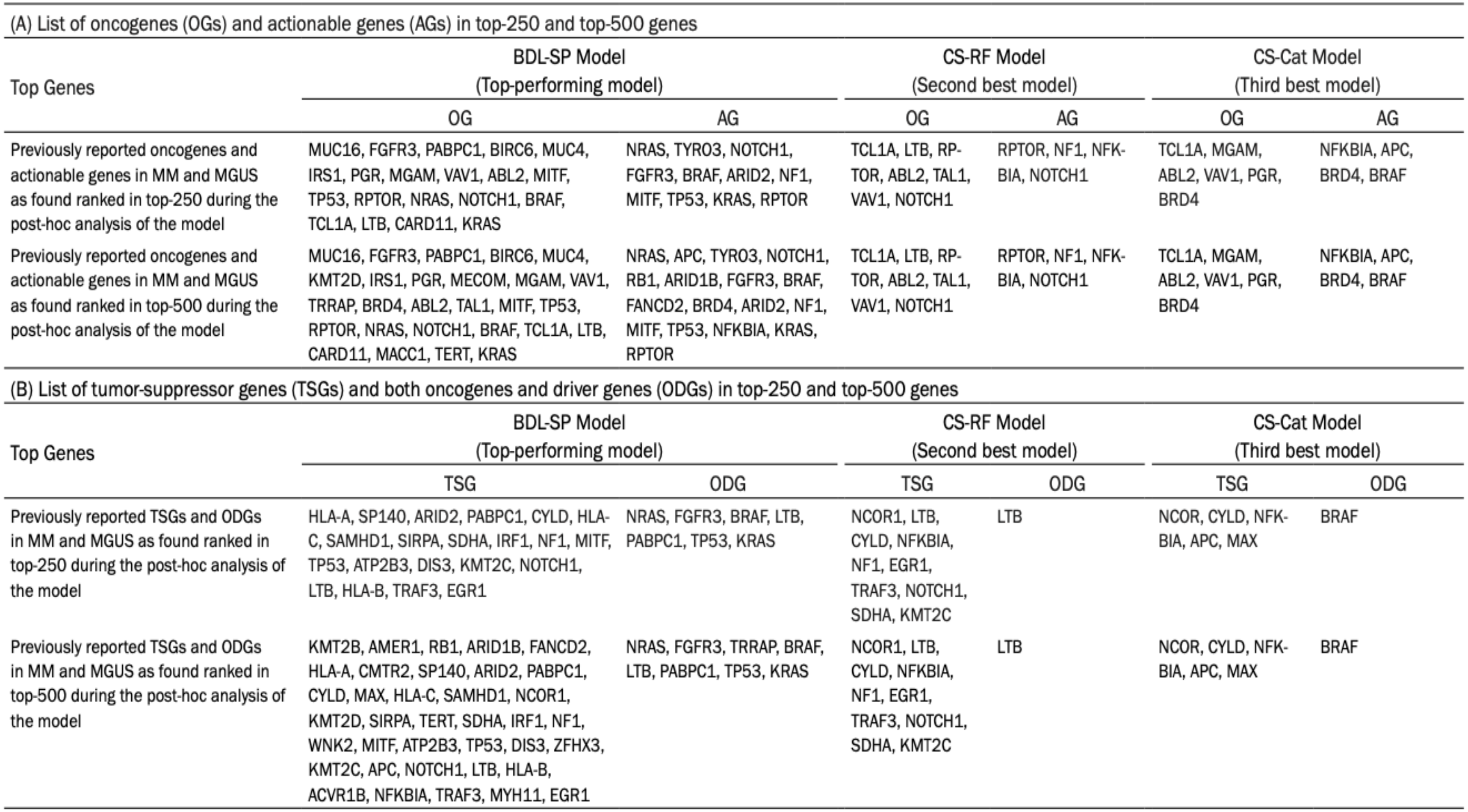
4.3.2 Pathway analysis on the top 500 genes obtained from the BDL-SP model

On comparing the top 500 significantly altered genes obtained from the BDL-SP model (that helped in differentiating MM from MGUS) to the significant gene list obtained for MM and MGUS using the dndscv tool, 301 genes were observed to be statistically significantly mutated only in the MM cohort, 101 genes were observed to be statistically significantly mutated only in the MGUS cohort, while 98 genes were observed to be statistically significantly mutated in both MM and MGUS cohorts. The set of 301 genes that were found to be significantly mutated only in the MM cohort included several important OGs, ODGs, TSGs, and AGs such as *BCL7A, BRAF, CARD11, CYLD, DIS3, EGR1, FAM46C, IGLL5, KRAS, KMT2D, NRAS, MECOM,* etc. Similarly, the set of 101 genes significantly mutated only in the MGUS cohort included *APC, FAM47B, MGAM, NOTCH1, TYRO3*, etc. The set of 98 common genes observed to be significantly mutated in MGUS and MM cohorts included *AMER1, FANCD2, HLA-B, KMT2C, PABPC1, TRRAP*, etc. The complete list of top significantly altered genes only in MM, only in MGUS, and common in both MM and MGUS is provided in Table S7 of supplementary material.

Enrichr and Reactome were used to infer the KEGG and Reactome pathways altered by 399 MM and 199 MGUS genes. A total of 5 KEGG pathways inferred from Enrichr were significantly altered in MGUS (Table S2 of supplementary material) and 108 KEGG pathways were significantly altered in MM (Table S3 of supplementary material). Similarly, a total of 10 Reactome pathways inferred from Enrichr were significantly altered in MGUS (Table S2 of supplementary material) and 134 Reactome pathways inferred from Enrichr were significantly altered in MM (Table S3 of supplementary material). Further, we grouped the significantly altered pathway into four categories based on the variations in their significance with disease progression from MGUS to MM as follows:

1. *Category-1*: Pathways that become more significant with disease progression from MGUS to MM.
2. *Category-2*: Pathways that become less significant with disease progression from MGUS to MM.
3. *Category-3*: Significantly altered pathways observed only in MM and not observed

Table 4.5: List of 4 categories of previously reported genes as found in the post-hoc analysis based top 250 and top 500 genes of the top 3 models (BDL-SP, CSRF, and CS-Cat)



(A) List of oncogenes (OGs) and actionable genes (AGs) in top-250 and top-500 genes

Top Genes	BDL-SP Model (Top-performing model)		CS-RF Model (Second best model)		CS-Cat Model (Third best model)	
	OG	AG	OG	AG	OG	AG
Previously reported oncogenes and actionable genes in MM and MGUS as found ranked in top-250 during the post-hoc analysis of the model	MUC16, FGFR3, PABPC1, BIRC6, MUC4, IRS1, PGR, MGAM, VAV1, ABL2, MITF, TP53, RPTOR, NRAS, NOTCH1, BRAF, TCL1A, LTB, CARD11, KRAS	NRAS, TYRO3, NOTCH1, FGFR3, BRAF, ARID2, NF1, MITF, TP53, KRAS, RPTOR	TCL1A, LTB, RPTOR, ABL2, TAL1, VAV1, NOTCH1	RPTOR, NF1, NFKBIA, NOTCH1	TCL1A, MGAM, ABL2, VAV1, PGR, BRD4	NFKBIA, APC, BRD4, BRAF
Previously reported oncogenes and actionable genes in MM and MGUS as found ranked in top-500 during the post-hoc analysis of the model	MUC16, FGFR3, PABPC1, BIRC6, MUC4, KMT2D, IRS1, PGR, MECOM, MGAM, VAV1, TRRAP, BRD4, ABL2, TAL1, MITF, TP53, RPTOR, NRAS, NOTCH1, BRAF, TCL1A, LTB, CARD11, MACC1, TERT, KRAS	NRAS, APC, TYRO3, NOTCH1, RB1, ARID1B, FGFR3, BRAF, FANCD2, BRD4, ARID2, NF1, MITF, TP53, NFKBIA, KRAS, RPTOR	TCL1A, LTB, RPTOR, ABL2, TAL1, VAV1, NOTCH1	RPTOR, NF1, NFKBIA, NOTCH1	TCL1A, MGAM, ABL2, VAV1, PGR, BRD4	NFKBIA, APC, BRD4, BRAF

(B) List of tumor-suppressor genes (TSGs) and both oncogenes and driver genes (ODGs) in top-250 and top-500 genes

Top Genes	BDL-SP Model (Top-performing model)		CS-RF Model (Second best model)		CS-Cat Model (Third best model)	
	TSG	ODG	TSG	ODG	TSG	ODG
Previously reported TSGs and ODGs in MM and MGUS as found ranked in top-250 during the post-hoc analysis of the model	HLA-A, SP140, ARID2, PABPC1, CYLD, HLA-C, SAMHD1, SIRPA, SDHA, IRF1, NF1, MITF, TP53, ATP2B3, DIS3, KMT2C, NOTCH1, LTB, HLA-B, TRAF3, EGR1	NRAS, FGFR3, BRAF, LTB, PABPC1, TP53, KRAS	NCOR1, LTB, CYLD, NFKBIA, NF1, EGR1, TRAF3, NOTCH1, SDHA, KMT2C	LTB	NCOR, CYLD, NFKBIA, APC, MAX	BRAF
Previously reported TSGs and ODGs in MM and MGUS as found ranked in top-500 during the post-hoc analysis of the model	KMT2B, AMER1, RB1, ARID1B, FANCD2, HLA-A, CMTR2, SP140, ARID2, PABPC1, CYLD, MAX, HLA-C, SAMHD1, NCOR1, KMT2D, SIRPA, TERT, SDHA, IRF1, NF1, WNK2, MITF, ATP2B3, TP53, DIS3, ZFHX3, KMT2C, APC, NOTCH1, LTB, HLA-B, ACVR1B, NFKBIA, TRAF3, MYH11, EGR1	NRAS, FGFR3, TRRAP, BRAF, LTB, PABPC1, TP53, KRAS	NCOR1, LTB, CYLD, NFKBIA, NF1, EGR1, TRAF3, NOTCH1, SDHA, KMT2C	LTB	NCOR, CYLD, NFKBIA, APC, MAX	BRAF

in MGUS.

4. *Category-4*: Significantly altered pathways observed only in MGUS and not observed in MM.

The complete list of significantly altered pathways for the above-mentioned four categories is provided in Tables S4 and Table S5 of supplementary material. In Category-1 of significantly altered pathways, 05 KEGG and 09 Reactome pathways became more significant as the disease progressed from MGUS to MM (Figure-4.6). In Category-2, no pathway became less significant with disease progression in KEGG and in Reactome. In Category-3, 103 KEGG pathways and 125 Reactome pathways were observed as significantly altered only in MM and not in MGUS (Figures-4.7 and 4.8).

We further observed that 14 out of 103 KEGG pathways and 14 out of 125 Reactome pathways had no overlapping genes with the set of 199 significantly altered genes in MGUS. Finally, in Category-4, no KEGG pathway, but one Reactome pathway was observed as significantly altered only in MGUS and not in MM (Figure-4.9). Further, we observed that several signaling pathways, such as Calcium signaling, B-cell receptor signaling, MAPK signaling pathway, regulation of actin cytoskeleton, etc., were significantly altered only in MM (adjusted p-value ≤ 0.05) and were not observed to be significantly altered in MGUS. The KEGG pathways that were significantly involved in disease progression from MGUS to MM with highlighted top-ranking genes identified by BDL-SP are shown in Figure-4.10.

4.3.3 Explainability of the BDL-SP model using ShAP algorithm

We utilized the ShAP algorithm for post-hoc model explainability and to rank genomic attributes based on their contribution to the model prediction. Here, each genomic attribute was assigned a ShAP score based on its contribution to each class (MM/MGUS) and has been ranked at the group-level (MM versus MGUS) and sample-level accordingly. We conducted the ShAP analysis for the post-hoc explainability of the trained model in three different ways, as explained in the subsequent sections.

4.3.4 Ranking of genes at the group-level from the explainability analysis of BDL-SP using ShAP

Based on the best ShAP score estimated for each genomic attribute using the algorithms shown in Figure-4.4 A and Figure-4.4B, we ranked all the significantly altered genes at the group-level (MM/MGUS) to identify the top genes that significantly contributed to the model's prediction. The gene ranking of all 824 genes at group-level is shown in the beeswarm plot in Table S6 of supplementary material.

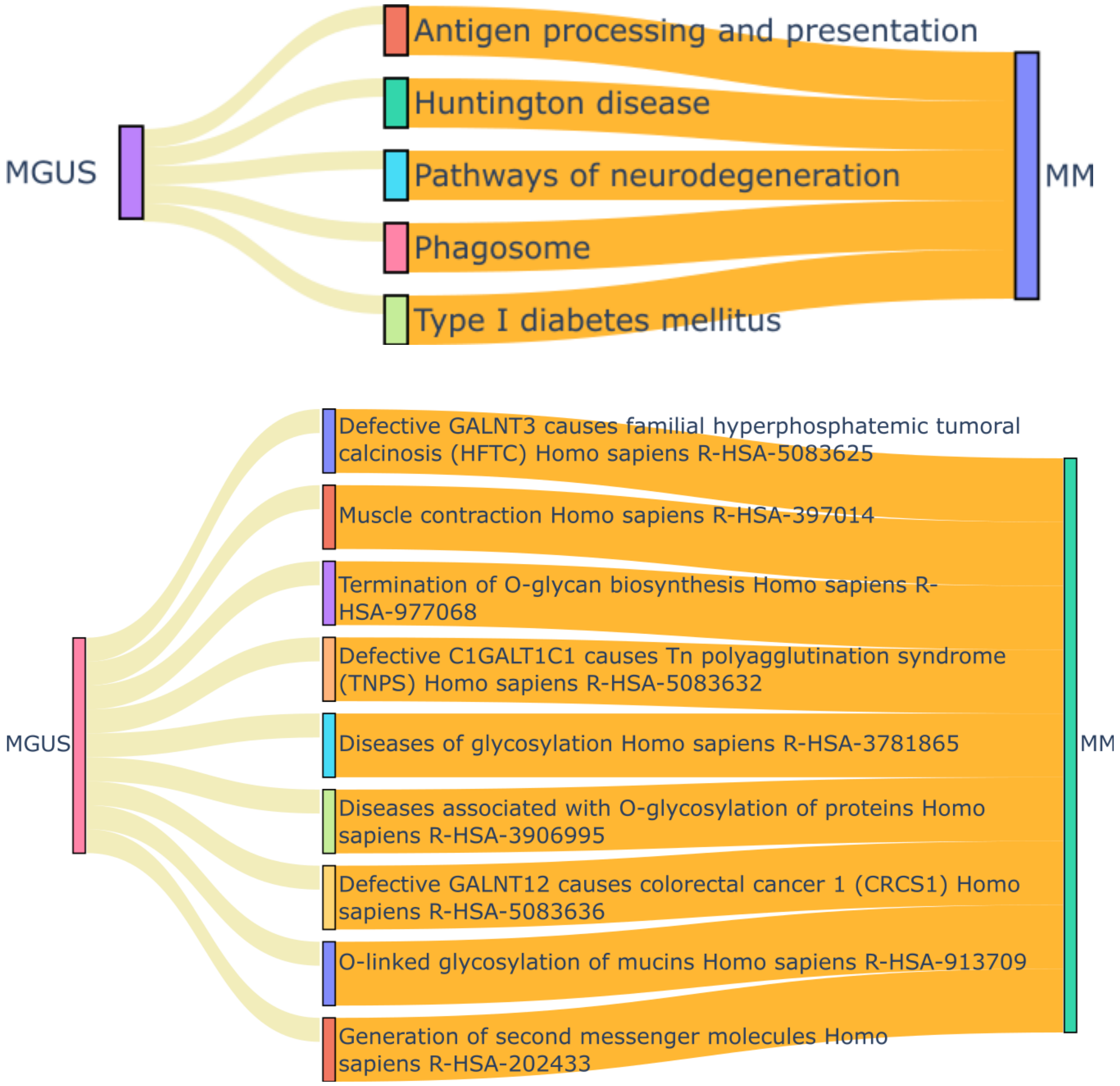
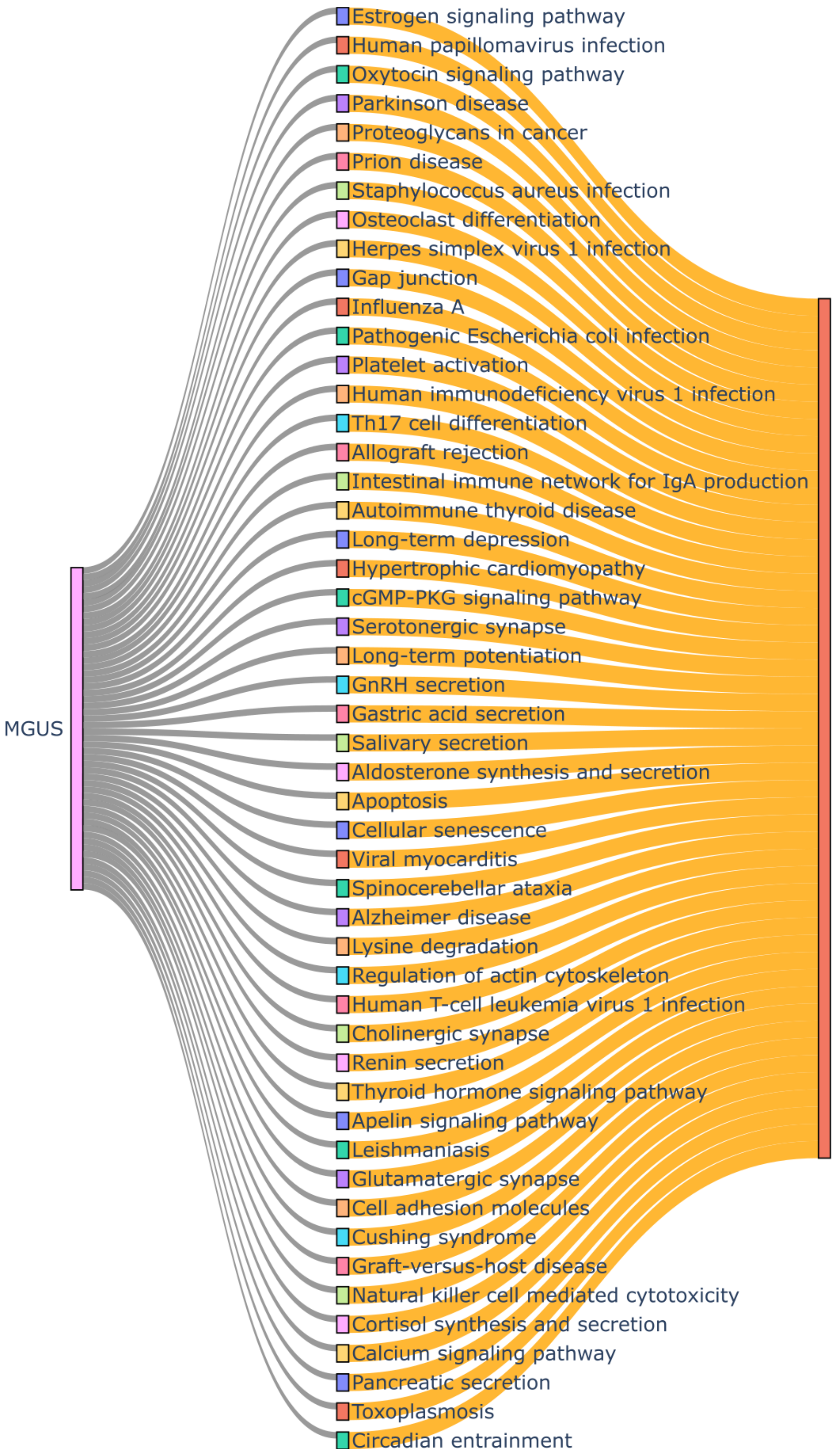

MGUS
Antigen processing and presentation
Huntington disease
Pathways of neurodegeneration
Phagosome
Type I diabetes mellitus
MM
MGUS
Defective GALNT3 causes familial hyperphosphatemic tumoral calcinosis (HFTC) Homo sapiens R-HSA-5083625
Muscle contraction Homo sapiens R-HSA-397014
Termination of O-glycan biosynthesis Homo sapiens R-HSA-977068
Defective C1GALT1C1 causes Tn polyagglutination syndrome (TNPS) Homo sapiens R-HSA-5083632
Diseases of glycosylation Homo sapiens R-HSA-3781865
Diseases associated with O-glycosylation of proteins Homo sapiens R-HSA-3906995
Defective GALNT12 causes colorectal cancer 1 (CRCS1) Homo sapiens R-HSA-5083636
O-linked glycosylation of mucins Homo sapiens R-HSA-913709
Generation of second messenger molecules Homo sapiens R-HSA-202433
MM

Figure 4.6: Pathway enrichment analysis of the top-genes obtained from BDL-SP model. A. KEGG Pathways that gained more significance during progression from MGUS to MM. B. Reactome Pathways gained more significance during the progression from MGUS to MM. Here, in both of the figures, pale golden and orange ribbon means significant p-adjusted value (≤ 0.05); orange refers to more significance, and pale golden color refers to less significance.



MGUS
Estrogen signaling pathway
Human papillomavirus infection
Oxytocin signaling pathway
Parkinson disease
Proteoglycans in cancer
Prion disease
Staphylococcus aureus infection
Osteoclast differentiation
Herpes simplex virus 1 infection
Gap junction
Influenza A
Pathogenic Escherichia coli infection
Platelet activation
Human immunodeficiency virus 1 infection
Th17 cell differentiation
Allograft rejection
Intestinal immune network for IgA production
Autoimmune thyroid disease
Long-term depression
Hypertrophic cardiomyopathy
cGMP-PKG signaling pathway
Serotonergic synapse
Long-term potentiation
GnRH secretion
Gastric acid secretion
Salivary secretion
Aldosterone synthesis and secretion
Apoptosis
Cellular senescence
Viral myocarditis
Spinocerebellar ataxia
Alzheimer disease
Lysine degradation
Regulation of actin cytoskeleton
Human T-cell leukemia virus 1 infection
Cholinergic synapse
Renin secretion
Thyroid hormone signaling pathway
Apelin signaling pathway
Leishmaniasis
Glutamatergic synapse
Cell adhesion molecules
Cushing syndrome
Graft-versus-host disease
Natural killer cell mediated cytotoxicity
Cortisol synthesis and secretion
Calcium signaling pathway
Pancreatic secretion
Toxoplasmosis
Circadian entrainment
MM

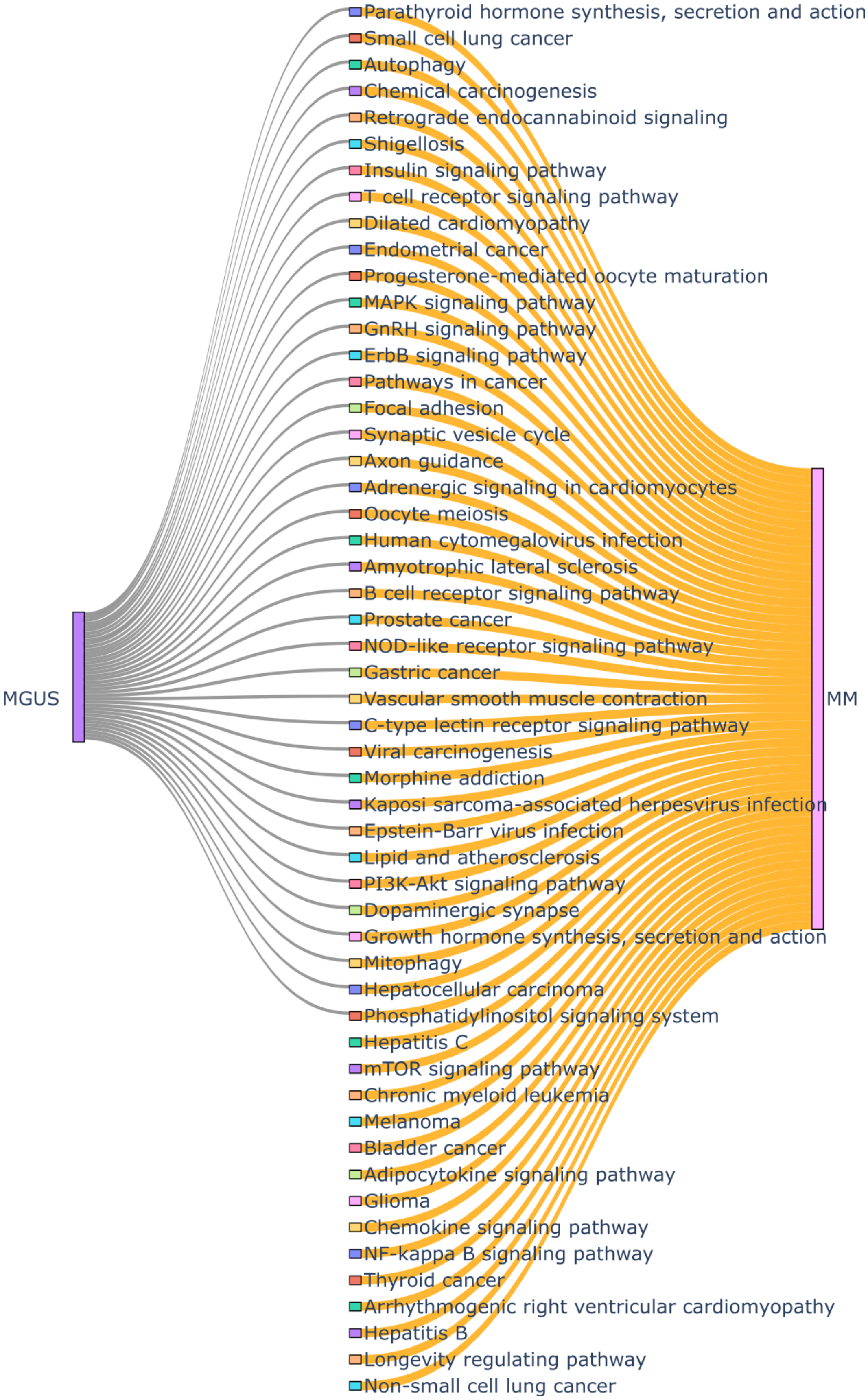

MGUS
MM
Parathyroid hormone synthesis, secretion and action
Small cell lung cancer
Autophagy
Chemical carcinogenesis
Retrograde endocannabinoid signaling
Shigellosis
Insulin signaling pathway
T cell receptor signaling pathway
Dilated cardiomyopathy
Endometrial cancer
Progesterone-mediated oocyte maturation
MAPK signaling pathway
GnRH signaling pathway
ErbB signaling pathway
Pathways in cancer
Focal adhesion
Synaptic vesicle cycle
Axon guidance
Adrenergic signaling in cardiomyocytes
Oocyte meiosis
Human cytomegalovirus infection
Amyotrophic lateral sclerosis
B cell receptor signaling pathway
Prostate cancer
NOD-like receptor signaling pathway
Gastric cancer
Vascular smooth muscle contraction
C-type lectin receptor signaling pathway
Viral carcinogenesis
Morphine addiction
Kaposi sarcoma-associated herpesvirus infection
Epstein-Barr virus infection
Lipid and atherosclerosis
PI3K-Akt signaling pathway
Dopaminergic synapse
Growth hormone synthesis, secretion and action
Mitophagy
Hepatocellular carcinoma
Phosphatidylinositol signaling system
Hepatitis C
mTOR signaling pathway
Chronic myeloid leukemia
Melanoma
Bladder cancer
Adipocytokine signaling pathway
Glioma
Chemokine signaling pathway
NF-kappa B signaling pathway
Thyroid cancer
Arrhythmogenic right ventricular cardiomyopathy
Hepatitis B
Longevity regulating pathway
Non-small cell lung cancer

Figure 4.7 *(previous page)*: A, B. Pathway enrichment analysis of the top-genes obtained from BDL-SP model: KEGG Pathways that are uniquely significant in MM. In the above figure, the orange ribbon means a significant p-adjusted value ($\leq$ 0.05), and the gray color refers to a non-significant p-adjusted value (> 0.05). A total of 108 KEGG pathways were observed to be significantly altered. Due to the large number of altered pathways, the above river plot was split into two parts to get more clarity.

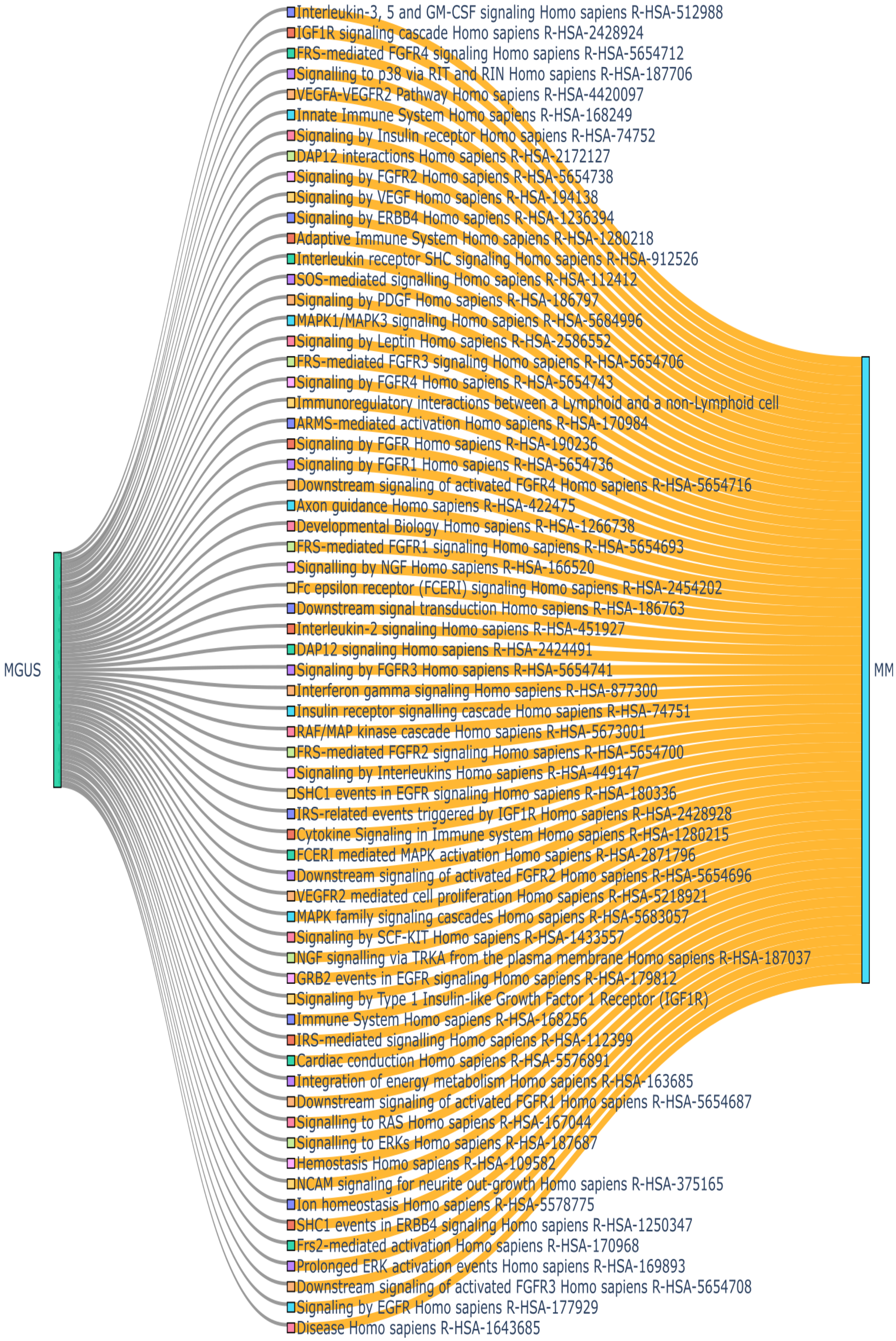
In the beeswarm plot, each sample is represented as a dot, and the color of each dot corresponds to the best ShAP score of the gene. We have also highlighted all the previously reported genes of high relevance in MM in the beeswarm plot. In our analysis, *FCGR2A, IGLL5, and KIR3DL2* are observed to be the top three genes based on their best ShAP scores in MGUS and MM samples from among the 824 significantly altered genes. Several previously reported driver genes in MM, such as *EGR1, FGFR3, HLA-A, IGLL5, IRF1, KRAS, LTB, NFKBIA, NRAS, TP53*, etc. are observed in these top-ranked genes. Similarly, the previously reported OGs such as *ABL2, CARD11, IRS1, MGAM, NOTCH1, VAV1* , etc., and TSGs such as *HLA-B, HLA-C, SDHA*, etc. are observed in the top-ranked genes in our analysis. Also, many AGs are observed among the top genes, such as *ARID1B, FGFR3, KRAS, NOTCH1, TP53* , etc.

4.3.5 Ranking of genes at the sample-level from the explainability analysis of BDL-SP using ShAP

In the sample-level analysis, we ranked genes found significantly altered in a sample according to their best ShAP scores estimated using the algorithm shown in Figure-4.4A and Algorithm A of Table-4.2. We observed that several previously reported OGs, TSGs, ODGs, and AGs were found in the top-ranked gene list of each sample. On assessing the ShAP scores of top significantly altered genes among all MM and MGUS samples, we observed that the mean $\pm$ standard deviation of the 100^{th} ranked gene's ShAP score for all MM and MGUS samples is 0.017 ± 0.0037 and 0.0171 ± 0.0040, respectively. Further, the ShAP score reduced to a considerably low value as we moved to a lower rank. Hence, we considered the top 100 significantly altered genes from all MM and MGUS samples based on their best ShAP scores for further analysis. The violin distribution plots for four gene groups of previously reported genes for all MM versus MGUS samples, only MGUS samples of EGA and AIIMS datasets, and only MM samples of MMRF and AIIMS datasets are shown in Figure-4.11A-C, respectively.

4.3.6 Analysis in MM & MGUS samples with ethnicity

We performed the statistical comparison of the disease stages (MM/MGUS) across American, European, and Indian populations (as mentioned in Section-4.2.2) on the basis


MGUS
MM
Interleukin-3, 5 and GM-CSF signaling Homo sapiens R-HSA-512988
IGF1R signaling cascade Homo sapiens R-HSA-2428924
FRS-mediated FGFR4 signaling Homo sapiens R-HSA-5654712
Signalling to p38 via RIT and RIN Homo sapiens R-HSA-187706
VEGFA-VEGFR2 Pathway Homo sapiens R-HSA-4420097
Innate Immune System Homo sapiens R-HSA-168249
Signaling by Insulin receptor Homo sapiens R-HSA-74752
DAP12 interactions Homo sapiens R-HSA-2172127
Signaling by FGFR2 Homo sapiens R-HSA-5654738
Signaling by VEGF Homo sapiens R-HSA-194138
Signaling by ERBB4 Homo sapiens R-HSA-1236394
Adaptive Immune System Homo sapiens R-HSA-1280218
Interleukin receptor SHC signaling Homo sapiens R-HSA-912526
SOS-mediated signalling Homo sapiens R-HSA-112412
Signaling by PDGF Homo sapiens R-HSA-186797
MAPK1/MAPK3 signaling Homo sapiens R-HSA-5684996
Signaling by Leptin Homo sapiens R-HSA-2586552
FRS-mediated FGFR3 signaling Homo sapiens R-HSA-5654706
Signaling by FGFR4 Homo sapiens R-HSA-5654743
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell
ARMS-mediated activation Homo sapiens R-HSA-170984
Signaling by FGFR Homo sapiens R-HSA-190236
Signaling by FGFR1 Homo sapiens R-HSA-5654736
Downstream signaling of activated FGFR4 Homo sapiens R-HSA-5654716
Axon guidance Homo sapiens R-HSA-422475
Developmental Biology Homo sapiens R-HSA-1266738
FRS-mediated FGFR1 signaling Homo sapiens R-HSA-5654693
Signalling by NGF Homo sapiens R-HSA-166520
Fc epsilon receptor (FCERI) signaling Homo sapiens R-HSA-2454202
Downstream signal transduction Homo sapiens R-HSA-186763
Interleukin-2 signaling Homo sapiens R-HSA-451927
DAP12 signaling Homo sapiens R-HSA-2424491
Signaling by FGFR3 Homo sapiens R-HSA-5654741
Interferon gamma signaling Homo sapiens R-HSA-877300
Insulin receptor signalling cascade Homo sapiens R-HSA-74751
RAF/MAP kinase cascade Homo sapiens R-HSA-5673001
FRS-mediated FGFR2 signaling Homo sapiens R-HSA-5654700
Signaling by Interleukins Homo sapiens R-HSA-449147
SHC1 events in EGFR signaling Homo sapiens R-HSA-180336
IRS-related events triggered by IGF1R Homo sapiens R-HSA-2428928
Cytokine Signaling in Immune system Homo sapiens R-HSA-1280215
FCERI mediated MAPK activation Homo sapiens R-HSA-2871796
Downstream signaling of activated FGFR2 Homo sapiens R-HSA-5654696
VEGFR2 mediated cell proliferation Homo sapiens R-HSA-5218921
MAPK family signaling cascades Homo sapiens R-HSA-5683057
Signaling by SCF-KIT Homo sapiens R-HSA-1433557
NGF signalling via TRKA from the plasma membrane Homo sapiens R-HSA-187037
GRB2 events in EGFR signaling Homo sapiens R-HSA-179812
Signaling by Type 1 Insulin-like Growth Factor 1 Receptor (IGF1R)
Immune System Homo sapiens R-HSA-168256
IRS-mediated signalling Homo sapiens R-HSA-112399
Cardiac conduction Homo sapiens R-HSA-5576891
Integration of energy metabolism Homo sapiens R-HSA-163685
Downstream signaling of activated FGFR1 Homo sapiens R-HSA-5654687
Signalling to RAS Homo sapiens R-HSA-167044
Signalling to ERKs Homo sapiens R-HSA-187687
Hemostasis Homo sapiens R-HSA-109582
NCAM signaling for neurite out-growth Homo sapiens R-HSA-375165
Ion homeostasis Homo sapiens R-HSA-5578775
SHC1 events in ERBB4 signaling Homo sapiens R-HSA-1250347
Frs2-mediated activation Homo sapiens R-HSA-170968
Prolonged ERK activation events Homo sapiens R-HSA-169893
Downstream signaling of activated FGFR3 Homo sapiens R-HSA-5654708
Signaling by EGFR Homo sapiens R-HSA-177929
Disease Homo sapiens R-HSA-1643685

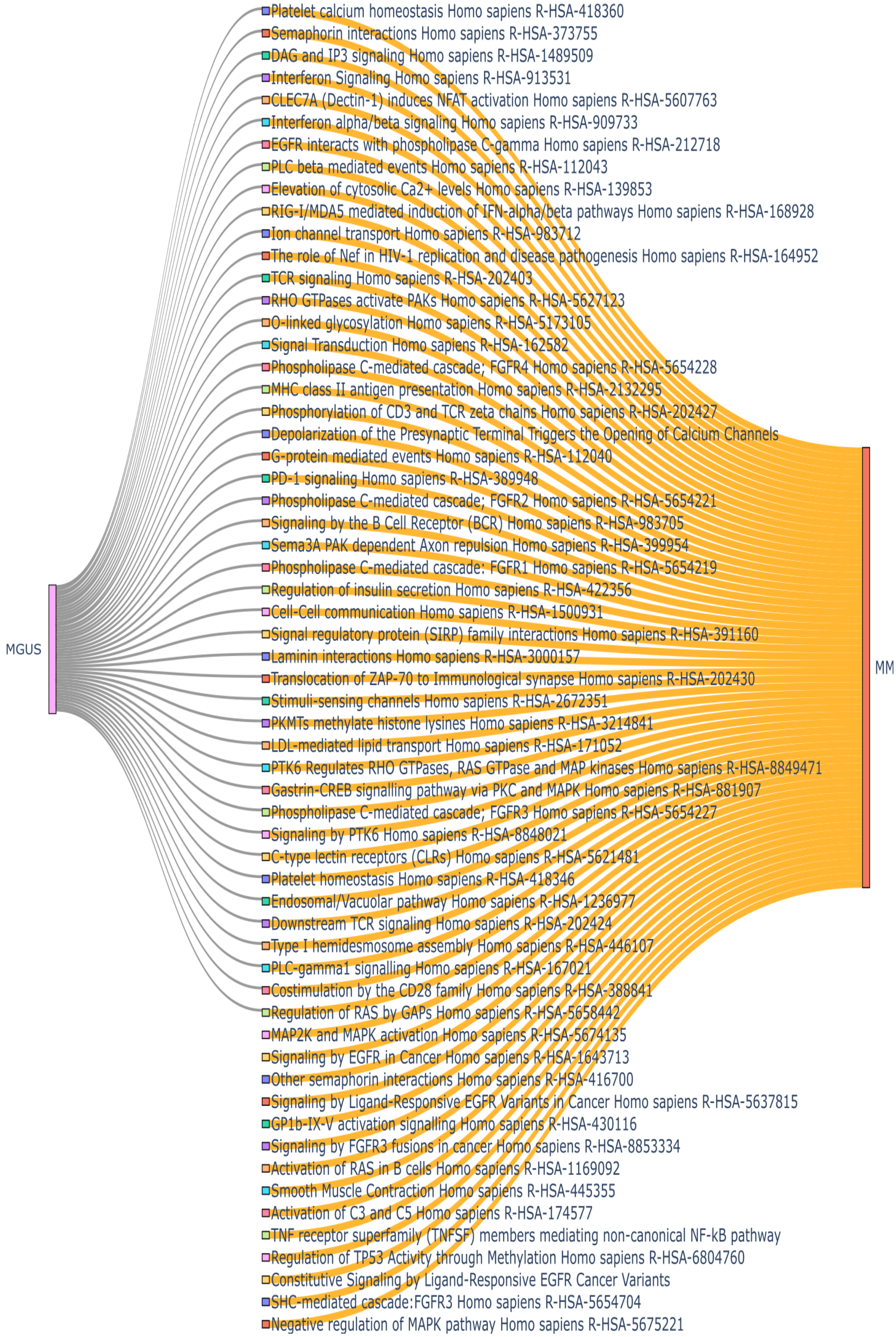

MGUS
MM
Platelet calcium homeostasis Homo sapiens R-HSA-418360
Semaphorin interactions Homo sapiens R-HSA-373755
DAG and IP3 signaling Homo sapiens R-HSA-1489509
Interferon Signaling Homo sapiens R-HSA-913531
CLEC7A (Dectin-1) induces NFAT activation Homo sapiens R-HSA-5607763
Interferon alpha/beta signaling Homo sapiens R-HSA-909733
EGFR interacts with phospholipase C-gamma Homo sapiens R-HSA-212718
PLC beta mediated events Homo sapiens R-HSA-112043
Elevation of cytosolic Ca2+ levels Homo sapiens R-HSA-139853
RIG-I/MDA5 mediated induction of IFN-alpha/beta pathways Homo sapiens R-HSA-168928
Ion channel transport Homo sapiens R-HSA-983712
The role of Nef in HIV-1 replication and disease pathogenesis Homo sapiens R-HSA-164952
TCR signaling Homo sapiens R-HSA-202403
RHO GTPases activate PAKs Homo sapiens R-HSA-5627123
O-linked glycosylation Homo sapiens R-HSA-5173105
Signal Transduction Homo sapiens R-HSA-162582
Phospholipase C-mediated cascade; FGFR4 Homo sapiens R-HSA-5654228
MHC class II antigen presentation Homo sapiens R-HSA-2132295
Phosphorylation of CD3 and TCR zeta chains Homo sapiens R-HSA-202427
Depolarization of the Presynaptic Terminal Triggers the Opening of Calcium Channels
G-protein mediated events Homo sapiens R-HSA-112040
PD-1 signaling Homo sapiens R-HSA-389948
Phospholipase C-mediated cascade; FGFR2 Homo sapiens R-HSA-5654221
Signaling by the B Cell Receptor (BCR) Homo sapiens R-HSA-983705
Sema3A PAK dependent Axon repulsion Homo sapiens R-HSA-399954
Phospholipase C-mediated cascade: FGFR1 Homo sapiens R-HSA-5654219
Regulation of insulin secretion Homo sapiens R-HSA-422356
Cell-Cell communication Homo sapiens R-HSA-1500931
Signal regulatory protein (SIRP) family interactions Homo sapiens R-HSA-391160
Laminin interactions Homo sapiens R-HSA-3000157
Translocation of ZAP-70 to Immunological synapse Homo sapiens R-HSA-202430
Stimuli-sensing channels Homo sapiens R-HSA-2672351
PKMTs methylate histone lysines Homo sapiens R-HSA-3214841
LDL-mediated lipid transport Homo sapiens R-HSA-171052
PTK6 Regulates RHO GTPases, RAS GTPase and MAP kinases Homo sapiens R-HSA-8849471
Gastrin-CREB signalling pathway via PKC and MAPK Homo sapiens R-HSA-881907
Phospholipase C-mediated cascade; FGFR3 Homo sapiens R-HSA-5654227
Signaling by PTK6 Homo sapiens R-HSA-8848021
C-type lectin receptors (CLRs) Homo sapiens R-HSA-5621481
Platelet homeostasis Homo sapiens R-HSA-418346
Endosomal/Vacuolar pathway Homo sapiens R-HSA-1236977
Downstream TCR signaling Homo sapiens R-HSA-202424
Type I hemidesmosome assembly Homo sapiens R-HSA-446107
PLC-gamma1 signalling Homo sapiens R-HSA-167021
Costimulation by the CD28 family Homo sapiens R-HSA-388841
Regulation of RAS by GAPs Homo sapiens R-HSA-5658442
MAP2K and MAPK activation Homo sapiens R-HSA-5674135
Signaling by EGFR in Cancer Homo sapiens R-HSA-1643713
Other semaphorin interactions Homo sapiens R-HSA-416700
Signaling by Ligand-Responsive EGFR Variants in Cancer Homo sapiens R-HSA-5637815
GP1b-IX-V activation signalling Homo sapiens R-HSA-430116
Signaling by FGFR3 fusions in cancer Homo sapiens R-HSA-8853334
Activation of RAS in B cells Homo sapiens R-HSA-1169092
Smooth Muscle Contraction Homo sapiens R-HSA-445355
Activation of C3 and C5 Homo sapiens R-HSA-174577
TNF receptor superfamily (TNFSF) members mediating non-canonical NF-kB pathway
Regulation of TP53 Activity through Methylation Homo sapiens R-HSA-6804760
Constitutive Signaling by Ligand-Responsive EGFR Cancer Variants
SHC-mediated cascade:FGFR3 Homo sapiens R-HSA-5654704
Negative regulation of MAPK pathway Homo sapiens R-HSA-5675221

Figure 4.8 *(previous page)*: A. Pathway enrichment analysis of the top-genes obtained from BDL-SP model: Reactome Pathways that are uniquely significant in MM. In the above figure, the orange ribbon means a significant p-adjusted value ($\leq$ 0.05), and the gray color refers to a non-significant p-adjusted value (> 0.05). A total of 134 Reactome pathways were observed to be significantly altered. Due to the large number of altered pathways, the above river plot was split into two parts to get more clarity. B. Pathway enrichment analysis of the top-genes obtained from BDL-SP model: Reactome Pathways that are uniquely significant in MM. In the above figure, the orange ribbon means a significant p-adjusted value ($\leq$ 0.05), and the gray color refers to a non-significant (p-adjusted value > 0.05). A total of 134 Reactome pathways were observed to be significantly altered. Due to the large number of altered pathways, the above river plot was split into two parts to get more clarity.

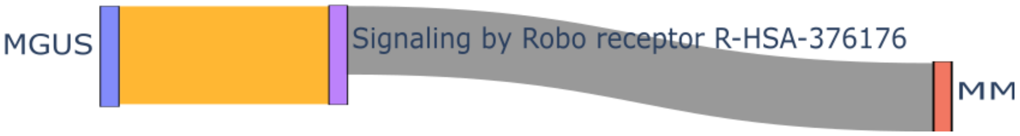
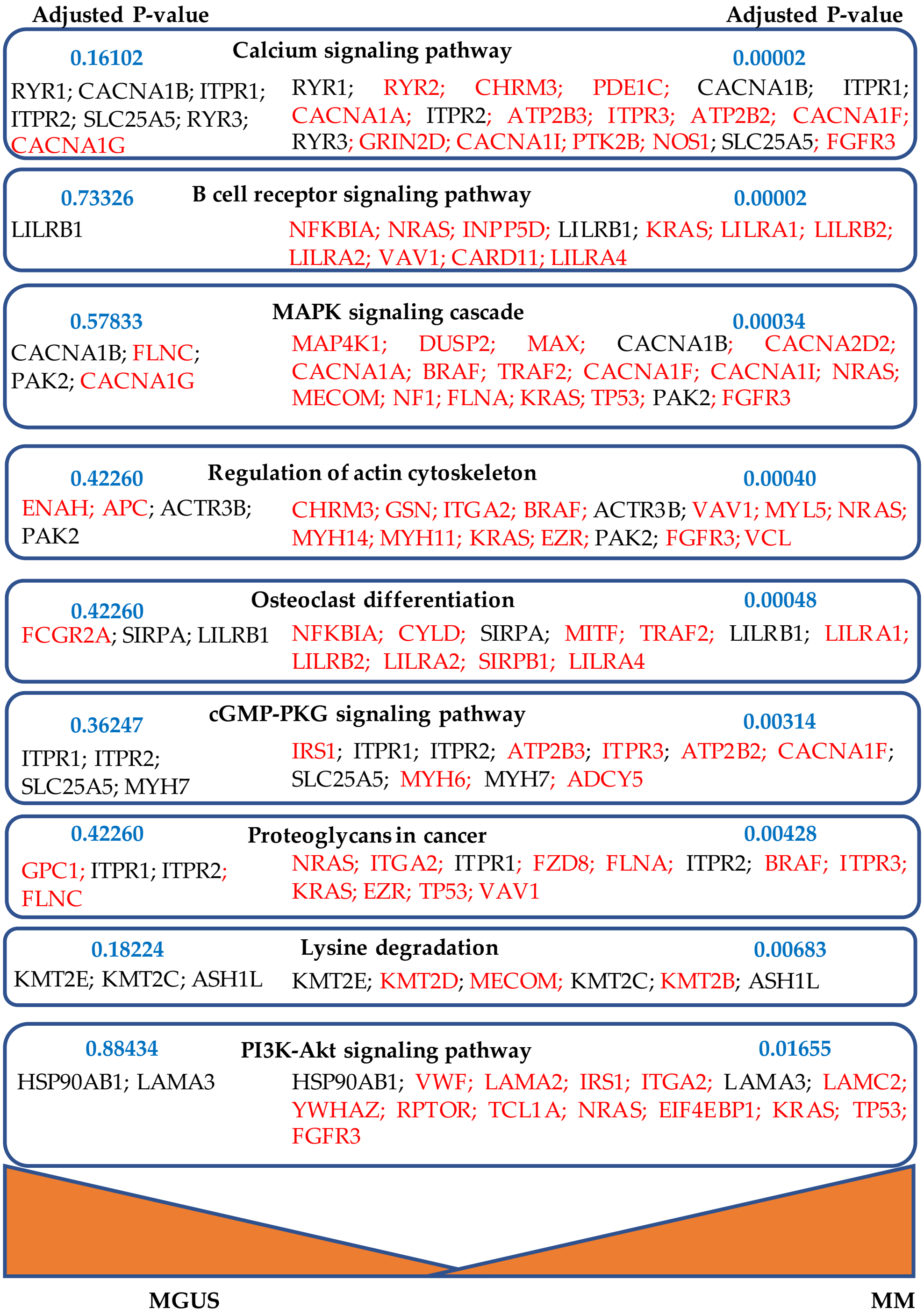


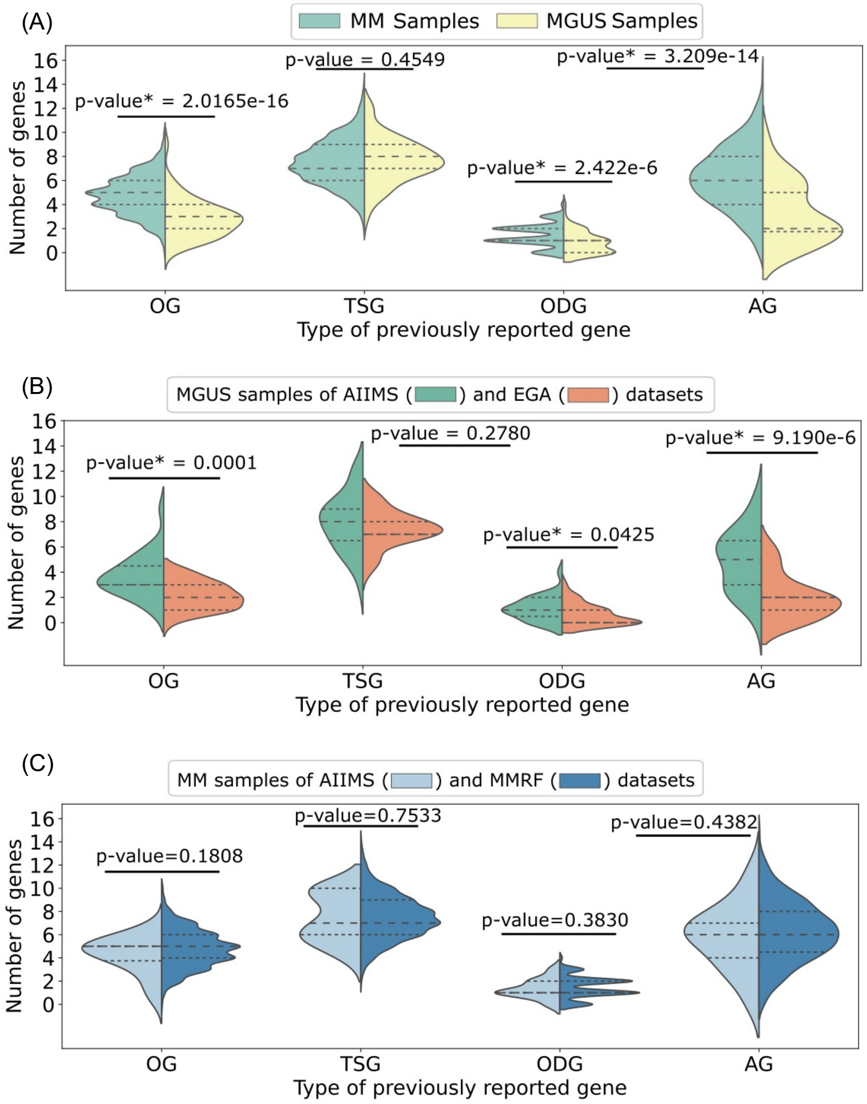
Figure 4.9: Pathway enrichment analysis of the top genes obtained from BDL-SP model: Reactome Pathways that are uniquely significant in MGUS. In the above figure, the orange ribbon means a significant p-adjusted value ($\leq$ 0.05), and the gray color refers to a non-significant (p-adjusted value > 0.05).

of the number of previously reported genes in four gene groups using unpaired Wilcoxon rank-sum test. We observed that the number of genes in OG, ODG, and AG gene groups is significantly different between the disease stages (MM and MGUS) in the analysis of combined data of different geographic populations (Figure-4.11A). Further, the medians of the number of genes in the OG, TSG, and AG gene groups were observed to be higher than the respective medians in the precursor stage (MGUS) (Figure-4.11A). Similarly, comparing the number of genes in all four gene groups between the MGUS samples of the Indian (AIIMS dataset) and European (EGA dataset) populations, the number of genes in OG, ODG, and AG gene groups were observed to be significantly different (Figure-4.11B). On the contrary, the number of genes in all four gene groups was not found to be statistically and significantly different in MM samples of the Indian and American populations (MMRF dataset) (Figure-4.11C). These observations indicate that ethnicity might play a significant role in disease development. Thus, ethnicity-specific analysis can be helpful in further gaining in-depth insights into the disease biology of the premalignant stage of MM (MGUS).



Adjusted P-value
Adjusted P-value
0.16102
Calcium signaling pathway
0.00002
RYR1; CACNA1B; ITPR1; ITPR2; SLC25A5; RYR3; CACNA1G
RYR1; RYR2; CHRM3; PDE1C; CACNA1B; ITPR1; CACNA1A; ITPR2; ATP2B3; ITPR3; ATP2B2; CACNA1F; RYR3; GRIN2D; CACNA1I; PTK2B; NOS1; SLC25A5; FGFR3
0.73326
B cell receptor signaling pathway
0.00002
LILRB1
NFKBIA; NRAS; INPP5D; LILRB1; KRAS; LILRA1; LILRB2; LILRA2; VAV1; CARD11; LILRA4
0.57833
MAPK signaling cascade
0.00034
CACNA1B; FLNC; PAK2; CACNA1G
MAP4K1; DUSP2; MAX; CACNA1B; CACNA2D2; CACNA1A; BRAF; TRAF2; CACNA1F; CACNA1I; NRAS; MECOM; NF1; FLNA; KRAS; TP53; PAK2; FGFR3
0.42260
Regulation of actin cytoskeleton
0.00040
ENAH; APC; ACTR3B; PAK2
CHRM3; GSN; ITGA2; BRAF; ACTR3B; VAV1; MYL5; NRAS; MYH14; MYH11; KRAS; EZR; PAK2; FGFR3; VCL
0.42260
Osteoclast differentiation
0.00048
FCGR2A; SIRPA; LILRB1
NFKBIA; CYLD; SIRPA; MITF; TRAF2; LILRB1; LILRA1; LILRB2; LILRA2; SIRPB1; LILRA4
0.36247
cGMP-PKG signaling pathway
0.00314
ITPR1; ITPR2; SLC25A5; MYH7
IRS1; ITPR1; ITPR2; ATP2B3; ITPR3; ATP2B2; CACNA1F; SLC25A5; MYH6; MYH7; ADCY5
0.42260
Proteoglycans in cancer
0.00428
GPC1; ITPR1; ITPR2; FLNC
NRAS; ITGA2; ITPR1; FZD8; FLNA; ITPR2; BRAF; ITPR3; KRAS; EZR; TP53; VAV1
0.18224
Lysine degradation
0.00683
KMT2E; KMT2C; ASH1L
KMT2E; KMT2D; MECOM; KMT2C; KMT2B; ASH1L
0.88434
PI3K-Akt signaling pathway
0.01655
HSP90AB1; LAMA3
HSP90AB1; VWF; LAMA2; IRS1; ITGA2; LAMA3; LAMC2; YWHAZ; RPTOR; TCL1A; NRAS; EIF4EBP1; KRAS; TP53; FGFR3
MGUS
MM

Figure 4.10: KEGG pathways were found to be significantly involved in the progression of MGUS to MM. Top genes that were identified by post-hoc analysis of BDL-SP using the ShAP algorithm as significantly mutated either in MGUS only or in MM only (acting as differentiators of MGUS and MM) are shown in red colored font.



Disease	Type of previously reported gene	Median of number of genes
MM/MGUS	OG	5/3
MM/MGUS	TSG	7/8
MM/MGUS	ODG	1/1
MM/MGUS	AG	6/2

Dataset	Type of previously reported gene	Median of number of genes
AIIMS/EGA	OG	3/2
AIIMS/EGA	TSG	8/7
AIIMS/EGA	ODG	1/0
AIIMS/EGA	AG	5/2

Dataset	Type of previously reported gene	Median of number of genes
AIIMS/MMRF	OG	5/5
AIIMS/MMRF	TSG	7/7
AIIMS/MMRF	ODG	1/1
AIIMS/MMRF	AG	6/6

Figure 4.11: The distribution of the number of previously reported genes in four gene groups (OGs, TSGs, ODGs, and AGs) was found significantly altered and ranked in the top 100 across all MM and MGUS samples (combined dataset of MMRF, EGA, and AIIMS samples). B. The distribution of the number of previously reported genes found significantly altered and ranked in the top 100 across all MGUS samples in EGA and AIIMS datasets. C. The distribution of the number of previously reported genes found significantly altered and ranked in top 100 across all MM samples in MMRF and AIIMS datasets. The P-value shown with each violin plot was estimated using an unpaired Wilcoxon rank-sum statistical test to check whether the number of genes in a particular gene group significantly differs from their respective counts in the other group. The gene group having a P-value with a superscript "*" (star) symbol represents that the number of genes in that gene group is significantly different compared to the other group. The table on the right of each figure shows the median of the number of genes in each gene group for disease stages (MM/MGUS) and datasets (MMRF, EGA, and AIIMS). Note: To have a better view of the violin plots, refer to the colored version of this figure.

4.3.7 Genomic feature ranking at a sample-level using ShAP analysis

Besides identifying the top significantly altered genes in MM and MGUS, we also ranked the genomic features based on their contribution to the model prediction. A set of 28 genomic features (Figure-4.1) was used to train the BDL-SP model. These genes were ranked on the basis of their ShAP scores. The algorithm for estimating the best ShAP score for each genomic feature is shown in Figure-4.4B. We observed that the total number of SNVs, the total number of SNVs in the Other SNV group (as shown in Figure-4.1), and VAF's standard deviation of SNVs in the Other SNV group were the top three genomic features, while VAF's standard deviation of SNVs in the nonsynonymous SNV group, VAF's standard deviation of SNVs in the Synonymous SNV group, and AD's standard deviation of SNVs in the nonsynonymous SNV group were the least contributing genomic features. The beeswarm plot for genomic feature ranking from BDL-SP model post-hoc analysis using ShAP is shown in Figure-4.11.

4.3.8 Significance of nonsynonymous SNVs in MM pathogenesis

To understand the role of nonsynonymous SNVs on MM pathogenesis, we meticulously analyzed the nonsynonymous SNVs identified using four variant callers. For this, we identified the genes having the most pathogenic nonsynonymous SNVs. Further, we analyzed the impact of nonsynonymous SNVs on protein function and alterations in the protein structure due to protein structure.

Distribution of nonsynonymous SNVs across critical genes and pathways

First, we identified the most pathogenic and deleterious SNVs. For this, first, filtered out the pathogenic nonsynonymous SNVs jointly identified by four variant callers. This yielded a total of 47,686 nonsynonymous SNVs identified collectively by all four variant callers. Next, we considered only those nonsynonymous SNVs classified as deleterious or damaging by all four deleteriousness scores, resulting in a subset of 9518 SNVs associated with 5508 genes. On comparing with 824 significantly altered genes, out of 6046 genes, we identified 903 nonsynonymous and deleterious SNVs associated with 325 significantly altered genes. Additionally, 760 nonsynonymous and deleterious SNVs were associated with 244 genes ranked within the top 500 ranked genes identified through the BDL-SP model post-hoc analysis (Table-4.6)

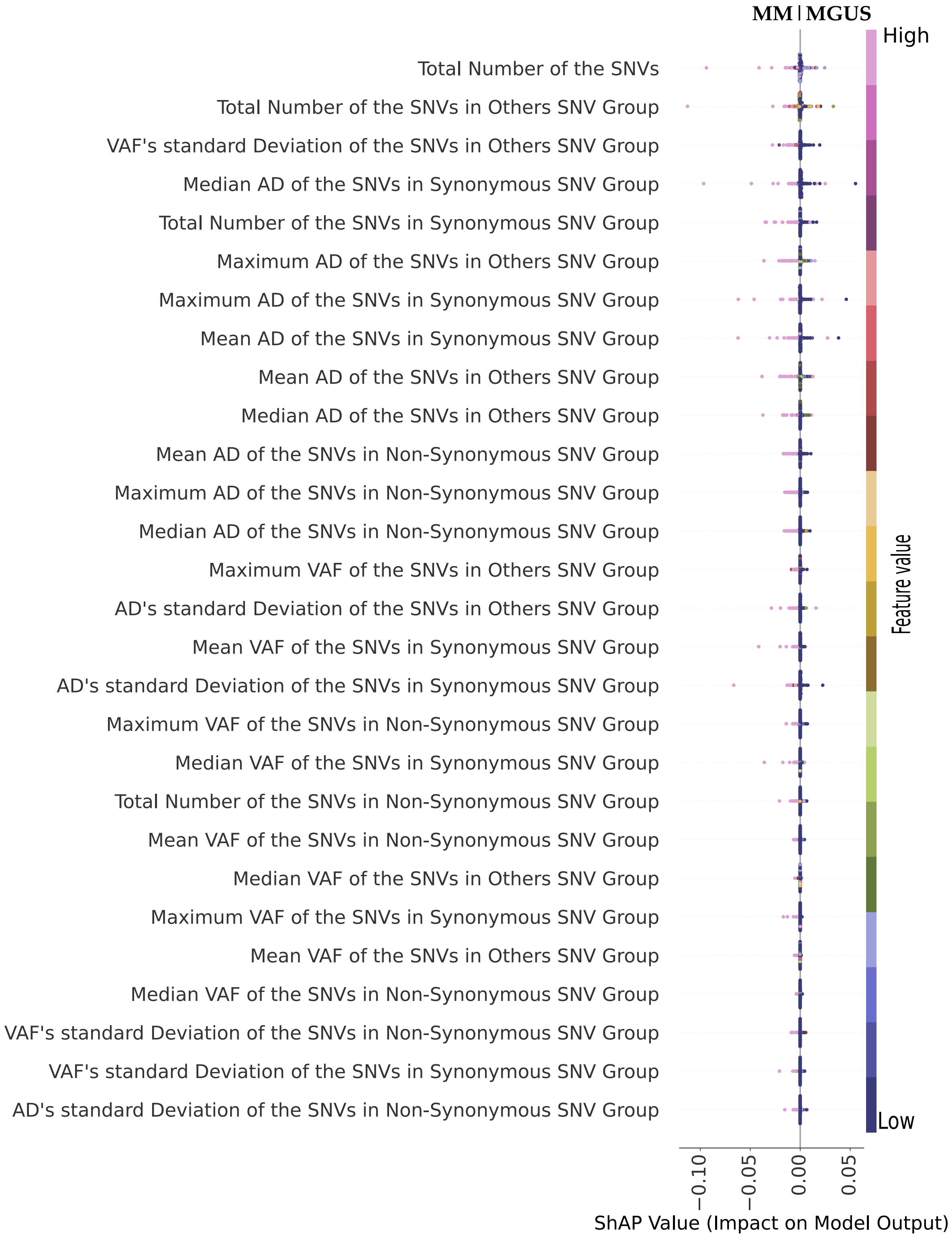

MM | MGUS
High
Low
Feature value
Total Number of the SNVs
Total Number of the SNVs in Others SNV Group
VAF's standard Deviation of the SNVs in Others SNV Group
Median AD of the SNVs in Synonymous SNV Group
Total Number of the SNVs in Synonymous SNV Group
Maximum AD of the SNVs in Others SNV Group
Maximum AD of the SNVs in Synonymous SNV Group
Mean AD of the SNVs in Synonymous SNV Group
Mean AD of the SNVs in Others SNV Group
Median AD of the SNVs in Others SNV Group
Mean AD of the SNVs in Non-Synonymous SNV Group
Maximum AD of the SNVs in Non-Synonymous SNV Group
Median AD of the SNVs in Non-Synonymous SNV Group
Maximum VAF of the SNVs in Others SNV Group
AD's standard Deviation of the SNVs in Others SNV Group
Mean VAF of the SNVs in Synonymous SNV Group
AD's standard Deviation of the SNVs in Synonymous SNV Group
Maximum VAF of the SNVs in Non-Synonymous SNV Group
Median VAF of the SNVs in Synonymous SNV Group
Total Number of the SNVs in Non-Synonymous SNV Group
Mean VAF of the SNVs in Non-Synonymous SNV Group
Median VAF of the SNVs in Others SNV Group
Maximum VAF of the SNVs in Synonymous SNV Group
Mean VAF of the SNVs in Others SNV Group
Median VAF of the SNVs in Non-Synonymous SNV Group
VAF's standard Deviation of the SNVs in Non-Synonymous SNV Group
VAF's standard Deviation of the SNVs in Synonymous SNV Group
AD's standard Deviation of the SNVs in Non-Synonymous SNV Group
−0.10
−0.05
0.00
0.05
ShAP Value (Impact on Model Output)

Figure 4.11 *(previous page)*: Genomic feature ranking based on the BDL-SP model's post-hoc explainability in MM and MGUS using the ShAP algorithm. Each genomic feature is ranked according to its best ShAP score estimated using the algorithm shown in Figure-4.4 and Table-4.2. The negative ShAP score represents the contribution of the genomic feature towards MM, while the positive ShAP score represents the contribution of the genomic feature towards MGUS. Further, each dot in the individual scatter plot of the genomic feature represents a sample, and the color of the dot represents the value of that genomic feature with the color codes as follows: the dark blue color represents low value, and the pink color represents the high value of the genomic feature. Note: Refer to the colored version of this figure for a clear view of the sample distribution for each genomic feature.

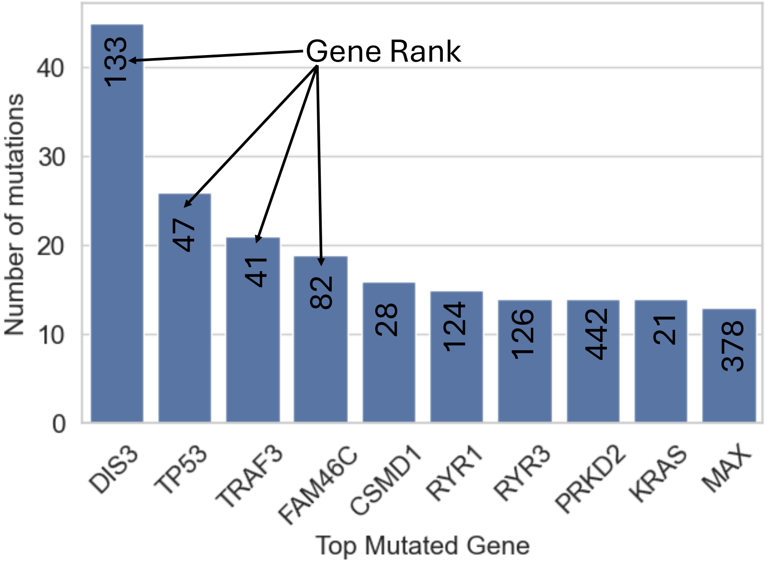


Figure 4.12: Frequency of nonsynonymous and deleterious mutations among the top 10 genes having the largest number of such mutations, as identified through the aforementioned strategy. The number highlighted in back color in each bar is the rank obtained through BDL-SP post-hoc analysis.

Intriguingly, several key MM driver genes, such as DIS3, KRAS, TRAF2, TRAF3, etc., emerged as the top genes harboring the most nonsynonymous and deleterious mutations, highlighting their relevance to multiple myeloma progression (Figure-4.12). On further analysis of the genes not included in the 232-gene set but ranked within the top 500, we observed their significance as pivotal biomarkers in distinguishing between MM and MGUS. Among these genes, there were several MM-relevant genes such as ARID2, BRAF, HLA-A, PRSS3, etc. These genes exhibited nonsynonymous and deleterious SNVs, identified by three or fewer variant callers and characterised as deleterious by three or fewer scoring methods.

Table 4.6: List of genes having both nonsynonymous jointly reported by all four variant callers (MuSE, Mutect2, SomaticSniper, and Varscan2) and deleterious mutations jointly declared by four deleteriousness scores (SIFT, PolyPhen2-HDIV, PolyPhen2-HVAR, and PROVEAN) and ranked in top 500 in the post-hoc analysis of BDL-SP model.

Range of number of nonsynonymous and deleterious mutations	Number of genes having nonsynonymous and deleterious SNVs in the given range	Number of genes in ranked in top 500 in BDL-SP post-hoc analysis	Name of genes in ranked in top 500 in BDL-SP post-hoc analysis	Names of missing genes
>40	1	1	DIS3	-
20 – 40	2	2	TP53, TRAF3	-
10 – 20	20	11	FAM46C, CSMD1, RYR1, KRAS, PRKD2, RYR3, MAX, DNAH9, DNAH5, FGFR3, TTN	FAT4, CSMD3, TTN-AS1, PCLO, ATM, FAT1, KLHL6, LRP1B, PCDHA4
7 – 10	28	7	RYR2, HUWE1, OBSCN, CACNA1B, CYLD, DNAH2, NF1	PRDM1, LRP2, MPDZ, CACNA1E, DDX3X, EHD1, ROBO1, PXDN, FBN2, SHANK1, ATP7B, KALRN, RP11-799N11.1, NOTCH3, COL4A1, NCKAP5, MAF, HIST1H2BD, MYH2, DST, SLIT2
1 – 7	5457	211	*	*

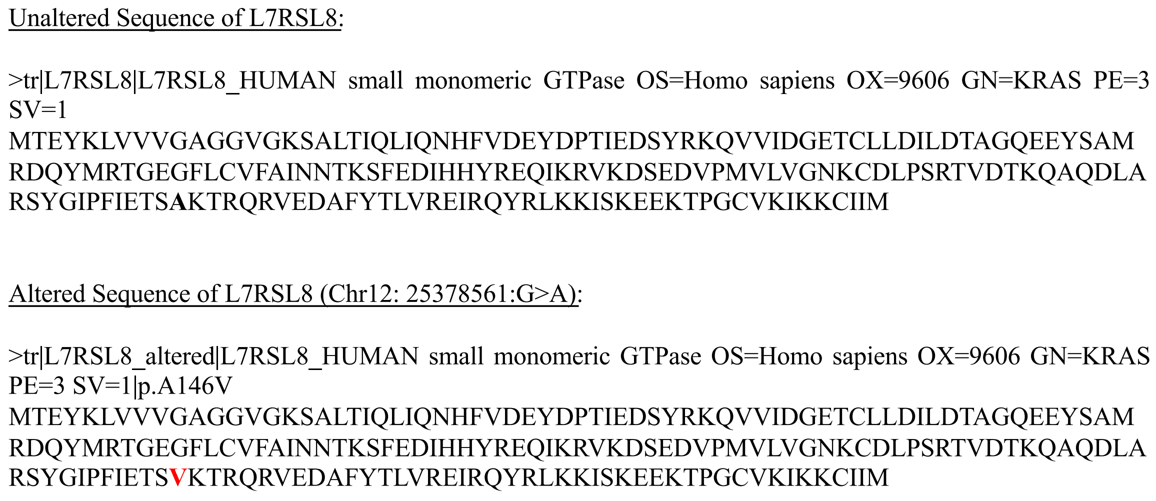
“*” indicates that the gene list is omitted due to its length.

The other top mutated genes included *CSMD1, FAM46C, KRAS, MAX, PRKD2, RYR1*, and *RYR3*. Notably, the pathways associated with these genes are implicated in various key cellular processes contributing to MM progressions, such as Apoptosis, Gap Junction, Pathway in Cancer, and the MAPK signaling pathway. To gain deeper insights into the impact of alterations on the corresponding protein function and structure, we focused on the SNVs associated with the two most reported MM driver genes, named KRAS and TP53, shown in the subsequent section.

Alterations in protein structure caused by nonsynonymous SNVs

1. Altrations in KRAS Genes Protein Structure:

Using the strategy shown in 4.2.2, we identified 14 nonsynonymous and deleterious SNVs in KRAS genes. These SNVs were consistently associated with the PF00071, PF08477, PF00025, and PF00009 Pfam domains. Further, we meticulously analyzed the structural change in the protein structure of the KRAS gene protein (LZRSL8) using SWISS-MODEL. To understand the impact of nonsynonymous SNV on protein structure alterations in KRAS protein (L7RSL8), first, we identified the protein associated with the KRAS gene mutation with the help of ANNOVAR, which was L7RSL8 protein. Next, we fetched the sequence of L7RSL8 protein from NCBI and compared the change in protein structure in both unaltered and altered states. The fasta sequence of unaltered and altered L7RSL8 protein are as follows (with highlighted nucleotide, black bold represents the actual nucleotide, and red bold represents the altered nucleotide) is shown in Figure-4.13.



Unaltered Sequence of L7RSL8:

>tr|L7RSL8|L7RSL8_HUMAN small monomeric GTPase OS=Homo sapiens OX=9606 GN=KRAS PE=3 SV=1
MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAM
RDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGNKCDLPSRTVDTKQAQDLA
RSYGIPFIETS**A**KTRQRVEDAFYTLVREIRQYRLKKISKEEKTPGCVKIKKCIIM

Altered Sequence of L7RSL8 (Chr12: 25378561:G>A):

>tr|L7RSL8_altered|L7RSL8_HUMAN small monomeric GTPase OS=Homo sapiens OX=9606 GN=KRAS PE=3 SV=1|p.A146V
MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAM
RDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGNKCDLPSRTVDTKQAQDLA
RSYGIPFIETS**V**KTRQRVEDAFYTLVREIRQYRLKKISKEEKTPGCVKIKKCIIM

Figure 4.13: Altered and unaltered fasta sequence of KRAS protein structure (L7RSL8). Here, the altered nucleotide is shown in **black bold** in the unaltered fasta sequence (top) and **red bold** in the altered fasta sequence (bottom)

The comparison of the change in the protein structure for the KRAS gene protein in both altered and unaltered states is shown below (Figure-4.14).

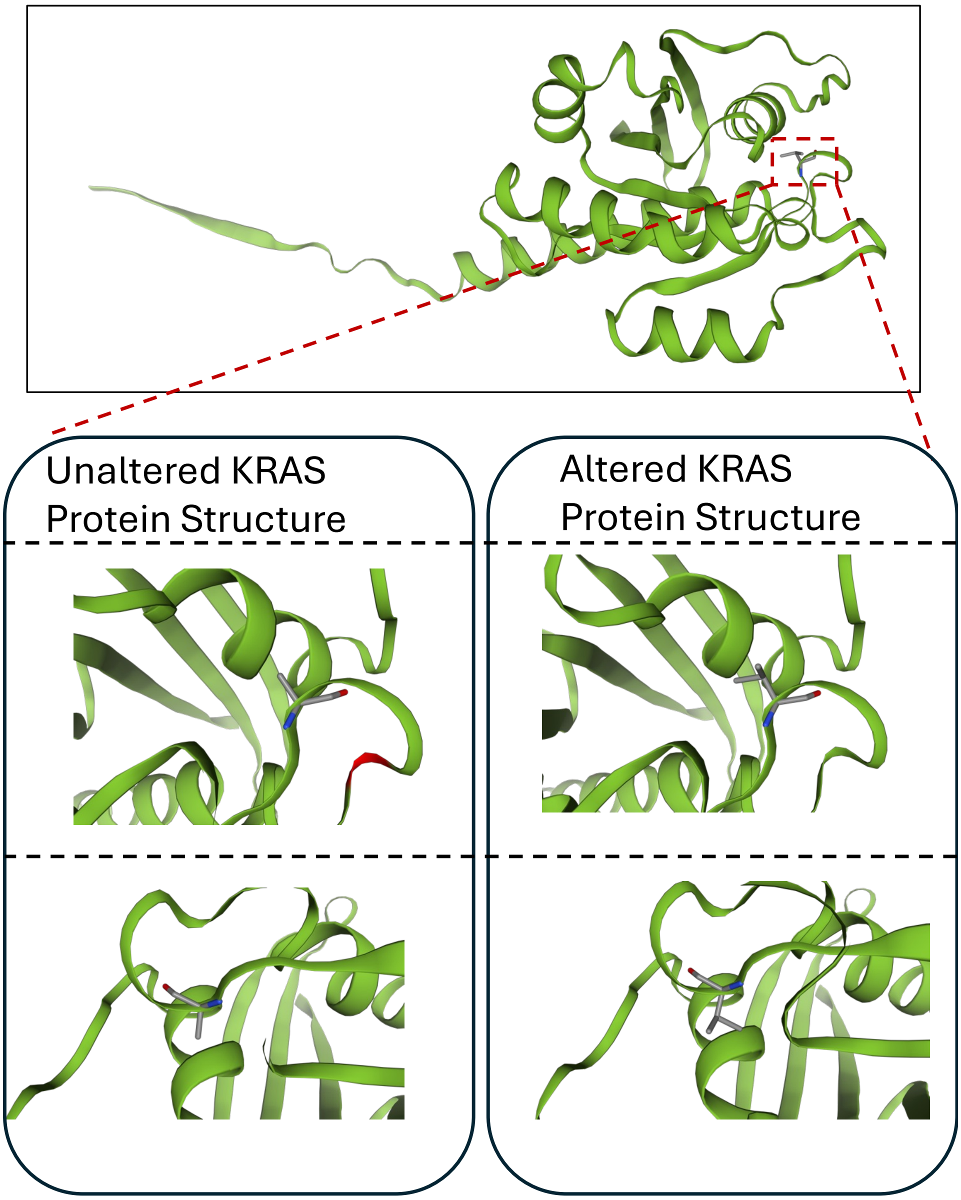
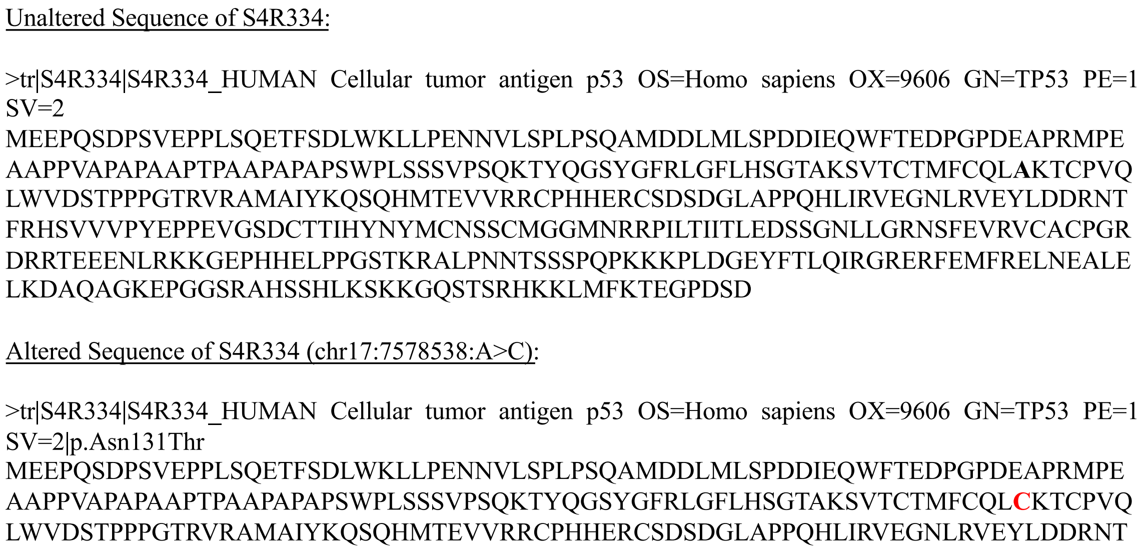


Figure 4.14: Comparison of unaltered and altered KRAS protein (L7RSL8) structure: Due to nonsynonymous mutations, there is an alteration in the protein sequence at location 146 where the nucleotide A is changed to V (p.A146V). This alteration causes a change in the binding sites at location 146.

2. Alterations in TP53 Genes Protein Structure:

We identified a total of 27 nonsynonymous and deleterious SNVs in the TP53 gene, as jointly reported by all variant callers and declared deleterious by four different deleteriousness scoring methods. To understand the impact of nonsynonymous SNVs on TP53 protein (S4R334) structure, first, we identified the associated protein, which is affected by the TP53 gene nonsynonymous SNV, with the help of ANNOVAR, which was the S4R334 protein. Next, we fetched the sequence of S4R334 protein from NCBI and compared the change in protein structure in both unaltered and altered states. The fasta sequence of unaltered and altered S4R334 protein are as follows (with highlighted nucleotide, black bold represents the actual nucleotide, and red bold represents the altered nucleotide) is shown in Figure-4.15.



Unaltered Sequence of S4R334:

>tr|S4R334|S4R334_HUMAN Cellular tumor antigen p53 OS=Homo sapiens OX=9606 GN=TP53 PE=1 SV=2
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPDEAPRMPE
AAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTMFCQL**A**KTCPVQ
LWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQHLIRVEGNLRVEYLDDRNT
FRHSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGR
DRRTEEENLRKKGEPHHELPPGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALE
LKDAQAGKEPGGSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD

Altered Sequence of S4R334 (chr17:7578538:A>C):

>tr|S4R334|S4R334_HUMAN Cellular tumor antigen p53 OS=Homo sapiens OX=9606 GN=TP53 PE=1 SV=2|p.Asn131Thr
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPDEAPRMPE
AAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTMFCQL**C**KTCPVQ
LWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQHLIRVEGNLRVEYLDDRNT

Figure 4.15: Altered and unaltered fasta sequence of TP53 protein structure (S4R334). Here, the altered nucleotide is shown in **black bold** in the unaltered fasta sequence (top) and **red bold** in the altered fasta sequence (bottom)

The protein structure of the TP53 protein (S4R334) obtained with the help of SWISS-MODEL is shown below (Figure-4.16).

It is evident from the preceding examples that nonsynonymous and deleterious SNVs have the potential to affect protein function and alter its structure profoundly. Similar analyses can be extended to other MM-relevant genes to gain deeper insights into the impact of such variants on protein function and structure. This broader examination holds promise for identifying effective therapeutic targets for targeted drug therapy. However, it's essential to note that understanding protein structure stability in altered states involves considering various criteria, including differences in free energy between folded and unfolded states, the impact of hydrogen bonding, and several thermodynamic stability parameters [47,48]. While these factors are crucial, delving into their complexities lies beyond the scope of this thesis. A researcher could explore these aspects further, leveraging intensive genomic analysis to pinpoint genes of interest and pave the way for enhanced therapeutic interventions.

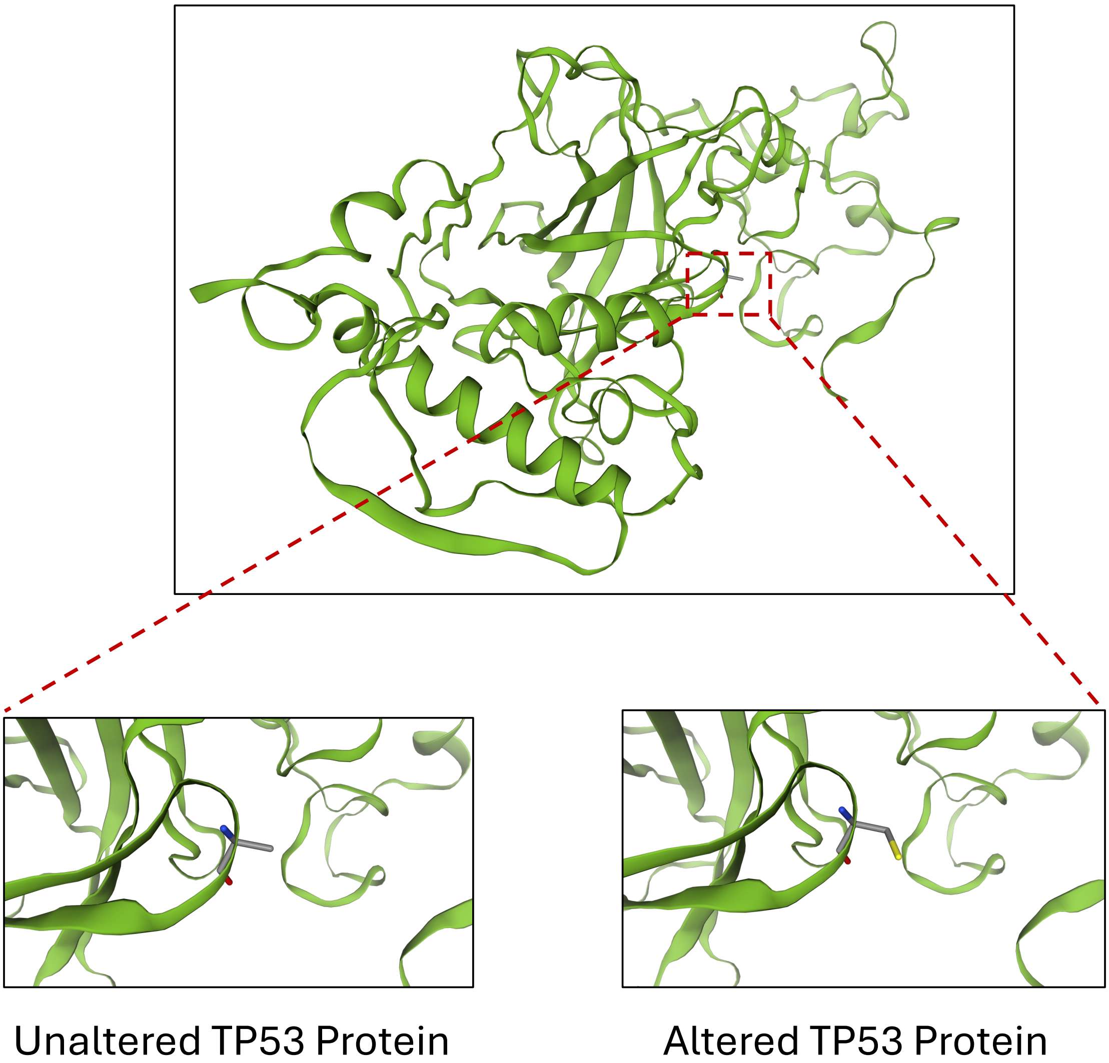


Figure 4.16: TP53 gene protein (S4R334) Structure (Left) before any nonsynonymous mutation, and (Right) after nonsynonymous mutation at chr17:7578538 : A>C.

4.4 Discussions

It is well established that MM evolves through premalignant stages driven by the acquisition of multiple genomic aberrations [229]. Though a few studies have analyzed the progression from MGUS to MM [90, 93], a limited amount of information is available on the notable biomarkers responsible for this transformation. However, if known apriori, appropriate treatment at the MGUS stage can help control the progression of MGUS to MM, thereby preventing the complications associated with MM, reducing morbidity, and increasing the overall survival of these patients. Thus, it is crucial to unravel the genomic features responsible for the malignant transformation of MGUS to MM.

In this work, we addressed the challenge of extracting relevant MM and MGUS differentiating genomic attributes from the large pool of mutational information available for each patient. Our proposed BDL-SP-based workflow has been successful in accomplishing this task. In the preprocessing of the data, we identified significantly mutated genes for each variant caller and then took their union so that we did not miss any important gene. Thus, large cohort size and an ensemble of four variant callers enabled us to obtain generalizable mutational information, driver genes, and altered pathway information. Recently, graph-based learning has been extensively explored in genomics for deciphering crucial information such as disease progression and identifying novel biomarkers for targeted drug therapy, etc. For instance, the graph-based model was used to learn the temporal graphs of diagnosis (Dx), procedure (Px) and prescription (Rx) of MM patients from the sequential electronic health records (EHR) and predict a patient's response to treatment [230]. Till now, graph-based learning approaches have not yet been explored to identify the underlying difference between MM and its precursor stage (MGUS). In our BDL-SP model, we have used the connectionist model of graph-based learning to learn genomic mutational profiles (as node features) that were extracted from the WES datasets of AIIMS, EGA, and MMRF. We additionally utilized the gene-gene interaction information from the PPI network to identify the pivotal biomarkers that can differentiate MM from MGUS.

Our proposed AI-based BDL-SP workflow is innovative in multiple ways, as explained below:

1. The identification of pivotal biomarkers using WES datasets of MM of three populations (American, European, and Indian) increases the robustness of the workflow by enhancing its ability to assess the granular-level insights of mutational profiles from multiple datasets of different geographic locations/ethnicities.
2. Because of the pathogenic nature of deleterious SNVs, only deleterious SNVs were considered for identifying the significantly altered genes in the proposed workflow. We observed that the total number of SNVs was reduced considerably after variant filtration of benign SNVs using the FATHMM-XF method.
3. An analysis of the genomic mutational profile and the gene-gene interaction information enables this workflow to look at interdependencies between genes,

making it a complete bio-inspired workflow.

4. The proposed workflow includes quantitative (using performance metrics) as well as an exhaustive qualitative (post-hoc interpretability analysis of the trained models) benchmarking. It also shows that multiple ML models behaving closely on the quantitative metrics may differ hugely in the qualitative analysis. Thus, application-aware interpretability analysis, as carried out in this workflow (ShAP on genes and genomic features), can help choose the right model and increase the confidence of doctors in the trained AI model.

4.4.1 Comparison of variant calling across four variant callers

We compared the number of pathogenic SNVs identified by each variant caller for the MM cohort (Figure-4.17). Also, the number of SNVs identified by each variant caller, the number of common SNVs reported by at least two variant callers, at least three variant callers and common to all variant callers are shown in Table-4.7. It is evident that the number of pathogenic SNVs identified in consensus by multiple variant callers decreases noticeably with an increase in the number of variant callers employed. This observation underscores the methodological disparities among these variant callers.

Table 4.7: Number of SNVs identified by four variant callers either individually or jointly in MM and MGUS cohort.

Categories of Pathogenic SNVs identified by four variant callers (either jointly or individually)	Number of pathogenic SNVs in MM cohort
Total number of pathogenic SNVs reported by all four variant callers	7100816
Number of pathogenic SNVs identified by four variant callers individually	Muse: 407813 Mutect2: 3220606 SomaticSniper: 4815051 Varscan2:440833
Number of pathogenic SNVs reported by at least two variant callers	1278846
Number of pathogenic SNVs reported by at least three variant callers	294603
Number of pathogenic SNVs reported by all four variant callers	210038

To evaluate the presence of inherent bias in any variant caller, we initially assess the number of pathogenic mutations identified by all four variant callers across each chromosome in the MM cohort (Figure-4.18). We observed that, in the MM cohort, SomaticSniper identified more pathogenic SNVs than any other variant caller. Additionally, chromosomes 1, 2, and 6 were the most altered chromosomes in the MM cohort. Subsequently, to delve deeper into the nature of SNVs identified by the variant callers, we conducted a comparative analysis based on the types of SNVs identified by each caller.

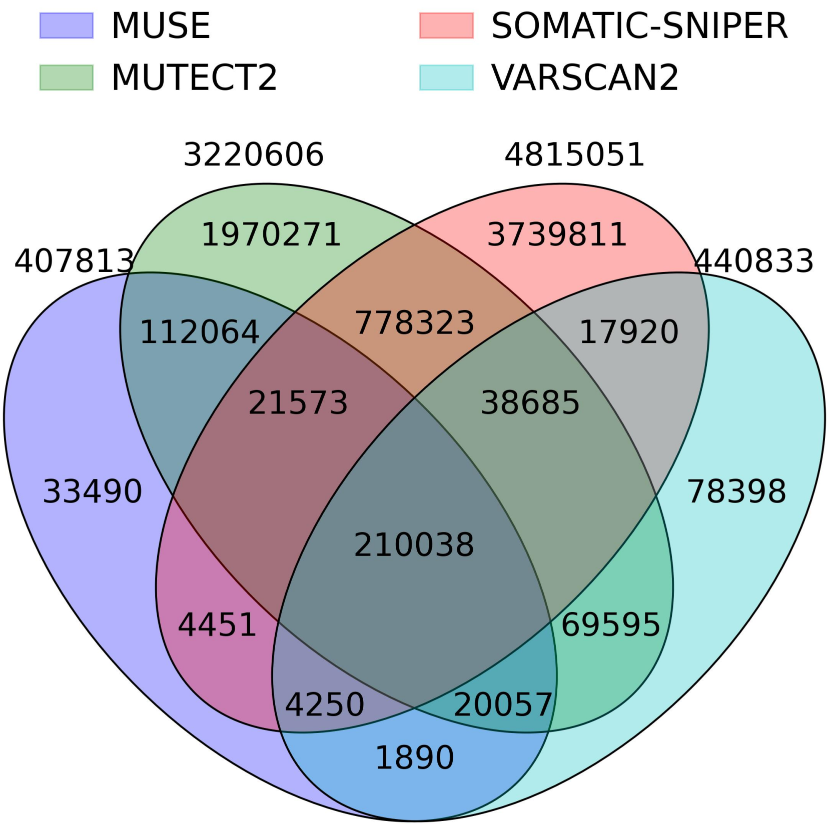


Figure 4.17: Venn diagram highlighting the pathogenic SNVs identified by four variant callers (MuSE, Mutect2, SomaticSniper and Varacsn2) in MM cohort.

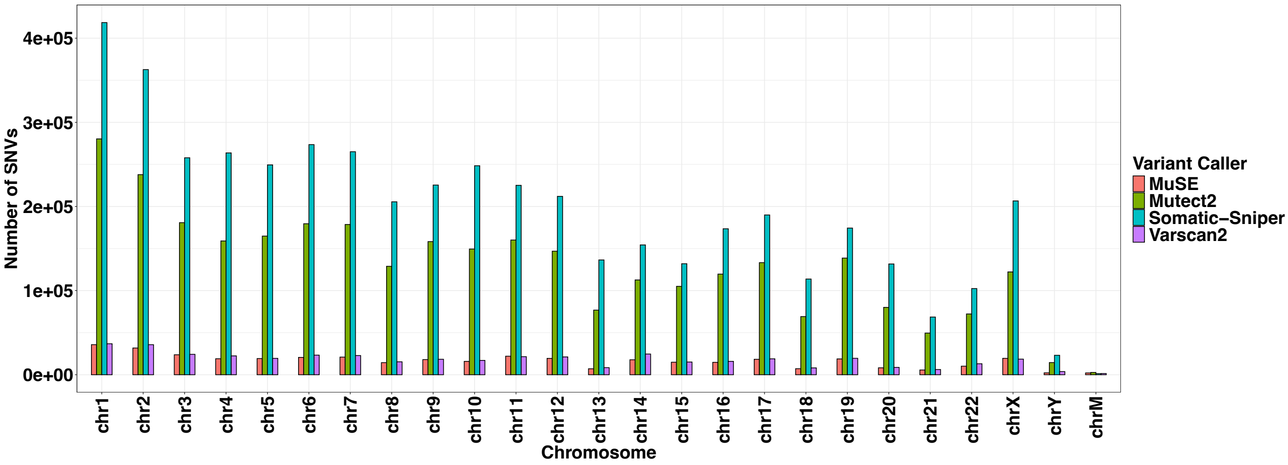


Figure 4.18: Distribution of the total number of pathogenic SNVs reported by four variant callers (MuSE, Mutect2, SomaticSniper, and Varscan2) at each chromosome in the MM cohort.

Table 4.8: Comparison of different types of SNVs identified by four variant callers (MuSE, Mutect2, SomaticSniper, and Varscan2) in the MM cohort.

SNV Category	SNV Type	Number of pathogenic SNVs of this type identified by MuSE variant caller in MM cohort	Number of pathogenic SNVs of this type identified by Mutect2 variant caller in MM cohort	Number of pathogenic SNVs of this type identified by SomaticSniper variant caller in MM cohort	Number of pathogenic SNVs of this type identified by Varscan2 variant caller in MM cohort
Nonsynonymous SNV group	Nonsynonymous SNVs	77603	225041	92916	67099
	ncRNA-exonic	8772	35621	32243	9573
	Stop gain	4704	15842	4688	3771
	Stop loss	125	403	168	127
	Start loss	142	381	162	121
	Exonic-splicing	42	193	92	63
	Splicing	1963	7826	2408	1836
	Frameshift insertion	0	5516	0	897
	Frameshift deletion	0	7315	0	1683
Synonymous SNV group	Synonymous SNVs	4308	12640	6071	3793
	UTR3	501210	192588	98954	64281
	UTR5	12353	47510	34215	11271
Others SNV group	Non-frameshift insertion	0	2527	0	208
	Non-frameshift deletion	0	6611	0	1286
	Non-frameshift substitution	0	1591	0	0
	Intronic	132121	826927	750167	142281
	Intergenic	85267	1610407	3486371	100025
	ncRNA-intronic	11741	104032	134386	14899
	Upstream	10587	64130	91062	9998
	Downstream	6778	49604	79060	6397
	ncRNA-Splicing	57	270	259	72

Table-4.8 compares four variant callers (MuSE, Mutect2, SomaticSniper, and Varscan2) based on the three SNV categories in the MM cohort. The SNV categories include nonsynonymous SNVs, synonymous SNVs, and other SNVs. Each category is further divided into specific SNV types based on their functional impact, as mentioned in Data pre-processing (Section 4.2.2). For each SNV type, we analyzed the number of pathogenic SNVs identified by each variant caller in the MM cohort.

Mutect2 identified the largest number of pathogenic nonsynonymous SNVs in the MM cohort, followed by SomaticSniper, Varscan2, and MuSE. Similarly, Mutect2 also performs well in identifying pathogenic in Synonymous and Other SNV groups compared to other callers. Other than Mutect2, SomaticSniper reported the most pathogenic SNVs but showed an inherent bias in identifying the frameshift and non-frameshift SNV types. Varscan2 identified all types of SNVS except non-frameshift SNVs. Lastly, MuSE, similar to SomaticSniper variant callers, showed inherent bias in identifying the frameshift and non-frameshift SNVs. These observations underscore the inherent bias due to methodological disparities among variant callers as there is a notable focus on nonsynonymous SNVs and SNVs in the splicing region. In contrast, other crucial SNVs types such as frameshift SNVs, non-frameshift SNVs, Start loss, and Stop loss are less focused by the variant callers.

It has been observed that a considerable portion of clinically actionable variants exhibits low VAFs, often attributed to factors like low tumour purity and treatment-induced mutations [231]. Ensuring accurate identification of SNVs requires variant callers with robust sensitivity, particularly in detecting SNVs with low VAFs found in impure tumors [232]. Notably, both SomaticSniper and Varscan2 demonstrate suboptimal sensitivity in recognizing variants with low allelic frequencies [233]. Although SomaticSniper exhibits the highest sensitivity for 100% pure tumor samples, Varscan2 can achieve enhanced sensitivity (up to 0.5) by lowering its minimum allele fraction threshold to 0.05, at the cost of much higher false positive rate (300 false positives per Mb) [233]. Additionally, Varscan2 often reports a high number of germline polymorphisms with relatively low confidence [234]. Conversely, both MuSE and Mutect2 exhibit comparable performance in low allelic variant calling, yet Mutect2 demonstrates superior sensitivity in detecting true positives while effectively managing false positives [235, 232].

Given the absence of a single somatic variant caller offering optimal performance across all scenarios, an ensemble approach combining results from multiple complementary callers may present the most balanced trade-off between sensitivity and specificity [236, 237]. Finally, considering the methodological disparities among variant callers, we adopted a comprehensive strategy by aggregating all pathogenic SNVs detected by each of the four variant callers. This approach serves to alleviate potential methodological and inherent biases. Subsequently, we subjected all SNVs identified by the variant callers to statistical analysis using the dndscv tool to pinpoint significantly altered genes.

Following the identification of these genes, we employed the BDL-SP model to further refine our selection, focusing on genes and their associated SNVs that serve as pivotal biomarkers for distinguishing MM from MGUS.

4.4.2 Basis for selecting six ML models and training strategy

For comprehensive benchmarking of the BDL-SP model, first, we listed the existing classical ML supervised algorithms in scikit-sklearn of all categories, such as the Support Vector Classifier (SVC) model and its variants, decision models and its variants (decision tree and random forest), boosting models (XGBoost, CatBoost, AdaBoost etc.), regression models (logistic regression, linear regression, etc.), etc. Next, we trained these ML models and selected the six models using the following basis:

1. We selected the model from each algorithm category based on their quantitative and qualitative performance to ensure methodological diversity in our analysis. The methodological diversity in ML algorithms is crucial in benchmarking the newly developed model as the method used in the newly developed model should perform better than all different ML algorithms. With this approach, we selected logistic regression from regression models, decision tree and random forest from decision models, XGBoost and CatBoost from boosting models and SVC model with RBF kernel in scikit-learn.
2. Including ML algorithms from decision models, regression models, boosting models, and SVC models ensures generalization and robustness as each algorithm handles the complexity of data structure differently.

In addition to these six ML models, we included deep learning models to enrich the quantitative and qualitative benchmarking process. This comprehensive approach ensures a thorough assessment of model performance and generalizability.

For the training of baseline ML models, our aim was to ensure that the ML model learns from all available features of the data, avoiding the potential bias that could arise from selecting only a subset of features. Therefore, instead of using feature selection before model training, we opted for PCA to extract the most informative components from the data. By using PCA, we transformed the original features into a set of orthogonal principal components (PCs) that capture the maximum variance in the data. These PCs were then utilized as the input features for training the machine learning model, enabling us to leverage the most relevant information while minimizing the risk of overfitting or bias. In our study, we included 1174 MM and 61 MGUS subjects with WES data from three global repositories (MMRF, EGA, and AIIMS). Due to the very low number of MGUS samples, it was not feasible to split the datasets into train, test and external validation datasets. We addressed this challenge by incorporating a five-fold cross-validation strategy for all ML model training. In this strategy, we divided the whole dataset into five stratified folds. Next, we trained the classifier using samples in 4 folds and evaluated the performance matrices using the one remaining fold. We repeated the above step five times to ensure the model was validated for all five folds as a validation dataset. Once we had five trained classifiers, we averaged the performance matrices (balanced accuracy and AUPRC) for all five classifiers to get the overall performance assessment. This strategy ensured that every data sample became an unseen test sample once. This allowed all the variability of the limited available dataset to be tested and harnessed fully. All metrics are eventually reported on the test samples.

4.4.3 Qualitative analysis of BDL-SP model

The complete list of top significantly altered genes identified by the best three performing models (BDL-SP, CSRF, and CS-Cat) is provided in Table-4.5.
Of all, our proposed BDL-SP model identified the largest number of previously reported OGs, TSGs, ODGs, and AGs compared to the other standard ML methods. This shows that our GCN-based BDL-SP workflow is indeed capable of robustly extracting the differentiating genomic features that are otherwise difficult to obtain. Many of the top-ranking genes in the present study included known cancer drivers (*HLA-A, IGLL5, KRAS, LTB*, etc.), OGs (*BRAF, FGFR3, KRAS, NRAS* , etc.), TSGs (*ARID2, CYLD, DIS3, EGR1, HLA-A, LTB, TRAF3, SAMHD1, SP140*, etc.) and AGs (*ARID2, NF1, NFKBIA, KRAS, TP53*, etc.) having high relevance in MM. Interestingly, some TSGs (*APC, ARID1B, CMTR2, FANCD2, HLA-B/C, KMT2D, MITF, NOTCH1, SDHA* and *AMER1*) and OGs (*CARD11, NOTCH1, VAV1, IRS1, MGAM, ABL2, TCL1A, PGR, MITF, RPTOR, TERT, BRD4, MECOM*, and *TAL1*) that are so far not reported as drivers in MM, were also listed in the top ranking genes of BDL-SP. Further focused studies are required to validate the above finding and to check the functional status and other characteristics of these genes before classifying them as MM drivers.

Pathway analysis on MM and MGUS genes revealed that the MM-related pathways, such as MAPK, cGMP-PKG, B-cell receptor, etc., were not significantly altered in MGUS (adjusted p-value > 0.05) and became significant in MM (adjusted p-value ≤ 0.05). We observed that several OGs, TSGs, ODGs, and AGs associated with the significantly altered pathways were found significantly altered only in the MM cohort and not in the MGUS cohort (See Figure-4.19). Here, the additional alterations in several previously reported genes, such as *BRAF, FGFR3, IRS1, MAX, KRAS*, etc., assisted the malignant progression of MGUS to MM. Our pathway analysis also demonstrated that some pathways that lost their statistical significance from MGUS to MM were actually related to the other cancer types.

Further, to identify the novel therapeutic targets and biomarkers, we have excluded the pathways that are already reported in MM or reported to any other non-cancer disease (for example, Alzheimer, obesity, diabetes, etc.). Using this criterion, we identified seven significantly altered MM-relevant pathways contributing to various biological processes.

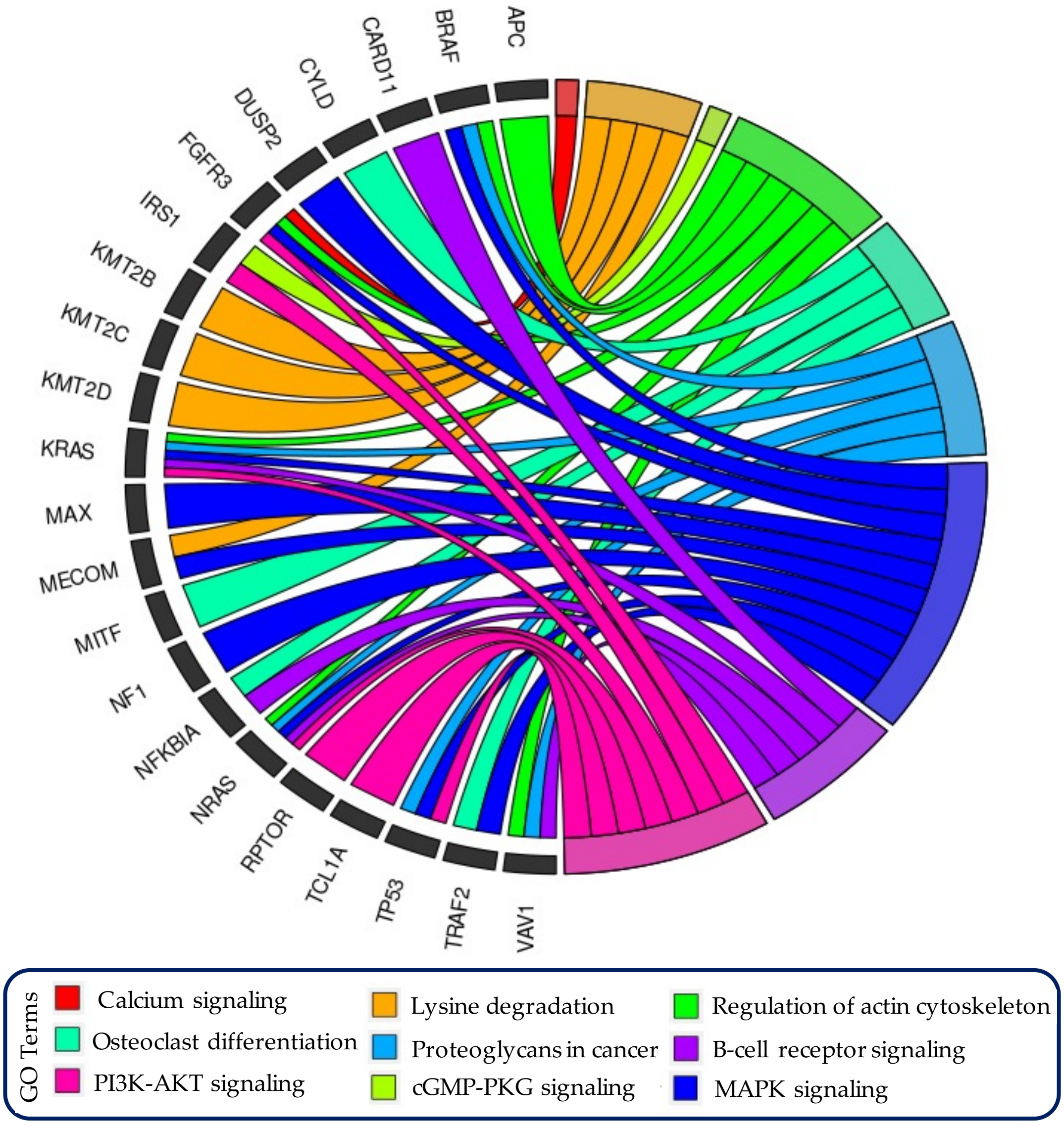


Figure 4.19: GOChord plot reveals the association of driver/TSG/Onco/Actionable genes associated with important pathways. The gene KMT2C was observed to be significantly mutated in MGUS and MM, while the gene APC was mutated only in MGUS. All other genes were observed to be significantly mutated in MM only.

The detailed examination of these seven significantly altered pathways is shown below:

1. Gap junction Pathway: Gap junctions (GJs) serve as vital intercellular channels facilitating molecular communication between adjacent cells. When the functionality of GJs is compromised, it can lead to pathological conditions, including cancer. In cancer, GJs play diverse roles, such as promoting cell invasion, facilitating dormancy of metastatic cells, enhancing nutrient exchange within tumors, and aiding immune evasion [238]. In the context of MM progression from MGUS, perturbations in the GJ pathway may disrupt normal cellular interactions, thus contributing to disease pathogenesis. Depending on the specific type of GJs and the tumor microenvironment, these junctions can exhibit both tumor-suppressive and tumor-promoting properties [239]. Notably, studies have implicated inositol 1,4,5-trisphosphate receptor type 1 (ITPR1) autoimmunity in various cancer types [240, 241]. However, the autoimmune aspect of MM remains poorly understood, and its pathogenesis is still elusive [242]. Therefore, further exploration into the impact of ITPR1 alterations on MM autoimmune status is suggested. Investigating the dysregulation of the GJ pathway in MM could provide valuable insights into its supportive role within the tumor microenvironment, thus shedding light on its significance in MM progression and pathogenesis.
2. Graft-versus-host disease (GVHD) pathway: The graft-versus-host disease (GVHD) pathway represents an amplified response of normal inflammatory mechanisms, where donor lymphocytes encounter foreign antigens in an environment conducive to inflammation [243]. Despite its significance, the role of the GVHD pathway in the progression of multiple myeloma (MM) remains understudied. Comprising several HLA family genes (HLA-A, HLA-B, HLA-C, HLA-DQA1, HLA-DQA2, HLA-DRB1, HLA-DRB5, etc.), this pathway is integral to T cell immune responses. Notably, HLA-A, HLA-B, and HLA-C genes serve as ligands for natural killer (NK) cell immunoglobulin (Ig)-like receptors, crucial components of innate immunity [244]. Functional investigations have revealed that NK cells retain their cytotoxic potential in MGUS but may exhibit diminished cytotoxicity as the disease progresses to MM [245]. Conversely, down-regulation of the HLA family gene HLA-DPA1 expression has been associated with poor outcomes in MM patients [246]. Moreover, studies across various populations have linked HLA class I and II genes to MM susceptibility and its pathogenesis [247]. Additionally, within the GVHD pathway, the killer cell immunoglobulin-like receptor (KIR) family genes play pivotal roles in regulating NK cell function and immune responses. Promising NK-cell-based immunotherapies, such as expanded and activated KIR-mismatched therapies, underscore the therapeutic potential of genes within this pathway [248]. Thus, the GVHD pathway harbors numerous genes with pivotal roles that can significantly influence MM pathogenesis and progression.
3. Human T-cell leukemia virus 1 (HTLV-1) infection pathway: Human T-cell leukemia virus type 1 (HTLV-1) is a retrovirus known for its association with adult T-cell leukemia/lymphoma (ATL) [249]. Within the HTLV-1 infection pathway lie several key driver genes of MM, including *KRAS, NRAS, TP53*, and *RB1*. Among these, *TP53* stands out for its pivotal role in safeguarding cells against DNA dam-

age and transformation. Deregulation within the HTLV-1 infection pathway can lead to TP53 activation, potentially resulting in the overexpression of c-MYC, as TP53 serves as a downstream target of c-MYC [250, 251]. Interestingly, while c-MYC expression has been linked to MM but not MGUS [252], the deregulation of the HTLV-1 infection pathway has been associated with the Tax protein, capable of activating transcription factors such as NF-kB and AP-1. These factors, in turn, regulate the expression of genes involved in inflammation and cell growth. Additionally, the deregulation of NF-kB and AKT pathways, pivotal for cellular survival, proliferation, and DNA damage response, further underscores the significance of the HTLV-1 infection pathway in various cellular processes potentially contributing to MM progression and pathogenesis [253, 254].

4. Kaposi sarcoma-associated herpesvirus (KSHV) pathway: Kaposi sarcoma herpesvirus (KSHV) is a human virus with a long evolutionary history that has co-evolved with human populations. While once widespread, it is now predominantly found in specific regions such as sub-Saharan Africa, the Mediterranean Sea area, parts of South America, and ethnic communities, where its seroprevalence exceeds 10% [255]. The role of the KSHV pathway in MM progression remains largely unexplored. Studies have indicated that KSHV has the capability to infect a wide range of cell types, including epithelial cells, monocytes, macrophages, dendritic cells, T cells, and fibroblasts [256, 257]. In a study examining a dataset comprising 8 MM and 2 MGUS samples, KSHV was detected in the bone marrow dendritic cells of MM patients, while no evidence of its presence was found in malignant plasma cells or bone marrow dendritic cells from normal individuals or patients with other malignancies. This finding suggests a potential association between the KSHV pathway and MM progression [258].
5. Natural killer (NK) cell-mediated cytotoxicity pathway: NK cells, a type of innate lymphoid cells, play a pivotal role in immune surveillance by identifying and eliminating aberrant cells, including tumor cells, through mechanisms like cytotoxic granule release and death receptor signaling. The NK cell-mediated cytotoxicity pathway is essential for immune surveillance against tumor cells, including those implicated in MM progression from MGUS. Frequent mutations in genes like *NRAS, KRAS*, and *BRAF* are observed in MM, disrupting intracellular signaling pathways crucial for cell proliferation and survival [83]. HLA class I genes (*HLA-A, HLA-B,* and *HLA-C*) are vital for presenting antigenic peptides to NK cells and cytotoxic T lymphocytes (CTLs), modulating their activation and cytotoxic activity. Polymorphisms in these genes can influence the recognition of MM cells by NK cells and impact NK cell-mediated cytotoxicity effectiveness [259]. Additionally, VAV1 gene regulation has been linked to NK cell-mediated cytotoxicity enhancement [260]. Consequently, dysregulation of the NK cell-mediated cytotoxicity pathway due to gene alterations may disrupt various biological processes, including cell proliferation, survival, and apoptosis. These findings underscore the importance of further investigating the roles of genes such as VAV1 and HLA class I genes, not traditionally considered MM driver genes, in MM progression and pathogenesis.
6. Viral carcinogenesis pathway: The Viral Carcinogenesis pathway encompasses

several key driver genes in MM, such as *KRAS, NRAS,* and *TRAF3*, alongside TSG like *RB1, TP53*, and *NFKBIA*. These genes can potentially influence MM progression from MGUS by modulating cellular signaling pathways and promoting oncogenic processes. For example, the *RB1* gene governs cell cycle progression and apoptosis. Dysregulation of *RB1*, whether through mutation or inactivation, can trigger uncontrolled cell proliferation, thereby contributing to MM pathogenesis. Similarly, *TP53*, acting as both an MM driver gene and a TSG, plays a critical role in safeguarding cells against DNA damage and inducing apoptosis in response to cellular stress. Mutations in *TP53* are frequently observed in MM and are linked to poorer survival outcomes [84]. Furthermore, alterations in HLA ligands can impact immune cell surveillance against pathogens [259, 261]. Thus, the cumulative effect of these gene alterations within the Viral Carcinogenesis pathway may fuel MM progression by fostering cell proliferation, enhancing cell survival, and facilitating immune evasion in the context of viral infections.

7. Allograft rejection pathway: The allograft rejection pathway orchestrates a complex interplay between the innate and adaptive immune systems, triggered when recipient T cells recognize antigens from the donor, a process known as allorecognition [262, 263]. This pathway assumes a pivotal role in mounting immune responses against foreign tissues or cells, ultimately leading to their rejection. Central to this mechanism is the HLA molecules, which serve as key players in antigen presentation to T cells, thus initiating immune responses. HLA class I molecules (*HLA-A, HLA-B*, and *HLA-C*) are responsible for presenting antigens derived from intracellular pathogens to CD8+ cytotoxic T cells, while HLA class II molecules (*HLA-DQA1, HLA-DQA2*) present antigens from extracellular pathogens to CD4+ helper T cells [264, 265]. Any alterations in the HLA family genes may compromise their ability to effectively present antigens to immune cells (CD4+/CD8+), thereby facilitating mechanisms of immune evasion that promote tumor growth and confer immune resistance [266]. It has been observed that the HLA class 1 and class II genes have susceptibility effects on MM. Studies of different populations have reported different HLA class I and II alleles that affect MM [247].

However, the results in our study are unique because we have demonstrated that these pathways are selectively and significantly dysregulated in MM compared to its precursor stage of MGUS due to a distinct set of genes that are differentially mutated in the two diseased stages. These observations warrant further investigations to decipher if any of these differentiating genes associated with these pathways could become druggable targets, especially during the early phase of MGUS. Some of the key genes and pathways that are selectively altered at the MM stage, such as *FGFR3, BRAF* and *MAP* kinase pathways, are actionable and hence, targeted therapy for them is under evaluation in clinical trials [267]. *FGFR3* is a partner gene in t(4;14) that has been observed as a significantly altered gene in all datasets of MM and MGUS. However, the poor prognostic impact of *FGFR3* has been linked to activating mutations in the *FGFR3* gene rather than the fusion event, which exerts its influence via activation of

WHSC/MMSET genes and is responsive to proteasome inhibitors [268, 269]. Besides single case reports demonstrating the efficacy of *BRAF* inhibitors in relapsed refractory MM with *BRAF* mutations, a recent report on NCI-Match trial shows promising results for BRAF inhibitors, Dabrafenib and Trametinib in patients harboring tumors with *BRAF V600E* mutations including MM [270]. Many SVs observed in MM, such as IgH translocations, 1q gain, and 1p del, are also observed at the MGUS stage. However, *C-MYC* alterations, which are predominantly structural variations, are secondary events and are seen in nearly 40% of MM patients [271]. The lack of analysis of SVs is one of the limitations of this study. However, we did observe mutations in *MAX* at the MM stage, which is known to dimerize with *C-MYC* and influence the transcription of multiple genes and, thus, the pathogenesis of MM [272].

The frequently observed complex genomic traits that can drive the disease progression from MGUS to MM can be 3'UTR/5'UTR mutations [92], CNVs, SVs [229, 81], and loss of the ability of the dysfunctional immune environment to control virulent cell clones [273]. Akin to higher levels of disease load in MM compared to its precursor states, measurable disease load, increased number of nonsynonymous mutations in MM compared with MGUS [89, 90, 91] and increased levels of deregulated cytokines in relapsed refractory MM compared to treatment naïve MM has been reported [274]. In addition, in MGUS, a positive correlation between the increasing chromosome changes and the somatic gene mutations, and the absence of *MYC* translocation and *TP53* deletions or mutations has been observed [91]. From the genomic profile analysis of paired MGUS-MM and SMM-MM samples, it has been observed that as the disease progressed, the number of nonsynonymous mutations actually decreased in 70% samples. This observation is in contrast to reports on unpaired samples, where an increase in the nonsynonymous mutations has been reported from MGUS to MM [89, 90, 91]. Further, the comparisons of unpaired MGUS/ SMM and MM samples have shown the mutational similarity of MGUS/SMM with MM [93]. Based on this observation, we hypothesize that the progression is associated with an altered landscape of acquired mutations rather than an increased total mutational burden.

The post-hoc explainability of the BDL-SP model using the ShAP algorithm revealed the top genomic attributes (genomic features and significantly altered genes) at both the group- and sample-levels. At group-level, all the 824 significantly altered genes were ranked according to their ShAP score using the algorithm shown in Figure-4.4 (Table S6 of supplementary material) and top 500 genes were further compared with the literature (Table S7 of supplementary material). Several significantly altered genes found in our analysis were previously reported as driver genes in [87, 222, 223], OGs and TSGs in [53], and AGs in [51, 224]. In contrast, some genes such as *KIR3DL2, FCGR2A, LILRB1/2, KIR2DL1/4* etc. were novel that contributed significantly in disease classification (See Table S6 of supplementary material). The *KIR* framework genes

(*KIR3DL2/2DL4*) were among the top significantly altered genes with the largest ShAP scores. The *KIR* gene complex on chromosome 19 encodes a series of inhibitory or activatory *KIR* genes expressed on *NK* cells [275, 276, 277]. These receptors serve as *HLA* ligands and modulate *NK* cell immune function against tumors [275]. A few activating genes in the *KIR* family (*KIR2DS4 and KIR2DS5*) have been shown to have a higher prevalence in MM patients than healthy people [275]. The *KIRs* have also been reported to influence the efficacy of therapies such as that of isatuximab [276]. The second topmost gene with the largest ShAP score was *IGLL5*. Again, the *IGLL5* gene undergoes point mutations and *IGLL5/IGH* translocations in MM [122]. Point mutations are largely mutually exclusive of *RAS* mutations and associated with a greater risk of disease progression. Similarly, other genes such as *HLA-A/B/C, FCGR2A* and *LILRB1/2* reported in previous studies are also shown to have a significant role in MM [278, 279, 280, 281]. The crucial role of these top immune-related genes highlighted by the ShAP ranking in our study suggests their potential role as drivers of progression and disease-stratifying biomarkers.

We have also highlighted the impact of ethnicity (Figure-4.11) among three groups of American (MMRF), European (EGA), and Indian (AIIMS) populations. The number of OGs, ODGs, and AGs significantly differed in the MM samples of the American and Indian populations and MGUS samples of the European and Indian populations (Figure-4.11A). Also, the median of the number of OGs and AGs increased with the disease progression from MGUS to MM. This increase could be due to the increasingly active participation of OGs and AGs in disease progression from MGUS to MM. Similarly, the number of OGs, ODGs, and AGs are significantly different in the MGUS samples of the Indian population and MGUS samples of the European population (Figure-4.11B). Here, we also observed the increasingly active participation of OGs, ODGs, and AGs in MGUS samples of the Indian and European populations. On the other hand, the number of previously reported genes (OGs, ODGs, TSGs, and AGs) present in the MM samples of the American and Indian population was not found to be statically different (Figure-4.11C). These observations indicate that the impact of ethnicity on disease biology can not be overlooked and might be an important factor during the initial phase or development phase of MM. Further analysis of ethnicity-specific information to infer the responsible prognostic factor for disease development and progression is strongly suggested.

The sample-wise gene ranking highlighted their contribution at the individual sample-level. The study in [91] showed that the transition from MGUS to MM is due to the acquisition of mutations in critical driver genes and OGs. Interestingly, we have observed that not only driver genes and OGs but several TSGs and AGs were also altered significantly in MGUS (Figure-4.11). Further, the role of OGs increased as the disease progressed from MGUS to MM. On comparing the top contributing genomic features in MGUS and MM samples, we observed that the genomic features related to the Synonymous SNVs group (a group of UTR3, synonymous, and UTR5 type SNVs) and the Other SNVs group (a group of Non-frameshift insertion/ deletion/ substitution, intronic, intergenic, ncRNA_intronic, upstream, downstream, unknown, and ncRNA_splicing SNVs) contributed largely in disease classification as compared to the genomic features of the nonsynonymous SNVs group (Figure-4.11).

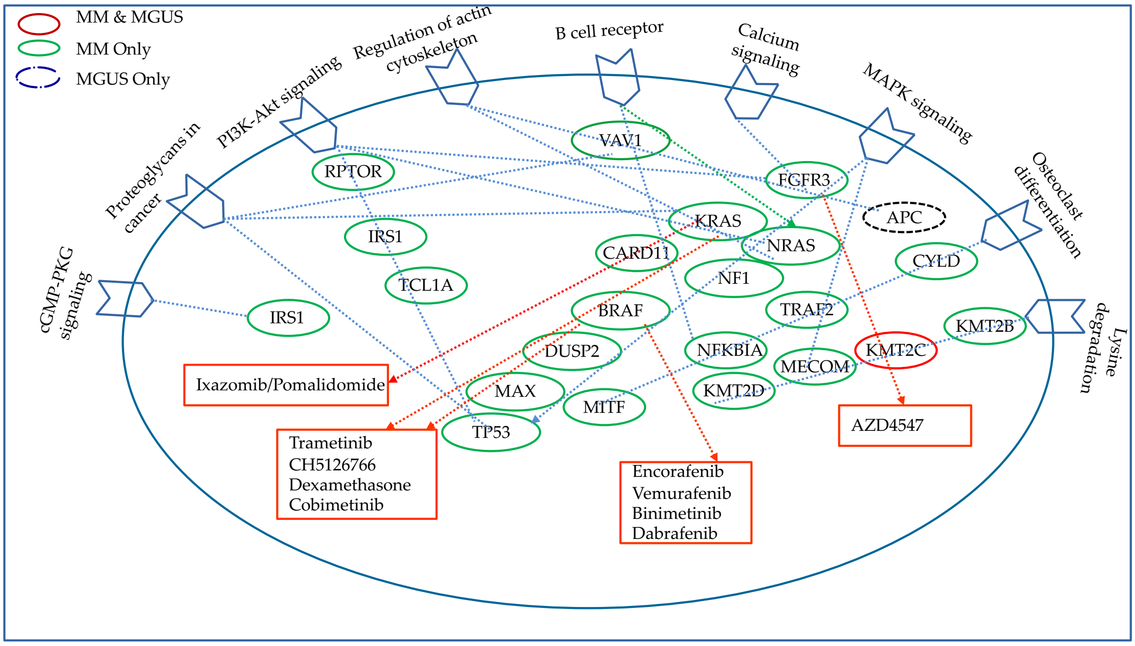


Figure 4.20: Important pathways significantly altered in MM. Drugs used for pathway-directed therapies associated with mutations in genes are also shown with red-colored text boxes and arrows.

Although the role of synonymous SNVs is unclear in MM, recent studies have observed these synonymous SNVs as significant contributors in multiple cancer types [282, 283, 284, 285, 286]. Further exploration of differentially affected biological pathways may provide the pathogenic link between MM, its precursor (MGUS or SMM), and overt disease stages so as to find appropriate targeted therapy to halt the progression from precursor to stage to MM (Figure-4.20). We have shown in the current study that the incremental accumulation of key mutations tilts the balance of biological pathways in favor of progression from the state of MGUS to MM in a large cohort of unpaired MGUS-MM samples. Some of these pathways are actionable, and targeting them may enable us to reverse the balance in favor of a controlled and relatively indolent clinical course. Further, AI-based workflow has successfully differentiated MGUS from MM. We have shown in our study that our trained ML classifiers are able to identify pivotal genomic biomarkers helpful in distinguishing MM and MGUS, thereby leading to a better understanding of the malignant transformation of MGUS to MM and prognostication.

4.4.4 Impact of nonsynonymous SNVs on protein function and protein structure

We analyzed the impact of nonsynonymous and deleterious SNVs on *KRAS* and *TP53* gene protein function and protein structure. We observed that all of the nonsynonymous and deleterious SNVs present in the *KRAS* gene protein (LZRSL8) were associated with the PF00071, PF08477, PF00025, and PF00009 Pfam domains. The Interpro-domain of these Pfam domains corresponds to Small GTP-Binding Proteins [287]. This superfamily comprises over 100 members and is structurally classified into five families: Ras, Rho, Rab, Sar1/Arf, and Ran. These families play crucial roles as biological timers, initiating and terminating specific cell functions [288]. Within the PF00071 Ras family are several sub-families, such as Ras, Rab, Rac, Ral, Ran, Rap, and Ypt1. The Rab GTPases are important regulators of vesicle formation, motility and fusion. PF08477, also known as the Ras of Complex (Roc) domain of DAPkinase, is implicated in mitochondrial homeostasis and apoptosis. These observations imply the key role of altered Pfam domains in the key cellular processes that may contribute to MM progression.

Similarly, all the nonsynonymous and deleterious SNVs in the *TP53* gene protein () were found associated with the PF00870 Pfam domain, which is commonly found in p53 transcription factors and plays a crucial role in DNA binding. The DNA-binding domain functions by securely clamping onto the DNA target, thereby stabilizing the protein-DNA complex [289]. Additionally, protein interactions within this domain can further contribute to stabilizing the complex [290]. The nonsynonymous and deleterious SNVs may impact the protein's structure and its ability to bind to DNA effectively.

4.5 Limitations of the study

Our study aimed to identify the pivotal distinguishing biomarkers between MM and MGUS using WES mutational profiles and protein-protein interactions between significantly altered genes. However, one potential limitation of our study is the absence of normal samples. To address this limitation and enable the identification of normal vs MGUS samples, researchers can leverage the same BDL-SP model for the analysis. In this scenario, researchers would need to re-train the BDL-SP model from scratch. Once properly re-trained and benchmarked against baseline models, both quantitatively and qualitatively, post-hoc analysis can facilitate the identification of pivotal biomarkers (genes and genomic features) distinguishing normal from MGUS samples.

4.6 Conclusion

MGUS and MM share many common features, such as genomic biomarkers and structural variants, although MGUS has a relatively less complex genomic profile than MM. Therefore, it is a challenging task to distinguish MM from MGUS. In our proposed work, we have presented an innovative, bio-inspired AI-based workflow BDL-SP to identify pivotal genomic biomarkers to distinguish MGUS from MM. The proposed graph convolutional network-based BDL-SP model is able to extract discriminative genomic biomarkers to identify MM and MGUS samples. BDL-SP outperformed all the baseline ML-based models. Further, using the application-aware interpretability analysis of the trained AI model, we have demonstrated a way to identify the best AI model from among the multiple machine learning or deep learning models that may have performed similarly on the quantitative metrics on the available data. In the post-hoc interpretability benchmarking, BDL-SP outperformed all the baseline models by identifying the largest number of previously reported genes such as *KRAS, BRAF, LTB, NRAS, FGFR3, NF1, NFKBIA, ARID2, RB1, HLA- A, TP53, SP140, TRAF3, EGR1, IRF1, SAMHD1, DIS3, CYLD, KMT2B/C, MAX, ZFHX3* and *NCOR1*, that are of high relevance in MM. Further, some of the genes that acted as differentiable biomarkers included TSGs (*HLA- B/C, NOTCH1, SDHA, MITF, ARID1B, FANCD2, KMT2D, APC, CMTR2*, and *AMER1*) and OGs (*CARD11, NOTCH1, VAV1, IRS1, MGAM, ABL2, TCL1A, PGR, MITF, RPTOR, TERT, BRD4, MECOM*, and *TAL1*) that have not yet been identified as MM drivers. These require validation by future studies before being declared as MM drivers. We further validated our findings by performing pathway analysis on the top mutated genes. It was inferred from the pathway analysis that several signaling pathways, such as the Calcium signaling pathway, B-cell receptor signaling pathway, PI3K-Akt signaling pathway, MAPK signaling pathway, etc., are selectively and more significantly dysregulated with disease progression. Additional mutations in driver genes, critical OGs, TSGs, and AGs fostered the transformation of benign MGUS to MM. Similarly, the genomic mutation associated with the Synonymous SNV group (synonymous SNVs, UTR3, and UTR5) was found to be the most significantly contributing biomarker differentiating MM from MGUS. These observations may hold great significance from a therapeutic point of view. We observed that the number of OGs, driver genes, and AGs in the MGUS samples of European and Indian populations was statistically different. Although no population-specific differences were observed in our analysis of the MM data, which consists of the American and Indian populations, the results of the MGUS data indicate that the impact of ethnicity on the disease biology of MM should be further explored.

Intriguingly, as we delved deeper into the post-hoc analysis of the BDL-SP model, a compelling observation emerged – gene-gene interactions appeared to play a pivotal role in the pathogenesis of MM. We noted that several of the top-ranked key distinguishing biomarkers were intricately interconnected with many known drivers or previously reported genes within the PPI network. Motivated by these findings, we extended our post-hoc analysis by incorporating gene-gene interactions, leading to the development of an enhanced version of the BDL-SP model, which we have aptly named Biological Network for Directed Gene-Gene Interaction Learning (BIO-DGI). This novel model is tailored to identify crucial distinguishing biomarkers based on directed gene-gene interactions, offering a more comprehensive perspective on MM pathogenesis. The details of the model's design, methodology, and post-hoc interpretability analysis are comprehensively elaborated in the upcoming chapter, shedding further light on our innovative approach and its implications.

Chapter 5

Directed Gene-Gene Interactions in Multiple Myeloma

5.1 Introduction

Multiple Myeloma is a haematological cancer marked by clonal plasma cell proliferation in the bone marrow. MGUS and MM represent different stages, with MGUS being a precursor condition. Around 1% of MGUS cases progress to MM yearly [291]. Advanced techniques like WES and WGS unveil genomic aberrations in MM and MGUS. Genomic studies reveal critical events like CNVs and SVs, such as del(1p), gain(1q), del(13q), t(4;14), and others, impacting MM prognosis and shedding light on their association with MM prognosis [292, 293, 294, 295, 296, 81, 83, 87, 128, 297]. Minor genomic changes also influence clinical outcomes. Recent findings highlight the significance of bi-allelic alterations in TP53 and DIS3 genes as high-risk markers in MM [298].

In this study, we designed a targeted sequencing panel for comprehensive genomic profiling of MM. To craft the sequencing panel, we designed a novel AI-based Biological Network for Directed Gene-Gene Interaction Learning (BIO-DGI) model employing the gene-gene interactions from nine PPI databases and exonic mutational profiles. Further, we delved deeper into gene-gene interactions by studying the learned adjacency matrices from trained BIO-DGI classifiers. On meticulous analysis of SNVs, CNVs, SVs and LOFs profiles of top-ranked genes obtained from post-hoc interpretability analysis of the BIO-DGI model, we introduced a clinically tailored 282-genes panel, aiming to capture the unique characteristics of MM and MGUS. Our study firmly establishes this panel as a promising novel strategy, particularly in identifying MGUS samples likely to progress to MM and pinpointing high-risk MM samples.

5.2 Materials & Methods

5.2.1 Whole-exome sequencing datasets of MM and MGUS patients

In this study, we included tumor-normal pairs of BM samples from a MM cohort of 1154 samples and an MGUS cohort of 61 samples sourced from three global repositories of WES data. For the MM cohort, 1072 samples were acquired through authorized access to the MMRF dbGaP study (phs000748; phs000348), predominantly comprising American population samples [134]. We also downloaded processed MMRF datasets (version IA12) containing copy number variations (CNVs), structural variations (SVs), and clinical data from the MMRF Research Gateway. Additionally, we included 82 MM samples from an AIIMS dataset representing the Indian population [5]. In the MGUS cohort, we incorporated 28 MGUS samples from the AIIMS dataset and 33 samples from the EGA data.

5.2.2 Methods

Whole exome sequencing data pre-processing for SNV identification

The WES data obtained from AIIMS and EGA contained the raw fastq files, and the MMRF dataset contained the processed VCF files. The computational workflow for the SNV identification, genomic annotation of SNVs, SNVs filtration, SNVs grouping, and the identification of significantly altered genes was taken from our previous related study [5]. Briefly, raw fastq files from AIIMS and EGA datasets were processed using the standard exome sequencing pipeline [219]. Similar to the MMRF data, the SNVs in AIIMS and EGA WES data were extracted using MuSE, Mutect2, VarScan2, and SomaticSniper variant callers. The SNVs in AIIMS, EGA and MMRF datasets were annotated using the ANNOVAR database. The annotated SNVs were categorized into three categories based on their functional significance, i.e. synonymous SNVs, nonsynonymous SNVs and other SNVs. The benign SNVs were filtered out using FATHMM-XF. Lastly, annotated SNVs were pooled for MM and MGUS separately and analyzed for identifying significantly altered genes using the 'dndscv' tool. The union of significantly altered genes from all four variant callers for the MM cohort of 1154 samples and the MGUS cohort of 61 samples led to 617 and 362 genes, respectively. The further union of the significantly altered genes in MM and MGUS mentioned above yielded a total of 824 genes.

For each significantly altered gene, their corresponding PPI were extracted from 9 PPI databases (BioGrid [299], BioPlex [300], FunCoup [301], HIPPIE [302], HumanNet [303], IntegratedAssociationCorrNet [304], ProteomHD [305], Reactome [306], and STRING [307]). Out of 824 genes, we filtered out 26 genes that had no interaction with any other significantly altered gene and merged all the interactions for the remaining 798 genes from all nine PPI databases to get the merged adjacency matrix (Table-S1, Supplementary File-1). Besides the PPI interactions, we extracted 26 genomic features that include distributive statistics (median and standard deviation) of VAF, AD, and four variant conservation scores (GERP [308], PhyloP [309], PhastCons [310], and Mutation Assessor [311]).

5.2.3 Proposed directed gene-gene interaction learning in biological network (BIO-DGI)

In this study, our objective was to pinpoint candidate driver genes and elucidate crucial gene-gene interactions responsible for the progression from MGUS to MM. We introduced an innovative GCN-based bio-inspired model named "Biological Network for Directed Gene-Gene Interaction Learning" (BIO-DGI). The BIO-DGI model, depicted in Figure-5.1, harnesses the power of GCN to grasp pivotal gene-gene interactions and forecast potential driver genes.

To empower the BIO-DGI model, we supplied two essential inputs: 1) an undirected PPI network adjacency matrix sourced from PPI interaction databases and 2) a sample feature matrix. Two versions of the PPI network adjacency matrix were considered in this study. The first version involved extracting PPI interactions solely from the STRING database, serving as the basis for training the BIO-DGI model, denoted as BIO-DGI (PPI-STRING). In the second version, we merged the PPI network adjacency matrix from nine distinct PPI databases. This merged adjacency matrix, referred to as BIO-DGI (PPI9), was then utilized for training purposes. In both versions of the adjacency matrix, each node corresponded to a significantly altered gene, while the links represented interactions between these genes. Additionally, each node was equipped with a feature vector of length 26, as illustrated in Figure-5.1. Consequently, the PPI network comprising 798 significantly altered genes, each associated with a feature vector of length 26, was integrated into the input layer of the BIO-DGI model.

The BIO-DGI model architecture contains two modules: 1. Multi-head attention module and 2. GCN Module. The multi-head attention modules contain three attention units to learn gene-gene interactions, which are followed by an attention consensus module for taking the consensus of all three attention unit weights to get the updated learned adjacency matrix. The purpose of the multi-head attention module was to learn and update the adjacency matrix to get a weighted PPI adjacency matrix. Similarly, in the GCN module, the input layer is followed by one hidden layer of GCN that is further followed by one fully connected layer of 798 neurons to 2 neurons, giving output through log-softmax activation function for sample class classification (MM vs MGUS).

In our study, there were 95% MM samples and 5% MGUS samples which made the data highly imbalanced. Hence, the BIO-DGI model was trained using a cost-sensitive negative log loss (NLL) function to account for the data imbalance. The BIO-DGI model was trained using a five-fold cross-validation technique that led to the training of five best-performing classifiers. All five classifiers with learned adjacency matrices were saved for further post-hoc analysis. We used the ShAP algorithm for post-hoc analysis of BIO-DGI model classifiers to get top-performing genes and genomic features that

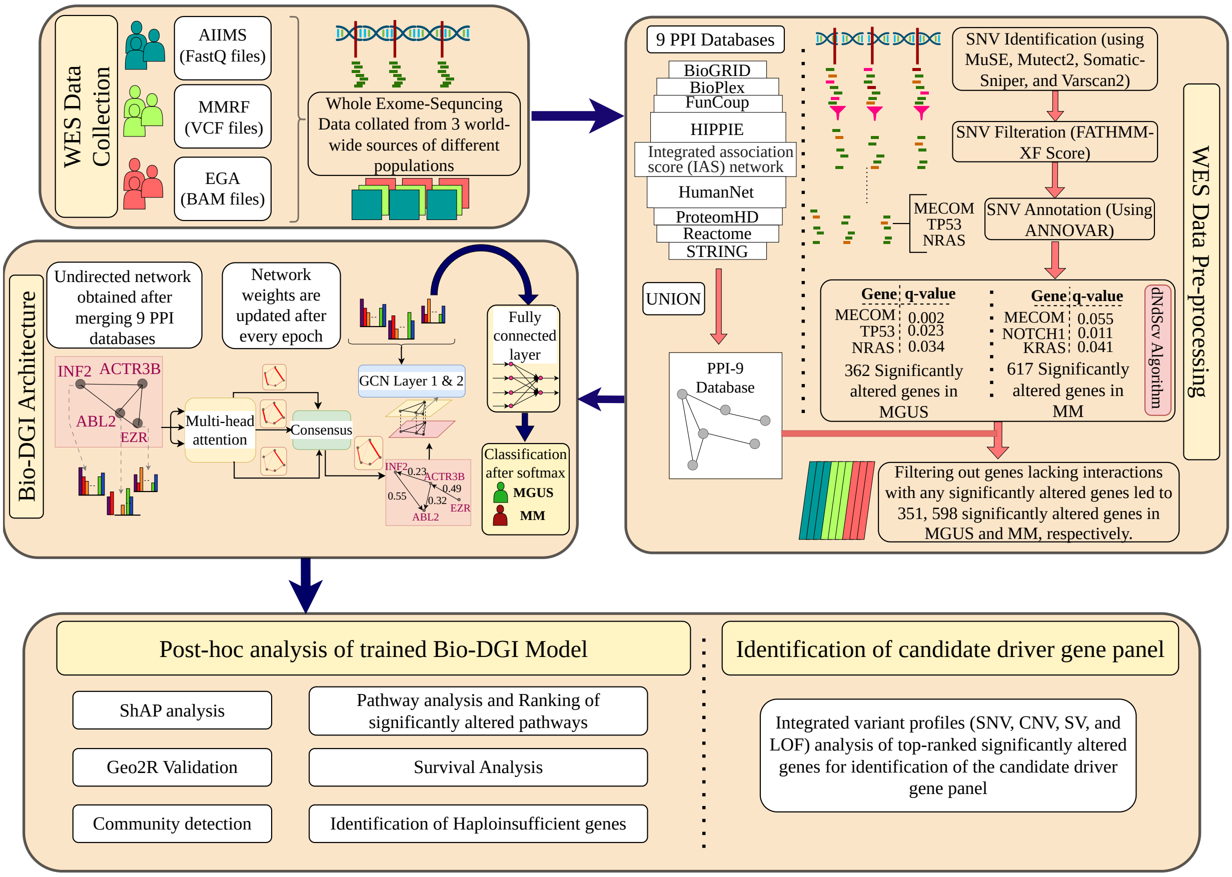

WES Data Collection
AIIMS (FastQ files)
MMRF (VCF files)
EGA (BAM files)
Whole Exome-Sequncing Data collated from 3 world-wide sources of different populations
9 PPI Databases
BioGRID
BioPlex
FunCoup
HIPPIE
Integrated association score (IAS) network
HumanNet
ProteomHD
Reactome
STRING
UNION
PPI-9 Database
SNV Identification (using MuSE, Mutect2, Somatic-Sniper, and Varscan2)
SNV Filteration (FATHMM-XF Score)
SNV Annotation (Using ANNOVAR)
MECOM
TP53
NRAS
Gene q-value
MECOM 0.002
TP53 0.023
NRAS 0.034
362 Significantly altered genes in MGUS
Gene q-value
MECOM 0.055
NOTCH1 0.011
KRAS 0.041
617 Significantly altered genes in MM
dNdScv Algorithm
WES Data Pre-processing
Filtering out genes lacking interactions with any significantly altered genes led to 351, 598 significantly altered genes in MGUS and MM, respectively.
Bio-DGI Architecture
Undirected network obtained after merging 9 PPI databases
Network weights are updated after every epoch
INF2
ACTR3B
ABL2
EZR
Multi-head attention
Consensus
GCN Layer 1 & 2
Fully connected layer
Classification after softmax
MGUS
MM
INF2
0.23
ACTR3B
0.49
0.55
0.32
EZR
ABL2
Post-hoc analysis of trained Bio-DGI Model
ShAP analysis
Pathway analysis and Ranking of significantly altered pathways
Geo2R Validation
Survival Analysis
Community detection
Identification of Haploinsufficient genes
Identification of candidate driver gene panel
Integrated variant profiles (SNV, CNV, SV, and LOF) analysis of top-ranked significantly altered genes for identification of the candidate driver gene panel

Figure 5.1: Infographic representation of proposed AI-based bio-inspired BIO-DGI model and post-hoc analysis for identifying pivotal genomic biomarkers that can distinguish MM from MGUS. In the proposed AI-based workflow, the BAM files sourced from EGA and AIIMS datasets, along with VCF files from the MMRF dataset, undergo processing to identify 798 notably altered genes utilizing the dndscv tool (as illustrated in the WES Data pre-processing block). Subsequently, interactions among these 798 genes are elucidated employing a PPI network from nine PPI databases (BioGRID, BioPlex, FunCoup, HIPPIE, IAS network, HumanNet, ProteomHD, Reactome, and STRING). A network is constructed with nodes representing the significantly altered genes and edges denoting interactions obtained after merging interactions from nine PPI databases. Each node is equipped with 26 genomic features specific to its corresponding gene. The architecture of the BIO-DGI model contains a multi-head attention unit and a GCN layer followed by a fully connected layer. The feature matrix and adjacency matrix are provided as input to the BIO-DGI model. The multi-head attention unit in the BIO-DGI model updates the weights of gene interactions in the adjacency matrix, which are then integrated with the sample feature matrix to gain insights on distinguishing biomarkers that can differentiate MM from MGUS. The output of the fully connected layer is converted into the classification probabilities using the softmax activation function. Consequently, the WES data of each subject is analyzed, and feature vectors for all 798 genes are derived. These feature vectors, in conjunction with the subjectś MM/MGUS target class label, constitute the input for supervised training of the GCN. Following the learning of the BIO-DGI model for distinguishing MGUS from MM, the top genomic features and significantly altered signaling pathways are extracted utilizing the ShAP algorithm and cross-referencing with the Enrichr pathway database.

were further used for pathway enrichment analysis, gene-community identification and candidate driver gene panel. The setting of layers, hyperparameters used to train the BIO-DGI model and mathematical description of the BIO-DGI model are available in Supplementary File-2.

5.2.4 Quantitative benchmarking of BIO-DGI model with traditional machine learning classifiers

In our quantitative benchmarking analysis, we conducted a comprehensive comparison of the BIO-DGI (PPI9) model involving three key performance metrics: balanced accuracy, area under the curve (AUC), and area under the precision and recall curve (AUPRC). This evaluation encompassed the five-fold cross-validation of six established baseline cost-sensitive machine learning models: random forest, decision tree, logistic regression, XGBoost, CatBoost, and SVM from the scikit-learn library [221]. Further, we also included two const-sensitive DL models: BDL-SP and BIO-DGI (PPI-STRING) model for quantitative benchmarking. To enhance the models' sensitivity to class imbalance, we incorporated a tailored cost-sensitive loss function. This function implements weighted penalization for sample misclassifications, with the weighting being directly proportional to the class imbalance ratio. This strategic implementation of weighted penalization ensures unbiased learning outcomes for major and minor classes, fostering a more equitable predictive capability.

5.2.5 Qualitative application-aware post-hoc benchmarking of BIO-DGI model using ShAP

The ShAP (SHapley Additive exPlanations) algorithm is a powerful tool for gauging the significance of attributes in a model's predictions. It achieves this by assigning scores to attributes based on their individual contributions. In this context, ShAP played a pivotal role in enhancing the post-hoc explainability of the BIO-DGI (PPI9) model. This process unearthed the most influential genomic features and the genes that experienced significant alterations, both at the group level (MGUS or MM) and at the level of individual samples. Since a rigorous five-fold validation process was executed during the model's training phase, the ShAP algorithm was subsequently applied to each trained classifier. This enabled the identification of significant genomic attributes (genes and features) for every individual sample. It's important to note that the ShAP score can encompass both positive and negative values. In this context, a positive ShAP score for a specific attribute highlights its contribution to the model's prediction for the MM class (positive class). Conversely, a negative score indicates its role in the model's prediction for the MGUS class (negative class). Consequently, the magnitude of the ShAP score directly correlates with the attribute's impact on the model's positive class outcome. Furthermore, the extraction of ShAP interpretability was limited to those samples that were correctly predicted by at least one of the five classifiers. This approach ensured a robust basis for deriving insights through ShAP analysis.

We employed the algorithm, as outlined in our previous study [5] (Figure 4A, Figure 4B, and Table 2A, Table 2B within [5]), to compute the best ShAP scores. The algorithm encompassed two key aspects: deriving the best ShAP scores: 1) for all 798 significantly altered genes and 2) for all 26 genomic features, both on a per-sample basis. For each sample in the MM and MGUS cohort, the five trained classifiers of the BIO-DGI (PPI9) model predicted the corresponding class, with the ShAP algorithm applied only by classifiers that made correct predictions. ShAP scores for all genomic attributes were collected at the classifier and sample levels. The best ShAP score for each attribute was initially computed at the classifier level and subsequently across all classifiers at the sample level. Regarding significantly altered genes, the ShAP scores of the 26 genomic features were grouped by their positive and negative signs. The best ShAP score for each gene was determined by comparing the absolute values of these grouped scores, considering the largest absolute value among all classifiers as the optimal score. Similarly, for each genomic feature, the ShAP scores of all 798 genes were grouped and assessed in a similar manner, resulting in the best ShAP score. Following this process, the most highly ranked genes and genomic attributes were identified at the group and sample levels.

We extended our analysis by comparing the BIO-DGI (PPI9) model's top-ranked significantly altered genes with those reported in previous studies, aiming to identify genes previously associated with disease progression or suppression. We validated and analyzed our model using information from multiple databases such as OncoKB, IntoGen, COSMIC, and TargetDB at the gene level. For model validation, we extracted 1064 cancer genes from the OncoKB database for OGs and TSGs. From the COSMIC database, we utilized 318 OGs and 320 TSGs.

We utilized the IntoGen database (https://www.intogen.org/) and MM-related studies [87, 222] to compile a catalogue of MM driver genes. Additionally, 180 AGs from COSMIC and 135 from TargetDB helped infer AGs. We systematically categorized the top-ranked significantly altered genes into distinct groups based on their biological characteristics, including OGs, TSGs, ODGs, and AGs. We meticulously compiled the top-ranking genes, both at the group level (MM/MGUS) and the sample level. This comprehensive approach facilitated a thorough exploration of genomic features in the post-hoc interpretability analysis of the BIO-DGI (PPI9) model, providing valuable insights into their roles in disease contexts. Additionally, we introduced a second layer of classification by assessing whether each gene exhibited significant alterations in MM or MGUS. Genes exclusively displaying significant changes in MM were designated as "Transformative" (or disease-progressing genes), while those altered in both MM and MGUS were labeled as "disease initiating". This dual categorization strategy deepened our understanding of these genes, shedding light on their biological functions and their specific relevance to MM and MGUS. We meticulously compiled the top-ranking genes within each category, both at the group level (MM/MGUS) and the sample level. This comprehensive approach facilitated a thorough exploration of genomic features in the post-hoc interpretability analysis of the BIO-DGI (PPI9) model, providing valuable insights into their roles in disease contexts.

5.2.6 Identification of CNVs, SVs and LOFs for top 500 significantly altered genes

Our exploration into significantly altered genes underwent expansion to encompass a broader array of genomic profiles, including copy number variants, structural variants, and loss-of-function events. This extended analysis allowed us to delve into the impact of these variants at the gene level, shedding light on their influence on disease progression. For the MMRF dataset, we leveraged the segment data obtained from MMRF CoMMpass to identify copy number variants (CNVs) using the CNVkit [312] tool. To ensure consistency in our CNV identification workflow, we applied CNVkit to detect CNVs in the WES samples of both AIIMS and EGA datasets. Within this framework, we filtered out genes with a copy number value of 2 across all samples, focusing on genes with copy number values that deviated from 2. Turning to SVs, we utilized the SVs identified through the in-house SV identification workflow developed by the Translational Genomics Research Institute (TGen) and Delly tool [40] to pinpoint SVs in WGS samples from the MMRF dataset. Our analysis centred on significantly altered genes ranked in the top 500 whose genomic regions were affected by SVs, spanning insertions, inversions, deletions, duplications, and translocations.

Furthermore, our investigation extended to encompass genes marked by LOF aberrations. The LOF denotes a disruption in the normal functioning of a gene, impeding the generation of the usual gene product or rendering it ineffectual. To identify the genes having LOF, we assessed all transcripts against the criteria as follows: deletion of over half of the coding sequence, deletion of the start codon, deletion of the first exon, deletion of a splice signal, or deletion causing a frameshift, it was considered to exhibit LOF [313]. This evaluation was conducted across all samples within the MM and MGUS cohorts to pinpoint genes featuring LOF anomalies.

5.2.7 Identification of Haploinsufficient genes

To assess the likelihood of genes exhibiting haploinsufficiency, we draw upon two previously published haploinsufficiency prediction scores: the genome-wide haploinsufficiency score (GHIS) [314] and the DECIPHER score [313]. The DECIPHER score amalgamates patient genomic data, evolutionary profiles, and functional and network properties to predict the likelihood of haploinsufficiency. Meanwhile, the GHIS score draws from diverse large-scale datasets, encompassing gene co-expression and genetic variation in over 6000 human exomes. These comprehensive methods enhance identifying haploinsufficient genes, revealing their crucial role in diseases. This deepens our understanding of genes that lack proper function when only one copy is present.

5.2.8 Identification of significantly altered pathways and pathway ranking

The noteworthy genes highlighted by the BIO-DGI (PPI9) model, instrumental in distinguishing between MM and MGUS, were cross-referenced with the significant gene lists derived separately for MM and MGUS. This mapping was facilitated by the dndscv tool. The outcomes revealed genes shared between both groups, as well as genes significantly mutated either exclusively in MGUS or solely in MM. Moving forward, a pathway analysis was executed on the top 500 genes identified by the BIO-DGI model. To uncover insights, we turned to the Enrichr gene set enrichment analysis web server, which aided in identifying significantly altered KEGG and Reactome pathways associated with the gene set. Subsequently, we proceeded to rank the significantly altered pathways in either the MM cohort, MGUS cohort, or both cohorts, employing their adjusted p-values as a metric. This pathway ranking mechanism provided us with a clear view of the foremost pathways that underwent the most significant alterations due to genomic aberrations in the significantly altered genes.

5.2.9 Identification of gene communities using learned adjacency matrices

We employed a five-fold cross-validation training strategy for our proposed BIO-DGI (PPI9) model. Subsequent to training the model, we retained the learned adjacency matrix from each classifier, yielding five distinct learned adjacency matrices. We proceeded to individually identify gene communities within each of these matrices using the Leiden algorithm. This process yielded 5, 5, 6, 5, and 6 gene communities across the respective learned adjacency matrices. These communities were meticulously curated, and from each fold, the top 3 sub-communities were selected based on the number of OGs, TSGs, ODGs, and AGs. Consequently, we generated a learned adjacency matrix for each fold, encompassing genes from these top 3 sub-communities. To achieve integration across the folds, we merged the learned adjacency matrices derived from genes within the top 3 sub-communities. This amalgamation was accomplished by calculating the mean weight of a gene across all folds. In cases where a gene wasn't present in the top 3 sub-community, its weight was treated as zero within that particular fold. Subsequent to merging the learned adjacency matrices across all five folds, we embarked on another round of community detection. This phase, again utilizing the Leiden algorithm, culminated in the final configuration of five gene communities.

5.2.10 Geo2R validation of top 500 significantly altered genes

We conducted a thorough validation using the Geo2R tool [315] to validate significantly altered genes that were ranked in top 500 in relation to multiple previously published studies focused on MM. In total, we utilized micro-array and miRNA data from 11 distinct MM-related studies for this validation process [99, 316, 317, 318, 319, 320, 321, 322, 323, 324, 325, 326, 327]. To ensure rigorous assessment, we exclusively considered genes that displayed significant dysregulation and maintained an adjusted p-value $\leq$ 0.05 in at least one MM-related dataset.

5.2.11 Workflow for targeted sequencing gene panel

We devised an innovative workflow to pinpoint potential driver genes within the top-ranked significantly altered genes. This comprehensive process incorporated all available genomic profiles, including gene SNVs, CNVs, SVs, LOF, and the count of validated datasets from Geo2R validation. For each of the genes within the top 500, we meticulously collected data from all these profiles. We applied individualized filtering criteria tailored to each variant profile to ensure precision. For instance, when filtering based on the gene SNV profile, criteria included the number of nonsynonymous mutations in the MM cohort, the presence of nonsynonymous or other category mutations in MM cohort samples, and the ratio of nonsynonymous mutations to the total mutations within a sample in the MM cohort. Similarly, for the CNV profile, we considered the presence of CNVs in the gene across MM cohort samples. Analogous criteria were applied for the SV and LOF profiles, focusing on the presence of structural variants and loss-of-function events within MM cohort genes.

Next, we retained genes that met the criteria in at least one variant profile. Finally, we selected genes that satisfied the aforementioned gene prioritization workflow and had at least one dataset validation in Geo2R analysis for further in-depth investigation. The genes that were either disease-transformative or disease-initiating were retained. Detailed workflow for the gene prioritization process is available in Figure-5.6(A), providing a comprehensive view of our approach to identifying potential driver genes. We checked the major molecular aberration for each gene in the 282 genes panel. Further, we checked the coding regions and their genomic locations for the altered regions of all the genes in the 282 genes panel using the UCSC Genome database [328].

5.2.12 Workflow for comprehensive survival analysis

In the end, we performed the survival analysis of 282 genes filtered out using the workflow shown above (Section-5.2.11). We devised a novel workflow for survival analysis to have a better understanding of the impact of gene variant profiles on MM patient survival. In order to identify the genes affecting the patients' survival outcomes significantly, we employed two distinct approaches, as shown below:

In the first approach, we performed the univariate survival analysis for all 282 genes for each variant profile individually (SNV, CNV, SV, and LOF). In this step, we considered one variant profile as a prognostic factor at once. For the SNV profile, we used the total number of (nonsynonymous + other) SNVs as a prognostic factor for univariate survival analysis. Similarly, for the CNV, SV, and LOF profiles, we generated a categorical vector (yes/no) representing whether the gene has CNV, SV, and LOF present in the MM sample or not and then performed univariate survival analysis individually. Finally, for each variant profile, we retained the genes having adjusted p-value $\leq$ 0.05 in univariate survival analysis of individual variant profiles for further analysis. Similarly, in the second approach, we combined the effect of all four variant profiles for each gene in order to get the overall effect of gene variant profiles on clinical outcomes. For this, we employed Factor Analysis of mixed data (FAMD) [329] approach for dimension reduction. Next, we performed the univariate survival analysis using the first FAMD component of all 282 genes as a prognostic factor. Lastly, we retained the genes having adjusted p-value <= 0.05 in univariate survival analysis of FAMD first component for further analysis.

5.2.13 Computational framework used for WES data analysis and survival analysis

We utilized the Python computational framework (version 3.9.13) for WES data analysis and visualization. For training all deep learning models in this study, we employed

PyTorch (version 1.12.0+cu113) [330]. Additionally, survival analysis was conducted using the statistical programming language R (version 4.3.1) with the "survival" package [331] (version 3.5.5).

5.3 Results

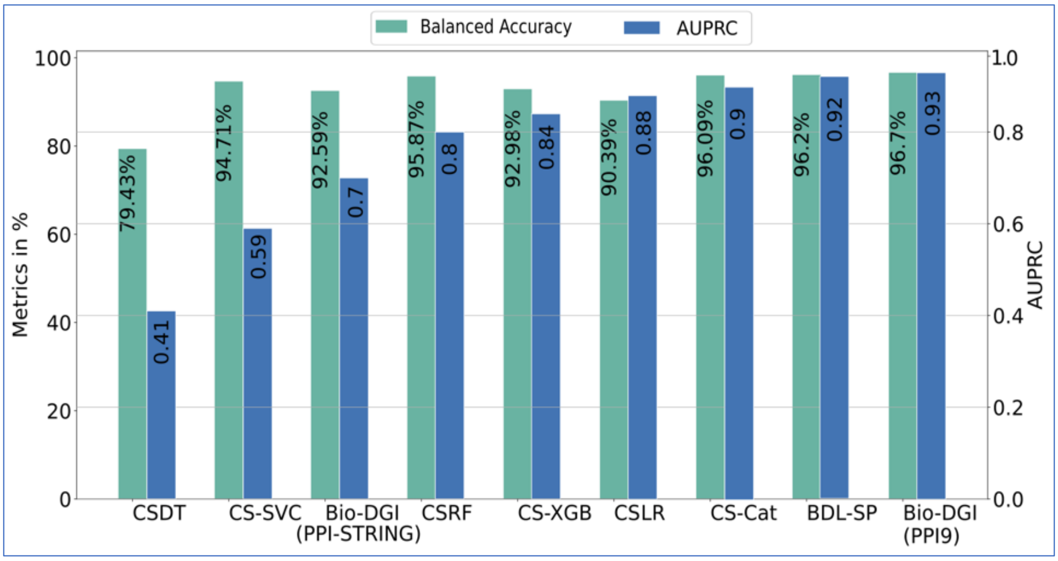
5.3.1 Cohort description

In this comprehensive study, we analyzed two distinct cohorts related to MM and MGUS, encompassing a total of 1154 MM samples and 61 MGUS samples sourced from three globally recognized datasets: AIIMS, EGA, and MMRF. Specifically, within the MM cohort, we examined 1072 samples from the MMRF dataset and 82 samples from the AIIMS dataset. Additionally, in the MGUS cohort, we examined 28 samples from the AIIMS repository and 33 samples from the EGA repository. Augmenting our analysis, we incorporated crucial clinical data, including OS time and OS event data for MM samples retrieved from the MMRF and AIIMS datasets. This enabled a thorough exploration of the clinical relevance of the proposed targeted sequencing panel, underlining the significance of our findings.

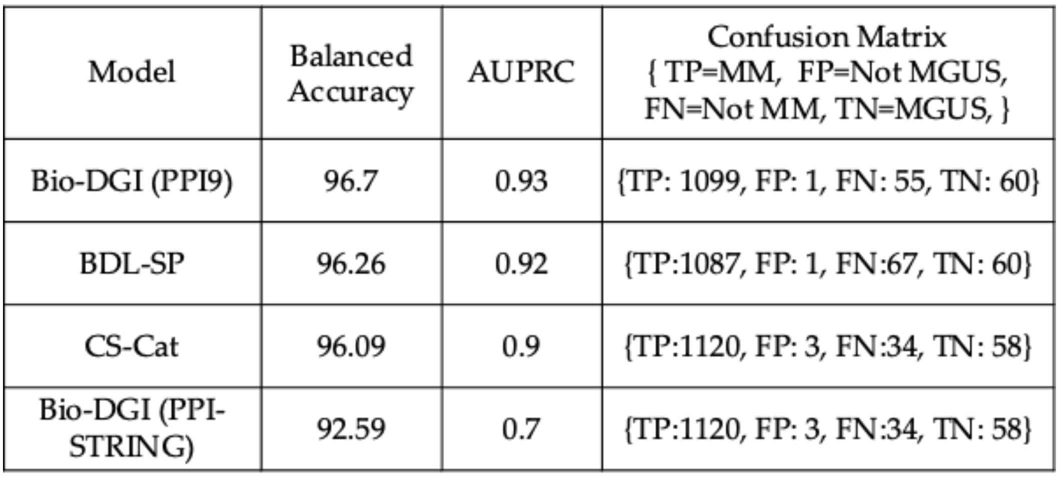
5.3.2 Identification of significantly altered genes

We utilized the dndscv tool (depicted in the pre-processing block of Figure-5.1) to identify significantly altered genes within the MM and MGUS cohorts. A total of 598 and 351 significantly altered genes were pinpointed in the MM and MGUS cohorts, respectively. Among these, 151 genes were found to be common to both MM and MGUS. Subsequently, we proceeded to infer the crucial genes and gene-gene interactions vital for distinguishing between MM and MGUS leveraging our innovative graph-based BIO-DGI (PPI9) model.

(a)



(b)



Model	Balanced Accuracy	AUPRC	Confusion Matrix { TP=MM, FP=Not MGUS, FN=Not MM, TN=MGUS, }
Bio-DGI (PPI9)	96.7	0.93	{TP: 1099, FP: 1, FN: 55, TN: 60}
BDL-SP	96.26	0.92	{TP:1087, FP: 1, FN:67, TN: 60}
CS-Cat	96.09	0.9	{TP:1120, FP: 3, FN:34, TN: 58}
Bio-DGI (PPI-STRING)	92.59	0.7	{TP:1120, FP: 3, FN:34, TN: 58}

Figure 5.2: Quantitative benchmarking of proposed BIO-DGI model. (a) Comparison of balanced accuracy and AUPRC score of BIO-DGI (PPI9) model with other baseline ML and DL models, and (b) Confusion matrix of top-performing models including BIO-DGI (PPI-STRING) model.

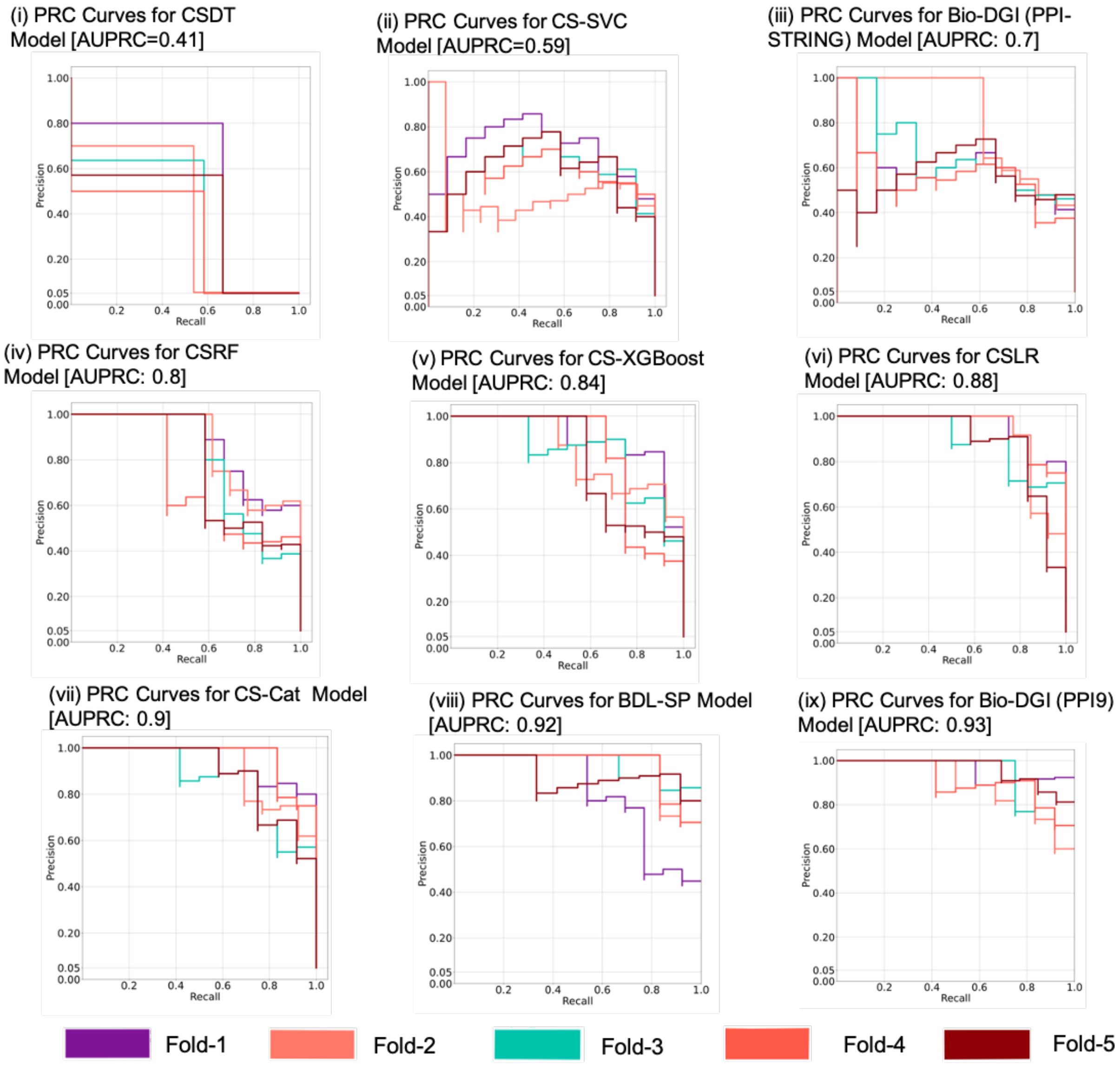


Figure 5.3: Precision-recall curves (PRC) for all five folds of (i) CSDT, (ii) CS-SVC, (iii) BIO-DGI (PPI-STRING), (iv) CSRF, (v) CS-XGB, (vi) CSLR, (vii) CS-Cat, (viii) BDL-SP, and (ix) BIO-DGI (PPI9).

5.3.3 Benchmarking of proposed BIO-DGI (PPI9) model

Employing our AI-driven BIO-DGI workflow (depicted in Figure-5.1), we trained the BIO-DGI (PPI9) models using 5-fold cross-validation and compared its performance with six standard cost-sensitive machine learning and two deep learning models. Remarkably, the proposed BIO-DGI (PPI9) model showcased superior performance in terms of balanced accuracy and AUPRC. Specifically, the BIO-DGI (PPI9) model achieved the largest balanced accuracy at 96.7%. Following closely, the BDL-SP model attained a balanced accuracy of 96.26%, and the cost-sensitive Catboost (CS-Cat) model achieved the third-best performance with a balanced accuracy of 96.09%. The BIO-DGI (PPI9) model also outperformed other models in AUPRC, securing the largest AUPRC score of 0.93, while the AUPRC score for BDL-SP and CS-Cat models stood at 0.92 and 0.9, respectively. Notably, the BIO-DGI (PPI9) model correctly identified 1099 out of 1154 MM samples and 60 out of 61 MGUS samples, showcasing its quantitative superiority.

Similarly, the second-best performing model, BDL-SP, identified 1087 out of 1154 MM samples and 60 out of 61 MGUS samples. Lastly, the third-best performing model, CS-Cat, identified 1120 out of 1154 MM samples and 58 out of 61 MGUS samples. These results affirm that, quantitatively, the BIO-DGI (PPI9) model demonstrated superior performance, with the BDL-SP model being the second best. For a comprehensive understanding of performance metrics (balanced accuracy and AUPRC scores), confusion matrices, and AUPRC curves, refer to Figure-5.2(A), (B), and Figure-5.3 respectively. Given the marginal difference in the balanced accuracy and AUPRC performance metrics among the top three models (BIO-DGI (PPI9), BDL-SP, and CS-Cat), we conducted post-hoc interpretability benchmarking. We applied the ShAP algorithm to identify the top-ranked genes for each of the top three performing models. Subsequently, we analyzed these genes to pinpoint previously reported OGs, TSG, ODG, and AG. Out of the total 798 genes, we identified 31 OGs (e.g., *ABL2, BIRC6, FUBP1, IRS1*), 43 TSGs (e.g., *APC, ARID1B, CYLD, PABPC1, ZFHX3*), 10 ODGs (e.g., *BRAF, FGFR3, TP53, TRRAP*), and 19 AGs (e.g., *ARID2, BRD4, MITF, NF1, TYRO3*) (Table-5.1).

Our analysis revealed that the proposed BIO-DGI (PPI9) model exhibited the largest count of identified OG, TSG, ODG, and AG in both the top 250 and top 500 gene lists (Table-5.2). Specifically, the BIO-DGI (PPI9) model detected 23 and 28 OGs in the top 250 and top 500 gene lists, respectively. Additionally, out of 43 known TSGs, the BIO-DGI (PPI9) model identified 26 in the top 250 and 41 in the top 500 gene lists. Similarly, out of the 10 known ODGs, the BIO-DGI (PPI9) model identified 8 and 9 in the top 250 and top 500 gene lists, respectively. Lastly, out of the 19 known AGs, the BIO-DGI (PPI9) model identified 14 and 19 in the top 250 and top 500 gene lists. The number of previously reported genes identified by other top-performing models (BDL-SP, CS-Cat, BIO-DGI (PPI-STRING)) is presented in Table-5.2. Furthermore, the lists of previously reported genes within the 798 significantly altered genes, ranked within the top 250 and top 500 by the top-performing models, are outlined in Table-5.3. Given the BIO-DGI (PPI9) model's superior identification of previously reported OGs, TSGs, ODGs, and AGs, it stands as the best-performing model and was subsequently used to infer the top significantly altered genes, gene-gene interactions, genomic features, and altered signaling pathways critical for distinguishing MM and MGUS. This analysis underscores the importance of model interpretability within the application domain, even when obtaining similar quantitative results with different machine learning models.

Table 5.1: Types of four different gene categories (OG, TSG, ODG, and AG) and their counts in 798 significantly altered genes

Gene type based on functionality	Total number of previously reported genes present in our list of 824 significantly altered genes
Oncogenes	31
Tumor-suppressor genes	43
Both oncogene and driver gene	10
Actionable genes	19

Table 5.2: Counts of previously reported 4 categories of genes as found in the post-hoc analysis based top 250 and top 500 genes of the top 3 models (BIO-DGI (PPI9), BDL-SP, CS-Cat, and BIO-DGI (PPI-STRING))

Top Genes	BIO-DGI (PPI9) (Top-performing model)				BDL-SP (Second best model)				CS-Cat (Third best model)				BIO-DGI (PPI-STRING) (Baseline version of BIO-DGI)			
	OG	TSG	ODG	AG	OG	TSG	ODG	AG	OG	TSG	ODG	AG	OG	TSG	ODG	AG
top 250	**23**	**26**	**8**	**14**	20	21	7	11	0	0	0	0	18	24	7	13
top 500	**28**	**41**	**9**	**19**	27	37	8	17	0	0	0	0	28	41	9	19

The number of previously reported genes (OG/TSG/ODG/AG) obtained in each category (top 250/top 500) using the best performing model is highlighted in bold.

models, we have considered only those genes in the top 250 or top 500 gene list that have a non-zero ShAP score in the post-hoc explainability analysis. The total counts of previously reported genes as found in the top 250 and top 500 genes of the top-three models is shown in Table-5.2.

5.3.4 Identification of significantly altered pathways and ranking of pathway

In comparing the top 500 significantly altered genes identified by the BIO-DGI (PPI9) model, crucial for distinguishing MM from MGUS, with the significant gene list obtained for MM and MGUS using the dndscv tool, 397 genes were significantly altered exclusively in the MM cohort, 197 genes were exclusively altered in the MGUS cohort, and 94 were found to be significantly altered in both MM and MGUS cohorts, as outlined in Table-S2 (Supplementary File-3).

Within the MM cohort, notable previously reported significantly altered genes were present, encompassing *ASH1L, BRAF, HLA-A/B/C, IGLL5, KMT2B/C/D/E, KRAS, TP53,*

Table 5.3: List of 4 categories of previously reported genes as found in the post-hoc analysis based on top 250 and top 500 genes of the top 3 models (BIO-DGI (PPI9), BDL-SP, and BIO-DGI (PPI-STRING))

Genes	Bio-DGI (PPI9)		BDL-SP Model		Bio-DGI (PPI-STRING)	
	OG	AG	OG	AG	OG	AG
Previously reported oncogenes and actionable genes in MM and MGUS as found ranked in top-250 during the post-hoc analysis of the model	MUC16, PABPC1, USP6, KRAS, NOTCH1, TRRAP, TP53, KMT2D, BIRC6, VAV1, CARD11, SETD1A, MUC4, BRD4, IRS1, FGFR3, ABL2, PGR, BRAF, FUBP1, NRAS, RPTOR, TERT	KRAS,NOTCH1,TP53,NF1,FANCD2,TYRO3,FGFR3,ARID1B,ARID2,BRAF,NRAS,RPTOR,NFKBIA,BRD4	MUC16, FGFR3, PABPC1, BIRC6, MUC4, IRS1, PGR, MGAM, VAV1, ABL2, MITF, TP53, RPTOR, NRAS, NOTCH1, BRAF, TCL1A, LTB, CARD11, KRAS	NRAS, TYRO3, NOTCH1, FGFR3, BRAF, ARID2, NF1, MITF, TP53, KRAS, RPTOR	PABPC1, KRAS, NOTCH1, TRRAP, BIRC6, SETD1A, TERT, IRS1, MUC16, TP53, KMT2D, NRAS, USP6, MUC4, BRD4, BRAF, RPTOR, FGFR3	ARID1B,ARID2,BRAF,BRD4,FANCD2,FGFR3,KRAS,NF1,NOTCH1,NRAS,RPTOR,TP53,TYRO3
Previously reported oncogenes and actionable genes in MM and MGUS as found ranked in top-500 during the post-hoc analysis of the model	MUC16, PABPC1, USP6, KRAS, NOTCH1, TRRAP, TP53, KMT2D, BIRC6, VAV1, CARD11, SETD1A, MUC4, BRD4, IRS1, FGFR3, ABL2, PGR, BRAF, FUBP1, NRAS, RPTOR, TERT, MGAM, MITF, LTB, MECOM, TAL1	KRAS,NOTCH1,TP53,NF1,FANCD2,TYRO3,FGFR3,ARID1B,ARID2,BRAF,NRAS,RPTOR,RAD54B,RB1,RTEL1,APC,NFKBIA,BRD4,MITF	MUC16, FGFR3, PABPC1, BIRC6, MUC4, KMT2D, IRS1, PGR, MECOM, MGAM, VAV1, TRRAP, BRD4, ABL2, TAL1, MITF, TP53, RPTOR, NRAS, NOTCH1, BRAF, TCL1A, LTB, CARD11, MACC1, TERT, KRAS	NRAS, APC, TYRO3, NOTCH1, RB1, ARID1B, FGFR3, BRAF, FANCD2, BRD4, ARID2, NF1, MITF, TP53, NFKBIA, KRAS, RPTOR	PABPC1, KRAS, NOTCH1, TRRAP, BIRC6, SETD1A, TERT, IRS1, MUC16, TP53, KMT2D, NRAS, USP6, MUC4, BRD4, BRAF, RPTOR, FGFR3, CARD11, ABL2, PGR, FUBP1, VAV1, MITF, MECOM, LTB, MGAM, TAL1	KRAS, NOTCH1, ARID2, NF1, TP53, NRAS, FANCD2, ARID1B, BRD4, BRAF, RPTOR, FGFR3, TYRO3, NFKBIA, RB1, RAD54B, MITF, RTEL1, APC

(B) List of tumor-suppressor genes (TSGs) and both oncogenes and driver genes (ODGs) in top-250 and top-500 genes

Genes	Bio-DGI (PPI9)		BDL-SP Model		Bio-DGI (PPI-STRING)	
	TSG	ODG	TSG	ODG	TSG	ODG
Previously reported oncogenes and actionable genes in MM and MGUS as found ranked in top-250 during the post-hoc analysis of the model	PABPC1, NCOR1, KMT2C, HLA-B, HLA-A, HLA-C, NOTCH1, TP53, EP400, KMT2D, NF1, FANCD2, SDHA, SIRPA, NFKBIA, EGR1, MYH11, MAX, TRAF3, KMT2B, ARID1B, DIS3, IRF1, ARID2, FUBP1, TERT	NRAS, BRAF, FUBP1, KRAS, PABPC1, FGFR3, TP53, TRRAP	HLA-A, SP140, ARID2, PABPC1, CYLD, HLA-C, SAMHD1, SIRPA, SDHA, IRF1, NF1, MITF, TP53, ATP2B3, DIS3, KMT2C, NOTCH1, LTB, HLA-B, TRAF3, EGR1	NRAS, FGFR3, BRAF, LTB, PABPC1, TP53, KRAS	NCOR1, KMT2C, MYH11, PABPC1, NOTCH1, HLA-A, SDHA, HLA-C, ARID2, NF1, TERT, EP400, TP53, KMT2D, HLA-B, DIS3, FANCD2, KMT2B, EGR1, SIRPA, ARID1B, TRAF3, RPL10, MAX	PABPC1, KRAS, TRRAP, TP53, NRAS, BRAF, FGFR3
Previously reported oncogenes and actionable genes in MM and MGUS as found ranked in top-500 during the post-hoc analysis of the model	PABPC1, NCOR1, KMT2C, HLA-B, HLA-A, HLA-C, NOTCH1, TP53, EP400, KMT2D, NF1, FANCD2, SDHA, SIRPA, NFKBIA, EGR1, MYH11, MAX, TRAF3, KMT2B, ARID1B, DIS3, IRF1, ARID2, FUBP1, TERT, ZFHX3, MITF, SAMHD1, RPL10, ACVR1B, ATP2B3, RB1, LTB, SP140, CYLD, DDX41, RTEL1, WNK2, AMER1, APC	NRAS, BRAF, FUBP1, KRAS, PABPC1, FGFR3, TP53, LTB, TRRAP	KMT2B, AMER1, RB1, ARID1B, FANCD2, HLA-A, CMTR2, SP140, ARID2, PABPC1, CYLD, MAX, HLA-C, SAMHD1, NCOR1, KMT2D, SIRPA, TERT, SDHA, IRF1, NF1, WNK2, MITF, ATP2B3, TP53, DIS3, ZFHX3, KMT2C, APC, NOTCH1, LTB, HLA-B, ACVR1B, NFKBIA, TRAF3, MYH11, EGR1	NRAS, FGFR3, TRRAP, BRAF, LTB, PABPC1, TP53, KRAS	NCOR1, KMT2C, MYH11, PABPC1, NOTCH1, HLA-A, SDHA, HLA-C, ARID2, NF1, TERT, EP400, TP53, KMT2D, HLA-B, DIS3, FANCD2, KMT2B, EGR1, SIRPA, ARID1B, TRAF3, RPL10, MAX, NFKBIA, RB1, ACVR1B, ATP2B3, FUBP1, MITF, ZFHX3, LTB, AMER1, RTEL1, SAMHD1, IRF1, APC, SP140, CYLD, DDX41, WNK2	PABPC1, KRAS, TRRAP, TP53, NRAS, BRAF, FGFR3, FUBP1, LTB

TRAF2/3, among others. Similarly, the MGUS cohort exhibited previously reported genes like *HLA-B, NOTCH1, PRSS3, TRRAP*, among others. Furthermore, several previously reported genes were found in both MM and MGUS cohorts, including *HLA-B, PRSS3, KMT2C, TRRAP*, among others, illustrating their potential role as shared genomic features in the progression from MGUS to MM. We utilized the Enrichr database to identify significantly altered KEGG and Reactome signaling pathways associated with 397 MM and 197 MGUS genes. In the MGUS group, 7 KEGG pathways and 10 Reactome pathways exhibited significant alterations (refer to Table-S7 in Supplementary File-4). Conversely, the MM group displayed more pronounced pathway alterations, with 105 KEGG and 84 Reactome pathways being significantly affected (refer to Table-S8 in Supplementary File-3). To categorize these alterations based on changing significance levels during the MGUS to MM transition, we employed a similar strategy to our previous study [5]. Specifically, we categorized the significantly altered pathways into four distinct groups according to their significance level changes during the MGUS to MM transition:

1. Category-1: Pathways increasingly significant during MM progression from MGUS.
2. Category-2: Pathways decreasing in significance during the MGUS to MM transition.
3. Category-3: Pathways significantly altered specifically in MM but not in MGUS.
4. Category-4: Pathways significantly altered specifically in MGUS but not in MM.

The complete list of significantly altered pathways for these categories is provided in Tables S9 and S10 of Supplementary File-4. In Category-1, 7 KEGG and 8 Reactome pathways became more significant as the disease progressed from MGUS to MM. In Category-2, no KEGG pathways but 1 Reactome pathway displayed reduced significance with disease progression from MGUS to MM. In Category-3, 98 KEGG pathways and 75 Reactome pathways were significantly altered only in MM and not in MGUS. Notably, 13 out of 98 KEGG pathways and 11 out of 75 Reactome pathways showed no overlapping genes with the set of 197 significantly altered genes in MGUS. Lastly, Category-4 revealed no significantly altered KEGG pathways, while 1 Reactome pathway was observed as significantly altered only in MGUS and not in MM.

To determine the top-ranked pathways, we conducted a ranking of significantly altered pathways in MM based on their adjusted p-values (refer to Table-S11, Supplementary File-4). This analysis revealed a selection of MM-related signaling pathways, notably encompassing the PI3K-AKT signaling pathways, antigen processing and presentation, and MAPK signaling pathways, prominently featured among the top-ranking pathways.

5.3.5 Interpretability of BIO-DGI (PPI9) model using ShAP algorithm

We utilized the ShAP algorithm for post-hoc model explainability and ranked genomic attributes based on their influence on the model prediction. Each genomic attribute received a ShAP score, representing its contribution to each class (MM/MGUS). Subsequently, the attributes were ranked at the group-level (MM versus MGUS) accordingly. This ShAP analysis served to provide post-hoc explainability of the trained model, following a methodology akin to that outlined in [5], enabling the ranking of genes and genomic features at both group and sample levels.

By evaluating the ShAP scores assigned to each gene, we identified *MUC6, LILRA1*, and *LILRB1* as the top three genes in MM and MGUS samples among the 798 significantly altered genes (Table-S2 in Supplementary File-3). Furthermore, several previously reported OGs (e.g., *MUC16, USP6, BIRC6, VAV1*), TSGs (e.g., *EP400, HLA-B/C, SDHA, MYH11*), ODGs (e.g., *PABPC1, KRAS, TRRAP, TP53, FGFR3, BRAF*), and AGs (e.g., *NOTCH1, FANCD2, TYRO3, ARID1B*) were highlighted as top-ranked genes.

Similarly, we ranked genomic features based on their impact on the model's prediction using their ShAP scores. In our model training for BIO-DGI (PPI9), a set of 26 genomic features was employed. Notably, the PhyloP score of nonsynonymous SNVs, allele depth of synonymous SNVs, and the total number of other SNVs emerged as the top three genomic features. Figure-5.7 presents the beeswarm plot illustrating the genomic feature ranking from BIO-DGI (PPI9) model post-hoc analysis using ShAP.

5.3.6 Identification of gene communities and most influencing community members

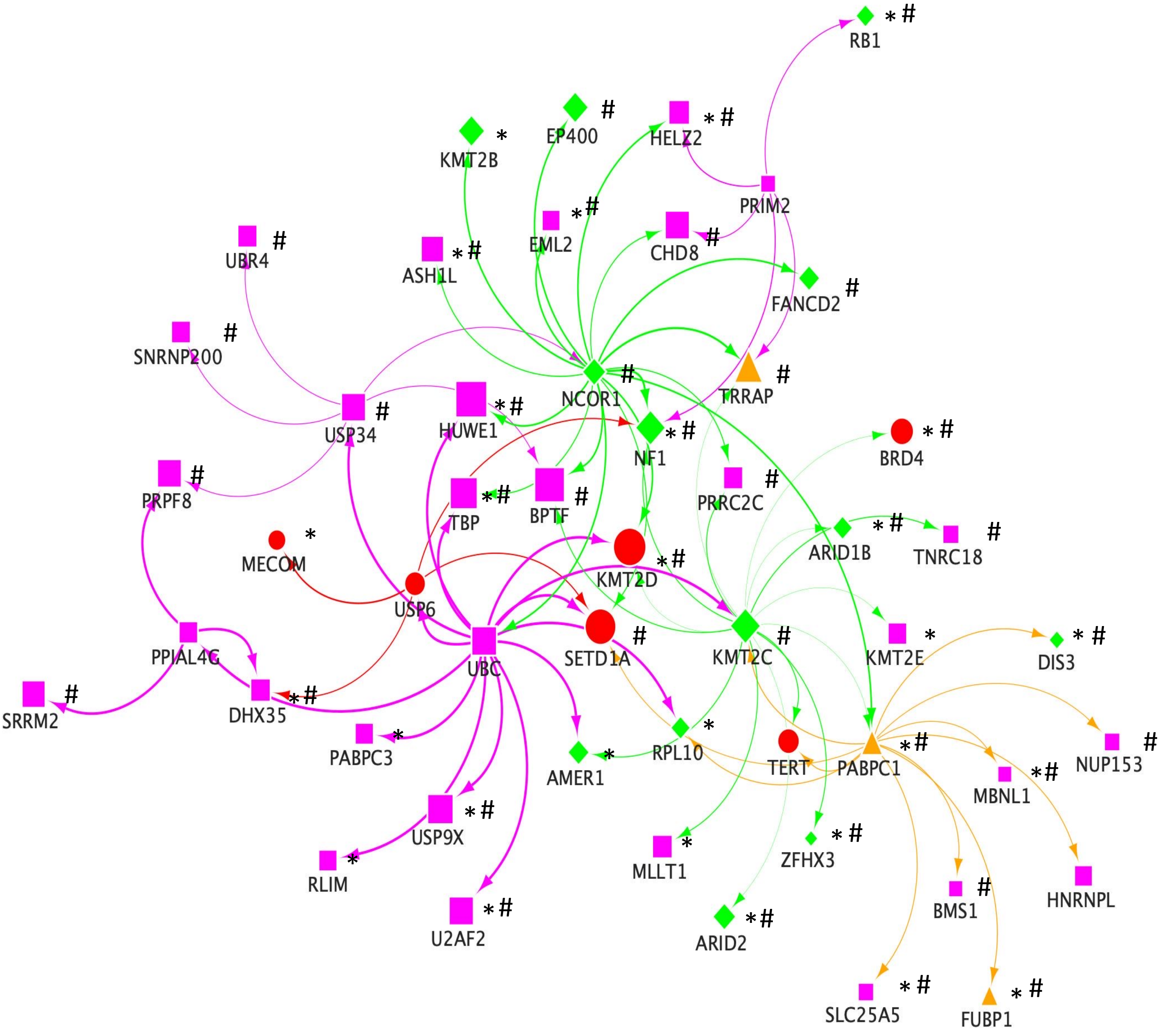
We employed a five-fold cross-validation training strategy to obtain five distinct learned adjacency matrices for five classifiers, each with a dimension of 798x798. We applied the Leiden algorithm to the respective learned adjacency matrix for each classifier to identify gene communities. Consequently, we derived 5, 5, 6, 5, and 6 gene communities using the learned adjacency matrices from the first, second, third, fourth, and fifth classifiers, respectively. Within each classifier, we ranked the communities based on the number of previously reported genes present within them and selected the top three gene communities for each. Subsequently, we merged these top three gene communities for each classifier, resulting in five new distinct learned adjacency matrices with dimensions of 500x500, 500x500, 539x539, 500x500, and 422x422.

In the following step, we merged these five distinct new learned adjacency matrices by computing the mean of gene-gene interactions across the five classifiers. In cases where a specific gene-gene interaction was absent in any fold, we assigned a weight of zero for the corresponding interaction in that fold. This process yielded a final adjacency matrix with dimensions of 690 x 690. Finally, we identified five gene communities from the final learned adjacency matrix using the Leiden algorithm, yielding communities having 202, 125, 122, 104, and 21 genes. The pseudo codes for community detection are provided in Supplementary File-5.

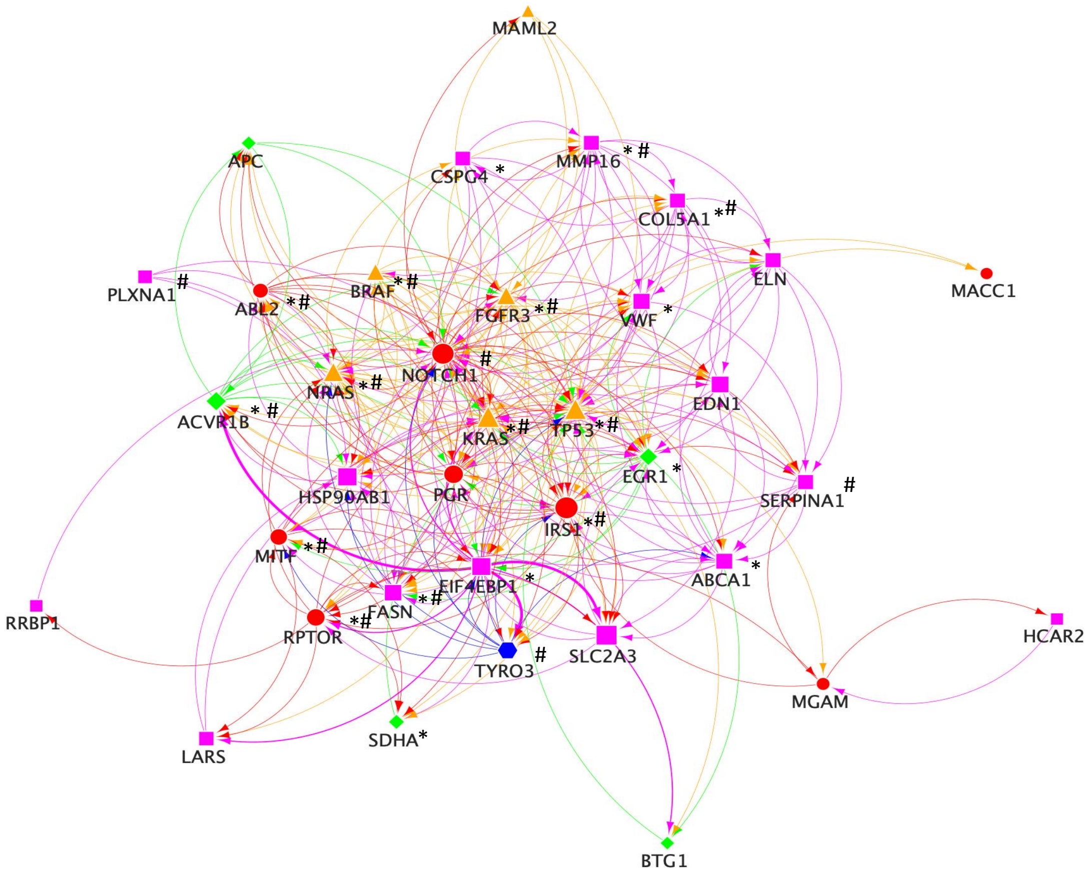
The first gene community, comprising 202 genes, contained 11 OGs, 21 TSGs, 3 ODGs, and 8 AGs. Similarly, the second gene community, with 125 genes, contained 14 OGs, 8 TSGs, 6 ODGs, and 10 AGs. The third gene community, comprising 122 genes, did not include any OGs, TSGs, ODGs, or AGs. The fourth gene community, with 104 genes, contained 4 OGs, 11 TSGs, 1 ODG, and 1 AG. Lastly, the fifth gene community, comprising 21 genes, contained 2 OGs and no TSGs, ODGs, or AGs. The list of genes present in all five gene communities and previously reported genes within each are provided in Table-S12 and Table-S13 of Supplementary File-6. Visualization of all five gene communities, including the top 250 genes and previously reported genes (regardless of their rank), is presented in Figure-5.4(A)-(E).

5.3.7 Geo2R validation of top 500 genes obtained from BIO-DGI (PPI9) model

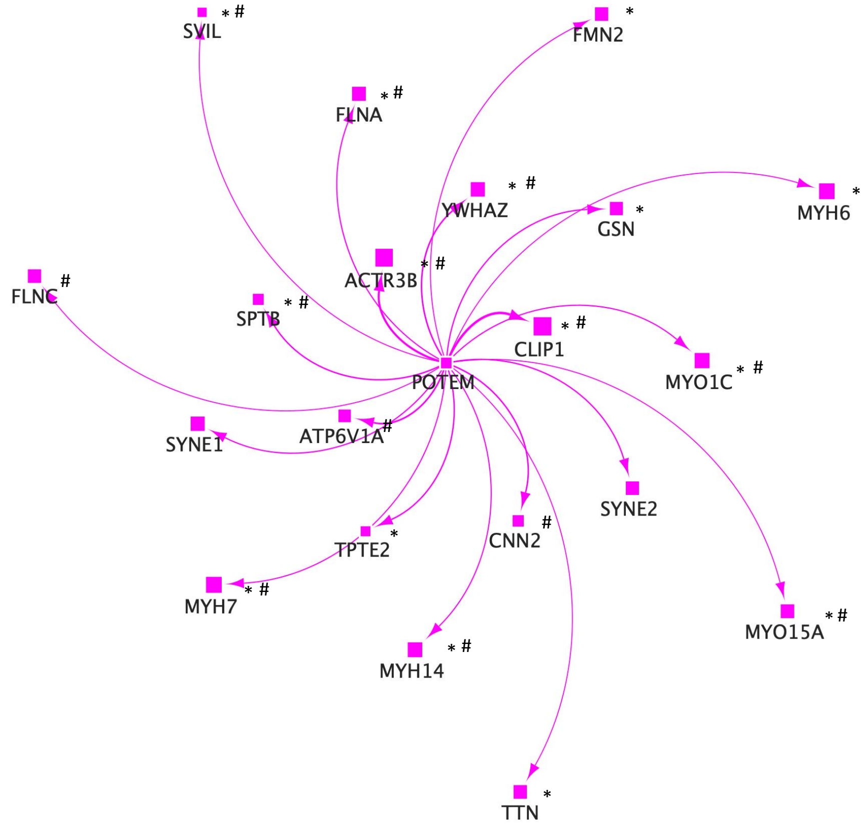
We validated the MM relevance of the top 500 genes obtained from the post-hoc analysis of the BIO-DGI (PPI9) model using datasets from MM-related studies, leveraging the Geo2R tool. We considered 11 MM-related studies for this validation, identifying significantly expressed genes with an adjusted p-value of <= 0.05 and compared them with our top-ranked genes. Remarkably, out of the top-ranked 500 genes, 488 genes were validated in at least one MM-related study. Moreover, within the top 500 genes, 127 (25.4%) and 111 (22.2%) genes were found to be significantly deregulated in MM across datasets related to four and five MM-related studies, as depicted in Figure-5.5(J). Comprehensive details of the Geo2R validation analysis for the top 500 genes are available in Table-S14, Supplementary File-7.



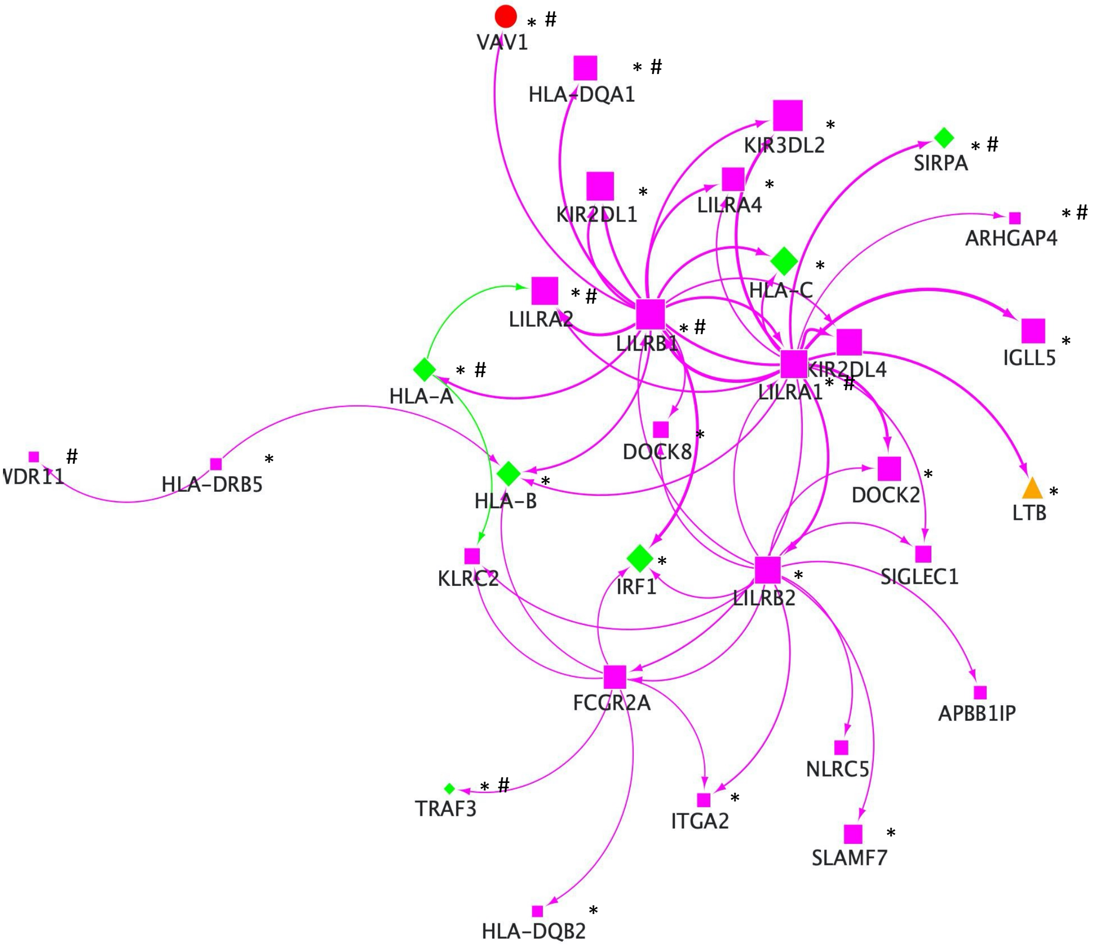
(a) 1st gene community



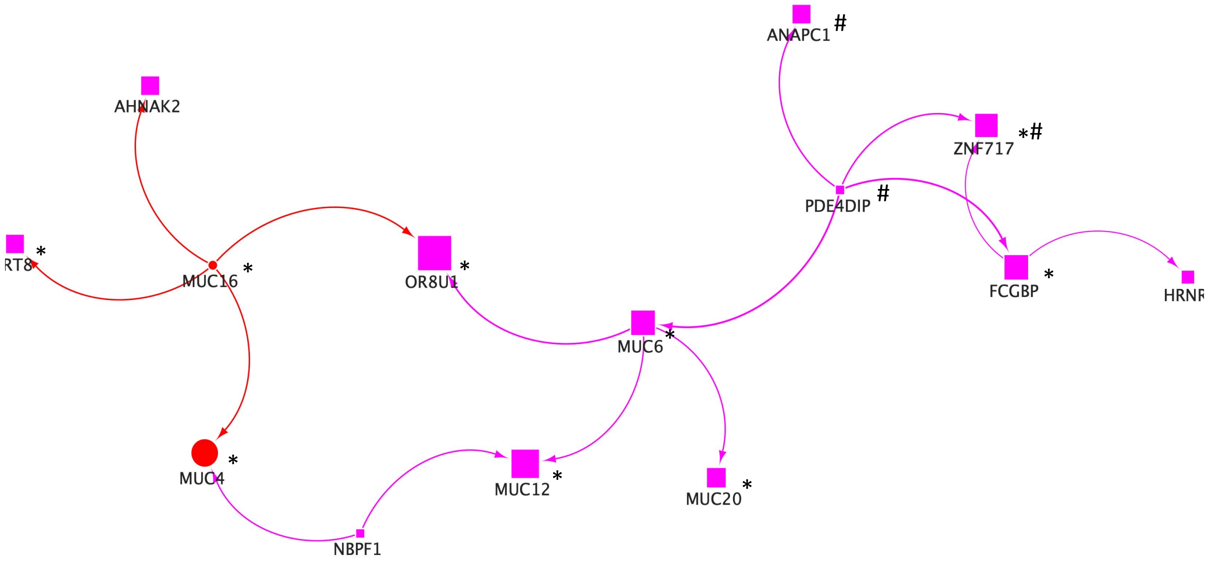
(b) 2nd gene community



(c) 3rd gene community



(d) 4th gene community

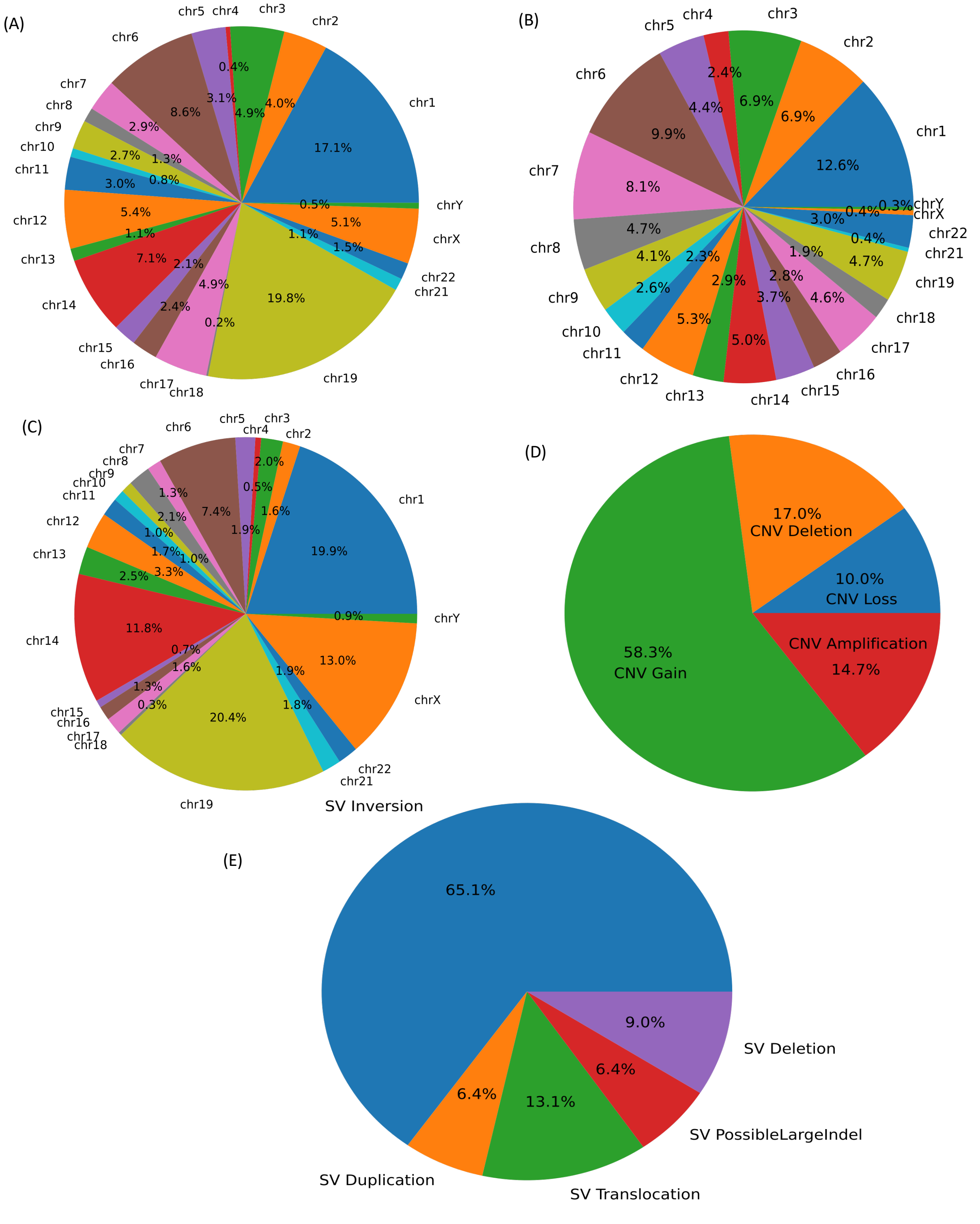


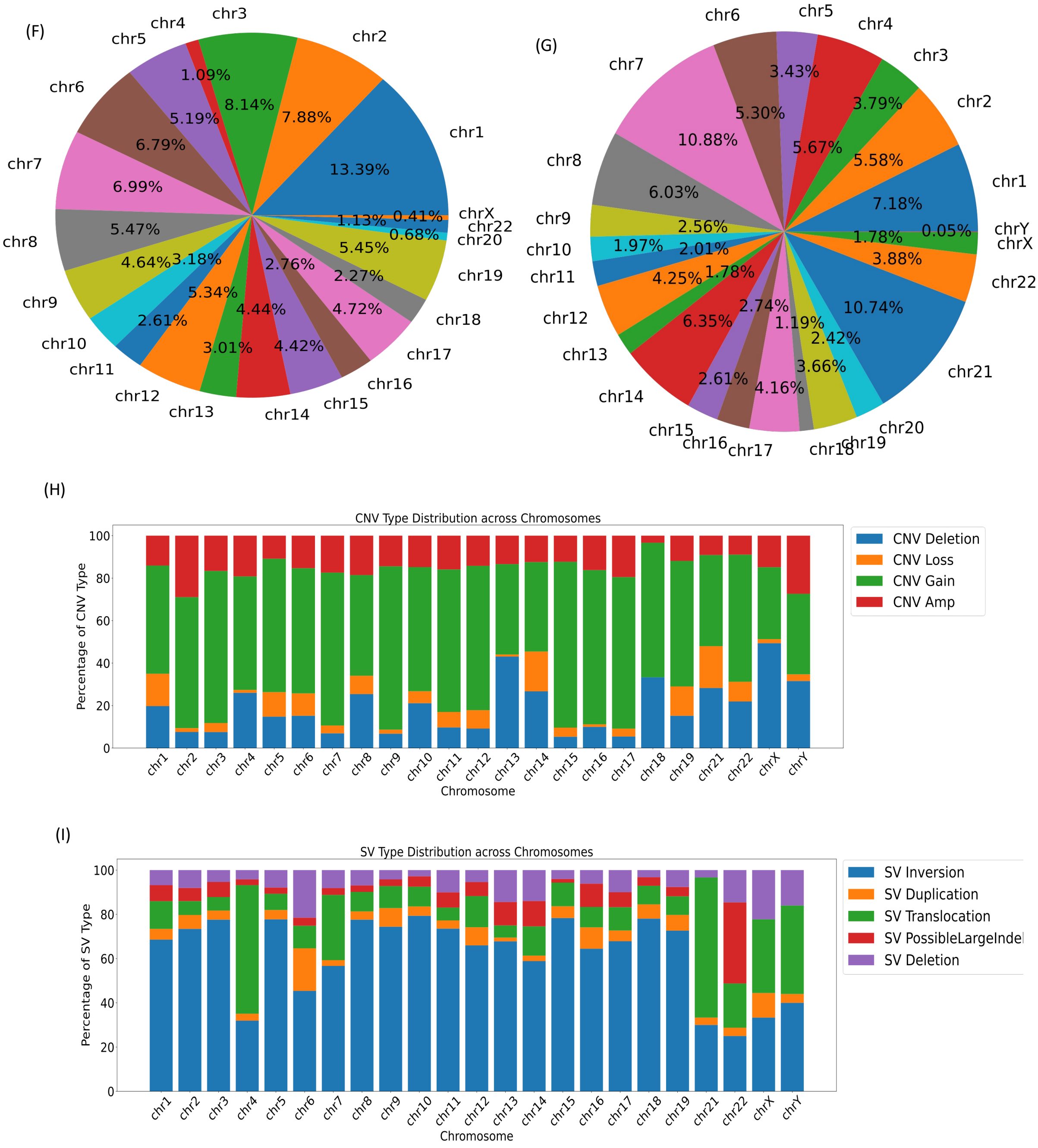
(e) 5th gene community

Figure 5.4 *(previous page)*: Gene community visualization using the learned adjacency matrix obtained from five trained BIO-DGI (PPI9) classifiers. In this figure, (a), (b), (c), (d), and (e) represent the top genes in the first, second, third, fourth and fifth gene communities, respectively. Here, These figures showcase the previously reported genes (OG, TSG, ODG, AG) regardless of their rank, alongside other non-reported genes (in magenta color) within the top 250 ranks, respectively. Genes marked with "*" are included in the 282-genes panel. Additionally, genes marked with "#" possess a high likelihood of being haploinsufficient, with a GHIS score > 0.52.

5.3.8 Analysis of CNVs, SVs and LOF for identifying the key genomic events in MM

In addition to analyzing SNV profiles, we conducted a comprehensive investigation of CNVs, SVs, and LOF in the MM cohort. CNV identification was performed using CNVkit on AIIMS MM samples and on exome segment data from MMRF CoMMpass for MMRF samples. Processed SV data from MMRF CoMMpass was utilized to identify key SVs in MM and 282-genes panel designing. For identifying genes having LOF within a sample, we employed established criteria, evaluating disruptions in gene transcripts due to deletion of essential coding segments, exons, splice signals, or frameshift-inducing deletions [313]. We scrutinized both CNVs and SNVs to identify genes with LOF within each sample. Analysis of CNVs, SVs, and LOF in the top 500 genes revealed crucial molecular aberrations in MM. Chromosome-wise distribution analysis indicated that chr19 (19%), chr1 (17%), chr6 (8.6%), and chr14 (7.1%) were notably affected by CNVs (Figure-5.5(A)). Similarly, chr1 (12.6%), chr6 (9.9%), chr12 (5.3%), and chr14 (5%) showed prominent SV involvement (Figure-5.5(B)), while chr19 (20%), chr1 (19.9%), chrX (13%), and chr14 (11.8%) were most affected by LOF (Figure-5.5(C)). The majority of CNVs were gains (58.3%) and deletions (17%) (Figure-5.5(D)), while inversions (65%) and translocations (13.1%) dominated the SV landscape (Figure-5.5(E)). Notable chromosomes impacted by inversion SV included chr1, chr3, chr2, and chr7 (Figure-5.5(F)), and translocations mainly affected chr7, chr21, chr1, and chr14 (Figure-5.5(G)). The distribution of CNV and SV types within each chromosome highlighted their relative abundance (Figure-5.5(H) and Figure-5.5(I)).


(A)
chr5
chr4
chr3
chr2
chr1
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr21
chr22
chrX
chrY
0.4%
3.1%
4.9%
4.0%
8.6%
2.9%
1.3%
2.7%
0.8%
3.0%
17.1%
0.5%
5.4%
5.1%
1.1%
1.1%
1.5%
7.1%
2.1%
4.9%
19.8%
2.4%
0.2%
(B)
chr4
chr3
chr5
chr2
chr6
chr1
2.4%
6.9%
4.4%
6.9%
9.9%
chr7
12.6%
8.1%
0.3%
chrY
0.4%
chrX
3.0%
chr22
4.7%
0.4%
chr21
chr8
1.9%
4.1%
2.3%
4.7%
chr19
2.8%
2.9%
2.6%
chr9
3.7%
4.6%
chr18
5.3%
chr10
5.0%
chr17
chr11
chr16
chr12
chr15
chr13
chr14
(C)
chr5
chr3
chr6
chr4
chr2
chr7
chr8
chr9
chr10
chr11
chr1
2.0%
0.5%
1.3%
7.4%
1.6%
2.1%
1.9%
chr12
1.0%
19.9%
1.7%
1.0%
chr13
3.3%
2.5%
0.9%
chrY
11.8%
13.0%
chr14
0.7%
1.6%
1.9%
1.3%
chrX
0.3%
1.8%
chr15
chr16
20.4%
chr17
chr18
chr22
chr21
chr19
SV Inversion
(D)
17.0%
CNV Deletion
10.0%
CNV Loss
58.3%
CNV Gain
CNV Amplification
14.7%
(E)
65.1%
9.0%
SV Deletion
6.4%
6.4%
13.1%
SV PossibleLargeIndel
SV Duplication
SV Translocation


(F)
chr3
chr4
chr5
chr2
chr6
chr1
chr7
chrX
chr22
chr20
chr8
chr19
chr9
chr18
chr10
chr17
chr11
chr16
chr12
chr15
chr13
chr14
1.09%
8.14%
7.88%
5.19%
6.79%
13.39%
6.99%
1.13%
0.41%
0.68%
5.47%
5.45%
4.64%
3.18%
2.76%
2.27%
5.34%
4.44%
4.72%
2.61%
3.01%
4.42%
(G)
chr6
chr5
chr4
chr3
chr7
chr2
chr8
chr1
chr9
chrY
chrX
chr10
chr11
chr22
chr12
chr13
chr21
chr14
chr20
chr15
chr16
chr17
chr18
chr19
3.43%
5.30%
3.79%
10.88%
5.67%
5.58%
6.03%
7.18%
2.56%
0.05%
1.97%
2.01%
1.78%
4.25%
1.78%
3.88%
2.74%
10.74%
6.35%
1.19%
2.42%
3.66%
2.61%
4.16%
(H)
CNV Type Distribution across Chromosomes
Percentage of CNV Type
0
20
40
60
80
100
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr21
chr22
chrX
chrY
Chromosome
CNV Deletion
CNV Loss
CNV Gain
CNV Amp
(I)
SV Type Distribution across Chromosomes
Percentage of SV Type
0
20
40
60
80
100
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr21
chr22
chrX
chrY
Chromosome
SV Inversion
SV Duplication
SV Translocation
SV PossibleLargeIndel
SV Deletion

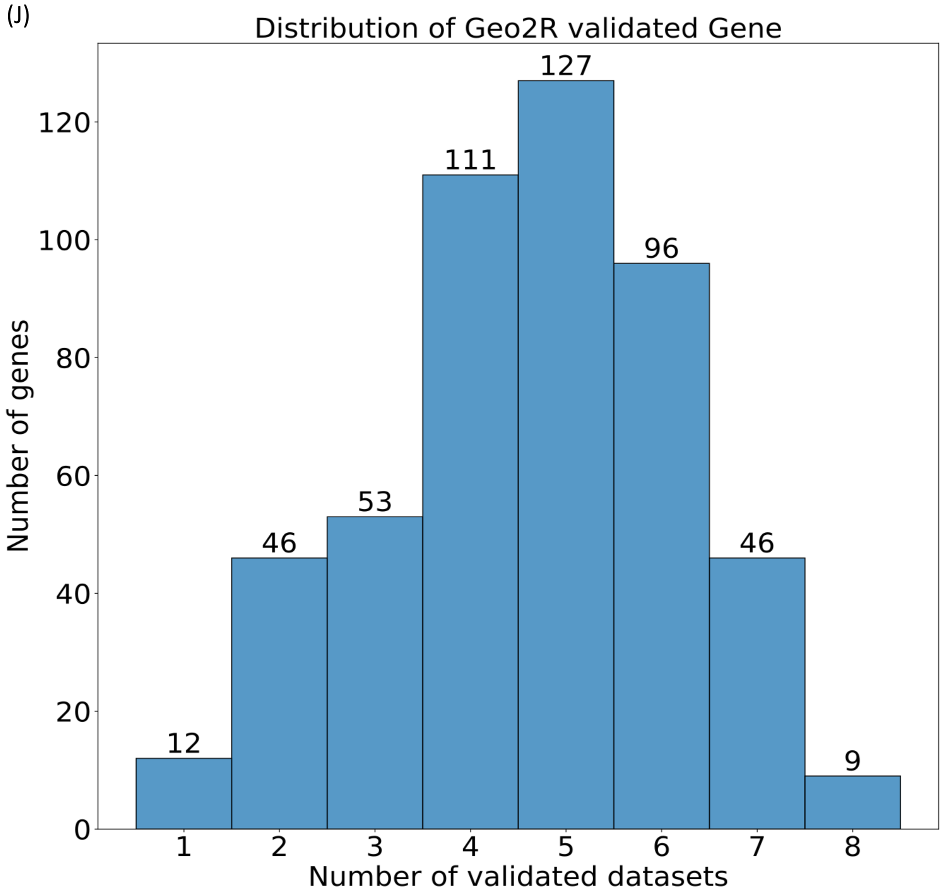
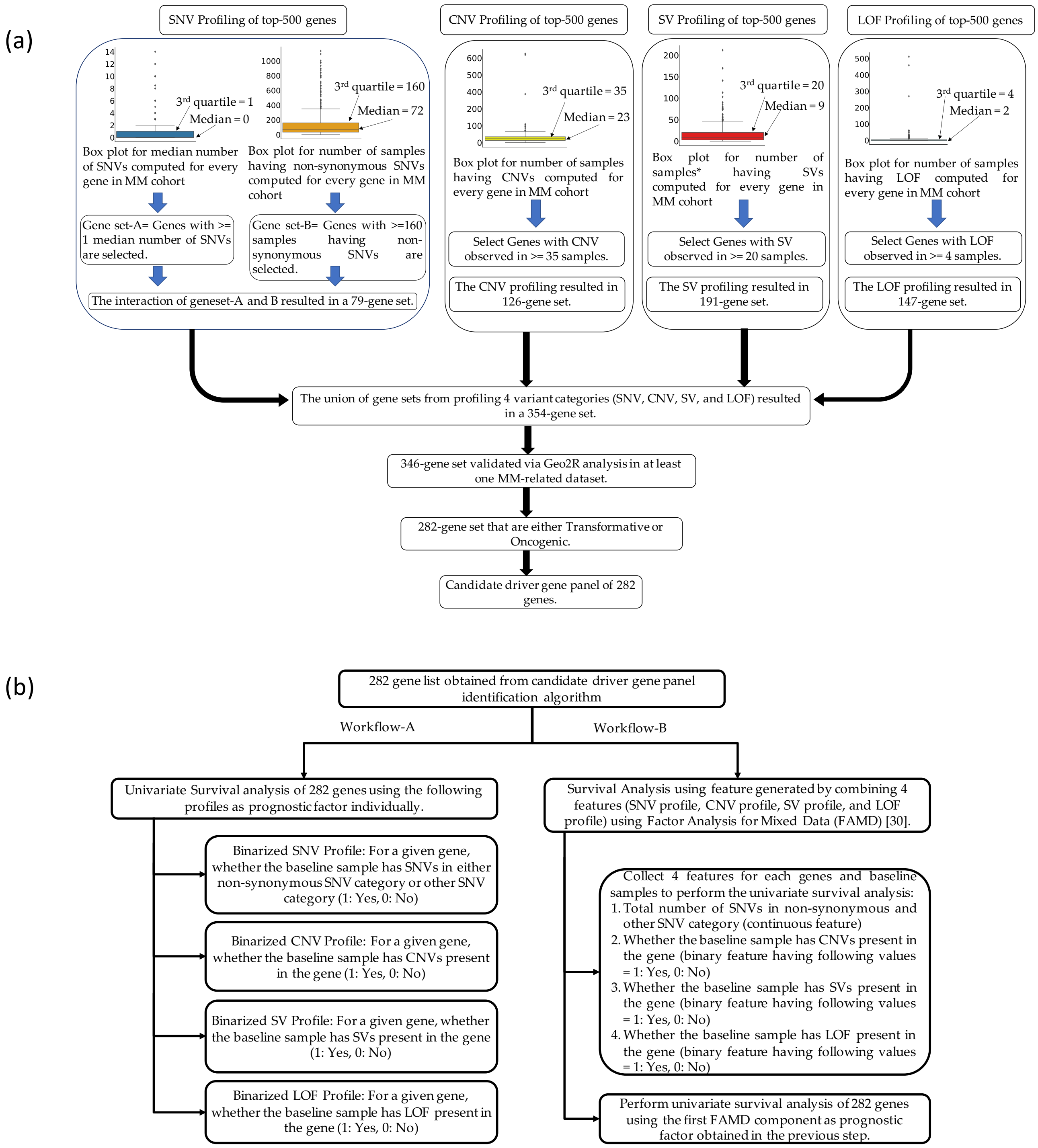


Figure 5.5: Genomic Aberrations Overview (CNVs, SVs, and LOF) in MM Samples from AIIMS and MMRF Repositories. In panels (A)-(C), the figure displays the chromosome-wise distribution of CNVs, SVs, and LOF, respectively. Panel (D) presents the distribution of CNV types identified in MM samples from both AIIMS and MMRF datasets. Similarly, panel (E) shows the distribution of SV types identified in MM samples from the MMRF dataset. Notably, SV analysis was conducted exclusively for MMRF samples due to the absence of WGS data in the AIIMS repository. Continuing SV analysis, panels (F) and (G) exhibit the chromosome-wise distribution of inversions and translocations found in MM samples, respectively. Panels (H) and (I) provide the individual distribution of CNV and SV types for each chromosome. Lastly, panel (J) portrays the distribution of genes validated through MM-related studies using the Geo2R tool. The x-axis represents the number of MM-related studies validating the gene, while the y-axis indicates the count of genes.


(a)
SNV Profiling of top-500 genes
CNV Profiling of top-500 genes
SV Profiling of top-500 genes
LOF Profiling of top-500 genes
14
12
10
8
6
4
2
0
3rd quartile = 1
Median = 0
1000
800
600
400
200
0
3rd quartile = 160
Median = 72
600
500
400
200
100
0
3rd quartile = 35
Median = 23
200
150
100
50
0
3rd quartile = 20
Median = 9
500
400
300
200
100
0
3rd quartile = 4
Median = 2
Box plot for median number of SNVs computed for every gene in MM cohort
Box plot for number of samples having non-synonymous SNVs computed for every gene in MM cohort
Box plot for number of samples having CNVs computed for every gene in MM cohort
Box plot for number of samples* having SVs computed for every gene in MM cohort
Box plot for number of samples having LOF computed for every gene in MM cohort
Gene set-A= Genes with >= 1 median number of SNVs are selected.
Gene set-B= Genes with >=160 samples having non-synonymous SNVs are selected.
Select Genes with CNV observed in >= 35 samples.
Select Genes with SV observed in >= 20 samples.
Select Genes with LOF observed in >= 4 samples.
The interaction of geneset-A and B resulted in a 79-gene set.
The CNV profiling resulted in 126-gene set.
The SV profiling resulted in 191-gene set.
The LOF profiling resulted in 147-gene set.
The union of gene sets from profiling 4 variant categories (SNV, CNV, SV, and LOF) resulted in a 354-gene set.
346-gene set validated via Geo2R analysis in at least one MM-related dataset.
282-gene set that are either Transformative or Oncogenic.
Candidate driver gene panel of 282 genes.
(b)
282 gene list obtained from candidate driver gene panel identification algorithm
Workflow-A
Workflow-B
Univariate Survival analysis of 282 genes using the following profiles as prognostic factor individually.
Binarized SNV Profile: For a given gene, whether the baseline sample has SNVs in either non-synonymous SNV category or other SNV category (1: Yes, 0: No)
Binarized CNV Profile: For a given gene, whether the baseline sample has CNVs present in the gene (1: Yes, 0: No)
Binarized SV Profile: For a given gene, whether the baseline sample has SVs present in the gene (1: Yes, 0: No)
Binarized LOF Profile: For a given gene, whether the baseline sample has LOF present in the gene (1: Yes, 0: No)
Survival Analysis using feature generated by combining 4 features (SNV profile, CNV profile, SV profile, and LOF profile) using Factor Analysis for Mixed Data (FAMD) [30].
Collect 4 features for each genes and baseline samples to perform the univariate survival analysis:
1. Total number of SNVs in non-synonymous and other SNV category (continuous feature)
2. Whether the baseline sample has CNVs present in the gene (binary feature having following values = 1: Yes, 0: No)
3. Whether the baseline sample has SVs present in the gene (binary feature having following values = 1: Yes, 0: No)
4. Whether the baseline sample has LOF present in the gene (binary feature having following values = 1: Yes, 0: No)
Perform univariate survival analysis of 282 genes using the first FAMD component as prognostic factor obtained in the previous step.

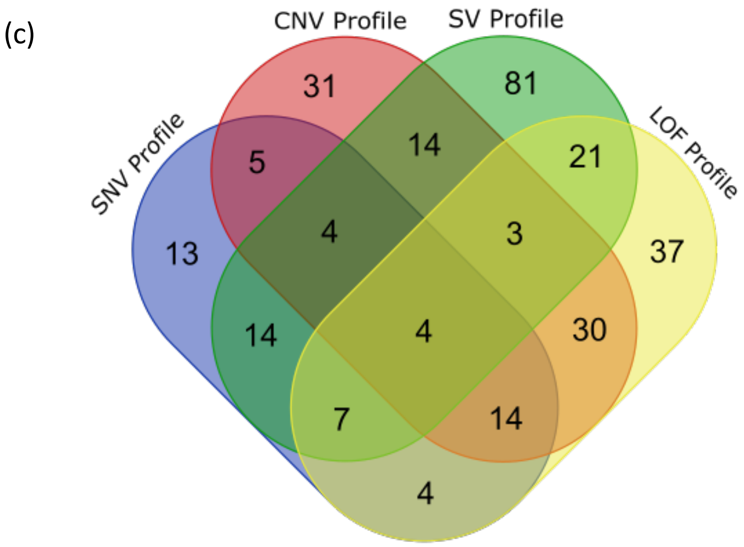


Figure 5.6: (a) Workflow Design for the Proposed 282-Gene Panel. The workflow incorporates variant profiles (SNVs, CNVs, SVs, and LOF) to identify MM-relevant genes. Specifically, 79, 126, 191, and 147 gene lists were generated based on SNVs, CNVs, SVs, and LOF variant profiles, respectively. The union of these lists resulted in a gene set of 354 genes. Following Geo2R validation, 346 genes were retained. Ultimately, disease-initiating (significantly altered in both MM and MGUS) and transformative (exclusively significantly altered in MM) genes were selectively chosen for inclusion in the targeted sequencing panel. (b) Venn diagram showing the overlapping of gene lists obtained after filtering top 500 genes based on their variant profiles. Out of 282, there were 4 genes, namely, *RYR3, HLA-A, HLA-B*, and *HLA-DRB5*, which were found heavily mutated in all four variant profiles. Further, a total of 32 genes were found to be heavily mutated in at least three variant profiles. (c) Workflow for two-fold survival analysis of proposed 282-gene panel. In this workflow, we estimated the clinical relevance of gene variant profiles on MM patient clinical outcomes using two distinct approaches. In the first approach, We individually assessed the impact of each variant profile (SNV, CNV, SV, and LOF) on clinical outcomes. Univariate survival analysis was performed for each variant profile, providing insights into their respective impact. Using this approach, 193 genes out of the 282 genes were found to significantly influence clinical outcomes in univariate survival analysis based on at least one prognostic factor. In the second approach, we amalgamated the four variant profiles (SNV, CNV, SV, and LOF) for each gene using Factor Analysis for Mixed Data (FAMD), enabling us to estimate a joint feature. Subsequently, we performed univariate survival analysis using the FAMD 1st component as a prognostic factor. In this approach, 185 out of 282 genes demonstrated significance in univariate survival analysis based on the FAMD 1st component (a combined feature generated by integrating the four variant profiles for each gene). Intriguingly, 139 genes out of these 185 were also identified as significant in univariate survival analysis. By combining both approaches, out of 282, a total of 239 genes were found to significantly influence the clinical outcomes of MM patients.

5.3.9 Design of 282-genes targeted sequencing panel

To design an effective targeted sequencing panel, we refined the initially identified top-ranked genes based on their significant alterations and the collective impact of their variant profiles in MM. Firstly, we considered four critical variant profiles to pinpoint the candidate driver gene panel: 1. SNV profile, 2. CNV profile, 3. SV profile, and 4. LOF profile. Additionally, we integrated the Geo2R validation profile to specifically include MM-relevant genes in the targeted sequencing panel. Finally, we excluded genes that were neither transformative nor disease-initiating. For the SNV profiling of the top 500 significantly altered genes, we filtered based on the median SNV count and the number of samples with nonsynonymous SNVs, resulting in a set of 79 genes. The features extracted for SNV profile analysis are detailed in Table-S3, Supplementary File-3. Moving on, the variant profiling for CNV, SV, and LOF involved filtering genes based on the number of samples exhibiting that particular variant type, yielding sets of 126, 191, and 147 genes, respectively. The features extracted for CNV, SV, and LOF profile analysis can be found in Table-S4-S6, Supplementary File-3.

By combining genes from SNV, CNV, SV, and LOF variant profiles, we arrived at a comprehensive set of 354 genes. To ensure relevance, we retained genes validated in at least one MM-related study using Geo2R validation. Out of the 354 genes, 346 genes were validated through Geo2R validation analysis. In the final selection, we focused on 212 transformative and 70 disease initiating genes, resulting in the 282-gene panel (Table-S15, Supplementary File-8). The workflow for designing the 282-genes panel is illustrated in Figure-5.6(A). In this panel, four genes, namely, HLA-A, HLA-B, HLA-DRB5, and RYR3, were found to be heavily mutated in all four variant profiles. Additionally, 120 and 32 genes were substantially mutated in at least two and three variant profiles, respectively (as shown in Figure-5.6(B)). For each gene, we determined the most prevalent molecular aberration, such as CNV gain, CNV loss, SV translocation, LOF, etc. We observed that CNV gain was the most frequent molecular aberration found in 188 out of the 282 genes, while LOF was the least common, identified in 12 out of the 282 genes. We assessed the most affected coding regions using the UCSC Genome database to further refine the targeted sequencing regions. The targeted sequencing panel of 282 genes covered a total of 9,272 coding regions in the human genome, spanning a genomic region with a total length of 2.577 Mb in the human genome (Table-S16, Supplementary File-8).

5.3.10 Comparison of proposed 282-genes panel with previously published MM targeted sequencing gene panel

We conducted a thorough evaluation of our proposed 282-gene panel, comparing it with five previously published targeted sequencing panels used for MM genomic profiling. These panels were thoughtfully crafted based on MM-related literature and underwent validation using diverse methods such as FISH and analysis of WGS data, etc. Upon scrutinizing the validated variant profiles, we noted that, alongside our proposed panel, Sudha et al. [120] also validated their panel on SNVs, CNVs, and SVs, encompassing translocations linked to IGH and MYC. However, Sudha et al.'s panel validation was carried out on WGS cohorts of MM samples and MM cell lines. It did not account for potentially distinguishing genomic biomarkers between MGUS and MM.

Moreover, our panel incorporated MM-relevant genes exhibiting loss-of-function (LOF), a critical consideration lacking in previous panels. Notably, Kortum et al.'s panel lacked validation for MM-related CNVs and translocations, Bolli et al.'s and White et al.'s panels lacked validation for translocations involving the MYC gene, and Cutler et al.'s panel lacked validation for any MM-related translocations. Comparing the genes across the previously published panels, we found that 16 out of 47 (34%) genes were common with Kortum et al.'s, 19 out of 182 (10.43%) with Bolli et al.'s, 39 out of 465 (8.38%) with White et al.'s, 15 out of 26 (57.69%) with Cutler et al.'s, and 33 out of 228 (14.5%) with Sudha et al.'s panels, respectively. The comprehensive gene list encompassing all genes from the five panels is provided in Table-S17, Supplementary File-9. Additionally, a detailed comparison of these panels is presented in Table-5.4 and Table-S18, Supplementary File-9.

5.3.11 Clinical relevance of targeted sequencing 282-genes panel

We performed a two-fold univariate survival analysis on a targeted sequencing panel comprising 282 genes to comprehend how gene variant profiles affect clinical outcomes in MM patients (Figure-5.6(C)). To gauge the effect of gene variant profiles on MM sample clinical outcomes, we utilized two distinct approaches. In the first approach, we individually assessed the impact of each variant profile (SNV, CNV, SV, and LOF) on clinical outcomes using univariate survival analysis. Notably, 193 out of the 282 genes significantly influenced clinical outcomes based on at least one variant profile. Out of these, 29, 137, 25, and 76 genes significantly impacted clinical outcomes based on SNV, CNV, SV, and LOF variant profiles as prognostic factors, respectively (Table-S15, Supplementary File-8).

In the second approach, we amalgamated all four variant profiles into a single feature

Table 5.4: Comparison of previously published targeted sequencing panels with our proposed 282-genes panel

S. No.	Panel Reference, Publication year	Total number of genes in the proposed gene panel	Number of samples used for panel validation	Data Type	Detected variant profiles	Overlapping with 282-genes panel
1	Kortum et al [119], 2015	47	22 NDMM, 3 pretreated MM samples	WES	SNVs, clonal evolution analysis	16
2	Bolli et al [121], 2016	182	5 MM samples	WGS	SNVs, CNVs, SVs(Ig)*	19
3	White et al [122], 2018	465	110 MM samples	WGS	SNVs, CNVs, SVs(Ig)*	39
4	Cutler et al. [123], 2021	26	76 (20 MGUS, 3 SMM, 52 MM, and 1 PCL) samples	WGS	SNVs, CNVs, Clinical validation using survival analysis	15
5	Sudha et al [120], 2022	228	185 MM samples	WGS	SNVs, CNVs, SVs(Ig)*	33
6	Vivek et al. (Current study)	282	1215 (1154 MM and 61 MGUS) samples + 11 MM-datasets	WES, microarray, mRNA	SNVs, CNVs, SVs, clinical validation using two-fold survival analysis	-

"*": SVs(Ig) represents the translocation structural variation involving IgH.

vector using the FAMD method, leveraging the FAMD first component as a prognostic factor for univariate survival analysis. Subsequently, we found that 185 out of the 282 genes significantly influenced clinical outcomes based on the first component of FAMD. Upon combining the clinically relevant genes obtained from the two approaches mentioned above, we discovered that 239 out of the 282 genes were clinically relevant for MM. To scrutinize the remaining 43 genes that did not show significance in any of the mentioned approaches, we meticulously examined them. We retained these genes in the proposed gene panel as these genes were heavily mutated in at least one variant profile (Table-S15, Supplementary File-8).

5.4 Discussions

MM is a malignancy that typically progresses from premalignant stages, often starting with MGUS[229]. A targeted sequencing panel is important for the precise characterization of genomic alterations to understand the risk of progression, enable timely interventions, and ultimately improve patient outcomes. Recent studies have shed light on the genomic events that drive the transformation from premalignant stages to MM [90, 91, 92, 93]. Moreover, a number of studies have proposed targeted sequencing panels for molecular profiling of MM patients based on previously identified genomic events in MM and MGUS [123, 120, 119, 121, 122]. However, none of these studies have taken into account the design of the panel using biomarkers and gene-gene interactions that have the potential to distinguish MM from MGUS.

In this study, we addressed this challenge by designing a targeted sequencing panel of 282 genes hosting key genomic biomarkers. For designing this panel, we designed an AI-based bio-inspired BIO-DGI (PPI9) model aimed at identifying the key genomic biomarkers and gene interactions. The BIO-DGI (PPI9) model is biologically inspired, learning to discern distinguishing patterns between MM and MGUS using gene-gene interactions and their corresponding genomic features. Genes with a higher number of interactions are deemed more biologically relevant. We specifically considered deleterious SNVs associated with MM and MGUS, resulting in highly MM-relevant, significantly altered genes being ranked at the top. The inclusivity of three global repositories housing MM and MGUS cohorts with diverse ethnicities, the ability of the AI-based workflow to comprehend gene interdependencies, extensive benchmarking aligned with the application's requirements, and rigorous post-hoc analysis collectively render the BIO-DGI (PPI9) model innovative and highly efficient.
In the post-hoc analysis for model interpretability, we utilized the ShAP algorithm to identify the top-ranked genes within the top-performing models. Table-5.1, Table-5.2, Table-5.3 provides an overview of the total number of previously reported genes present in the 798 significantly altered genes and those identified by top-performing models, presenting complete gene lists under top 250 and top 500 ranks. Notably, the BIO-DGI (PPI9) model outperformed by identifying the largest number of previously reported genes, encompassing known OGs such as *BIRC6, MUC4, NOTCH1, PGR, SETD1A, VAV1*, TSGs like *DIS3, EP400, MYH11, SDHA,* ODGs such as *KRAS, NRAS, TP53, TRRAP,* and AGs including *APC, ARID1B, MITF, NFKBIA, TYRO3*. Intriguingly, most of these genes (except ODGs) display high relevance to MM despite not being explicitly reported as MM driver genes. Additionally, our analysis identified *MUC6, LILRA1,* and *LILRB1* as the top three genes contributing significantly to the classification of MM and MGUS, none of which have been previously categorized as OGs, TSGs, ODGs, or AGs in the literature.

Interestingly, the MUC6 gene is associated with the immune system pathway, playing a crucial role in MM development and progression [332]. Similarly, the other two genes, *LILRA1* and *LILRB1*, were found to be associated with the innate immune system pathway. Notably, *LILRB1* has been associated with MM pathogenesis as an inhibitory immune checkpoint for B-cell function in prior studies [278, 279]. We have employed the Geo2R tool to validate the top-ranked genes obtained from post-hoc analysis of the BIO-DGI (PPI9) model. The Geo2R tool is the most widely used tool for identifying significantly dysregulated genes using gene expression or microarray data from previously published studies. We have included 11 MM-related studies for validation and observed that 488 out of 500 genes were found to be disrupted in MM.

This finding ensures the relevance of top-ranked genes in MM.

Notably, the functional significance of nonsynonymous SNVs, as quantified by Phylop scores, emerged as the most prominent genomic feature contributing to the classification. Following closely, the allele depth of synonymous SNVs and the overall count of other SNVs (encompassing non-frameshift insertions/deletions/substitutions, intronic, intergenic, ncRNA_intronic, upstream, downstream, unknown, and ncRNA_splicing SNVs) ranked as the second and third most influential genomic features, respectively (Figure-5.7). The substantial impact of synonymous SNVs across various cancer types has been well-documented [282, 283, 284, 285, 286]. Additionally, we conducted pathway analysis utilizing the Enrichr database to identify significantly altered pathways associated with the top 500 genes. These pathways were then ranked based on their statistical significance (adjusted p-value) to pinpoint the top significantly altered pathways. We observed a noteworthy pattern after comparing the significance of altered pathways with disease progression. Pathways linked to other cancer types were significantly altered in MGUS and became statistically insignificant as the disease progressed from MGUS to MM. In contrast, MM-related pathways, including the immune system, neurodegeneration, PI3K-AKT, MAPK, and NFKBIA pathways, exhibited significant alterations as the disease advanced from MGUS to MM. These intriguing findings prompt further investigation to ascertain if the significantly altered genes associated with these pathways can potentially serve as valuable biomarkers during the early phases of the disease, particularly in MGUS.

We comprehensively analysed CNVs, SVs, and LOF identified in MM samples from both AIIMS and MMRF datasets. Our analysis highlighted chr1, chr14, and chr19 as the most affected chromosomes, displaying various genomic alterations, including CNVs, SVs, and LOFs. Notably, chr1 exhibited significant alterations, such as amp(1q), associated with disease aggressiveness [128, 333], and del(1p), frequently observed in MGUS [128, 334]. Furthermore, chr14 revealed prevalent translocations involving IGH, such as t(4;14), t(14;16), t(14;20), established as biomarkers in MM [128]. Additionally, CNVs linked to chr19, such as gain(19p) and gain(19q), were significantly more prevalent in MM than in MGUS [229]. The intricate interplay between alterations in these chromosomes and other genetic events contributes to increased genomic instability, facilitating the acquisition of additional mutations that promote MM aggressiveness [83].

Moreover, we meticulously curated a 282-gene panel by rigorously analysing variant profiles (SNVs, CNVs, SVs, and LOF) from the top 500 genes. Our focus was solely on MM-relevant genes found disrupted in at least one previously published MM study. To identify pivotal genomic events responsible for MM development and progression, we categorized these events into two groups based on their occurrence at specific disease stages (MGUS or MM or both). Genomic events observed in both MGUS and MM, such as translocations associated with the IGH and MYC genes [91, 128, 335, 336, 337, 130], and amp(1q) [128] and exclusive to MM, including del(13q), del(16q), del(17p), etc. [128] are shown in Table-5.5, Table-5.6.

Table 5.5: List of previously reported genomic events observed in both MM and MGUS and the overlapping of their associated genes with our proposed 282-genes panel

S.No.	A. Genomic Events in MM and MGUS	B. Genes associated with the event (Column-A)	C. References for the genes shown in column-B	D. Whether the gene associated with the genomic event in column-A is present in 282-genes panel	E. If yes, list of associated genes from 282-genes panel	F. If no, associated gene-gene interactions that are present in 282 gene panel
1	t(11;14)	CCND1, BCL-2	[91, 335, 336, 337]	No		BRD4, IRS1, ITPR1, KRAS, KRT8, NRAS, TP53, RB1, SLC25A5, and TAF1
2	t(4;14)	FGFR3	[91, 128, 336, 337, 130]	Yes	FGFR3	
3	t(14;16)	MAF	[336, 337, 338, 91]	No		FLNA
4	t(14;20)	MAFB	[91, 337, 338]	No		HUWE1, USP9X
5	t(6;14)	CCND3	[337, 128]	No		RB1
6	Amp(1q21)	MCL1, CKS1B, ANP32E or BCL9	[128]	No		KPRP, BRD4, PLEC, USP9X
7	Del(17p13)	TP53	[129]	Yes	TP53	
8	KRAS mutations	KRAS	[91]	Yes	KRAS	
9	NRAS Mutations	NRAS	[91]	Yes	NRAS	
10	LTB Mutations	LTB	[91]	Yes	LTB	
11	DIS3 mutations	DIS3	[91]	Yes	DIS3	
12	EGR1 mutations	EGR1	[91]	Yes	EGR1	
13	MYC Rearrangement	IGH, IGL, IGK, NSMCE2, TXNDC5, FAM46C, FOXO3, IGJ, PRDM1	[339]	Yes	FAM46C	

Table 5.6: List of previously reported genomic events in MM but not in MGUS and the overlapping of their associated genes with our proposed 282 genes panel.

S.No.	A. Transformative Genomic Events	B. Genes associated with the event (Column-A)	C. References for the genes shown in column-B	D. Whether the gene associated with the genomic event in column-A is present in 282-genes panel	E. If yes, list of associated genes from 282-genes panel	F. If no, the associated gene-gene alterations present in 282-genes panel
1	Del(13q14)	RB1	[340, 128]	Yes	RB1, DIS3	
3	Del(16q23)	CYLD	[128]	Yes	CYLD	
4	Del(1p21)	CDC14A	[128, 334]	Yes	FAM46C	
5	Del(12p13)	CD27	[337]	No		TRAF2, TRAF3, ATP2B3
6	TP53 Mutations	TP53	[99, 341]	Yes	TP53	
7	BRAF Mutations	BRAF	[342, 343]	Yes	BRAF	
8	Gain(9q)	ABCA1, KCNT1, TRAF2, VPS13A	[344, 345]	Yes	ABCA1, KCNT1, TRAF2, VPS13A	
9	del(14q)	TRAF3	[128]	Yes	TRAF3	
10	del(17p)	TP53	[128]	Yes	TP53	
11	del(8p)	PTK2B, TP53	[346, 347]	Yes	PTK2B, TP53	

The inclusion of key genomic events in the 282 genes panel helps in the early diagnosis of MM in the following ways:

1. Selection of distinguishing biomarkers: Including genes in the gene panel is solely based on their mutational patterns (SNVs, CNVs, SVs, and LOFs) in pivotal differentiating biomarker genes in MM and MGUS. Comparing the significantly altered genes from the MM patient samples with the the mutations in genes associated with the gene panel may help assess whether the sample will progress to MM.
2. MM-associated disease-initiating genomic events: The disease-initiating genomic events are the key genomic events observed in both the MM and MGUS stages. Identifying disease-initiating genomic events suggests the progression of MGUS towards MM and can pinpoint the disease development in the sample. Hence, genes associated with these genomic events can be used as a diagnostic biomarker, indicating the early onset of MM. For instance, the gene *FGFR3* is associated with translocation t(4;14). In addition to this, the *FGFR3* gene has PPI with several critical genes such as *KRAS, NRAS, DIS3, CYLD*, etc. The alterations in *FGFR3* indicate the presence of a primary genomic event (t(4;14)), which is also associated with *IGH* and several key MM-driver genes that may help in triggering this genomic event. Therefore, the presence of translocation associated with *FGFR3* in the sample at the stage of the MGUS suggests the early diagnosis of MM. Identifying mutational patterns in the genes involved in the disease-initiating genomic events paves the way for early detection of MM. Examples of genes involved in disease-initiating genomic events in MM and MGUS are shown in

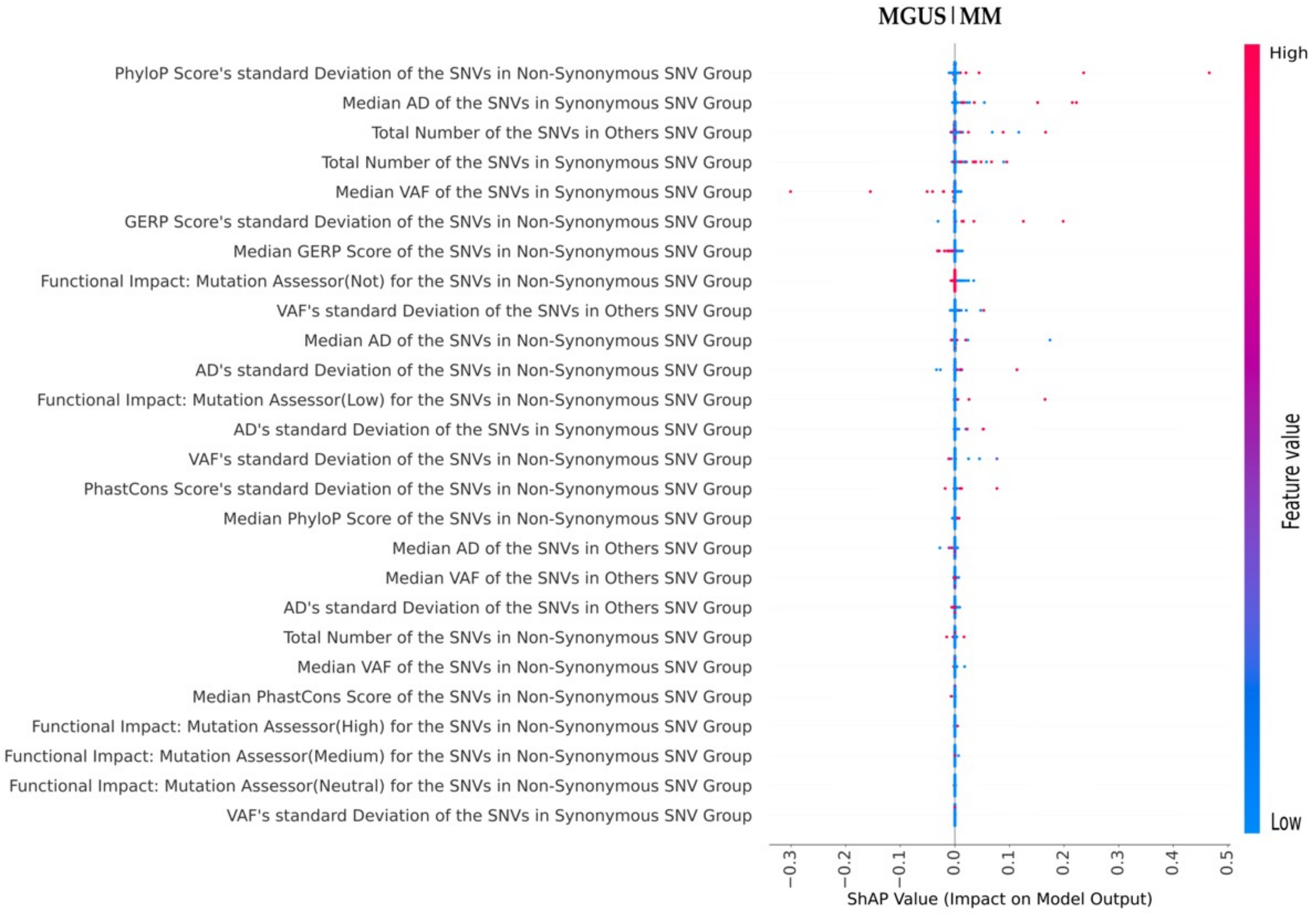


Figure 5.7: Genomic feature ranking using the ShAP algorithm in MM and MGUS based on post-hoc explainability by the BIO-DGI (PPI9) model. Genomic features are ranked according to their ShAP scores, as outlined in our prior research (Figure 4 and Table 2 of reference [5]). A positive ShAP score indicates the feature's contribution to MM, while a negative score represents its contribution to MGUS. Each dot in the scatter plot represents a sample color-coded to reflect genomic feature values—dark blue for low and red for high values.

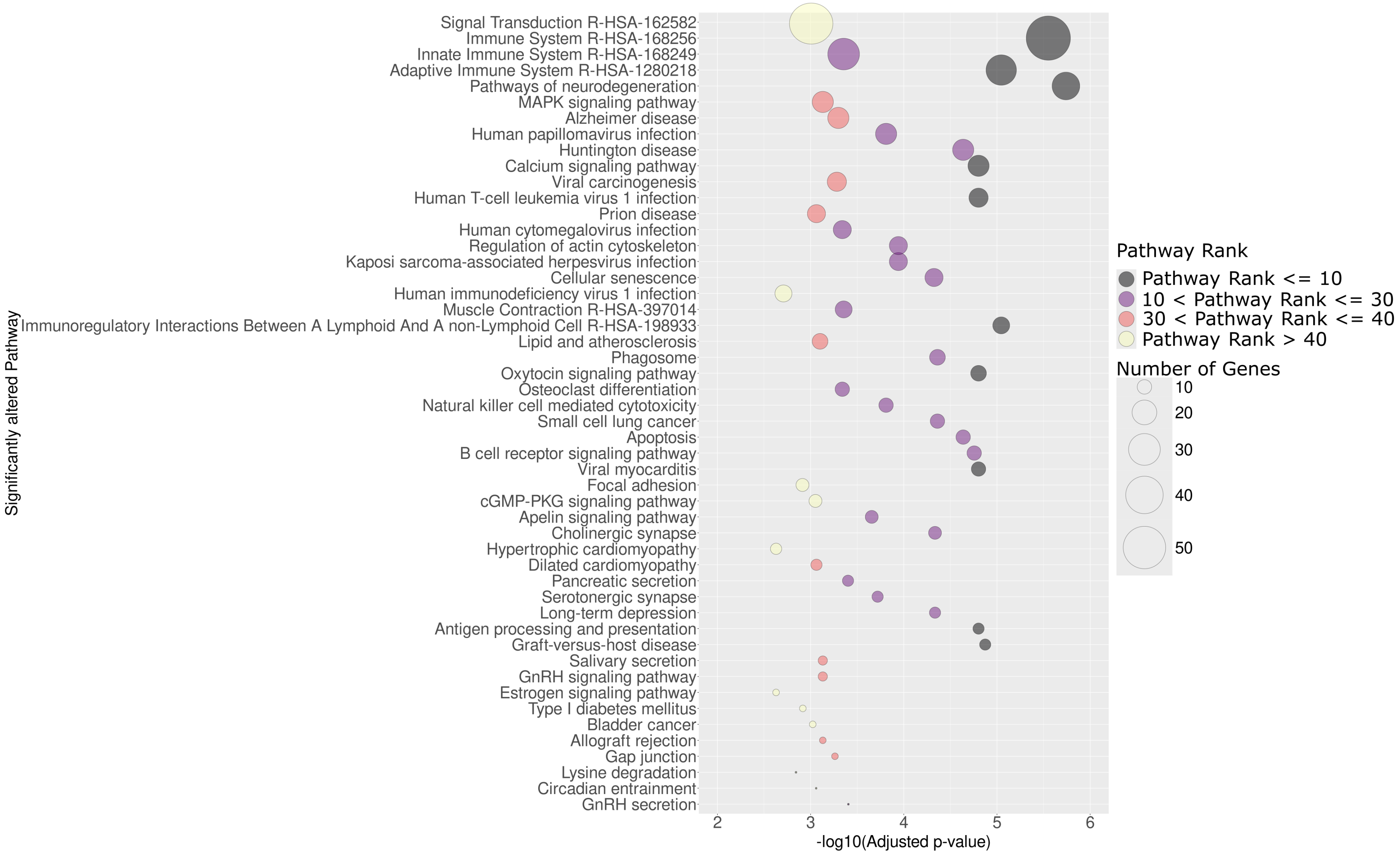


Figure 5.8: A bubble plot illustrating the top 50 significantly altered signaling pathways in MM associated with genes included in the proposed 282-gene panel. The bubble size indicates the number of significantly altered genes linked to the 282-gene panel, and color signifies the pathway rank. The x-axis represents the -log10 (adjusted p-value) score of the pathway, and the y-axis displays the pathway names.

Table Table-5.4.

3. MM-associated disease transformative genomic events: The disease-transformative genomic events are observed in MM only and not in MGUS. Including the genes associated with the disease-transformative genomic events can help predict the progression of MM from MGUS. Hence, genes linked to these genomic events can be used as a prognostic biomarker, reflecting the disease's advancement. For instance, CNV deletion in *RB1, TP53*, and *TRAF3* genes is observed in MM only and not in MGUS. The example of some disease-transformative genomic events and their associated genes are shown in Table Table-5.5.
4. Clinical relevance of the 282 genes panel: The clinical relevance was established through a two-fold survival analysis, ensuring that the genes included in the 282 gene panel are significant for understanding MM progression. The comparison of the significantly altered genes representing the significant association with the clinical parameters (for example, OS time, progression-free survival (PFS), etc.) in a sample with the clinically relevant genes involved in the key disease-initiating and disease-transformative genomic events can help get more insights on the progression of MGUS into MM in that sample.

Notably, out of the 282 genes, 184 were associated with significantly altered pathways. Pathways such as the immune system, signal transduction, gene expression, and RNA polymerase pathways harbored the highest number of significantly altered genes. Additionally, neurodegeneration, immune system, calcium signaling, and antigen processing and presentation pathways were the most significantly altered based on their adjusted p-value. Beyond these pathways, MM-relevant pathways like MAPK signaling, NF-kappaB, PI3K-AKT, and apoptosis pathways ranked prominently among the significantly altered pathways. The bubble plot in Figure-5.8 showcases the top 50 pathways linked with the proposed 282-genes panel and the number of significantly altered genes associated with each pathway and their respective rankings.

We employed the weights acquired from the BIO-DGI (PPI9) model to delineate gene communities. Within each BIO-DGI (PPI9) model classifier, we preserved the learned adjacency matrix. Employing the algorithm outlined in Section-5.2.9 and Supplementary file-5, we identified five gene communities encompassing 202, 125, 122, 104, and 21 genes, respectively. To enhance the information for each node within a gene community, we integrated two additional aspects: node influence determined by the Katz centrality score and likelihood of haploinsufficiency gauged through the GHIS score. We estimated the median GHIS score and highlighted genes surpassing this threshold (= 0.52). Interestingly, numerous previously reported genes exhibited high node influence and GHIS scores. Notably, in the first gene community (Figure-5.4(A)), genes like *UBC, USP6, PRIM2,* and *USP34*; *POTEM* in the third gene community (Figure-5.4(C)); *LILRA1, LILRB1* in fourth gene community (Figure-5.4(D)) acted as central genes and may play a significant role in MM pathogenesis. We strongly recommend further analysis of these central genes to unveil their role in disease progression.

Furthermore, as we explored gene influences within each gene community (depicted in Figure-5.4(A)-(E)), we meticulously examined genes associated with the proposed targeted sequencing panel, revealing high node influence and a high likelihood of haploinsufficiency. Out of the 282 genes, 67 displayed substantial node influence within the gene community, encompassing various previously reported MM-relevant genes like *BRAF, HLA-A/B, FGFR, IRS1, NRAS*, and *SDHA*. Additionally, 74 genes exhibited a high likelihood of haploinsufficiency, including several previously reported MM-relevant genes such as *ARID1B, FGFR, NRAS, TRAF2*, and *ZNF717*. Moreover, 32 genes displayed both high node influence and a high likelihood of haploinsufficiency, including *FGFR, HUWE1, KRAS, KMT2C/D, TP53*, and *ZNF717*.

In examining the gene communities and their involvement in key genomic events of MM, we noted several genes with substantial node influence and likelihood of actively participating in these events. For instance, in the first gene community (Figure-5.3(A)), seven genes (*BRD4, DIS3, HUWE1, RB1, SLC25A5, RB1*, and *USP9X*) were associated with genomic events observed in both MM and MGUS. Similarly, the second gene community (Figure-5.3(B)) included five genes (*EGR1, IRS1, KRAS, NRAS,* and *TP53*) involved in genomic events observed in both MM and MGUS, with one gene (*BRAF*) observed in genomic events observed in MM only. In the third gene community, *FLNA* was found to be associated with genomic events observed in both MM and MGUS. The fourth community featured *LTB* associated with genomic events observed in both MM and MGUS, while *TRAF3* was associated with genomic events observed in MM only. Finally, the fifth gene community had no genes linked to the key genomic events shown in Table-5.5 and Table-5.6. The presence of genes actively participating in MM-related key genomic events, displaying high node influence within the community, and a high likelihood of haploinsufficiency underscores the relevance of our proposed targeted sequencing panel in MM and MGUS.

5.5 Limitations of the study

In this study, we aimed to design the 282 gene panel hosting the key biomarkers in MM. However, the current study has some potential limitations. Firstly, we meticulously analyzed the multiple mutational profiles, such as SNVs, CNVs, SVs, and LOFs, of MM and MGUS cohorts to craft the gene panel. Further clinical validation on diverse patient cohorts is required to ensure panel reliability, accuracy and generalizability. Here, the diverse patient cohort may include transcriptional profiles of MM and MGUS patients, epigenomic data, and longitudinal study comprising multiple time point data of MM and MGUS patients. Including genomic and clinical information from different populations worldwide may uncover crucial clinical findings and support translating findings into clinical practice. Next, our findings and panel design rely on the genomic profiles of

MM and MGUS datasets considered in our current study. Considering the genomic complexity and heterogeneous nature of the disease, similar studies should be continued to account for the dynamic nature of the disease and add biomarkers significantly contributing to the MM pathogenesis and progression. The genomic and clinical profile analysis of newly generated cohorts may shed light on the evolving dynamics of MM. Lastly, consideration of multi-omics analysis, environmental factors, and ethnicity may contribute to identifying the generalized patterns of disease evolution and help design a generalized diagnosis strategy for diverse populations. These factors are crucial contributors to the disease process and warrant further exploration to understand MM pathogenesis comprehensively.

5.6 Conclusion

Distinguishing MM from its precursor stage, MGUS, at the genomic level and identifying those at risk of progression to overt MM presents a formidable challenge due to overlapping genomic characteristics. Unveiling MM underlying pathogenesis necessitates identifying pivotal biomarkers that set MM apart from MGUS. To address this, we propose a clinically oriented targeted sequencing panel of 282 genes aiming for early detection of MM. For the 282-genes panel design, we introduced the novel AI-based bio-inspired BIO-DGI (PPI9) model, encompassing gene interactions from nine PPI databases and exonic mutational profiles from three global MM and MGUS repositories (AIIMS, EGA, and MMRF). The BIO-DGI (PPI9) model demonstrated quantitative and qualitative superior performance, ensuring application-aware interpretability. Notably, the model identified the most previously reported genes, including OGs, TSGs, ODGs, and AGs, which are known for their high relevance in MM. Further exploration of these genes is recommended to unveil novel driver genes. The validation of the top 500 genes set against MM-related datasets using Geo2R confirmed disruption in 488 out of 500 genes, underscoring their pertinence to MM. Similarly, pathway analysis of top-ranked genes further corroborated the relevance of top-ranked genes, revealing a shift in pathway deregulation from MGUS to MM. Key pathways like PI3K-AKT, NFKBIA, and MAPK were prominently altered, emphasizing their role in MM progression. Moreover, in the post-hoc analysis, the functional significance of nonsynonymous mutations, allele depth of synonymous SNVs and total number of other SNVs were found to be the most contributing genomic biomarkers in distinguishing MM from MGUS. Subsequently, the significant alterations on chromosomes 1, 14, and 19 in the MM cohort suggest major inclusion of the genes associated with these chromosomes in MM progression. Notably, CNV gain and SV inversion emerged as prevalent genomic aberrations, with CNV deletion and SV translocation being the second most common molecular aberrations. Through meticulous analysis of variant profiles and validation using Geo2R, we curated a targeted sequencing panel comprising 282 MM-relevant genes. Within this panel, we highlighted genes exhibiting substantial node influence and prominent gene-gene interactions from five gene communities, shed light on crucial gene biomarkers and their interactions pivotal to MM pathogenesis. These observations hold immense potential for informed therapeutic interventions and may facilitate early detection and interception of disease progression in MM.

Chapter 6

Concluding Remarks and Future Works

In this dissertation, we proposed robust and efficient solutions to address challenges in cancer genomics. We successfully validated the significance of our proposed methods qualitatively and quantitatively.

In Chapter 2, we aimed to gain a deeper understanding of the global transcriptional landscape of small non-coding RNAs (sncRNAs) and their potential impact on clinical outcomes. This exploration had the potential to enhance patient stratification, introduce sncRNAs as additional molecular biomarkers for improved prognosis, and open doors to future therapeutic target investigations. To achieve this, we developed a novel method for identifying sncRNAs, encompassing known miRNAs, novel miRNAs, and piRNAs. Our analysis revealed a distinctive pattern of eight dysregulated miRNAs and seven novel miRNAs, which were identified as piRNAs, tRNAs, and snoRNAs using sequence homology with ncRNAs. Subsequent multivariate survival analysis confirmed the clinical significance of these dysregulated miRNAs. Notably, hsa-mir-4524a and hsa-mir-744 were found to be significantly associated with risk and time to first treatment. We also observed that some of the newly identified miRNAs shared common gene targets due to seed sequence similarity. Notably, there were no existing workflows in the literature for this purpose. Additionally, our observation highlighted the substantial influence of configuration parameters on the accuracy of sncRNA identification pipelines, prompting us to further refine the workflow for more robust sncRNA identification. This work has the following possible future directions:

1. Our workflow can be utilized for bulk-RNA Seq data analysis for other diseases such as MM and MGUS. For instance, if bulk-RNA Seq data of MM and MGUS are provided, it would be intriguing to explore how transcriptional patterns evolve with disease progression. By applying our proposed workflow, the researcher could compare the fold change of DEMs, investigate the pathway alterations, and elucidate the regulatory networks involved in the transition from MGUS to MM.
2. In addition to examining sequence homology, several avenues can be pursued for validating novel DEMs. This includes conducting RT-qPCR validation to confirm the expression patterns of identified DEMs. Additionally, expanding the analysis to include larger population-level cohorts can strengthen the identification of novel DEMs and provide insights into their broader relevance. Furthermore, investigating whether novel DEMs are dysregulated in multiple diseases can shed light on their potential roles across different pathological conditions.

In Chapter 3, we continued our pursuit by enhancing the workflow introduced in the previous chapter to more accurately identify miRNAs, including functionally similar miRNAs, also known as paralogues, and piRNAs. Our improved workflow, named

miRPipe, offers several advantages, including parallel batch processing and availability as a Docker image, making it faster and easier to deploy. To comprehensively benchmark miRPipe, we created a synthetic sequence simulator called miRSim, which generates synthetic RNA-Seq data with associated ground truth information. We conducted a rigorous performance assessment of miRPipe at four levels: 1, using synthetic datasets generated by miRSim, 2. using three real cancer datasets (CLL, Lung cancer, and Breast cancer), 3. using literature validation of identified DEMs, and 4. using the workflow ability to identify the reverse complement of known miRNA as known miRNA. miRPipe outperformed existing state-of-the-art pipelines in terms of accuracy (95.23%) and F1-score (94.17%) when benchmarked with synthetic RNA-Seq data. Analysis of all three cancer datasets further demonstrated miRPipe's superior ability to extract a higher number of known dysregulated miRNAs and piRNAs compared to existing pipelines. Furthermore, miRPipe identified the most literature-validated miRNAs as DEMS and the most number of the reverse complement of miRNAs as known miRNAs.
This work has the following possible future directions:

1. We have validated miRPipe on three publicly available datasets. Meanwhile, miRPipe applicability can be expected for other disease-related datasets.
2. Within miRPipe, we incorporated Bowtiw 1 sequence aligner. However, Bowtie 1 is deprecated now, and Bowtie 2 can be replaced in the miRPipe workflow. In future, our focus will be on releasing the next version of the miRPipe docker, incorporating the recent versions of the tools used within the miRPipe workflow.
3. miRPipe is a generic workflow applicable to both human and non-human datasets. Currently optimized for the human genome, miRPipe offers default parameters tailored to human genomic data. Nonetheless, researchers can readily extend miRPipe's utility to non-human genomes by providing the respective reference genome and its index to the sequence aligner, miRDeep*, utilized in Step 3 of the miRPipe pipeline. In the future, our aim will be to release the next version of miRpipe docker with enhanced functionality to process non-human genome datasets efficiently.
4. Akin to miRPipe, miRSim has been tested on generating the synthetic sncRNA seq data for the human genome. The applicability of the miRSim simulator to the non-human genome can be extended by providing the non-human genome reference sequences and adjusting the seed and xseed region location within the miRSim tool. In the future, as time permits, we aim to release the next version, designed explicitly for non-human genomes, ensuring broader applicability and versatility across diverse research domains.

In Chapter 4, we undertook the challenging task of identifying the pivotal biomarkers that can distinguish MM from MGUS. To address this, we designed an AI-based workflow, Bio-inspired Deep Learning architecture for the identification of altered Signaling Pathways (BDL-SP) to identify pivotal genomic biomarkers for distinguishing MGUS from MM. The proposed graph convolutional network-based BDL-SP model can extract discriminative genomic biomarkers for identifying MM and MGUS samples, outperforming all baseline ML-based models. Furthermore, by applying application-aware interpretability analysis to the trained AI model, we demonstrated a method for selecting the best AI model from among multiple machine learning or deep learning models that may have performed similarly in terms of quantitative metrics on the available data. In post-hoc interpretability benchmarking, BDL-SP excelled over all baseline models by identifying the largest number of previously reported genes, including genes not yet identified as MM drivers. Additionally, we conducted pathway analysis on the top-ranked genes and observed that several signaling pathways are selectively and significantly dysregulated with disease progression. Additional mutations in driver genes, critical OGs, TSGs, and AGs contributed to transforming benign MGUS into MM. Similarly, genomic mutations associated with the Synonymous SNV group (synonymous SNVs, UTR3, and UTR5) were found to be the most significantly contributing biomarker distinguishing MM from MGUS. These observations may hold great therapeutic significance. We also noted that the number of OGs, driver genes, and AGs in MGUS samples from European and Indian populations differed statistically, highlighting the impact of ethnicity during the development of MM. Although we did not observe population-specific differences in our analysis of MM data from the American and Indian populations, the results with MGUS data indicate that the impact of ethnicity on the disease biology of MM should be further explored. Further exploration of gene-gene interactions among the top-ranked genes may provide a better understanding of MM pathogenesis.

This work has the following possible future directions:

1. In our study, the BDL-SP model, initially trained on MM vs. MGUS datasets, identifies pivotal biomarkers distinguishing MM from MGUS. While the workflow is generic, the model can be retrained from scratch using normal and MGUS samples. This enables the same workflow to identify pivotal biomarkers distinguishing normal and MGUS samples. Further, for three class classifications, the BDL-SP model can be retrained for three classes, including normal, MGUS, and MM samples, the pivotal biomarkers differentiating normal vs. MGUS samples. Post hoc analysis, benchmarked against baseline machine learning and deep learning models, facilitates the identification of pivotal biomarkers distinguishing normal from MGUS, normal from MM, and MGUS from MM.
2. In this study, we highlighted the impact of nonsynonymous SNVs on protein functions and protein structure. Analyzing protein structure stability in altered states involves considering several structure-related criteria, such as differences in free energy between folded and unfolded states, the impact of hydrogen bonding, etc. In future, the work can be extended to the impact of pathogenic SNVs identified using the BDL-SP model on protein function and protein structure alterations to gain more insights into MM pathogenesis.

In Chapter 5, we introduced a clinically focused targeted sequencing panel consisting of 282 genes with the goal of early detection of MM. For the crafting of this 282-gene panel, we designed an innovative AI-driven Biological Network for Directed Gene-Gene Interaction Learning (BIO-DGI) model. This model incorporates gene interactions from nine PPI databases and utilizes exonic mutational profiles from three global MM and MGUS repositories (AIIMS, EGA, and MMRF). The BIO-DGI model demonstrated exceptional performance both quantitatively and qualitatively, ensuring application-specific interpretability. To validate the relevance of the top-ranked genes, we confirmed their pertinence to MM using the Geo2R tool. Furthermore, pathway analysis of these top-ranked genes reinforced their importance, highlighting a shift in pathway deregulation from MGUS to MM. Through post-hoc analysis of the BIO-DGI model, we determined that the functional significance of nonsynonymous mutations, allele depth of synonymous SNVs, and the total number of other SNVs played significant roles in distinguishing MM from MGUS. This comprehensive analysis allowed us to identify the top-ranked genes and gene communities using five learned adjacency matrices. The meticulous analysis of SNVs, CNVs, SVs, and LOFs profiles of the top-ranked genes led us to create a targeted sequencing panel comprising 282 genes specifically relevant to MM. Within this panel, we identified genes with substantial node influence and prominent interactions within five gene communities, shedding light on crucial gene biomarkers and their pivotal roles in MM pathogenesis. These findings offer great potential for informed therapeutic interventions, as well as the possibility of early disease detection and intervention for MM. Further exploration of the influential genes and their interactions with MM-relevant genes may unveil additional key players in MM pathogenesis beyond well-established driver genes. This work has the following possible future directions:

1. In this study, we designed a 282 genes panel using the gene-gene interactions and multiple mutational profiles (SNVs, CNVs, SVs, and LOFs). To ensure the clinical utility and generalizability of the proposed gene panel and BIO-DGI model predictions, further clinical validation and genomic studies are necessary. These studies should involve diverse patient cohorts representing various populations and ethnicities. We can ensure their effectiveness in real-world clinical settings by assessing the accuracy, reliability, and performance of the panel and model predictions across different demographic groups. This validation step is essential for translating research findings into clinical practice, ultimately benefiting patients with multiple myeloma (MM) and its precursor stages.
2. Recognizing the dynamic nature of MM progression, future efforts should focus on adapting the gene panel to accommodate these changes. The progression from MGUS to MM involves complex genetic alterations that evolve over time. Therefore, continuous updates to the gene panel are necessary to incorporate new genetic discoveries and insights into disease mechanisms. By dynamically adjusting the panel based on emerging genetic data and evolving understanding of MM pathogenesis, we can ensure that it remains relevant and effective in capturing the heterogeneity of the disease.
3. While the study has primarily focused on genetic interactions and SNVs, it is important to acknowledge the significant contributions of environmental factors, epigenetics, and non-genetic variations to MM progression. Future research should explore the interplay between genetic and environmental factors, as well as epigenetic modifications, in shaping MM development and progression. By

incorporating comprehensive analyses that consider these overlooked factors, we can gain a more holistic understanding of MM pathogenesis and identify novel therapeutic targets and prognostic markers. This integrative approach will provide valuable insights into the multifaceted nature of MM and inform personalized treatment strategies.

Finally, it is important to recognize that both sncRNAs and genes can be used as prognostic and diagnostic biomarkers depending on the study hypothesis as both the gene and sncRNAs shed light on disease biology from distinct perspectives. Exploring both perspectives is essential for a comprehensive understanding of disease pathogenesis and biology, ultimately leading to more effective diagnostic and prognostic strategies.

References

[1] Gurvinder Kaur, Vivek Ruhela, Lata Rani, Anubha Gupta, Krishnamachari Sriram, Ajay Gogia, Atul Sharma, Lalit Kumar, and Ritu Gupta. RNA-Seq profiling of deregulated miRs in CLL and their impact on clinical outcome. *Blood cancer journal*, 10(1):1–9, 2020.

[2] Annelynn Wallaert, Wouter Van Loocke, Lucie Hernandez, Tom Taghon, Frank Speleman, and Pieter Van Vlierberghe. Comprehensive miRNA expression profiling in human T-cell acute lymphoblastic leukemia by small RNA-sequencing. *Scientific reports*, 7(1):1–8, 2017.

[3] Simon Andrews et al. FastQC: a quality control tool for high throughput sequence data, 2010.

[4] Ana Kozomara, Maria Birgaoanu, and Sam Griffiths-Jones. miRBase: from microRNA sequences to function. *Nucleic acids research*, 47(D1):D155–D162, 2019.

[5] Vivek Ruhela, Lingaraja Jena, Gurvinder Kaur, Ritu Gupta, and Anubha Gupta. Bdl-sp: A bio-inspired dl model for the identification of altered signaling pathways in multiple myeloma using wes data. *American Journal of Cancer Research*, 13 (4):1155, 2023.

[6] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[7] Diana Sarfati and Jason Gurney. Preventing cancer: the only way forward. *The Lancet*, 400(10352):540–541, 2022.

[8] Lina van der Straten, Mark-David Levin, Otto Visser, Hidde Posthuma, Jeanette Doorduijn, Arnon Kater, and Avinash Dinmohamed. Survival continues to increase in chronic lymphocytic leukaemia: a population-based analysis among 20 468 patients diagnosed in the netherlands between 1989 and 2016. *British Journal of Haematology*, 2020.

[9] Sigurdur Y Kristinsson, Paul W Dickman, Wyndham H Wilson, Neil Caporaso, Magnus Björkholm, and Ola Landgren. Improved survival in chronic lymphocytic leukemia in the past decade: a population-based study including 11,179 patients diagnosed between 1973–2003 in sweden. *haematologica*, 94(9):1259, 2009.

[10] Lau Caspar Thygesen, Ove Juul Nielsen, and Christoffer Johansen. Trends in adult leukemia incidence and survival in denmark, 1943–2003. *Cancer causes & control*, 20:1671–1680, 2009.

[11] Graça M Dores, William F Anderson, Rochelle E Curtis, Ola Landgren, Evgenia Ostroumova, Elizabeth C Bluhm, Charles S Rabkin, Susan S Devesa, and Martha S Linet. Chronic lymphocytic leukaemia and small lymphocytic lymphoma: overview of the descriptive epidemiology. *British journal of haematology*, 139(5):809–819, 2007.

[12] A Smith, D Howell, R Patmore, A Jack, and E Roman. Incidence of haematological malignancy by sub-type: a report from the haematological malignancy research network. *British journal of cancer*, 105(11):1684–1692, 2011.

[13] Christian Maurer, Petra Langerbeins, Jasmin Bahlo, Paula Cramer, Anna Maria Fink, Natali Pflug, Anja Engelke, Julia von Tresckow, Gabor Kovacs, Stephan Stilgenbauer, et al. Effect of first-line treatment on second primary malignancies and richter's transformation in patients with cll. *Leukemia*, 30(10):2019–2025, 2016.

[14] David A Bond, Ying Huang, James L Fisher, Amy S Ruppert, Dwight H Owen, Erin M Bertino, Kerry A Rogers, Seema A Bhat, Michael R Grever, Samantha M Jaglowski, et al. Second cancer incidence in cll patients receiving btk inhibitors. *Leukemia*, 34(12):3197–3205, 2020.

[15] Ohad Benjamini, Preetesh Jain, Long Trinh, Wei Qiao, Sara S Strom, Susan Lerner, Xuemei Wang, Jan Burger, Alessandra Ferrajoli, Hagop Kantarjian, et al. Second cancers in patients with chronic lymphocytic leukemia who received frontline fludarabine, cyclophosphamide and rituximab therapy: distribution and clinical outcomes. *Leukemia & lymphoma*, 56(6):1643–1650, 2015.

[16] Michie Hisada, Robert J Biggar, Mark H Greene, Joseph F Fraumeni Jr, and Lois B Travis. Solid tumors after chronic lymphocytic leukemia. *Blood, The Journal of the American Society of Hematology*, 98(6):1979–1981, 2001.

[17] Claudia Schöllkopf, Ditte Rosendahl, Klaus Rostgaard, Christian Pipper, and Henrik Hjalgrim. Risk of second cancer after chronic lymphocytic leukemia. *International journal of cancer*, 121(1):151–156, 2007.

[18] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

[19] Jacques Ferlay, Isabelle Soerjomataram, Rajesh Dikshit, Sultan Eser, Colin Mathers, Marise Rebelo, Donald Maxwell Parkin, David Forman, and Freddie Bray. Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. *International journal of cancer*, 136(5):E359–E386, 2015.

[20] A Palumbo and K Anderson. Multiple myeloma n engl j med. *2011*, 1046:10, 2011.

[21] Dickran Kazandjian. Multiple myeloma epidemiology and survival: A unique malignancy. In *Seminars in oncology*, volume 43, pages 676–681. Elsevier, 2016.

[22] Felix Krueger, Frankie James, Phil Ewels, Ebrahim Afyounian, Michael Weinstein, Benjamin Schuster-Boeckler, Gert Hulselmans, and Sclamons. Felixkrueger/trimgalore: v0. 6.10-add default decompression path. *Zenodo*, 2023.

[23] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, 25(14):1754–1760, 2009.

[24] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):R25, 2009.

[25] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359, 2012.

[26] Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, and Heng Li. Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2), 02 2021. ISSN 2047-217X. doi: 10.1093/gigascience/giab008. URL https://doi.org/10.1093/gigascience/giab008. giab008.

[27] Yuk Yee Leung, Pavel P Kuksa, Alexandre Amlie-Wolf, Otto Valladares, Lyle H Ungar, Sampath Kannan, Brian D Gregory, and Li-San Wang. Dashr: database of small human noncoding rnas. *Nucleic acids research*, 44(D1):D216–D222, 2016.

[28] Danny Bergeron, Hermes Paraqindes, Étienne Fafard-Couture, Gabrielle Deschamps-Francoeur, Laurence Faucher-Giguère, Philia Bouchard-Bourelle, Sherif Abou Elela, Frédéric Catez, Virginie Marcel, and Michelle S Scott. snodb 2.0: an enhanced interactive database, specializing in human snornas. *Nucleic Acids Research*, 51(D1):D291–D296, 2023.

[29] Patricia P Chan and Todd M Lowe. Gtrnadb: a database of transfer rna genes detected in genomic sequence. *Nucleic acids research*, 37(suppl_1):D93–D97, 2009.

[30] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. Htseq—a python framework to work with high-throughput sequencing data. *bioinformatics*, 31(2): 166–169, 2015.

[31] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15 (12):550, 2014.

[32] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics*, 26(1):139–140, 2010.

[33] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

[34] Le Chang, Guangyan Zhou, Othman Soufan, and Jianguo Xia. mirnet 2.0: network-based visual analytics for mirna functional analysis and systems biology. *Nucleic acids research*, 48(W1):W244–W251, 2020.

[35] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

[36] Yu Fan, Liu Xi, Daniel ST Hughes, Jianjun Zhang, Jianhua Zhang, P Andrew Futreal, David A Wheeler, and Wenyi Wang. Muse: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome biology*, 17(1):1–11, 2016.

[37] David Benjamin, Takuto Sato, Kristian Cibulskis, Gad Getz, Chip Stewart, and Lee Lichtenstein. Calling somatic snvs and indels with mutect2. *BioRxiv*, page 861054, 2019.

[38] David E Larson, Christopher C Harris, Ken Chen, Daniel C Koboldt, Travis E Abbott, David J Dooling, Timothy J Ley, Elaine R Mardis, Richard K Wilson, and Li Ding. Somaticsniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, 28(3):311–317, 2012.

[39] Daniel C Koboldt, Qunyuan Zhang, David E Larson, Dong Shen, Michael D McLellan, Ling Lin, Christopher A Miller, Elaine R Mardis, Li Ding, and Richard K Wilson. Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, 22(3):568–576, 2012.

[40] Tobias Rausch, Thomas Zichner, Andreas Schlattl, Adrian M Stütz, Vladimir Benes, and Jan O Korbel. Delly: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, 2012.

[41] Xian Fan, Travis E Abbott, David Larson, and Ken Chen. Breakdancer: Identification of genomic structural variation from paired-end read mapping. *Current protocols in bioinformatics*, 45(1):15–6, 2014.

[42] Kai Ye, Li Guo, Xiaofei Yang, Eric-Wubbo Lamijer, Keiran Raine, and Zemin Ning. Split-read indel and structural variant calling using pindel. *Copy Number Variants: Methods and Protocols*, pages 95–105, 2018.

[43] Kai Wang, Mingyao Li, and Hakon Hakonarson. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164–e164, 2010.

[44] William McLaren, Bethan Pritchard, Daniel Rios, Yuan Chen, Paul Flicek, and Fiona Cunningham. Deriving the consequences of genomic variants with the ensembl api and snp effect predictor. *Bioinformatics*, 26(16):2069–2070, 2010.

[45] Stephen T Sherry, M-H Ward, M Kholodov, J Baker, Lon Phan, Elizabeth M Smigielski, and Karl Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic acids research*, 29(1):308–311, 2001.

[46] Pauline C Ng and Steven Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–3814, 2003.

[47] Ivan Adzhubei, Daniel M Jordan, and Shamil R Sunyaev. Predicting functional effect of human missense mutations using polyphen-2. *Current protocols in human genetics*, 76(1):7–20, 2013.

[48] Mark F Rogers, Hashem A Shihab, Matthew Mort, David N Cooper, Tom R Gaunt, and Colin Campbell. Fathmm-xf: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics*, 34(3):511–513, 2018.

[49] Philipp Rentzsch, Daniela Witten, Gregory M Cooper, Jay Shendure, and Martin Kircher. Cadd: predicting the deleteriousness of variants throughout the human genome. *Nucleic acids research*, 47(D1):D886–D894, 2019.

[50] Yongwook Choi, Gregory E Sims, Sean Murphy, Jason R Miller, and Agnes P Chan. Predicting the functional effect of amino acid substitutions and indels. 2012.

[51] John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, et al. Cosmic: the catalogue of somatic mutations in cancer. *Nucleic acids research*, 47(D1):D941–D947, 2019.

[52] Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, et al. Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic acids research*, 44(D1):D862–D868, 2016.

[53] Debyani Chakravarty, Jianjiong Gao, Sarah Phillips, Ritika Kundra, Hongxin Zhang, Jiaojiao Wang, Julia E Rudolph, Rona Yaeger, Tara Soumerai, Moriah H Nissan, et al. Oncokb: a precision oncology knowledge base. *JCO precision oncology*, 1:1–16, 2017.

[54] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.

[55] Gregory Lazarian, Romain Guièze, and Catherine J Wu. Clinical implications of novel genomic discoveries in chronic lymphocytic leukemia. *Journal of Clinical Oncology*, 35(9):984, 2017.

[56] Anna Puiggros, Gonzalo Blanco, Blanca Espinet, et al. Genetic abnormalities in chronic lymphocytic leukemia: where we are and where we go. *BioMed research international*, 2014, 2014.

[57] Dan A Landau and Catherine J Wu. Chronic lymphocytic leukemia: molecular heterogeneity revealed by high-throughput genomics. *Genome medicine*, 5(5): 1–13, 2013.

[58] Patricia Rojas-Ríos and Martine Simonelig. pirnas and piwi proteins: regulators of gene expression in development and stem cells. *Development*, 145(17):dev161786, 2018.

[59] Witold Świątowy and Paweł P Jagodziński. Molecules derived from trna and snorna: entering the degradome pool. *Biomedicine & Pharmacotherapy*, 108: 36–42, 2018.

[60] George Adrian Calin, Manuela Ferracin, Amelia Cimmino, Gianpiero Di Leva, Masayoshi Shimizu, Sylwia E Wojcik, Marilena V Iorio, Rosa Visone, Nurettin Ilfer Sever, Muller Fabbri, et al. A microrna signature associated with prognosis and progression in chronic lymphocytic leukemia. *New England Journal of Medicine*, 353(17):1793–1801, 2005.

[61] Valerio Fulci, Sabina Chiaretti, Marina Goldoni, Gianluca Azzalin, Nicoletta Carucci, Simona Tavolaro, Leandro Castellano, Armando Magrelli, Franca Citarella, Monica Messina, et al. Quantitative technologies establish a novel microrna profile of chronic lymphocytic leukemia. *Blood, The Journal of the American Society of Hematology*, 109(11):4944–4951, 2007.

[62] S Marton, MR Garcia, C Robello, Helena Persson, F Trajtenberg, O Pritsch, Carlos Rovira, H Naya, G Dighiero, and A Cayota. Small rnas analysis in cll reveals a deregulation of mirna expression and novel mirna candidates of putative relevance in cll pathogenesis. *Leukemia*, 22(2):330–338, 2008.

[63] Dan-Xia Zhu, Kou-Rong Miao, Cheng Fang, Lei Fan, Wei Zhu, Hua-Yuan Zhu, Yun Zhuang, Ming Hong, Peng Liu, Wei Xu, et al. Aberrant microrna expression in chinese patients with chronic lymphocytic leukemia. *Leukemia research*, 35(6): 730–734, 2011.

[64] Nikos Papakonstantinou, Stavroula Ntoufa, Elisavet Chartomatsidou, Giorgio Papadopoulos, Artemis Hatzigeorgiou, Achiles Anagnostopoulos, Katerina Chlichlia, Paolo Ghia, Marta Muzio, Chrysoula Belessi, et al. Differential microrna profiles and their functional implications in different immunogenetic subsets of chronic lymphocytic leukemia. *Molecular medicine*, 19(1):115–123, 2013.

[65] Veronica Balatti, Mario Acunzo, Yuri Pekarky, and Carlo M Croce. Novel mechanisms of regulation of mirnas in cll. *Trends in cancer*, 2(3):134–143, 2016.

[66] Rosa Visone, Laura Z Rassenti, Angelo Veronese, Cristian Taccioli, Stefan Costinean, Baltazar D Aguda, Stefano Volinia, Manuela Ferracin, Jeff Palatini,

Veronica Balatti, et al. Karyotype-specific microrna signature in chronic lymphocytic leukemia. *Blood, The Journal of the American Society of Hematology*, 114 (18):3872–3879, 2009.

[67] Yvonne Kiefer, Christoph Schulte, Markus Tiemann, and Joern Bullerdiek. Chronic lymphocytic leukemia-associated chromosomal abnormalities and mirna deregulation. *The application of clinical genetics*, 5:21, 2012.

[68] Tatiane Vieira Braga, Fernanda Cristina Gontijo Evangelista, Lorena Caixeta Gomes, Sérgio Schusterschitz da Silva Araújo, Maria das Graças Carvalho, and Adriano de Paula Sabino. Evaluation of mir-15a and mir-16-1 as prognostic biomarkers in chronic lymphocytic leukemia. *Biomedicine & Pharmacotherapy*, 92:864–869, 2017.

[69] Massimo Negrini, Giovanna Cutrona, Cristian Bassi, Sonia Fabris, Barbara Zagatti, Monica Colombo, Manuela Ferracin, Lucilla D'Abundo, Elena Saccenti, Serena Matis, et al. micrornaome expression in chronic lymphocytic leukemia: Comparison with normal b-cell subsets and correlations with prognostic and clinical parametersmicrorna expression in cll. *Clinical Cancer Research*, 20(15): 4141–4153, 2014.

[70] Marek Mráz, Karla Malinova, J Kotaskova, S Pavlova, B Tichy, J Malcikova, K Stano Kozubik, J Smardova, Yvona Brychtová, Michael Doubek, et al. mir-34a, mir-29c and mir-17-5p are downregulated in cll patients with tp53 abnormalities. *Leukemia*, 23(6):1159–1163, 2009.

[71] Basile Stamatopoulos, Nathalie Meuleman, Cécile De Bruyn, Karlien Pieters, Géraldine Anthoine, Philippe Mineur, Dominique Bron, and Laurence Lagneaux. A molecular score by quantitative pcr as a new prognostic tool at diagnosis for chronic lymphocytic leukemia patients. *PLoS One*, 5(9):e12780, 2010.

[72] Simona Rossi, Masayoshi Shimizu, Elisa Barbarotto, Milena S Nicoloso, Federica Dimitri, Deepa Sampath, Muller Fabbri, Susan Lerner, Lynn L Barron, Laura Z Rassenti, et al. microrna fingerprinting of cll patients with chromosome 17p deletion identify a mir-21 score that stratifies early survival. *Blood, The Journal of the American Society of Hematology*, 116(6):945–952, 2010.

[73] Chao-Po Lin and Lin He. Noncoding RNAs in cancer development. *Annual Review of Cancer Biology*, 1(1):163–184, 2017.

[74] Marc R Friedländer, Sebastian D Mackowiak, Na Li, Wei Chen, and Nikolaus Rajewsky. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic acids research*, 40(1):37–52, 2012.

[75] Jiyuan An, John Lai, Melanie L Lehman, and Colleen C Nelson. miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. *Nucleic acids research*, 41(2):727–737, 2013.

[76] Jieming Shi, Min Dong, Lei Li, Lin Liu, Agustin Luz-Madrigal, Panagiotis A Tsonis, Katia Del Rio-Tsonis, and Chun Liang. mirPRo–a novel standalone program for differential expression and variation analysis of miRNAs. *Scientific reports*, 5:14617, 2015.

[77] Dimitrios M Vitsios, Elissavet Kentepozidou, Leonor Quintais, Elia Benito-Gutiérrez, Stijn van Dongen, Matthew P Davis, and Anton J Enright. Mirnovo: genome-free prediction of microRNAs from small RNA sequencing data and single-cells using decision forests. *Nucleic acids research*, 45(21):e177–e177, 2017.

[78] Yin Lu, Alexander S Baras, and Marc K Halushka. miRge 2.0 for comprehensive analysis of microRNA sequencing data. *BMC bioinformatics*, 19(1):275, 2018.

[79] Ernesto Aparicio-Puerta, Ricardo Lebrón, Antonio Rueda, Cristina Gómez-Martín, Stavros Giannoukakos, David Jaspez, José María Medina, Andreja Zubkovic, Igor Jurak, Bastian Fromm, et al. srnabench and srnatoolbox 2019: intuitive fast small rna profiling and differential expression. *Nucleic acids research*, 47(W1): W530–W535, 2019.

[80] Enrico Gaffo, Michele Bortolomeazzi, Andrea Bisognin, Piero Di Battista, Federica Lovisa, Lara Mussolin, and Stefania Bortoluzzi. MiR&moRe2: A Bioinformatics Tool to Characterize microRNAs and microRNA-Offset RNAs from Small RNA-Seq Data. *International journal of molecular sciences*, 21(5):1754, 2020.

[81] Salomon Manier, Karma Salem, Siobhan V Glavey, Aldo M Roccaro, and Irene M Ghobrial. Genomic aberrations in multiple myeloma. *Plasma Cell Dyscrasias*, pages 23–34, 2016.

[82] Michael A Chapman, Michael S Lawrence, Jonathan J Keats, Kristian Cibulskis, Carrie Sougnez, Anna C Schinzel, Christina L Harview, Jean-Philippe Brunet, Gregory J Ahmann, Mazhar Adli, et al. Initial genome sequencing and analysis of multiple myeloma. *Nature*, 471(7339):467–472, 2011.

[83] Niccolo Bolli, Hervé Avet-Loiseau, David C Wedge, Peter Van Loo, Ludmil B Alexandrov, Inigo Martincorena, Kevin J Dawson, Francesco Iorio, Serena Nik-Zainal, Graham R Bignell, et al. Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nature communications*, 5(1):1–13, 2014.

[84] Brian A Walker, Eileen M Boyle, Christopher P Wardell, Alex Murison, Dil B Begum, Nasrin B Dahir, Paula Z Proszek, David C Johnson, Martin F Kaiser, Lorenzo Melchor, et al. Mutational spectrum, copy number changes, and outcome: results of a sequencing study of patients with newly diagnosed myeloma. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 33(33):3911, 2015.

[85] Jens G Lohr, Petar Stojanov, Scott L Carter, Peter Cruz-Gordillo, Michael S Lawrence, Daniel Auclair, Carrie Sougnez, Birgit Knoechel, Joshua Gould, Gordon Saksena, et al. Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. *Cancer cell*, 25(1):91–101, 2014.

[86] Akanksha Farswan, Lingaraja Jena, Gurvinder Kaur, Anubha Gupta, Ritu Gupta, Lata Rani, Atul Sharma, and Lalit Kumar. Branching clonal evolution patterns predominate mutational landscape in multiple myeloma. *American journal of cancer research*, 11(11):5659, 2021.

[87] Brian A Walker, Konstantinos Mavrommatis, Christopher P Wardell, T Cody Ashby, Michael Bauer, Faith E Davies, Adam Rosenthal, Hongwei Wang, Pingping Qu, Antje Hoering, et al. Identification of novel mutational drivers reveals oncogene dependencies in multiple myeloma. *Blood, The Journal of the American Society of Hematology*, 132(6):587–597, 2018.

[88] Gurvinder Kaur, Lingaraja Jena, Ritu Gupta, Akanksha Farswan, Anubha Gupta, and K Sriram. Correlation of changes in subclonal architecture with progression in the mmrf commpass study. *Translational oncology*, 23:101472, 2022.

[89] Aneta Mikulasova, Jan Smetana, Marketa Wayhelova, Helena Janyskova, Viera Sandecka, Zuzana Kufova, Martina Almasi, Jiri Jarkovsky, Evzen Gregora, Petr Kessler, et al. Genomewide profiling of copy-number alteration in monoclonal gammopathy of undetermined significance. *European journal of haematology*, 97 (6):568–575, 2016.

[90] Brian A Walker, Christopher P Wardell, Lorenzo Melchor, Annamaria Brioli, David C Johnson, Martin F Kaiser, Fabio Mirabella, Lucia Lopez-Corral, Sean Humphray, Lisa Murray, et al. Intraclonal heterogeneity is a critical early event in the development of myeloma and precedes the development of clinical symptoms. *Leukemia*, 28(2):384–390, 2014.

[91] Aneta Mikulasova, Christopher P Wardell, Alexander Murison, Eileen M Boyle, Graham H Jackson, Jan Smetana, Zuzana Kufova, Ludek Pour, Viera Sandecka, Martina Almasi, et al. The spectrum of somatic mutations in monoclonal gammopathy of undetermined significance indicates a less complex genomic landscape than that in multiple myeloma. *Haematologica*, 102(9):1617, 2017.

[92] Akanksha Farswan, Anubha Gupta, Lingaraja Jena, Vivek Ruhela, Gurvinder Kaur, and Ritu Gupta. Characterizing the mutational landscape of mm and its precursor mgus. *American journal of cancer research*, 12(4):1919, 2022.

[93] Ankit K Dutta, J Lynn Fink, John P Grady, Gareth J Morgan, Charles G Mullighan, Luen B To, Duncan R Hewett, and Andrew CW Zannettino. Subclonal evolution in disease progression from mgus/smm to multiple myeloma is characterised by clonal stability. *Leukemia*, 33(2):457–468, 2019.

[94] Adrián Mosquera Orgueira, Marta Sonia González Pérez, José Ángel Díaz Arias, Beatriz Antelo Rodríguez, Natalia Alonso Vence, Ángeles Bendaña López, Aitor Abuín Blanco, Laura Bao Pérez, Andrés Peleteiro Raíndo, Miguel Cid López, et al. Survival prediction and treatment optimization of multiple myeloma patients using machine-learning models based on clinical and gene expression data. *Leukemia*, 35(10):2924–2935, 2021.

[95] Lucas Venezian Povoa, Carlos Henrique Costa Ribeiro, and Israel Tojal da Silva. Machine learning predicts treatment sensitivity in multiple myeloma based on molecular and clinical information coupled with drug response. *PloS one*, 16(7): e0254596, 2021.

[96] Akanksha Farswan, Anubha Gupta, Ritu Gupta, Saswati Hazra, Sadaf Khan, Lalit Kumar, and Atul Sharma. Ai-supported modified risk staging for multiple myeloma cancer useful in real-world scenario. *Translational oncology*, 14(9): 101157, 2021.

[97] Akanksha Farswan, Anubha Gupta, Krishnamachari Sriram, Atul Sharma, Lalit Kumar, and Ritu Gupta. Does ethnicity matter in multiple myeloma risk prediction in the era of genomics and novel agents? evidence from real-world data. *Frontiers in oncology*, 11, 2021.

[98] Akanksha Farswan and Anubha Gupta. Tv-dct: Method to impute gene expression data using dct based sparsity and total variation denoising. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1244–1248. IEEE, 2019.

[99] Akanksha Farswan, Anubha Gupta, Ritu Gupta, and Gurvinder Kaur. Imputation of gene expression data in blood cancer and its significance in inferring biological pathways. *Frontiers in oncology*, 9:1442, 2020.

[100] Irantzu Anzar, Angelina Sverchkova, Richard Stratford, and Trevor Clancy. Neomutate: an ensemble machine learning framework for the prediction of somatic mutations in cancer. *BMC medical genomics*, 12(1):1–14, 2019.

[101] Yu-Chin Hsu, Yu-Ting Hsiao, Tzu-Yuan Kao, Jan-Gowth Chang, and Grace S Shieh. Detection of somatic mutations in exome sequencing of tumor-only samples. *Scientific reports*, 7(1):1–9, 2017.

[102] Vijay Kumar Pounraja, Gopal Jayakar, Matthew Jensen, Neil Kelkar, and Santhosh Girirajan. A machine-learning approach for accurate detection of copy number variants from exome sequencing. *Genome research*, 29(7):1134–1143, 2019.

[103] Tihao Huang, Junqing Li, Baoxian Jia, and Hongyan Sang. Cnv-meann: A neural network and mind evolutionary algorithm-based detection of copy number variations from next-generation sequencing data. *Frontiers in Genetics*, 12, 2021.

[104] Tom Hill and Robert L Unckless. A deep learning approach for detecting copy number variation in next-generation sequencing data. *G3: Genes, Genomes, Genetics*, 9(11):3575–3582, 2019.

[105] Yawei Li and Yuan Luo. Performance-weighted-voting model: an ensemble machine learning method for cancer type classification using whole-exome sequencing mutation. *Quantitative Biology*, 8(4):347–358, 2020.

[106] Olivier Collier, Véronique Stoven, and Jean-Philippe Vert. Lotus: A single-and multitask machine learning algorithm for the prediction of cancer driver genes. *PLoS computational biology*, 15(9):e1007381, 2019.

[107] Yi Han, Juze Yang, Xinyi Qian, Wei-Chung Cheng, Shu-Hsuan Liu, Xing Hua, Liyuan Zhou, Yaning Yang, Qingbiao Wu, Pengyuan Liu, et al. Driverml: a machine learning algorithm for identifying driver genes in cancer sequencing studies. *Nucleic acids research*, 47(8):e45–e45, 2019.

[108] Zexian Zeng, Chengsheng Mao, Andy Vo, Xiaoyu Li, Janna Ore Nugent, Seema A Khan, Susan E Clare, and Yuan Luo. Deep learning for cancer type classification and driver gene identification. *BMC bioinformatics*, 22(4):1–13, 2021.

[109] Ping Luo, Yulian Ding, Xiujuan Lei, and Fang-Xiang Wu. deepdriver: predicting cancer driver genes based on somatic mutations using deep convolutional neural networks. *Frontiers in genetics*, 10:13, 2019.

[110] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[111] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013.

[112] Roman Schulte-Sasse, Stefan Budach, Denes Hnisz, and Annalisa Marsico. Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. *Nature Machine Intelligence*, 3(6):513–526, 2021.

[113] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature reviews genetics*, 12(1):56–68, 2011.

[114] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[115] Damian Szklarczyk, Annika L Gable, Katerina C Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T Doncheva, Marc Legeay, Tao Fang, Peer Bork, et al. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, 49(D1):D605–D612, 2021.

[116] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[117] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[118] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32, 2019.

[119] KM Kortüm, C Langer, J Monge, L Bruins, YX Zhu, CX Shi, P Jedlowski, JB Egan, J Ojha, Lars Bullinger, et al. Longitudinal analysis of 25 sequential sample-pairs using a custom multiple myeloma mutation sequencing panel (m 3 p). *Annals of hematology*, 94:1205–1211, 2015.

[120] Parvathi Sudha, Aarif Ahsan, Cody Ashby, Tasneem Kausar, Akhil Khera, Mohammad H Kazeroun, Chih-Chao Hsu, Lin Wang, Evelyn Fitzsimons, Outi Salminen, et al. Myeloma genome project panel is a comprehensive targeted genomics panel for molecular profiling of patients with multiple myeloma. *Clinical Cancer Research*, 28(13):2854–2864, 2022.

[121] Niccolò Bolli, Y Li, V Sathiaseelan, Keiran Raine, D Jones, Peter Ganly, Federica Cocito, G Bignell, Mike A Chapman, AS Sperling, et al. A dna target-enrichment approach to detect mutations, copy number changes and immunoglobulin translocations in multiple myeloma. *Blood Cancer Journal*, 6(9):e467–e467, 2016.

[122] Brian S White, Irena Lanc, Julie O'Neal, Harshath Gupta, Robert S Fulton, Heather Schmidt, Catrina Fronick, Edward A Belter, Mark Fiala, Justin King, et al. A multiple myeloma-specific capture sequencing platform discovers novel translocations and frequent, risk-associated point mutations in igll5. *Blood cancer journal*, 8(3):1–10, 2018.

[123] Samuel D Cutler, Philipp Knopf, Clinton JV Campbell, Andrea Thoni, Mohamed Abou El Hassan, Nicholas Forward, Darrell White, Julie Wagner, Marissa Goudie, Jeanette E Boudreau, et al. Dmg26: A targeted sequencing panel for mutation profiling to address gaps in the prognostication of multiple myeloma. *The Journal of Molecular Diagnostics*, 23(12):1699–1714, 2021.

[124] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):5233, 2019.

[125] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[126] Jason YH Chang and Sylvain Ladame. Diagnostic, prognostic, and predictive biomarkers for cancer. In *Bioengineering Innovative Solutions for Cancer*, pages 3–21. Elsevier, 2020.

[127] Maarten Van Smeden, Johannes B Reitsma, Richard D Riley, Gary S Collins, and Karel GM Moons. Clinical prediction models: diagnosis versus prognosis. *Journal of clinical epidemiology*, 132:142–145, 2021.

[128] Salomon Manier, Karma Z Salem, Jihye Park, Dan A Landau, Gad Getz, and Irene M Ghobrial. Genomic complexity of multiple myeloma and its clinical implications. *Nature reviews Clinical oncology*, 14(2):100–113, 2017.

[129] Kai Neben, Anna Jauch, Thomas Hielscher, Jens Hillengass, Nicola Lehners, Anja Seckinger, Martin Granzow, Marc S Raab, Anthony D Ho, Hartmut Goldschmidt, et al. Progression in smoldering myeloma is independently determined by the chromosomal abnormalities del (17p), t (4; 14), gain 1q, hyperdiploidy, and tumor load. *Journal of clinical oncology*, 31(34):4325–4332, 2013.

[130] Scott A Van Wier, Gregory J Ahmann, Kimberly J Henderson, Philip R Greipp, S Vincent Rajkumar, Dirk M Larson, Angela Dispenzieri, Morie A Gertz, Robert A Kyle, and Rafael Fonseca. The t (4; 14) is present in patients with early stage plasma cell proliferative disorders including mgus and smoldering multiple myeloma (smm). *Blood*, 106(11):1545, 2005.

[131] Dan Chen, Xinhong Yang, Min Liu, Zhihua Zhang, and Enhong Xing. Roles of mirna dysregulation in the pathogenesis of multiple myeloma. *Cancer gene therapy*, 28(12):1256–1268, 2021.

[132] Zhongqing Li, Lanting Liu, Chenxing Du, Zhen Yu, Yuanyuan Yang, Jie Xu, Xiaojing Wei, Fenghuang Zhan, Yongrong Lai, Lugui Qiu, et al. Therapeutic effects of oligo-single-stranded dna mimicking of hsa-mir-15a-5p on multiple myeloma. *Cancer Gene Therapy*, 27(12):869–877, 2020.

[133] Tomasz Sewastianik, Juerg R Straubhaar, Jian-Jun Zhao, Mehmet K Samur, Keith Adler, Helen E Tanton, Vignesh Shanmugam, Omar Nadeem, Peter S Dennis, Vinodh Pillai, et al. mir-15a/16-1 deletion in activated b cells promotes plasma cell and mature b-cell neoplasms. *Blood, The Journal of the American Society of Hematology*, 137(14):1905–1919, 2021.

[134] Jonathan J Keats, David W Craig, Winnie Liang, Yellapantula Venkata, Ahmet Kurdoglu, Jessica Aldrich, Daniel Auclair, Kristi Allen, Beverly Harrison, Scott Jewell, et al. Interim analysis of the mmrf commpass trial, a longitudinal study in multiple myeloma relating clinical outcomes to genomic and immunophenotypic profiles, 2013.

[135] Ilkka Lappalainen, Jeff Almeida-King, Vasudev Kumanduri, Alexander Senf, John Dylan Spalding, Gary Saunders, Jag Kandasamy, Mario Caccamo, Rasko Leinonen, Brendan Vaughan, et al. The european genome-phenome archive of human data consented for biomedical research. *Nature genetics*, 47(7):692–695, 2015.

[136] Aiims data submitted to biorepository prjna685283 and prjan694218.

[137] Michael Hallek, Bruce D Cheson, Daniel Catovsky, Federico Caligaris-Cappio, Guillaume Dighiero, Hartmut Döhner, Peter Hillmen, Michael J Keating, Emili Montserrat, Kanti R Rai, et al. Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia: a report from the international workshop on chronic lymphocytic leukemia updating the national cancer institute–working group 1996 guidelines. *Blood, The Journal of the American Society of Hematology*, 111(12):5446–5456, 2008.

[138] Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13): 1658–1659, 2006.

[139] Pavel P Kuksa, Alexandre Amlie-Wolf, Živadin Katanić, Otto Valladares, Li-San Wang, and Yuk Yee Leung. DASHR 2.0: integrated database of human small non-coding RNA genes and mature products. *Bioinformatics*, 35(6):1033–1039, 2019.

[140] Yannan Fan, Keith Siklenka, Simran K Arora, Paula Ribeiro, Sarah Kimmins, and Jianguo Xia. mirnet-dissecting mirna-target interactions and functional associations through network-based visual analysis. *Nucleic acids research*, 44 (W1):W135–W141, 2016.

[141] Jason P Fine and Robert J Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association*, 94 (446):496–509, 1999.

[142] International CLL-IPI Working Group et al. An international prognostic index for patients with chronic lymphocytic leukaemia (cll-ipi): a meta-analysis of individual patient data. *The Lancet Oncology*, 17(6):779–790, 2016.

[143] Natalia Ruiz-Lafuente, Maria-Jose Alcaraz-Garcia, Silvia Sebastian-Ruiz, Azahara-Maria Garcia-Serna, Joaquin Gomez-Espuch, Jose-Maria Moraleda, Alfredo Minguela, Ana-Maria Garcia-Alonso, and Antonio Parrado. Il-4 upregulates mir-21 and the mirnas hosted in the clcn5 gene in chronic lymphocytic leukemia. *PloS one*, 10(4):e0124936, 2015.

[144] Agata A Filip, Anna Grenda, Sylwia Popek, Dorota Koczkodaj, Małgorzata Michalak-Wojnowska, Michał Budzyński, Ewa Wąsik-Szczepanek, Szymon Zmorzyński, Agnieszka Karczmarczyk, and Krzysztof Giannopoulos. Expression of circulating mirnas associated with lymphocyte differentiation and activation in cll—another piece in the puzzle. *Annals of hematology*, 96(1):33–50, 2017.

[145] Ruixue Tang, Lu Liang, Dianzhong Luo, Zhenbo Feng, Qiuxia Huang, Rongquan He, Tingqing Gan, Lihua Yang, and Gang Chen. Downregulation of mir-30a is associated with poor prognosis in lung cancer. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, 21:2514, 2015.

[146] Su-Jin Yang, Su-Yu Yang, Dan-Dan Wang, Xiu Chen, Hong-Yu Shen, Xiao-Hui Zhang, Shan-Liang Zhong, Jin-Hai Tang, and Jian-Hua Zhao. The mir-30 family: Versatile players in breast cancer. *Tumor Biology*, 39(3):1010428317692204, 2017.

[147] Marek Mraz, Liguang Chen, Laura Z Rassenti, Emanuela M Ghia, Hongying Li, Kristen Jepsen, Erin N Smith, Karen Messer, Kelly A Frazer, and Thomas J Kipps. mir-150 influences b-cell receptor signaling in chronic lymphocytic leukemia by

regulating expression of gab1 and foxp1. *Blood, The Journal of the American Society of Hematology*, 124(1):84–95, 2014.

[148] Shuqiang Li, Howell F Moffett, Jun Lu, Lillian Werner, Hao Zhang, Jerome Ritz, Donna Neuberg, Kai W Wucherpfennig, Jennifer R Brown, and Carl D Novina. Microrna expression profiling identifies activated b cell status in chronic lymphocytic leukemia cells. *PloS one*, 6(3):e16956, 2011.

[149] J Wang, X Tian, R Han, X Zhang, X Wang, H Shen, L Xue, Y Liu, X Yan, J Shen, et al. Downregulation of mir-486-5p contributes to tumor progression and metastasis by targeting protumorigenic arhgap5 in lung cancer. *Oncogene*, 33(9): 1181–1189, 2014.

[150] Reza Ghanbari, Sama Rezasoltani, Javad Hashemi, Ashraf Mohamadkhani, Arash Tahmasebifar, Ehsan Arefian, Naser Mobarra, Jahanbakhsh Asadi, Ehsan Nazemalhosseini Mojarad, Yaghoub Yazdani, et al. Expression analysis of previously verified fecal and plasma down-regulated micrornas (mir-4478, 1295-3p, 142-3p and 26a-5p), in ffpe tissue samples of crc patients. *Archives of Iranian Medicine (AIM)*, 20(2), 2017.

[151] Weixin Wang, Meghan Corrigan-Cummins, Justin Hudson, Irina Maric, Olga Simakova, Sattva S Neelapu, Larry W Kwak, John E Janik, Barry Gause, Elaine S Jaffe, et al. Microrna profiling of follicular lymphoma identifies micrornas related to cell proliferation and tumor response. *haematologica*, 97(4):586, 2012.

[152] Yan Li, Min Mao, Hong Liu, Xiaomin Wang, Zhen Kou, Yuling Nie, Yichun Wang, Zengsheng Wang, Qin Huang, Tao Lang, et al. mir-34a and mir-29b as indicators for prognosis of treatment-free survival of chronic lymphocytic leukemia patients in chinese uygur and han populations. *Molecular and cellular probes*, 47:101436, 2019.

[153] Hua Chen, Hong Pan, Yi Qian, Wenbin Zhou, and Xiaoan Liu. Mir-25-3p promotes the proliferation of triple negative breast cancer by targeting btg2. *Molecular cancer*, 17(1):1–11, 2018.

[154] Jian Wang, Hui Wang, Aifen Liu, Changge Fang, Jianguo Hao, and Zhenghui Wang. Lactate dehydrogenase a negatively regulated by mirnas promotes aerobic glycolysis and is increased in colorectal cancer. *Oncotarget*, 6(23):19456, 2015.

[155] Fabien Dupuis-Sandoval, Mikaël Poirier, and Michelle S Scott. The emerging landscape of small nucleolar rnas in cell biology. *Wiley Interdisciplinary Reviews: RNA*, 6(4):381–397, 2015.

[156] Wenhao Weng, Hanhua Li, and Ajay Goel. Piwi-interacting rnas (pirnas) and cancer: Emerging biological concepts and potential clinical implications. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1871(1):160–169, 2019.

[157] Fengbiao Zhou, Yi Liu, Christian Rohde, Cornelius Pauli, Dennis Gerloff, Marcel Köhn, Danny Misiak, Nicole Bäumer, Chunhong Cui, Stefanie Göllner, et al.

Aml1-eto requires enhanced c/d box snorna/rnp formation to induce self-renewal and leukaemia. *Nature cell biology*, 19(7):844–855, 2017.

[158] Domenica Ronchetti, Laura Mosca, Giovanna Cutrona, Giacomo Tuana, Massimo Gentile, Sonia Fabris, Luca Agnelli, Gabriella Ciceri, Serena Matis, Carlotta Massucco, et al. Small nucleolar rnas as new biomarkers in chronic lymphocytic leukemia. *BMC medical genomics*, 6(1):1–11, 2013.

[159] Laure Berquet, Wilfried Valleron, Srdana Grgurevic, Cathy Quelen, Ouafa Zaki, Anne Quillet-Mary, Frederic Davi, Pierre Brousset, Marina Bousquet, and Loïc Ysebaert. Small nucleolar rna expression profiles refine the prognostic impact of ighv mutational status on treatment-free survival in chronic lymphocytic leukaemia. *British Journal of Haematology*, 172(5):819–823, 2015.

[160] Vivek Ruhela, Ritu Gupta, Sriram Krishnamachari, Gaurav Ahuja, and Anubha Gupta. miRSim: Seed-based Synthetic Small Non-coding RNA Sequence Simulator. *Zenodo. https://doi.org/10.5281/zenodo.6546356*, Feb 2021.

[161] Chiao-Yi Lin, Wen-Ting Tseng, Yao-Yin Chang, Mong-Hsun Tsai, Eric Y Chuang, Tzu-Pin Lu, and Liang-Chuan Lai. Lidocaine and bupivacaine downregulate myb and dancr lncrna by upregulating mir-187-5p in mcf-7 cells. *Frontiers in Medicine*, 8, 2021.

[162] N.A. Nogueira Jorge, G. Wajnberg, C.G. Ferreira, B. de Sa Carvalho, and F. Passetti. snorna and pirna expression levels modified by tobacco use in women with lung adenocarcinoma. *PloS one*, 12(8), 2017.

[163] Weichun Huang, Leping Li, Jason R Myers, and Gabor T Marth. ART: a next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594, 2012.

[164] Xuesong Hu, Jianying Yuan, Yujian Shi, Jianliang Lu, Binghang Liu, Zhenyu Li, Yanxiang Chen, Desheng Mu, Hao Zhang, Nan Li, et al. pIRS: Profile-based Illumina pair-end reads simulator. *Bioinformatics*, 28(11):1533–1535, 2012.

[165] Thasso Griebel, Benedikt Zacher, Paolo Ribeca, Emanuele Raineri, Vincent Lacroix, Roderic Guigó, and Michael Sammeth. Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic acids research*, 40 (20):10073–10083, 2012.

[166] Alyssa C Frazee, Andrew E Jaffe, Ben Langmead, and Jeffrey T Leek. Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics*, 31(17):2778–2784, 2015.

[167] Bo Li and Colin N Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12:1–16, 2011.

[168] Nicholas F Lahens, Thomas G Brooks, Dimitra Sarantopoulou, Soumyashant Nayak, Cris Lawrence, Antonijo Mrčela, Anand Srinivasan, Jonathan Schug, John B Hogenesch, Yoseph Barash, et al. Camparee: a robust and configurable rna expression simulator. *BMC genomics*, 22:1–12, 2021.

[169] Thomas G Brooks, Nicholas F Lahens, Antonijo Mrčela, Dimitra Sarantopoulou, Soumyashant Nayak, Amruta Naik, Shaon Sengupta, Peter S Choi, and Gregory R Grant. Beers2: Rna-seq simulation through high fidelity in silico modeling. *bioRxiv*, 2023.

[170] Zachary D Stephens, Matthew E Hudson, Liudmila S Mainzer, Morgan Taschuk, Matthew R Weber, and Ravishankar K Iyer. Simulating next-generation sequencing datasets from empirical mutation and sequencing models. *PloS one*, 11(11): e0167047, 2016.

[171] N Homer. wgsim-Read simulator for next generation sequencing. *https://github.com/nh13/DWGSIM*, 2011.

[172] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. *bioinformatics*, 25(16):2078–2079, 2009.

[173] T Massingham. Software for simulating next-generation sequencing data. *https://github.com/timmassingham/simNGS*, 2011.

[174] Sam Benidt and Dan Nettleton. Simseq: a nonparametric approach to simulation of rna-sequence datasets. *Bioinformatics*, 31(13):2131–2140, 2015.

[175] Hadrien Gourlé, Oskar Karlsson-Lindsjö, Juliette Hayer, and Erik Bongcam-Rudloff. Simulating illumina metagenomic data with insilicoseq. *Bioinformatics*, 35(3):521–522, 2019.

[176] Manuel Holtgrewe. Mason–a read simulator for second generation sequencing data. *Technical Report FU Berlin*, 2010.

[177] Gregory R Grant, Michael H Farkas, Angel D Pizarro, Nicholas F Lahens, Jonathan Schug, Brian P Brunk, Christian J Stoeckert, John B Hogenesch, and Eric A Pierce. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*, 27(18):2518–2528, 2011.

[178] Tim Kehl, Christina Backes, Fabian Kern, Tobias Fehlmann, Nicole Ludwig, Eckart Meese, Hans-Peter Lenhof, and Andreas Keller. About miRNAs, miRNA seeds, target genes and target pathways. *Oncotarget*, 8(63):107167, 2017.

[179] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[180] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14(4):1–13, 2013.

[181] Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[182] Victor Ambros, Bonnie Bartel, David P Bartel, Christopher B Burge, James C Carrington, Xuemei Chen, Gideon Dreyfuss, Sean R Eddy, SAM Griffiths-Jones, Mhairi Marshall, et al. A uniform system for microRNA annotation. *Rna*, 9(3): 277–279, 2003.

[183] Xavier Bofill-De Ros, Wojciech K Kasprzak, Yuba Bhandari, Lixin Fan, Quinn Cavanaugh, Minjie Jiang, Lisheng Dai, Acong Yang, Tie-Juan Shao, Bruce A Shapiro, et al. Structural differences between pri-mirna paralogs promote alternative drosha cleavage and expand target repertoires. *Cell reports*, 26(2):447–459, 2019.

[184] C Glenn Begley and Lee M Ellis. Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533, 2012.

[185] Zohreh Rahimi, Zahra Ghorbani, Hajar Motamed, and Nazanin Jalilian. Aberrant expression profile of mir-32, mir-98 and mir-374 in chronic lymphocytic leukemia. *Leukemia Research*, 111:106691, 2021.

[186] Veronica Balatti, Yuri Pekarky, Lara Rizzotto, and Carlo M Croce. mir deregulation in cll. *Advances in Chronic Lymphocytic Leukemia*, pages 309–325, 2013.

[187] Andrea Sitlinger, Michael A Deal, Erwin Garcia, Dana K Thompson, Tiffany Stewart, Grace A MacDonald, Nicolas Devos, David Corcoran, Janet S Staats, Jennifer Enzor, et al. Higher physical fitness regulates in vitro tumor cell growth in older adults with treatment naive chronic lymphocytic leukemia (cll). *medRxiv*, 2021.

[188] Ehsan Farzadfard, Tahereh Kalantari, and Gholamhossein Tamaddon. Serum expression of seven micrornas in chronic lymphocytic leukemia patients. *Journal of Blood Medicine*, 11:97, 2020.

[189] Tianxiang Hu, Yating Chong, Sumin Lu, Rebecca Wang, Haiyan Qin, Jeane Silva, Eiko Kitamura, Chang-Sheng Chang, LesleyAnn Hawthorn, and John K Cowell. mir-339 promotes development of stem cell leukemia/lymphoma syndrome via downregulation of the bcl2l11 and bax proapoptotic genes. *Cancer research*, 78 (13):3522–3531, 2018.

[190] Nahla Mohamed Gamal Farahat, Dalal Mohamed Nasr El Din Elkaffash, Ashraf Hussein Alghandour, Rania Shafik Swelem, and Reham Abdel Haleem Abo El-Wafa. Study of microrna profile as a molecular biomarker in egyptian chronic lymphocytic leukemia. *Indian Journal of Hematology and Blood Transfusion*, 35(1):89–99, 2019.

[191] Jeong Mi Yang, Ji-Young Jang, Yoon Kyung Jeon, and Jin Ho Paik. Clinicopathologic implication of microrna-197 in diffuse large b cell lymphoma. *Journal of translational medicine*, 16(1):1–14, 2018.

[192] Chundi Gao, Chao Zhou, Jing Zhuang, Lijuan Liu, Junyu Wei, Cun Liu, Huayao Li, and Changgang Sun. Identification of key candidate genes and mirna-mrna target pairs in chronic lymphocytic leukemia by integrated bioinformatics analysis. *Molecular Medicine Reports*, 19(1):362–374, 2019.

[193] JIAN-ZHEN SHEN, YUAN-YUAN ZHANG, HAI-YING FU, DAN-SEN WU, and HUA-RONG ZHOU. Overexpression of microRNA-143 inhibits growth and induces apoptosis in human leukemia cells. *Oncology Reports*, 31(5):2035–2042, March 2014. doi: 10.3892/or.2014.3078. URL https://doi.org/10.3892/or.2014.3078.

[194] Yuping Mei, Yuyan Wang, Priti Kumari, Amol Carl Shetty, David Clark, Tyler Gable, Alexander D MacKerell, Mark Z Ma, David J Weber, Austin J Yang, et al. A pirna-like small rna interacts with and modulates p-erm proteins in human somatic cells. *Nature communications*, 6(1):1–12, 2015.

[195] Yafei Shi, Min Qiu, Yanyan Wu, and Lu Hai. Mir-548-3p functions as an anti-oncogenic regulator in breast cancer. *Biomedicine & Pharmacotherapy*, 75: 111–116, 2015.

[196] Preethi Krishnan, Sunita Ghosh, Bo Wang, Dongping Li, Ashok Narasimhan, Richard Berendt, Kathryn Graham, John R Mackey, Olga Kovalchuk, and Sambasivarao Damaraju. Next generation sequencing profiling identifies mir-574-3p and mir-660-5p as potential novel prognostic markers for breast cancer. *BMC genomics*, 16(1):1–17, 2015.

[197] Xiaohui Tan, Yebo Fu, Liang Chen, Woojin Lee, Yinglei Lai, Katayoon Rezaei, Sana Tabbara, Patricia Latham, Christine B Teal, Yan-Gao Man, et al. mir-671-5p inhibits epithelial-to-mesenchymal transition by downregulating foxm1 expression in breast cancer. *Oncotarget*, 7(1):293, 2016.

[198] Bethany N Hannafon, Yvonne D Trigoso, Cameron L Calloway, Y Daniel Zhao, David H Lum, Alana L Welm, Zhizhuang J Zhao, Kenneth E Blick, William C Dooley, and WQ5016889 Ding. Plasma exosome micrornas are indicative of breast cancer. *Breast cancer research*, 18(1):1–14, 2016.

[199] Xiu Juan Li, Zhao Jun Ren, Jin Hai Tang, and Qiao Yu. Exosomal microrna mir-1246 promotes cell proliferation, invasion and drug resistance by targeting ccng2 in breast cancer. *Cellular Physiology and Biochemistry*, 44(5):1741–1748, 2017.

[200] Wei Xia, JueYu Zhou, HaiBo Luo, YunZhou Liu, CanCan Peng, WenLing Zheng, and WenLi Ma. Microrna-32 promotes cell proliferation, migration and suppresses apoptosis in breast cancer cells by targeting fbxw7. *Cancer cell international*, 17 (1):1–11, 2017.

[201] David J Schultz, Penn Muluhngwi, Negin Alizadeh-Rad, Madelyn A Green, Eric C Rouchka, Sabine J Waigel, and Carolyn M Klinge. Genome-wide mirna

response to anacardic acid in breast cancer cells. *PLoS One*, 12(9):e0184471, 2017.

[202] Lakshmi Sripada, Kritarth Singh, Anastasiya V Lipatova, Aru Singh, Paresh Prajapati, Dhanendra Tomar, Khyati Bhatelia, Milton Roy, Rochika Singh, Madan M Godbole, et al. hsa-mir-4485 regulates mitochondrial functions and inhibits the tumorigenicity of breast cancer cells. *Journal of Molecular Medicine*, 95(6): 641–651, 2017.

[203] Abu Musa Md Reza, Yun-Jung Choi, Yu-Guo Yuan, Joydeep Das, Hideyo Yasuda, Jin-Hoi Kim, et al. Microrna-7641 is a regulator of ribosomal proteins and a promising targeting factor to improve the efficacy of cancer therapy. *Scientific reports*, 7(1):1–11, 2017.

[204] Mirelle Lagendijk, Sepideh Sadaatmand, Linetta B Koppert, Madeleine MA Tilanus-Linthorst, Vanja de Weerd, Raquel Ramírez-Moreno, Marcel Smid, Anieta M Sieuwerts, and John WM Martens. Microrna expression in pre-treatment plasma of patients with benign breast diseases and breast cancer. *Oncotarget*, 9 (36):24335, 2018.

[205] Qipeng Xie, Caiyi Chen, Haiying Li, Jiheng Xu, Lei Wu, Yuan Yu, Shuwei Ren, Hongyan Li, Xiaohui Hua, Huiying Yan, et al. mir-3687 overexpression promotes bladder cancer cell growth by inhibiting the negative effect of foxp1 on cyclin e2 transcription. *Molecular Therapy*, 27(5):1028–1038, 2019.

[206] Linge Li, Juan Feng, Dinghao Zhang, Jun Yong, Yan Wang, Jianfeng Yao, and Rongfu Huang. Differential expression of mir-4492 and il-10 is involved in chronic rhinosinusitis with nasal polyps. *Experimental and therapeutic medicine*, 18(5):3968–3976, 2019.

[207] Jianguo Lai, Hongli Wang, Zihao Pan, and Fengxi Su. A novel six-microrna-based model to improve prognosis prediction of breast cancer. *Aging (Albany NY)*, 11 (2):649, 2019.

[208] WM Sun, W Tao, JC Li, DM Zhu, and Y Miao. Microrna-296 functions as a tumor suppressor in breast cancer by targeting fgfr1 and regulating the wnt/beta-catenin signaling pathway. *Eur Rev Med Pharmacol Sci*, 23(23):10422–10432, 2019.

[209] Shelly Mahlab-Aviv, Keren Zohar, Yael Cohen, Ayelet R Peretz, Tsiona Eliyahu, Michal Linial, and Ruth Sperling. Spliceosome-associated micrornas signify breast cancer cells and portray potential novel nuclear targets. *International journal of molecular sciences*, 21(21):8132, 2020.

[210] Ke-Jing Zhang, Yu Hu, Na Luo, Xin Li, Fei-Yu Chen, Jia-Qi Yuan, and Lei Guo. mir-574-5p attenuates proliferation, migration and emt in triple-negative breast cancer cells by targeting bcl11a and sox2 to inhibit the skil/taz/ctgf axis. *International journal of oncology*, 56(5):1240–1251, 2020.

[211] Songjie Shen, Yu Song, Bin Zhao, Yali Xu, Xinyu Ren, Yidong Zhou, and Qiang Sun. Cancer-derived exosomal mir-7641 promotes breast cancer progression and metastasis. *Cell Communication and Signaling*, 19(1):1–13, 2021.

[212] Soudeh Ghafouri-Fard, Ali Khanbabapour Sasi, Atefe Abak, Hamed Shoorei, Ali Khoshkar, and Mohammad Taheri. Contribution of mirnas in the pathogenesis of breast cancer. *Frontiers in Oncology*, 11, 2021.

[213] Barbara Zellinger, Ulrich Bodenhofer, Immanuela A Engländer, Cornelia Kronberger, Brane Grambozov, Elvis Ruznic, Markus Stana, Josef Karner, Gerd Fastner, Karl Sotlar, et al. Hsa-mir-3651 could serve as a novel predictor for in-breast recurrence via frmd3. *Breast Cancer*, pages 1–13, 2021.

[214] Qianxi Yang, Shaorong Zhao, Zhendong Shi, Lixia Cao, Jingjing Liu, Teng Pan, Dongdong Zhou, and Jin Zhang. Chemotherapy-elicited exosomal mir-378a-3p and mir-378d promote breast cancer stemness and chemoresistance via the activation of ezh2/stat3 signaling. *Journal of Experimental & Clinical Cancer Research*, 40(1):1–18, 2021.

[215] Sunyoung Park, Jungho Kim, Yoonjung Cho, Sungwoo Ahn, Geehyuk Kim, Dasom Hwang, Yunhee Chang, Sunmok Ha, Yeonim Choi, Min Ho Lee, et al. Promotion of tumorigenesis by mir-1260b–targeting caps8: Potential diagnostic and prognostic marker for breast cancer. *Cancer Science*, 2022.

[216] Michael Hackenberg, Naiara Rodríguez-Ezpeleta, and Ana M Aransay. miranalyzer: an update on the detection and analysis of micrornas in high-throughput sequencing experiments. *Nucleic acids research*, 39(suppl_2):W132–W138, 2011.

[217] Fatima Heinicke, Xiangfu Zhong, Manuela Zucknick, Johannes Breidenbach, Arvind YM Sundaram, Siri T. Flåm, Magnus Leithaug, Marianne Dalland, Andrew Farmer, Jordana M Henderson, et al. Systematic assessment of commercially available low-input mirna library preparation kits. *RNA biology*, 17(1):75–86, 2020.

[218] L. Lorenzi, H.S. Chiu, F. Avila Cobos, S. Gross, P.J. Volders, R. Cannoodt, J. Nuytens, K. Vanderheyden, J. Anckaert, S. Lefever, and A.P. Tay. The rna atlas expands the catalog of human non-coding rnas. *Nature Biotechnology*, 39(11): 1453–1465, 2021.

[219] Geraldine A Van der Auwera, Mauricio O Carneiro, Christopher Hartl, Ryan Poplin, Guillermo Del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, et al. From fastq data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, 43(1):11–10, 2013.

[220] Iñigo Martincorena, Keiran M Raine, Moritz Gerstung, Kevin J Dawson, Kerstin Haase, Peter Van Loo, Helen Davies, Michael R Stratton, and Peter J Campbell. Universal patterns of selection in cancer and somatic tissues. *Cell*, 171(5):1029–1041, 2017.

[221] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12, 2011.

[222] Francesco Maura, Niccoló Bolli, Nicos Angelopoulos, Kevin J Dawson, Daniel Leongamornlert, Inigo Martincorena, Thomas J Mitchell, Anthony Fullam, Santiago Gonzalez, Raphael Szalat, et al. Genomic landscape and chronological reconstruction of driver events in multiple myeloma. *Nature communications*, 10 (1):1–12, 2019.

[223] Francisco Martínez-Jiménez, Ferran Muiños, Inés Sentís, Jordi Deu-Pons, Iker Reyes-Salazar, Claudia Arnedo-Pac, Loris Mularoni, Oriol Pich, Jose Bonet, Hanna Kranas, et al. A compendium of mutational cancer driver genes. *Nature Reviews Cancer*, 20(10):555–572, 2020.

[224] Stephane De Cesco, John B Davis, and Paul E Brennan. Targetdb: A target information aggregation tool and tractability predictor. *PloS one*, 15(9):e0232644, 2020.

[225] Maxim V Kuleshov, Matthew R Jones, Andrew D Rouillard, Nicolas F Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L Jenkins, Kathleen M Jagodnik, Alexander Lachmann, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1):W90–W97, 2016.

[226] Zhuorui Xie, Allison Bailey, Maxim V Kuleshov, Daniel JB Clarke, John E Evangelista, Sherry L Jenkins, Alexander Lachmann, Megan L Wojciechowicz, Eryk Kropiwnicki, Kathleen M Jagodnik, et al. Gene set knowledge discovery with enrichr. *Current protocols*, 1(3):e90, 2021.

[227] Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R Clark, and Avi Ma'ayan. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC bioinformatics*, 14 (1):1–14, 2013.

[228] Andrew Waterhouse, Martino Bertoni, Stefan Bienert, Gabriel Studer, Gerardo Tauriello, Rafal Gumienny, Florian T Heer, Tjaart A P de Beer, Christine Rempfer, Lorenza Bordoli, et al. Swiss-model: homology modelling of protein structures and complexes. *Nucleic acids research*, 46(W1):W296–W303, 2018.

[229] L Lopez-Corral, María Eugenia Sarasquete, Silvia Beà, Ramón García-Sanz, Maria Victoria Mateos, LA Corchete, JM Sayagués, EM García, J Bladé, A Oriol, et al. Snp-based mapping arrays reveal high genomic complexity in monoclonal gammopathies, from mgus to myeloma status. *Leukemia*, 26(12):2521–2529, 2012.

[230] Guanhao Wei, Li Zhou, Lynn L Lu, and Marc Romano. Sequential ehr-based dynamic graph network for multiple myeloma detection and feature interaction investigation., 2022.

[231] Hyun-Tae Shin, Yoon-La Choi, Jae Won Yun, Nayoung KD Kim, Sook-Young Kim, Hyo Jeong Jeon, Jae-Yong Nam, Chung Lee, Daeun Ryu, Sang Cheol Kim, et al. Prevalence and detection of low-allele-fraction variants in clinical cancer samples. *Nature communications*, 8(1):1377, 2017.

[232] Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*, 31(3):213–219, 2013.

[233] Huilei Xu, John DiCarlo, Ravi Vijaya Satya, Quan Peng, and Yexun Wang. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC genomics*, 15:1–10, 2014.

[234] Nicola D Roberts, R Daniel Kortschak, Wendy T Parker, Andreas W Schreiber, Susan Branford, Hamish S Scott, Garique Glonek, and David L Adelson. A comparative analysis of algorithms for somatic snv detection in cancer. *Bioinformatics*, 29(18):2223–2230, 2013.

[235] Lau K Vestergaard, Douglas NP Oliveira, Claus K Høgdall, and Estrid V Høgdall. Next generation sequencing technology in the clinic and its challenges. *Cancers*, 13(8):1751, 2021.

[236] Maurizio Callari, Stephen-John Sammut, Leticia De Mattos-Arruda, Alejandra Bruna, Oscar M Rueda, Suet-Feung Chin, and Carlos Caldas. Intersect-then-combine approach: improving the performance of somatic variant calling in whole exome sequencing data using multiple aligners and callers. *Genome medicine*, 9: 1–11, 2017.

[237] Li Tai Fang, Pegah Tootoonchi Afshar, Aparna Chhibber, Marghoob Mohiyuddin, Yu Fan, John C Mu, Greg Gibeling, Sharon Barr, Narges Bani Asadi, Mark B Gerstein, et al. An ensemble approach to accurately detect somatic mutations using somaticseq. *Genome biology*, 16:1–13, 2015.

[238] Randall Ruch. Gap junctions and connexins in cancer formation, progression, and therapy, 2020.

[239] Maria C Oliveira, Hanne Verswyvel, Evelien Smits, Rodrigo M Cordeiro, Annemie Bogaerts, and Abraham Lin. The pro-and anti-tumoral properties of gap junctions in cancer and their role in therapeutic strategies. *Redox Biology*, 57: 102503, 2022.

[240] Nora Alfugham, Avi Gadoth, Vanda A Lennon, Lars Komorowski, Madeleine Scharf, Shannon Hinson, Andrew McKeon, and Sean J Pittock. Itpr1 autoimmunity: frequency, neurologic phenotype, and cancer association. *Neurology: Neuroimmunology & Neuroinflammation*, 5(1):e418, 2017.

[241] Sven Jarius, Marius Ringelstein, Jürgen Haas, Irina I Serysheva, Lars Komorowski, Kai Fechner, Klaus-Peter Wandinger, Philipp Albrecht, Harald Hefter, Andreas Moser, et al. Inositol 1, 4, 5-trisphosphate receptor type 1 autoantibodies in paraneoplastic and non-paraneoplastic peripheral neuropathy. *Journal of Neuroinflammation*, 13:1–17, 2016.

[242] Ana María Ávila and Sergio Giralt. Autoimmune disorders and multiple myeloma-two illustrative case reports and a literature review. *Revista Colombiana de Cancerología*, 22(2):76–83, 2018.

[243] James LM Ferrara and Pavan Reddy. Pathophysiology of graft-versus-host disease. In *Seminars in hematology*, volume 43, pages 3–10. Elsevier, 2006.

[244] Effie W Petersdorf. Genetics of graft-versus-host disease: the major histocompatibility complex. *Blood reviews*, 27(1):1–12, 2013.

[245] G Famularo, A D'Ambrosio, F Quintieri, S Di Giovanni, I Parzanese, F Pizzuto, R Giacomelli, O Pugliese, and G Tonietti. Natural killer cell frequency and function in patients with monoclonal gammopathies. *Journal of clinical & laboratory immunology*, 37(3):99–109, 1992.

[246] Jie Yang, Fei Wang, and Baoan Chen. Hla-dpa1 gene is a potential predictor with prognostic values in multiple myeloma. *BMC cancer*, 20:1–10, 2020.

[247] Arezou Sayad, Mohammad Taghi Akbari, Mahshid Mehdizadeh, Elham Roshandel, Soheila Abedinpour, and Abbas Hajifathali. The association of hla class 1 and class 2 antigens with multiple myeloma in iranian patients. 2014.

[248] Marie Thérèse Rubio, Adèle Dhuyser, and Stéphanie Nguyen. Role and modulation of nk cells in multiple myeloma. *hemato*, 2(2):167–181, 2021.

[249] Minoru Kanehisa, Miho Furumichi, Yoko Sato, Masayuki Kawashima, and Mari Ishiguro-Watanabe. Kegg for taxonomy-based analysis of pathways and genomes. *Nucleic acids research*, 51(D1):D587–D592, 2023.

[250] Megan Romeo, Tetiana Hutchison, Aditi Malu, Averi White, Janice Kim, Rachel Gardner, Katie Smith, Katherine Nelson, Rachel Bergeson, Ryan McKee, et al. The human t-cell leukemia virus type-1 p30ii protein activates p53 and induces the tigar and suppresses oncogene-induced oxidative stress during viral carcinogenesis. *Virology*, 518:103–115, 2018.

[251] Cynthia A Pise-Masison, Renaud Mahieux, Hua Jiang, Margaret Ashcroft, Michael Radonovich, Janet Duvall, Claire Guillerm, and John N Brady. Inactivation of p53 by human t-cell lymphotropic virus type 1 tax requires activation of the nf-κb pathway and is dependent on p53 phosphorylation. *Molecular and Cellular Biology*, 20(10):3377–3386, 2000.

[252] Ruobing Xiao, Jan Cerny, Katherine Devitt, Karen Dresser, Rajneesh Nath, Muthalagu Ramanathan, Scott J Rodig, Benjamin J Chen, Bruce A Woda, and Hongbo

Yu. Myc protein expression is detected in plasma cell myeloma but not in monoclonal gammopathy of undetermined significance (mgus). *The American journal of surgical pathology*, 38(6):776–783, 2014.

[253] Mitsuaki Yoshida. Multiple viral strategies of htlv-1 for dysregulation of cell growth control. *Annual review of immunology*, 19(1):475–496, 2001.

[254] Masao Matsuoka and Kuan-Teh Jeang. Human t-cell leukemia virus type 1 (htlv-1) and leukemic transformation: viral infectivity, tax, hbz and therapy. *Oncogene*, 30(12):1379–1389, 2011.

[255] Giuseppe Mariggio, Sandra Koch, and Thomas F Schulz. Kaposi sarcoma herpesvirus pathogenesis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1732):20160275, 2017.

[256] Ying Zhang, Wei Guo, Zhumei Zhan, and Ou Bai. Carcinogenic mechanisms of virus-associated lymphoma. 2024.

[257] Stephen J Dollery. Towards understanding kshv fusion and entry. *Viruses*, 11(11): 1073, 2019.

[258] MB Rettig, HJ Ma, RA Vescio, M Pold, G Schiller, D Belson, A Savage, C Nishikubo, C Wu, J Fraser, et al. Kaposi's sarcoma-associated herpesvirus infection of bone marrow dendritic cells from multiple myeloma patients. *Science*, 276(5320):1851–1854, 1997.

[259] Gianfranco Pittari, Luca Vago, Moreno Festuccia, Chiara Bonini, Deena Mudawi, Luisa Giaccone, and Benedetto Bruno. Restoring natural killer cell immunity against multiple myeloma in the era of new drugs. *Frontiers in immunology*, 8: 262014, 2017.

[260] Gordon Chan, Thomas Hanke, and Klaus-Dieter Fischer. Vav-1 regulates nk t cell development and nk cell cytotoxicity. *European journal of immunology*, 31(8): 2403–2410, 2001.

[261] Pankaj Trivedi, Sandesh Kumar Patel, Diana Bellavia, Elena Messina, Rocco Palermo, Simona Ceccarelli, Cinzia Marchese, Eleni Anastasiadou, Lisa M Minter, and Maria Pia Felli. When viruses cross developmental pathways. *Frontiers in Cell and Developmental Biology*, 9:691644, 2021.

[262] Alain Le Moine, Michel Goldman, and Daniel Abramowicz. Multiple pathways to allograft rejection. *Transplantation*, 73(9):1373–1381, 2002.

[263] Aurélie Moreau, Emilie Varey, Ignacio Anegon, and Maria-Cristina Cuturi. Effector mechanisms of rejection. *Cold Spring Harbor perspectives in medicine*, 3(11): a015461, 2013.

[264] Paola Cruz-Tapias, John Castiblanco, and Juan-Manuel Anaya. Major histocompatibility complex: Antigen processing and presentation. In *Autoimmunity: From Bench to Bedside [Internet]*. El Rosario University Press, 2013.

[265] Alexandre Couture, Anthony Garnier, Fabian Docagne, Olivier Boyer, Denis Vivien, Brigitte Le-Mauff, Jean-Baptiste Latouche, and Olivier Toutirais. Hla-class ii artificial antigen presenting cells in cd4+ t cell-based immunotherapy. *Frontiers in immunology*, 10:447508, 2019.

[266] Simona Pagliuca, Carmelo Gurnari, Marie Thérèse Rubio, Valeria Visconte, and Tobias L Lenz. Individual hla heterogeneity and its implications for cellular immune evasion in cancer and beyond. *Frontiers in Immunology*, 13:944872, 2022.

[267] NCI National Cancer Institute. Targeted therapy directed by genetic testing in treating patients with advanced refractory solid tumors, lymphomas, or multiple myeloma (the match screening trial). *NLM Identifier: NCT02465060*, 2020.

[268] Ritu Gupta, Gurvinder Kaur, Lalit Kumar, Lata Rani, Nitin Mathur, Atul Sharma, Meetu Dahiya, Varun Shekhar, Sadaf Khan, Anjali Mookerjee, et al. Nucleic acid based risk assessment and staging for clinical practice in multiple myeloma. *Annals of Hematology*, 97(12):2447–2454, 2018.

[269] Brooks Benard, Austin Christofferson, Christophe Legendre, Jessica Aldrich, Sara Nasser, Jennifer Yesil, Daniel Auclair, Winnie Liang, Sagar Lonial, and Jonathan J Keats. Fgfr3 mutations are an adverse prognostic factor in patients with t (4; 14)(p16; q32) multiple myeloma: An mmrf commpass analysis. *Blood*, 130:3027, 2017.

[270] April KS Salama, Shuli Li, Erin R Macrae, Jong-In Park, Edith P Mitchell, James A Zwiebel, Helen X Chen, Robert J Gray, Lisa M McShane, Larry V Rubinstein, et al. Dabrafenib and trametinib in patients with tumors with brafv600e mutations: Results of the nci-match trial subprotocol h. *Journal of Clinical Oncology*, 38(33):3895, 2020.

[271] Kristine Misund, Niamh Keane, Caleb K Stein, Yan W Asmann, Grady Day, Seth Welsh, Scott A Van Wier, Daniel L Riggs, Greg Ahmann, Marta Chesi, et al. Myc dysregulation in the progression of multiple myeloma. *Leukemia*, 34(1):322–326, 2020.

[272] Santiago Barrio Garcia, Yanira Ruiz-Heredia, Matteo Da Via, Miguel Gallardo, Andoni Garitano-Trojaola, Josip Zovko, Marc S Raab, Pieter Sonneveld, Esteban Braggio, A Keith Stewart, et al. Role of max as a tumor suppressor driver gene in multiple myeloma. *Blood*, 130:4347, 2017.

[273] Ken Ohmine and Ryosuke Uchibori. Novel immunotherapies in multiple myeloma. *International Journal of Hematology*, pages 1–12, 2022.

[274] Shivali Jasrotia, Ritu Gupta, Atul Sharma, Ashutosh Halder, and Lalit Kumar. Cytokine profile in multiple myeloma. *Cytokine*, 136:155271, 2020.

[275] R Hoteit, A Bazarbachi, A Antar, Z Salem, D Shammaa, and R Mahfouz. Kir genotype distribution among patients with multiple myeloma: Higher prevalence of kir 2ds4 and kir 2ds5 genes. *Meta gene*, 2:730–736, 2014.

[276] Haibo Sun, Thomas G Martin, John Marra, Denice Kong, Jonathon Keats, Sandrine Macé, Marielle Chiron, Jeffrey L Wolf, Jeffrey M Venstrom, and Raja Rajalingam. Individualized genetic makeup that controls natural killer cell function influences the efficacy of isatuximab immunotherapy in patients with multiple myeloma. *Journal for immunotherapy of cancer*, 9(7), 2021.

[277] Niken M Mahaweni, Femke AI Ehlers, Gerard MJ Bos, and Lotte Wieten. Tuning natural killer cell anti-multiple myeloma reactivity by targeting inhibitory signaling via kir and nkg2a. *Frontiers in immunology*, 9:2848, 2018.

[278] Ester Lozano, Tania Díaz, Mari-Pau Mena, Guillermo Suñe, Xavier Calvo, Marcos Calderón, Lorena Pérez-Amill, Vanina Rodríguez, Patricia Pérez-Galán, Gaël Roué, et al. Loss of the immune checkpoint cd85j/lilrb1 on malignant plasma cells contributes to immune escape in multiple myeloma. *The Journal of Immunology*, 200(8):2581–2591, 2018.

[279] Xunlei Kang, Jaehyup Kim, Mi Deng, Samuel John, Heyu Chen, Guojin Wu, Hiep Phan, and Cheng Cheng Zhang. Inhibitory leukocyte immunoglobulin-like receptors: immune checkpoint proteins and tumor sustaining factors. *Cell cycle*, 15(1):25–40, 2016.

[280] Meral Beksac, Loren Gragert, Stephanie Fingerson, Martin Maiers, Mei-Jie Zhang, Mark Albrecht, Xiaobo Zhong, Wendy Cozen, Angela Dispenzieri, Sagar Lonial, et al. Hla polymorphism and risk of multiple myeloma. *Leukemia*, 30(11): 2260–2264, 2016.

[281] Sahar Kassem, Béré K Diallo, Nizar El-Murr, Nadège Carrié, Alexandre Tang, Alain Fournier, Hélène Bonnevaux, Céline Nicolazzi, Marine Cuisinier, Isabelle Arnould, et al. Sar442085, a novel anti-cd38 antibody with enhanced antitumor activity against multiple myeloma. *Blood, The Journal of the American Society of Hematology*, 139(8):1160–1176, 2022.

[282] Duan Chu and Lai Wei. Nonsynonymous, synonymous and nonsense mutations in human cancer-related genes undergo stronger purifying selections than expectation. *BMC cancer*, 19(1):1–12, 2019.

[283] Yogita Sharma, Milad Miladi, Sandeep Dukare, Karine Boulay, Maiwen Caudron-Herger, Matthias Groß, Rolf Backofen, and Sven Diederichs. A pan-cancer analysis of synonymous mutations. *Nature communications*, 10(1):1–14, 2019.

[284] Thierry Soussi, Peter EM Taschner, and Yardena Samuels. Synonymous somatic variants in human cancer are not infamous: a plea for full disclosure in databases and publications. *Human mutation*, 38(4):339–342, 2017.

[285] Huajing Teng, Wenqing Wei, Qinglan Li, Meiying Xue, Xiaohui Shi, Xianfeng Li, Fengbiao Mao, and Zhongsheng Sun. Prevalence and architecture of posttranscriptionally impaired synonymous mutations in 8,320 genomes across 22 cancer types. *Nucleic acids research*, 48(3):1192–1205, 2020.

[286] Fran Supek, Belén Miñana, Juan Valcárcel, Toni Gabaldón, and Ben Lehner. Synonymous mutations frequently act as driver mutations in human cancers. *Cell*, 156(6):1324–1335, 2014.

[287] Typhaine Paysan-Lafosse, Matthias Blum, Sara Chuguransky, Tiago Grego, Beatriz Lázaro Pinto, Gustavo A Salazar, Maxwell L Bileschi, Peer Bork, Alan Bridge, Lucy Colwell, et al. Interpro in 2022. *Nucleic acids research*, 51(D1):D418–D427, 2023.

[288] Yoshimi Takai, Takuya Sasaki, and Takashi Matozaki. Small gtp-binding proteins. *Physiological reviews*, 81(1):153–208, 2001.

[289] James C Stroud, Cristina Lopez-Rodriguez, Anjana Rao, and Lin Chen. Structure of a tonebp–dna complex reveals dna encircled by a transcription factor. *Nature structural biology*, 9(2):90–94, 2002.

[290] Xiaomin Chen, Uwe Vinkemeier, Yanxiang Zhao, David Jeruzalmi, James E Darnell, and John Kuriyan. Crystal structure of a tyrosine phosphorylated stat-1 dimer bound to dna. *Cell*, 93(5):827–839, 1998.

[291] Robert A Kyle, Terry M Therneau, S Vincent Rajkumar, Janice R Offord, Dirk R Larson, Matthew F Plevak, and L Joseph Melton III. A long-term study of prognosis in monoclonal gammopathy of undetermined significance. *New England Journal of Medicine*, 346(8):564–569, 2002.

[292] Antonio Palumbo, Hervé Avet-Loiseau, Stefania Oliva, Henk M Lokhorst, Hartmut Goldschmidt, Laura Rosinol, Paul Richardson, Simona Caltagirone, Juan José Lahuerta, Thierry Facon, et al. Revised international staging system for multiple myeloma: a report from international myeloma working group. *Journal of clinical oncology*, 33(26):2863, 2015.

[293] Sarah A Holstein and Philip L McCarthy. Immunomodulatory drugs in multiple myeloma: mechanisms of action and clinical experience. *Drugs*, 77:505–520, 2017.

[294] Vallari Shah, David C Johnson, Amy L Sherborne, Sidra Ellis, Frances M Aldridge, Julie Howard-Reeves, Farzana Begum, Amy Price, Jack Kendall, Laura Chiecchio, et al. Subclonal tp53 copy number is associated with prognosis in multiple myeloma. *Blood, The Journal of the American Society of Hematology*, 132(23):2465–2469, 2018.

[295] Aneta Mikulasova, Cody Ashby, Ruslana G Tytarenko, Pingping Qu, Adam Rosenthal, Judith A Dent, Katie R Ryan, Michael A Bauer, Christopher P Wardell, Antje Hoering, et al. Microhomology-mediated end joining drives complex rearrangements and overexpression of myc and pvt1 in multiple myeloma. *Haematologica*, 105(4):1055, 2020.

[296] Nadine Abdallah, Linda B Baughn, S Vincent Rajkumar, Prashant Kapoor, Morie A Gertz, Angela Dispenzieri, Martha Q Lacy, Suzanne R Hayman, Francis K Buadi, David Dingli, et al. Implications of myc rearrangements in newly diagnosed multiple myeloma. *Clinical Cancer Research*, 26(24):6581–6588, 2020.

[297] Jan B Egan, Chang-Xin Shi, Waibhav Tembe, Alexis Christoforides, Ahmet Kurdoglu, Shripad Sinari, Sumit Middha, Yan Asmann, Jessica Schmidt, Esteban Braggio, et al. Whole-genome sequencing of multiple myeloma from diagnosis to plasma cell leukemia reveals genomic initiating events, evolution, and clonal tides. *Blood, The Journal of the American Society of Hematology*, 120(5):1060–1066, 2012.

[298] Brian A Walker, Konstantinos Mavrommatis, Christopher P Wardell, T Cody Ashby, Michael Bauer, Faith Davies, Adam Rosenthal, Hongwei Wang, Pingping Qu, Antje Hoering, et al. A high-risk, double-hit, group of newly diagnosed myeloma identified by genomic analysis. *Leukemia*, 33(1):159–170, 2019.

[299] Rose Oughtred, Chris Stark, Bobby-Joe Breitkreutz, Jennifer Rust, Lorrie Boucher, Christie Chang, Nadine Kolas, Lara O'Donnell, Genie Leung, Rochelle McAdam, et al. The biogrid interaction database: 2019 update. *Nucleic acids research*, 47 (D1):D529–D541, 2019.

[300] Edward L Huttlin, Raphael J Bruckner, Jose Navarrete-Perea, Joe R Cannon, Kurt Baltier, Fana Gebreab, Melanie P Gygi, Alexandra Thornock, Gabriela Zarraga, Stanley Tam, et al. Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell*, 184(11):3022–3040, 2021.

[301] Emma Persson, Miguel Castresana-Aguirre, Davide Buzzao, Dimitri Guala, and Erik LL Sonnhammer. Funcoup 5: functional association networks in all domains of life, supporting directed links and tissue-specificity. *Journal of Molecular Biology*, 433(11):166835, 2021.

[302] Gregorio Alanis-Lobato, Miguel A Andrade-Navarro, and Martin H Schaefer. Hippie v2. 0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic acids research*, page gkw985, 2016.

[303] Chan Yeong Kim, Seungbyn Baek, Junha Cha, Sunmo Yang, Eiru Kim, Edward M Marcotte, Traver Hart, and Insuk Lee. Humannet v3: an improved database of human gene networks for disease research. *Nucleic acids research*, 50(D1): D632–D639, 2022.

[304] Fan Zheng, Marcus R Kelly, Dana J Ramms, Marissa L Heintschel, Kai Tao, Beril Tutuncuoglu, John J Lee, Keiichiro Ono, Helene Foussard, Michael Chen, et al. Interpretation of cancer mutations using a multiscale map of protein systems. *Science*, 374(6563):eabf3067, 2021.

[305] Georg Kustatscher, Piotr Grabowski, Tina A Schrader, Josiah B Passmore, Michael Schrader, and Juri Rappsilber. Co-regulation map of the human proteome enables identification of protein functions. *Nature biotechnology*, 37(11): 1361–1371, 2019.

[306] Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, et al. The reactome pathway knowledgebase 2022. *Nucleic acids research*, 50(D1):D687–D692, 2022.

[307] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al. The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research*, 51(D1):D638–D646, 2023.

[308] Christian D Huber, Bernard Y Kim, and Kirk E Lohmueller. Population genetic models of gerp scores suggest pervasive turnover of constrained sites across mammalian evolution. *PLoS genetics*, 16(5):e1008827, 2020.

[309] Katherine S Pollard, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*, 20(1):110–121, 2010.

[310] Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W Hillier, Stephen Richards, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):1034–1050, 2005.

[311] Boris Reva, Yevgeniy Antipin, and Chris Sander. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research*, 39 (17):e118–e118, 2011.

[312] Eric Talevich, A Hunter Shain, Thomas Botton, and Boris C Bastian. Cnvkit: genome-wide copy number detection and visualization from targeted dna sequencing. *PLoS computational biology*, 12(4):e1004873, 2016.

[313] Ni Huang, Insuk Lee, Edward M Marcotte, and Matthew E Hurles. Characterising and predicting haploinsufficiency in the human genome. *PLoS genetics*, 6(10): e1001154, 2010.

[314] Julia Steinberg, Frantisek Honti, Stephen Meader, and Caleb Webber. Haploinsufficiency predictions without study bias. *Nucleic acids research*, 43(15):e101–e101, 2015.

[315] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. Ncbi geo: archive for functional genomics data sets—update. *Nucleic acids research*, 41(D1):D991–D995, 2012.

[316] Antonio Garcia-Gomez, Tianlu Li, Carlos de la Calle-Fabregat, Javier Rodríguez-Ubreva, Laura Ciudad, Francesc Català-Moll, Gerard Godoy-Tena, Montserrat Martín-Sánchez, Laura San-Segundo, Sandra Muntión, et al. Targeting aberrant dna methylation in mesenchymal stromal cells as a treatment for myeloma bone disease. *Nature communications*, 12(1):421, 2021.

[317] Norma Carmen Gutiérrez, María Eugenia Sarasquete, I Misiewicz-Krzeminska, M Delgado, J De Las Rivas, FV Ticona, E Ferminan, P Martin-Jimenez, C Chillon, A Risueno, et al. Deregulation of microrna expression in the different genetic subtypes of multiple myeloma and correlation with gene expression profiling. *Leukemia*, 24(3):629–637, 2010.

[318] Yiming Zhou, Lijuan Chen, Bart Barlogie, Owen Stephens, Xiaosong Wu, David R Williams, Marie-Astrid Cartron, Frits van Rhee, Bijay Nair, Sarah Waheed, et al. High-risk myeloma is associated with global elevation of mirnas and overexpression of eif2c2/ago2. *Proceedings of the National Academy of Sciences*, 107(17):7904–7909, 2010.

[319] Marta Lionetti, Marta Biasiolo, Luca Agnelli, Katia Todoerti, Laura Mosca, Sonia Fabris, Gabriele Sales, Giorgio Lambertenghi Deliliers, Silvio Bicciato, Luigia Lombardi, et al. Identification of microrna expression patterns and definition of a microrna/mrna regulatory network in distinct molecular groups of multiple myeloma. *Blood, The Journal of the American Society of Hematology*, 114(25): e20–e26, 2009.

[320] Marta Biasiolo, Gabriele Sales, Marta Lionetti, Luca Agnelli, Katia Todoerti, Andrea Bisognin, Alessandro Coppe, Chiara Romualdi, Antonino Neri, and Stefania Bortoluzzi. Impact of host genes and strand selection on mirna and mirna* expression. *PloS one*, 6(8):e23854, 2011.

[321] N Amodio, MT Di Martino, U Foresta, E Leone, M Lionetti, M Leotta, AM Gullà, MR Pitari, F Conforti, MJCD Rossi, et al. mir-29b sensitizes multiple myeloma cells to bortezomib-induced apoptosis through the activation of a feedback loop with the transcription factor sp1. *Cell death & disease*, 3(11):e436–e436, 2012.

[322] Marina Bolzoni, Paola Storti, Sabrina Bonomini, Katia Todoerti, Daniela Guasco, Denise Toscani, Luca Agnelli, Antonino Neri, Vittorio Rizzoli, and Nicola Giuliani. Immunomodulatory drugs lenalidomide and pomalidomide inhibit multiple myeloma-induced osteoclast formation and the rankl/opg ratio in the myeloma microenvironment targeting the expression of adhesion molecules. *Experimental hematology*, 41(4):387–397, 2013.

[323] Domenica Ronchetti, Katia Todoerti, Giacomo Tuana, Luca Agnelli, Laura Mosca, Marta Lionetti, Sonia Fabris, Patrizia Colapietro, Monica Miozzo, Manlio Ferrarini, et al. The expression pattern of small nucleolar and small cajal body-specific rnas characterizes distinct molecular subtypes of multiple myeloma. *Blood cancer journal*, 2(11):e96–e96, 2012.

[324] Ivyna Pau Ni Bong, Ching Ching Ng, Norodiyah Othman, and Ezalia Esa. Gene expression profiling and in vitro functional studies reveal rad54l as a potential therapeutic target in multiple myeloma. *Genes & Genomics*, 44(8):957–966, 2022.

[325] Jayakumar R Nair, Justin Caserta, Krista Belko, Tyger Howell, G Fetterly, Carmen Baldino, and Kelvin P Lee. Novel inhibition of pim2 kinase has significant antitumor efficacy in multiple myeloma. *Leukemia*, 31(8):1715–1726, 2017.

[326] Antonio Sacco, Cinzia Federico, Katia Todoerti, Bachisio Ziccheddu, Valentina Palermo, Arianna Giacomini, Cosetta Ravelli, Federica Maccarinelli, Giada Bianchi, Angelo Belotti, et al. Specific targeting of the kras mutational landscape in myeloma as a tool to unveil the elicited antitumor activity. *Blood, The Journal of the American Society of Hematology*, 138(18):1705–1720, 2021.

[327] Debora Soncini, Claudia Martinuzzi, Pamela Becherini, Elisa Gelli, Samantha Ruberti, Katia Todoerti, Luca Mastracci, Paola Contini, Antonia Cagnetta, Antonella Laudisi, et al. Apoptosis reprogramming triggered by splicing inhibitors sensitizes multiple myeloma cells to venetoclax treatment. *Haematologica*, 107 (6):1410, 2022.

[328] Jairo Navarro Gonzalez, Ann S Zweig, Matthew L Speir, Daniel Schmelter, Kate R Rosenbloom, Brian J Raney, Conner C Powell, Luis R Nassar, Nathan D Maulding, Christopher M Lee, et al. The ucsc genome browser database: 2021 update. *Nucleic acids research*, 49(D1):D1046–D1057, 2021.

[329] Jérôme Pagès. *Multiple factor analysis by example using R*. CRC Press, 2014.

[330] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[331] Terry Therneau et al. A package for survival analysis in s. *R package version*, 2 (7), 2015.

[332] Kyohei Nakamura, Mark J Smyth, and Ludovic Martinet. Cancer immunoediting and immune dysregulation in multiple myeloma. *Blood, The Journal of the American Society of Hematology*, 136(24):2731–2740, 2020.

[333] Timothy M Schmidt, Rafael Fonseca, and Saad Z Usmani. Chromosome 1q21 abnormalities in multiple myeloma. *Blood cancer journal*, 11(4):83, 2021.

[334] H Chang, X Qi, A Jiang, W Xu, T Young, and D Reece. 1p21 deletions are strongly associated with 1q21 gains and are an independent adverse prognostic factor for the outcome of high-dose chemotherapy in patients with multiple myeloma. *Bone marrow transplantation*, 45(1):117–121, 2010.

[335] Lucía López-Corral, Norma C Gutiérrez, Maria Belén Vidriales, Maria Victoria Mateos, Ana Rasillo, Ramón García-Sanz, Bruno Paiva, and Jesús F San Miguel. The progression from mgus to smoldering myeloma and eventually to multiple myeloma involves a clonal expansion of genetically abnormal plasma cells. *Clinical cancer research*, 17(7):1692–1700, 2011.

[336] Joan Bladé. Monoclonal gammopathy of undetermined significance. *New England Journal of Medicine*, 355(26):2765–2770, 2006.

[337] Neha Korde, Sigurdur Y Kristinsson, and Ola Landgren. Monoclonal gammopathy of undetermined significance (mgus) and smoldering multiple myeloma (smm): novel biological insights and development of early treatment strategies. *Blood, The Journal of the American Society of Hematology*, 117(21):5573–5581, 2011.

[338] Fiona M Ross, Laura Chiecchio, GianPaolo Dagrada, Rebecca KM Protheroe, David M Stockley, Christine J Harrison, Nicholas CP Cross, Alex J Szubert, Mark T Drayson, and Gareth J Morgan. The t (14; 20) is a poor prognostic factor in myeloma but is associated with long-term stable disease in monoclonal gammopathies of undetermined significance. *haematologica*, 95(7):1221, 2010.

[339] Maurizio Affer, Marta Chesi, WD Chen, Jonathan J Keats, Yulia N Demchenko, K Tamizhmani, VM Garbitt, DL Riggs, LA Brents, AV Roschke, et al. Promiscuous myc locus rearrangements hijack enhancers but mostly super-enhancers to dysregulate myc expression in multiple myeloma. *Leukemia*, 28(8):1725–1735, 2014.

[340] Zhiwen He, Julie O'Neal, William C Wilson, Nitin Mahajan, Jun Luo, Yinan Wang, Mack Y Su, Lan Lu, James B Skeath, Deepta Bhattacharya, et al. Deletion of rb1 induces both hyperproliferation and cell death in murine germinal center b cells. *Experimental hematology*, 44(3):161–165, 2016.

[341] Katarina K Jovanović, Guillaume Escure, Jordane Demonchy, Alexandre Willaume, Zoe Van de Wyngaert, Meryem Farhat, Paul Chauvet, Thierry Facon, Bruno Quesnel, and Salomon Manier. Deregulation and targeting of tp53 pathway in multiple myeloma. *Frontiers in oncology*, 8:665, 2019.

[342] Shehab Fareed Mohamed, Maliha Khan, Andres Quesada, Junsheng Ma, Pei Lin, C Cameron Yin, Koji Sasaki, Gautam Borthakur, Naveen Pemmaraju, Qaiser Bashir, et al. Disease characteristics of multiple myeloma involving braf mutations. *Blood*, 138:4755, 2021.

[343] Sergiu Pasca, Ciprian Tomuleasa, Patric Teodorescu, Gabriel Ghiaur, Delia Dima, Vlad Moisoiu, Cristian Berce, Cristina Stefan, Aaron Ciechanover, and Herman Einsele. Kras/nras/braf mutations as potential targets in multiple myeloma. *Frontiers in Oncology*, 9:1137, 2019.

[344] P Liebisch, C Wendl, A Wellmann, A Kröber, G Schilling, H Goldschmidt, H Einsele, C Straka, M Bentz, S Stilgenbauer, et al. High incidence of trisomies

1q, 9q, and 11q in multiple myeloma: results from a comprehensive molecular cytogenetic analysis. *Leukemia*, 17(12):2535–2537, 2003.

[345] Peter Liebisch, Daniel Scheck, Seiichi Alvise Erné, Alexander Wellmann, Christiane Wendl, Sibylle Janczik, Sonja Kolmus, Alexander Kröber, Hermann Einsele, Christian Straka, et al. Duplication of chromosome arms 9q and 11q: Evidence for a novel, 14q32 translocation–independent pathogenetic pathway in multiple myeloma. *Genes, Chromosomes and Cancer*, 42(1):78–81, 2005.

[346] Anil Aktas Samur, Stephane Minvielle, Masood Shammas, Mariateresa Fulciniti, Florence Magrangeas, Paul G Richardson, Philippe Moreau, Michel Attal, Kenneth C Anderson, Giovanni Parmigiani, et al. Deciphering the chronology of copy number alterations in multiple myeloma. *Blood cancer journal*, 9(4):39, 2019.

[347] Adil Doganay Duru, Tolga Sutlu, Ann Wallblom, Katarina Uttervall, Johan Lund, Birgitta Stellan, Gösta Gahrton, Hareth Nahi, and Evren Alici. Deletion of chromosomal region 8p21 confers resistance to bortezomib and is associated with upregulated decoy trail receptor expression in patients with multiple myeloma. *PLoS One*, 10(9):e0138248, 2015.