A Comprehensive Targeted Panel of 282 Genes: Unveiling Key Biomarkers in Multiple Myeloma

Vivek Ruhela 0000-0003-4652-2376[1,2], Rupin Oberoi 0009-0006-9541-1903[1], Ritu Gupta 0000-0001-5364-4086[3,✉], and Anubha Gupta 0000-0002-7752-1926[1,✉]

[1]SBILab, Deptt. of ECE & Centre of Excellence in Healthcare, Indraprastha Institute of Information Technology-Delhi (IIIT-D), India
[2]Dept. of Computational Biology, Indraprastha Institute of Information Technology-Delhi (IIIT-D), India
[3]Laboratory Oncology Unit, Dr.B.R.A.IRCH, All India Institute of Medical Sciences

Multiple myeloma (MM) is a hematological cancer that evolves from the benign precursor stage termed monoclonal gammopathy of undetermined significance (MGUS). Understanding the pivotal biomarkers, genomic events, and gene interactions distinguishing MM from MGUS can significantly contribute to early detection and an improved understanding of MM's pathogenesis. In this study, we present 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 *bio-inspired graph network learning based gene gene interactions (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 protein-protein interaction (PPI) networks and analyzes the genomic features from 1154 MM and 61 MGUS samples, respectively. The proposed 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. 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 both 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.

AI in Cancer | Haematological malignancy | Multiple Myeloma | MGUS | Genomic Aberrations | ShAP | FAMD | PPI

Correspondence: ***anubha@iiitd.ac.in, drritugupta@gmail.com***

Introduction

Multiple Myeloma (MM) is a hematological malignancy characterized by the clonal proliferation of plasma cells within the bone marrow. Monoclonal Gammopathy of Undetermined Significance (MGUS) and MM are both plasma cell disorders, representing distinct stages along the disease progression continuum. MM entails malignant plasma cell proliferation, organ damage, and clinical symptoms, whereas MGUS is a precursor condition with no apparent clinical manifestations. Progression

from MGUS to MM occurs at a rate of 1% per year; thus, all MGUS patients do not transition to overt MM during their lifetime [1]. In this context, identifying MGUS individuals likely to progress to MM is crucial for timely intervention and improved outcomes. The clinical distinction between the two conditions primarily relies on tumor load, reflected in monoclonal proteinemia, % bone marrow plasma cell infiltration, and end organ damage [2]. This emphasizes the need to delve into genomic markers and gene-gene interactions to enhance diagnostic accuracy.

The advanced genomic profiling techniques like whole-exome sequencing (WES) and whole-genome sequencing (WGS) enable thorough examination of genomic aberrations in MM and MGUS. They have proven crucial in identifying key events, including copy number variations (CNVs) and structural variations (SVs) [3]. Numerous MM-related genomic studies have reported significant CNVs, such as del(1p), gain(1q), del(13q), and del(17p), alongside key SVs like translocation involving IgH (e.g. t(4;14), t(11;14), t(14;16), t(8;14)), MYC rearrangement, etc., shedding light on their association with MM prognosis [4–9]. Moreover, recent studies have highlighted the impact of minor genomic alterations on MM patients' clinical outcomes [10–13]. Recently, biallelic alterations in TP53 and DIS3 gene have been reported as high-risk markers in MM [14].

Integrating protein-protein interactions (PPI) with WES and WGS variant profiles can provide crucial insights into genomic biomarkers essential for MGUS to MM progression. Distinguishing MM from MGUS via gene-gene interactions can revolutionize clinical approaches, aiding early detection. This can enable proactive monitoring and intervention for high-risk MGUS patients, while sparing low-risk MM cases from aggressive treatments [15].

Several targeted sequencing panels have been devised to comprehensively profile the genomics complexity of MM [16–18]. These panels encompass critical genomic aberrations related to MM. For instance, a 26-gene panel focused on prevalent mutations in previously published MM-relevant genes [16], 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 [19]. A more extensive 228-gene panel covered various alterations, including SNVs, CNVs, translocations involving IgH and MYC genes [17]. 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 [18]. Lastly, the largest gene panel of 465 genes was designed and validated for MM-related SNVs, CNVs, and translocations related to IGH gene only [20]. However, these panels were designed using only MM samples and hence, lacked markers and interactions distinguishing MM from MGUS that can give insights on MM pathogenesis.

Advancements in machine learning (ML) and deep learning (DL) have revolutionized bioinformatics, enabling precise biomarker discovery for early disease detection. Researchers utilize these tools to predict Protein-Protein Interactions (PPIs) and unravel crucial gene interactions in cancer. Notably, models like DeepPPI (Deep neural networks for Protein–Protein Interactions prediction) predict gene interactions based on shared protein descriptors [21]. Additionally, ML and DL-driven approaches facilitate inferring semantic similarity of gene ontology terms using PPIs [22–24]. Despite these strides, no computational model was tailored to pinpoint pivotal biomarkers and gene interactions distinguishing Multiple Myeloma (MM) from Monoclonal Gammopathy of Undetermined Significance (MGUS).

The fusion of genomic mutation profiles with Protein-Protein Interaction (PPI) data remains an underexplored domain. Graph Deep Learning (GDL) emerges as a potent tool in genomics, promising profound insights from intricate biological graph

data structures. Genomic data inherently manifests a graph-like structure, where nodes embody biological entities (genes, proteins) and edges depict relationships. In this realm, Graph Convolutional Networks (GCNs) play a pivotal role. GCNs are instrumental in processing and analyzing graph-structured genomic data, where biological entities and their relationships are effectively modeled as a graph, particularly for disease classification [25]. In our previous work, we seamlessly integrated exonic mutation profiles with PPI to unveil key distinguishing biomarkers in MM and MGUS through the bio-inspired BDL-SP model [26]. Expanding on this, we pursued heightened precision in identifying biomarkers and gene interactions by integrating gene interactions from nine diverse PPI databases.

Motivated to bridge this gap, we envisioned a targeted sequencing panel for a thorough genomic profiling of MM, aiming to capture the unique characteristics of MM and MGUS. To address this challenge, we introduced a novel AI-based bio-inspired Graph network learning based directed gene-gene interactions (Bio-DGI), 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 exomic 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 graph convolution network layers and multi-head attention modules compared to traditional machine learning (ML) or deep learning (DL) models that relied solely on exomic mutational profiles. Moreover, our proposed Bio-DGI model outperformed exhaustive benchmarking against several baseline ML and DL models, including both quantitative and qualitative evaluations. We further identified top-ranked genes and genomic features utilizing the ShAP (SHapley Additive exPlanations) algorithm [27].

To delve deeper, we identified five distinct gene communities using Leiden algorithm [28] by utilizing the adjacency matrices derived from five trained Bio-DGI (PPI9) classifiers. This analysis shed light on the influential genes within these communities, quantified through Katz centrality scores [29]. Interestingly, we have highlighted the genes that are sitting in the central position in the gene community and may play significant roles in MM pathogenesis.

We conducted a meticulous analysis of 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 comprehensive genomic profiling of Multiple Myeloma (MM). The inclusion of genes from our proposed panel in MM-related pathways strongly underlines their pivotal roles in the disease progression from MGUS to 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.

Materials & Methods

Whole-exome sequencing datasets of MM and MGUS patients.

In this study, we included tumor-normal pairs of bone marrow (BM) samples from a MM cohort of 1154 samples and a MGUS cohort of 61 samples sourced from three global repositories of whole-exome sequencing (WES) data. For the MM cohort, 1072 samples were acquired through authorized access to the MMRF dbGaP study (phs000748; phs000348) predominantly compris-

ing American population samples [30]. We also downloaded processed MMRF datasets (version IA12) containing copy number variations (CNVs), structural variations (SVs), and clinical data from MMRF Research Gateway. Additionally, we included 82 MM samples from an AIIMS dataset representing the Indian population [26]. In the MGUS cohort, we incorporated 28 MGUS samples from the AIIMS dataset and 33 samples from the European Genome-phenome Archive (EGA) data. Methods

Whole exome sequencing Data Pre-Processing for SNV identification.

The WES data obtained from AIIMS and EGA contained the raw fastq files and from MMRF dataset contained the processed VCF (Variant Call Format) files. The computational workflow for the single nucleotide variations (SNVs) identification, genomic annotation of SNVs, SNVs filtration, SNVs grouping, and the identification of significantly altered genes was taken from our previous related study [26]. Briefly, raw fastq files from AIIMS and EGA datasets were processed using the standard exome sequencing pipeline [31]. Similar to the MMRF data, the SNVs in AIIMS and EGA WES dara were extracted using MuSE [32], Mutect2 [33], VarScan2 [34], and Somatic- Sniper [35] variant callers. The SNVs in AIIMS, EGA and MMRF datasets were annotated using ANNOVAR database [36]. The annotated SNVs were categorized into three categories based on their functional significance i.e. synonymous SNVs, non-synonymous SNVs and other SNVs. The benign SNVs were filtered out using FATHMM-XF [37]. Lastly, annotated SNVs were pooled for MM and MGUS separately and analyzed for identifying significantly altered genes using the 'dndscv' tool [38]. 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 protein-protein interactions (PPI) were extracted from 9 PPI databases (BioGrid [39], Bio-Plex [40], FunCoup [41], HIPPIE [42], HumanNet [43], IntegratedAssociationCorrNet [44], ProteomHD [45], Reactome [46], and STRING [47]). 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 9 PPI databases to get the merged adjacency matrix (Table-S2, Supplementary File-1). Besides the PPI interactions, we extracted 26 genomic features that include distributive statistics (median and standard deviation) of variant allele frequency (VAF), allele depth (AD), and four variant conservation scores (GERP [48], PhyloP [49], PhastCons [50], and Mutation assessor [51]). The details of pre-processing workflow available in Supplementary File-2.

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 Monoclonal Gammopathy of Undetermined Significance (MGUS) to Multiple Myeloma (MM). We introduced an innovative Graph Convolutional Network (GCN)-based bio-inspired model named "Biological Network for Directed Gene-Gene Interaction Learning" (Bio-DGI). The Bio-DGI model, depicted in Figure-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 Protein-Protein Interaction (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 was then utilized for training purposes of the model, referred to as Bio-DGI (PPI9). 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-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.

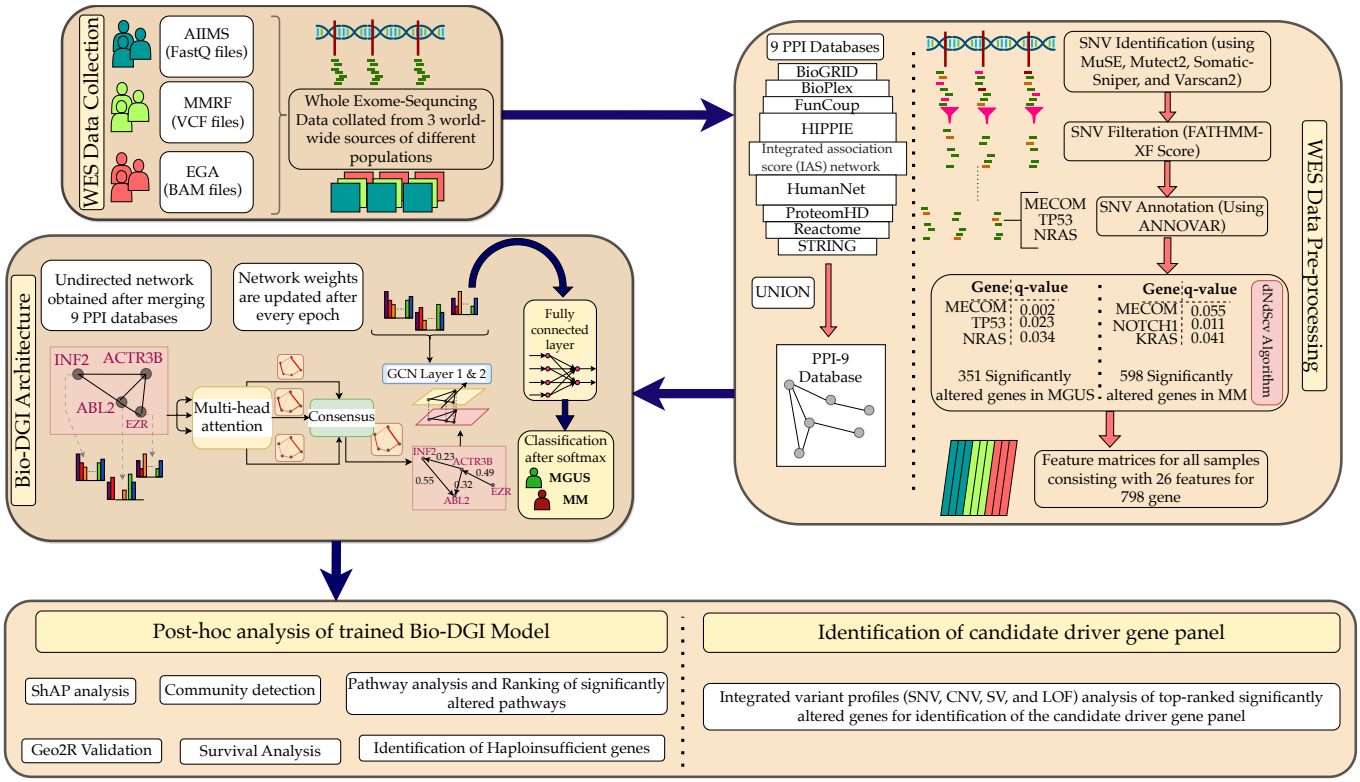


Fig. 1. Infographic representation of proposed AI-based bio-inspired Bio-DGI model and post-hoc analysis for identifying pivotal genomic biomarkers that can distinguish MM from MGUS. In the proposed AI-based workflow, the BAM files sourced from EGA and AIIMS datasets, along with VCF files from MMRF dataset, 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 protein-protein interaction (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 is 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.

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, that are followed by an attention consensus module for taking 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 that 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 the training of five best performing classifiers. All the five classifiers with learned adjacency matrices were saved for further post-hoc analysis. We used ShAP algorithm for post-hoc analysis of Bio-DGI model classifiers to get top-performing genes and genomic features that 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.

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 [52]. 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 both major and minor classes, fostering a more equitable predictive capability.

Qualitative application-aware post-hoc benchmarking of Bio-DGI model using ShAP.

The ShAP (SHapley Additive exPlanations) algorithm serves as 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 [26] (Figure 4A, Figure 4B, and Table 2A, Table 2B within [26]), 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 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 were collected both at the classifier and sample levels for all genomic attributes. 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 highest 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 both the group and sample levels.

We extended our analysis by comparing 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 [53], , IntoGen [54], COSMIC [55], and TargetDB [56] at the gene level. For model validation, we extracted 1064 cancer genes from the OncoKB database for oncogenes and tumor-suppressor genes. From the COSMIC database, we utilized 318 oncogenes and 320 tumor-suppressor genes.

We utilized the IntoGen database (https://www.intogen.org/) and MM-related studies [11, 57] to compile a catalog of MM driver genes. Additionally, 180 actionable genes from COSMIC and 135 from TargetDB helped infer actionable genes. We systematically categorized the top-ranked significantly altered genes into distinct groups based on their biological characteristics, including Oncogenes (OGs), Tumor-Suppressor genes (TSGs), Onco-driver genes (ODGs), and Actionable genes (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.

2.2.5 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 [58] 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 (CN) value of 2 across all samples, focusing on genes with CN values that deviated from 2. Turning to structural variants (SVs), we utilized the structural variants identified through the in-house SV identification workflow developed by the Translational Genomics Research Institute (TGen) and Delly tool [59] to pinpoint SVs in WGS samples from the MMRF dataset. Our analysis centered on significantly altered genes ranked in the top-500 whose genomic regions were affected by structural variants, spanning insertions, inversions, deletions, duplications, and translocations.

Furthermore, our investigation extended to encompass genes marked by loss-of-function aberrations. The loss-of-function 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 loss-of-function, we assessed all transcripts against specific criteria. If a transcript satisfied any of the following conditions: 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 loss-of-function [60]. This evaluation was conducted across all samples within the MM and MGUS cohorts to pinpoint genes featuring loss-of-function anomalies.

Identification of Haploinsufficient genes.

To assess the likelihood of genes exhibiting haplo-insufficiency, we draw upon two previously published haploinsufficiency prediction scores: the genome-wide haploinsufficiency score (GHIS) [61] and the DECIPHER score [60]. The DECIPHER score amalgamates patient genomic data, evolutionary profiles, functional and network properties to predict the likelihood of the haplo-insufficiency. 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 haplo-insufficient genes, revealing their crucial role in diseases. This deepens our understanding of genes that lack proper function when only one copy is present.

Identification of significantly altered pathways and pathway ranking.

The noteworthy genes highlighted by 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 [62–64], which aided in identifying 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.

Identification of gene communities using learned adjacency matrices.

We employed a five-fold cross-validation training strategy to 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.

Geo2R Validation of top-500 significantly altered genes.

We conducted a thorough validation using the Geo2R tool [65] to validate significantly altered genes that were ranked in top-500 in relation to multiple previously published studies focused on multiple myeloma (MM). In total, we utilized micro-array and miRNA data from 11 distinct MM-related studies for this validation process [66–78]. To ensure rigorous assessment, we exclusively considered genes that displayed significant deregulation and maintained an adjusted p-value ≤ 0.05 in each validation instance.

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 Single Nucleotide Variants, Copy Number Variants, Structural Variants, loss-of-function, 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. To ensure precision, we applied individualized filtering criteria tailored to each variant profile. For instance, when filtering based on the gene SNV profile, criteria included the number of non-synonymous mutations in the MM cohort, the presence of non-synonymous or other category mutations in MM cohort samples, and the ratio of non-synonymous 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 afore-

mentioned gene prioritization workflow and also had at least one dataset validation in Geo2R analysis for further in-depth investigation. The genes that were either transformative or disease initiating were retained. Detailed workflow for the gene prioritization process are available in Figure-5, 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 UCSC Genome database [79].

Workflow for comprehensive survival analysis.

In the end, we performed the survival analysis of 282 genes filtered out using workflow shown above (Section 2.2.10). 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 (non-synonymous + other) SNVs as a prognostic factor for univariate survival analysis. Similarly, for CNV, SV, and LOF profile, 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 ≤ 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) [80] 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.

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 (DL) models in this study, we employed PyTorch (version 1.12.0+cu113) [81]. Additionally, survival analysis was conducted using the statistical programming language R (version 4.3.1) with the "survival" package [82] (version 3.5.5).

Results

Cohort Description.

In this comprehensive study, we analyzed two distinct cohorts related to Multiple Myeloma (MM) and Monoclonal Gammopathy of Undetermined Significance (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 overall survival (OS) time and event data for MM samples retrieved from both 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.

Identification of significantly altered genes.

We utilized the dndscv tool (depicted in the pre-processing block of Figure-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.

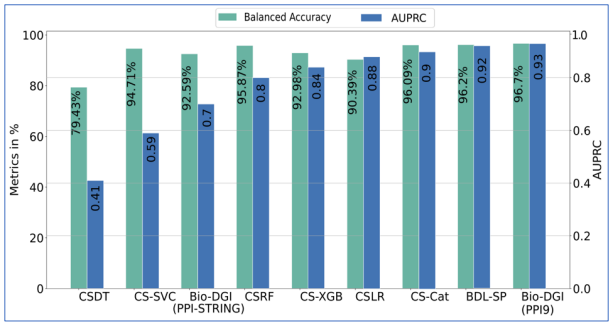
Benchmarking of proposed Bio-DGI (PPI9) model.

Employing our AI-driven Bio-DGI workflow (depicted in Figure-1), we trained the Bio-DGI (PPI9) model 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 (area under the precision-recall curve). Specifically, the Bio-DGI (PPI9) model achieved the highest 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 highest 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-2 (A), (B), and (C) 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 oncogenes (OG), tumor-suppressor genes (TSG), both oncogenes and driver genes (ODG), and actionable genes (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-1).

Our analysis revealed that the proposed Bio-DGI (PPI9) model exhibited the highest count of identified oncogenes (OG), tumor-suppressor genes (TSG), both oncogene and driver genes (ODG), and actionable genes (AG) in both the top-250 and top-500 gene lists (Table-2). Specifically, the Bio-DGI (PPI9) model detected 23 and 28 oncogenes 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

(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}

(c)

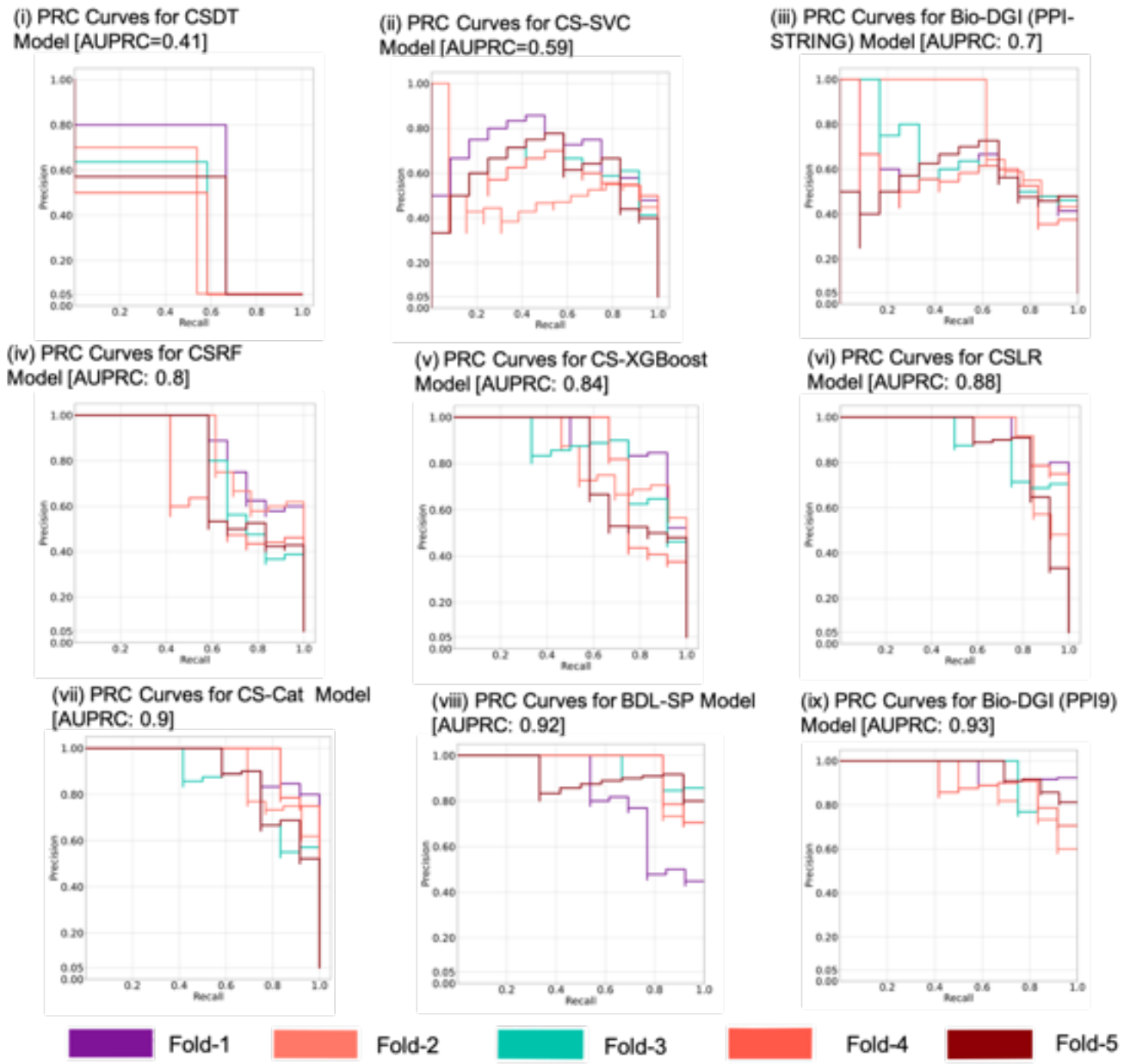


Fig. 2 *(previous page)*. Quantitative benchmarking of proposed Bio-DGI model. (a) Comparison of balanced accuracy and AUPRC score of Bio-DGI (PPI9) model with other baseline ML and DL models, (b) Confusion matrix of top-performing models including Bio-DGI (PPI-STRING) model, and (c) Precision-recall curves (PRC) for all five folds of (i) CSDT, (ii) CS-SVC, (iii) Bio-DGI (PPI-STRING), (iv) CS-RF, (v) CS-XGB, (vi) CSLR, (vii) CS-Cat, (viii) BDL-SP, and (ix) Bio-DGI (PPI9).

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-1(B). 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-1(C) and Table-1(D). 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 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 (OGs)	31
Tumor-suppressor genes (TSGs)	43
Both oncogene and driver gene (ODGs)	10
Actionable genes (AGs)	19

Table 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 are 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-2.

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

Table 3. 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 (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

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-S1 (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, 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 [26]. 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.

Interpretability of Bio-DGI (PPI9) model using ShAP algorithm.

We utilized the ShAP algorithm for post-hoc model explainability and rank 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 [26], 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-S3 in Supplementary File-3). Furthermore, several previously reported oncogenes (e.g., *MUC16, USP6, BIRC6, VAV1*), tumor-suppressor genes (e.g., *EP400, HLA-B/C, SDHA, MYH11*), both oncogenes and driver genes (e.g., *PABPC1, KRAS, TRRAP, TP53, FGFR3, BRAF*), and actionable genes (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 non-synonymous SNVs, allele depth of synonymous SNVs, and the total number of other SNVs emerged as the top three genomic features. Figure-6 presents the beeswarm plot illustrating the genomic feature ranking from Bio-DGI (PPI9) model post-hoc analysis using ShAP.

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. For each classifier, we applied the Leiden algorithm to the respective learned adjacency matrix 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-3(A)-(E).

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

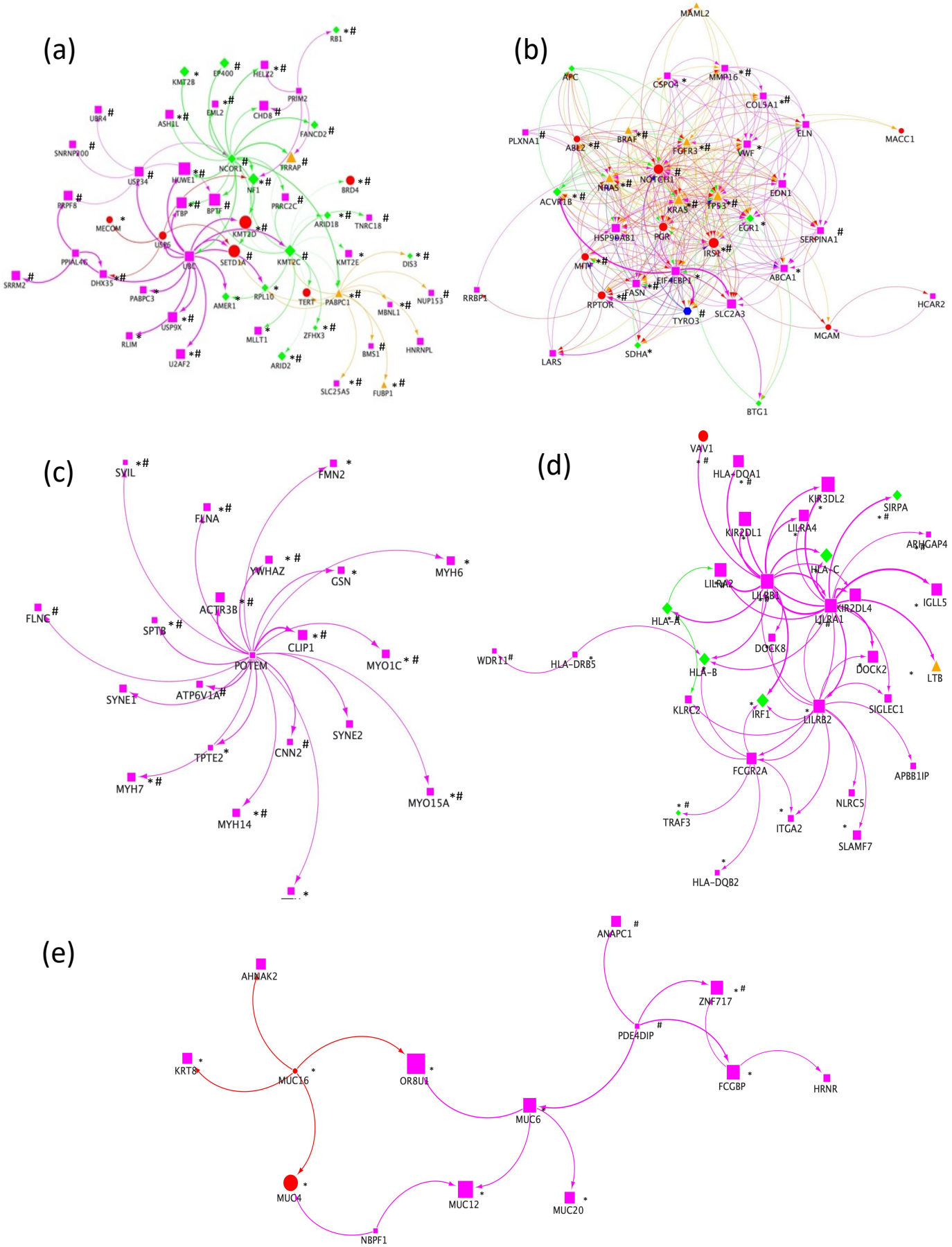

(a)
RB1
KMT2B
EP400
HELZ2
PRIM2
EML2
CHD8
UBR4
ASH1L
FANCD2
SNRNP200
NCOR1
TRRAP
USP34
HUWE1
NF1
BRD4
PRPF8
BPTF
TBP
PRRC2C
MECOM
ARID1B
TNRC18
KMT2D
USP6
SETD1A
KMT2C
KMT2E
DIS3
PPIAL4G
SRRM2
DHX35
UBC
PABPC3
AMER1
RPL10
TERT
PABPC1
NUP153
MBNL1
USP9X
RLIM
MLLT1
ZFHX3
BMS1
HNRNPL
U2AF2
ARID2
SLC25A5
FUBP1
(b)
MAML2
APC
CSPG4
MMP16
COL3A1
ELN
MACC1
PLXNA1
ABL2
BRAF
FGFR3
VWF
NRAS
NOTCH1
EDN1
ACVR1B
KRAS
TP53
HSP90AB1
PGR
EGR1
SERPINA1
MTF
IRS1
ABCA1
EIF4EBP1
FASN
RPTOR
RRBP1
TYRO3
SLC2A3
HCAR2
MGAM
LARS
SDHA
BTG1
(c)
SVIL
FMN2
FLNA
YWHAZ
GSN
MYH6
ACTR3B
FLNC
SPTB
CLIP1
POTEM
MYO1C
SYNE1
ATP6V1A
SYNE2
TPTE2
CNN2
MYH7
MYH14
MYO15A
(d)
VAV1
HLA-DQA1
KIR3DL2
SIRPA
KIR2DL1
LILRA4
ARHGAP4
HLA-C
LILRA2
LILRB1
KIR2DL4
IGLL5
HLA-A
LILRA1
DOCK8
WDR11
HLA-DRB5
HLA-B
DOCK2
LTB
KLRC2
IRF1
LILRB2
SIGLEC1
FCGR2A
APBB1IP
NLRC5
TRAF3
ITGA2
SLAMF7
HLA-DQB2
(e)
ANAPC1
AHNAK2
ZNF717
PDE4DIP
KRT8
MUC16
OR8U1
FCGBP
HRNR
MUC6
MUC4
MUC12
MUC20
NBPF1

Fig. 3 *(previous page).* Directed 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 community, 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.

datasets from MM-related studies, leveraging the Geo2R tool. For this validation, we considered 11 MM-related studies, 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-4(J). Comprehensive details of the Geo2R validation analysis for the top 500 genes are available in Table-S14, Supplementary File-7.

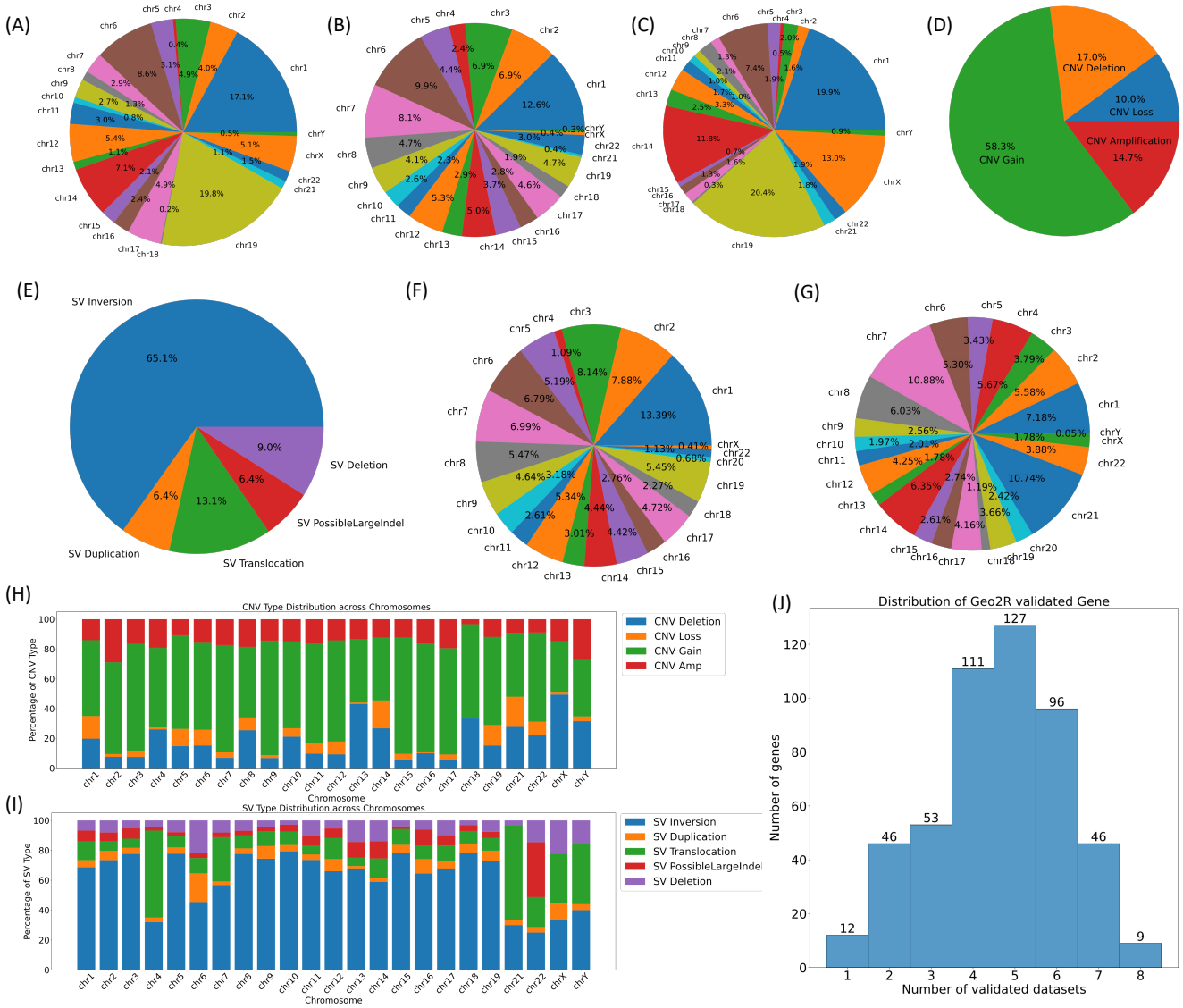


Fig. 4. Genomic Aberrations Overview (CNVs, SVs, and LOF) in MM Samples from AIIMS and MMRF Repositories. In panels (A)-(C), the figure displays the chromosome-wise distribution of CNVs, SVs, and LOF, respectively. Panel (D) presents the distribution of CNV types identified in MM samples from both AIIMS 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 distribution of CNV and SV types for each chromosome individually. 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.

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 for identifying 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 [60]. 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-4(A)). Similarly, chr1 (12.6%), chr6 (9.9%), chr12 (5.3%), and chr14 (5%) showed prominent SV involvement (Figure-4(B)), while chr19 (20%), chr1 (19.9%), chrX (13%), and chr14 (11.8%) were most affected by LOF (Figure-4(C)). The majority of CNVs were gains (58.3%) and deletions (17%) (Figure-4(D)), while inversions (65%) and translocations (13.1%) dominated the SV landscape (Figure-4(E)). Notable chromosomes impacted by inversion SV included chr1, chr3, chr2, and chr7 (Figure-4(F)), and translocations mainly affected chr7, chr21, chr1, and chr14 (Figure-4(G)). The distribution of CNV and SV types within each chromosome highlighted their relative abundance (Figure-4(H) and Figure-4(I)).

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 median SNV count and the number of samples with non-synonymous 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(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(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. To further refine the targeted sequencing regions, we assessed the most affected coding regions using the UCSC Genome database. 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).

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 whole-genome sequencing (WGS) data, etc. Upon scrutinizing the validated variant profiles, we noted that, alongside our proposed panel, Sudha et al. [17] 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 and 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-6 and Table-S18, Supplementary File-9.

Table 4. 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 to 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	[83–86]	No		BRD4, IRS1, ITPR1, KRAS, KRT8, NRAS, TP53, RB1, SLC25A5, and TAF1
2	t(4;14)	FGFR3	[12, 83, 85–87]	Yes	FGFR3	
3	t(14;16)	MAF	[83, 85, 86, 88]	No		FLNA
4	t(14;20)	MAFB	[83, 86, 88]	No		HUWE1, USP9X
5	t(6;14)	CCND3	[12, 86]	No		RB1
6	Amp(1q21)	MCL1, CKS1B, ANP32E or BCL9	[12]	No		KPRP, BRD4, PLEC, USP9X
7	Del(17p13)	TP53	[89]	Yes	TP53	
8	KRAS mutations	KRAS	[83]	Yes	KRAS	
9	NRAS Mutations	NRAS	[83]	Yes	NRAS	
10	LTB Mutations	LTB	[83]	Yes	LTB	
11	DIS3 mutations	DIS3	[83]	Yes	DIS3	
12	EGR1 mutations	EGR1	[83]	Yes	EGR1	
13	MYC Rear-rangement	IGH, IGL, IGK, NSMCE2, TXNDC5, FAM46C, FOXO3, IGJ, PRDM1	[90]	Yes	FAM46C	

Table 5. 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 to 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	[12, 91]	Yes	RB1, DIS3	
3	Del(16q23)	CYLD	[12]	Yes	CYLD	
4	Del(1p21)	FAM46C	[12, 92]	Yes	FAM46C	
5	Del(12p13)	CD27	[86]	No		TRAF2, TRAF3, ATP2B3
6	TP53 Mutations	TP53	[66, 93]	Yes	TP53	
7	BRAF Mutations	BRAF	[94, 95]	Yes	BRAF	
8	Gain(9q)	ABCA1, KCNT1, TRAF2, VPS13A	[96, 97]	Yes	ABCA1, KCNT1, TRAF2, VPS13A	
9	del(14q)	TRAF3	[12]	Yes	TRAF3	
10	del(17p)	TP53	[12]	Yes	TP53	
11	del(8p)	PTK2B, TP53	[98, 99]	Yes	PTK2B, TP53	

Table 6. 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 [18], 2015	47	22 NDMM, 3 pretreated MM samples	WES	SNVs, clonal evolution analysis	16
2	Bolli et al [19], 2016	182	5 MM samples	WGS	SNVs, CNVs, SVs*	19
3	White et al [20], 2018	465	110 MM samples	WGS	SNVs, CNVs, SVs*	39
4	Cutler et al [16], 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 [17], 2022	228	185 MM samples	WGS	SNVs, CNVs, SVs*	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	-

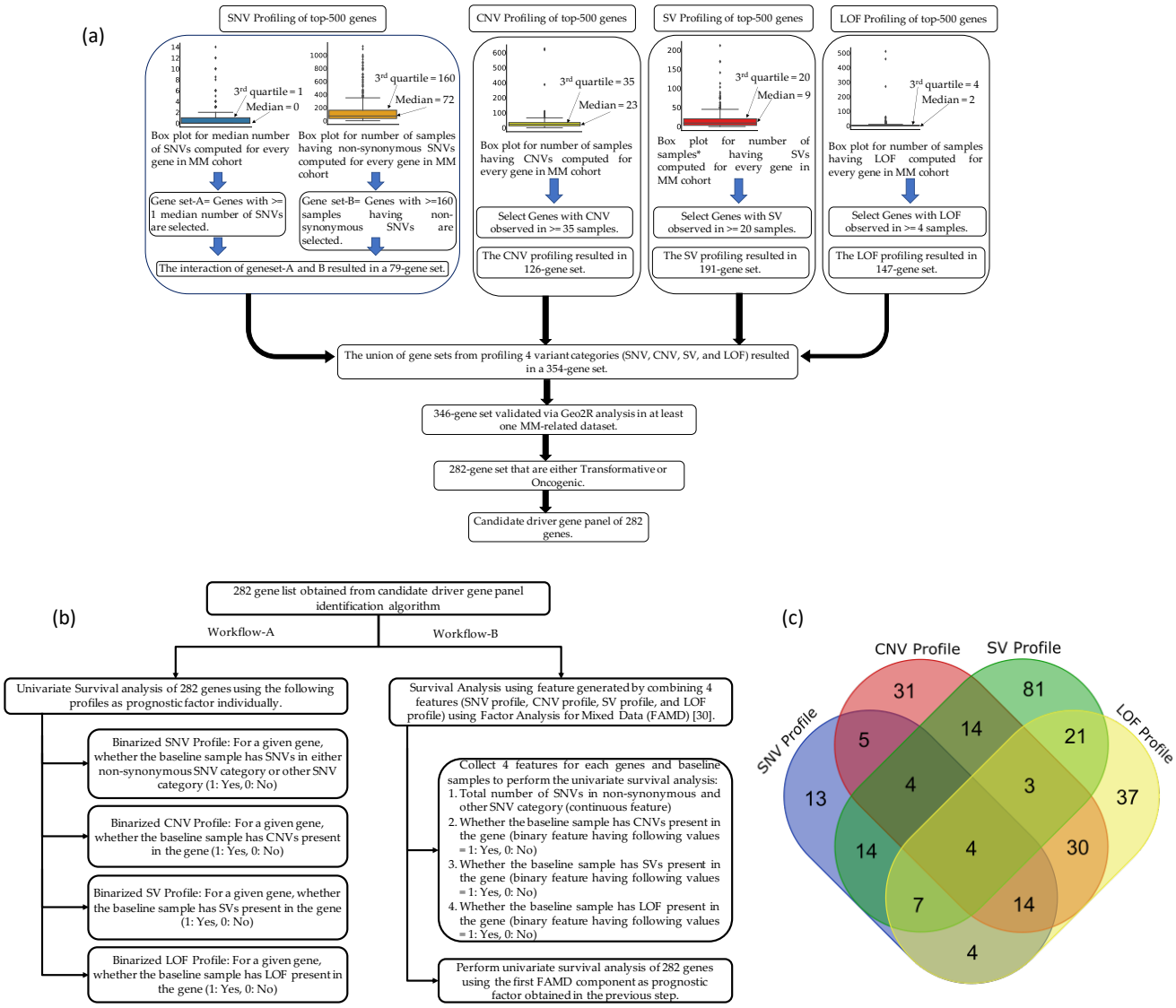


Fig. 5. (a) Workflow Design for the Proposed 282-Gene Panel. The workflow incorporates variant profiles (SNVs, CNVs, SVs, and LOF) to identify MM-relevant genes. Specifically, 79, 126, 191, and 147 gene lists were generated based on SNVs, CNVs, SVs, and LOF variant profiles, respectively. The union of 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 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 profile 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 using. By combining both approaches, out of 282, a total of 239 genes were found to significantly influence clinical outcomes of MM patients.

Clinical relevance of targeted sequencing 282-genes panel.

We performed 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(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 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 FAMD first component. 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 and 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).

Discussion

Multiple Myeloma (MM) is a malignancy that typically progresses from premalignant stages, often starting with Monoclonal Gammopathy of Undetermined Significance (MGUS)[100]. A targeted sequencing panel is important for precise characterization of the genomic alterations to understand risk of progression, enabling timely interventions and ultimately improving patient outcomes. Recent studies have shed light on the genomic events that drive the transformation from premalignant stages to MM [83, 101–103]. 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 [16–20]. 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-1, Table-2, Table-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 highest number of previously reported genes, encompassing known oncogenes (OGs) such as BIRC6, MUC4, NOTCH1, PGR, SETD1A, and VAV1, tumor suppressor genes (TSGs) like DIS3, EP400, MYH11, SDHA, both oncogenes and driver genes (ODGs) such as KRAS, NRAS, TP53, TRRAP, and actionable genes (AGs) including APC, ARID1B, MITF, NFKBIA, and 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 [104]. 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 [105, 106]. We have employed the Geo2R tool for the validation of 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 the identification of significantly dysregulated genes using gene expression or microarray data of previously published studies. We have included 11 MM-related studies for the validation and observed that 488 out of 500 genes were found as 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-6(A)). The substantial impact of synonymous SNVs across various cancer types has been well-documented [107–111].

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. Upon comparing the significance of altered pathways with disease progression, we observed a noteworthy pattern. 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 conducted a comprehensive analysis of 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 [12, 112], and del(1p), frequently observed in MGUS [12, 92]. Furthermore, chr14 revealed prevalent translocations involving IGH, such as t(4;14), t(14;16), t(14;20), established as biomarkers in MM [12]. Additionally,

(a)

MGUS | MM
PhyloP Score's standard Deviation of the SNVs in Non-Synonymous SNV Group
Median AD of the SNVs in Synonymous SNV Group
Total Number of the SNVs in Others SNV Group
Total Number of the SNVs in Synonymous SNV Group
Median VAF of the SNVs in Synonymous SNV Group
GERP Score's standard Deviation of the SNVs in Non-Synonymous SNV Group
Median GERP Score of the SNVs in Non-Synonymous SNV Group
Functional Impact: Mutation Assessor(Not) for the SNVs in Non-Synonymous SNV Group
VAF's standard Deviation of the SNVs in Others SNV Group
Median AD of the SNVs in Non-Synonymous SNV Group
AD's standard Deviation of the SNVs in Non-Synonymous SNV Group
Functional Impact: Mutation Assessor(Low) for the SNVs in Non-Synonymous SNV Group
AD's standard Deviation of the SNVs in Synonymous SNV Group
VAF's standard Deviation of the SNVs in Non-Synonymous SNV Group
PhastCons Score's standard Deviation of the SNVs in Non-Synonymous SNV Group
Median PhyloP Score of the SNVs in Non-Synonymous SNV Group
Median AD of the SNVs in Others SNV Group
Median VAF of the SNVs in Others SNV Group
AD's standard Deviation of the SNVs in Others SNV Group
Total Number of the SNVs in Non-Synonymous SNV Group
Median VAF of the SNVs in Non-Synonymous SNV Group
Median PhastCons Score of the SNVs in Non-Synonymous SNV Group
Functional Impact: Mutation Assessor(High) for the SNVs in Non-Synonymous SNV Group
Functional Impact: Mutation Assessor(Medium) for the SNVs in Non-Synonymous SNV Group
Functional Impact: Mutation Assessor(Neutral) for the SNVs in Non-Synonymous SNV Group
VAF's standard Deviation of the SNVs in Synonymous SNV Group
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
ShAP Value (Impact on Model Output)
High
Feature value
Low

(b)

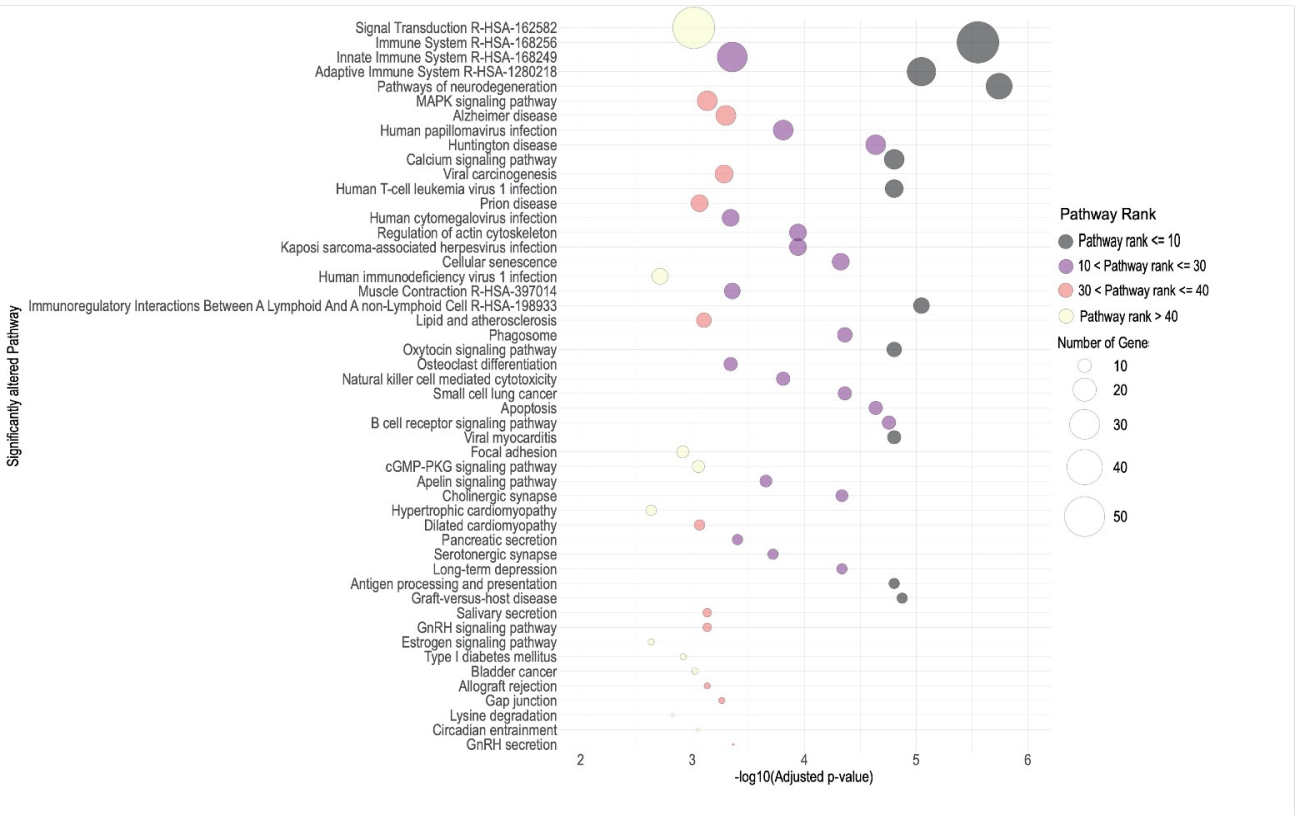

Signal Transduction R-HSA-162582
Immune System R-HSA-168256
Innate Immune System R-HSA-168249
Adaptive Immune System R-HSA-1280218
Pathways of neurodegeneration
MAPK signaling pathway
Alzheimer disease
Human papillomavirus infection
Huntington disease
Calcium signaling pathway
Viral carcinogenesis
Human T-cell leukemia virus 1 infection
Prion disease
Human cytomegalovirus infection
Regulation of actin cytoskeleton
Kaposi sarcoma-associated herpesvirus infection
Cellular senescence
Human immunodeficiency virus 1 infection
Muscle Contraction R-HSA-397014
Immunoregulatory Interactions Between A Lymphoid And A non-Lymphoid Cell R-HSA-198933
Lipid and atherosclerosis
Phagosome
Oxytocin signaling pathway
Osteoclast differentiation
Natural killer cell mediated cytotoxicity
Small cell lung cancer
Apoptosis
B cell receptor signaling pathway
Viral myocarditis
Focal adhesion
cGMP-PKG signaling pathway
Apelin signaling pathway
Cholinergic synapse
Hypertrophic cardiomyopathy
Dilated cardiomyopathy
Pancreatic secretion
Serotonergic synapse
Long-term depression
Antigen processing and presentation
Graft-versus-host disease
Salivary secretion
GnRH signaling pathway
Estrogen signaling pathway
Type I diabetes mellitus
Bladder cancer
Allograft rejection
Gap junction
Lysine degradation
Circadian entrainment
GnRH secretion
Significantly altered Pathway
2
3
4
5
6
-log10(Adjusted p-value)
Pathway Rank
Pathway rank <= 10
10 < Pathway rank <= 30
30 < Pathway rank <= 40
Pathway rank > 40
Number of Gene
10
20
30
40
50

Fig. 6 *(previous page)*. (a) Genomic Feature Ranking using the ShAP Algorithm in MM and MGUS based on post-hoc explainability by the Bio-DGI model. Genomic features are ranked according to their ShAP scores, as outlined in our prior research (Figure 4 and Table 2 of reference [26]). 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. (c) Bubble Plot illustrating the Top 50 Significantly Altered Signaling Pathways in MM associated with genes included in the proposed 282-gene panel. 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.

CNVs linked to chr19, such as gain(19p) and gain(19q), were significantly more prevalent in MM than in MGUS [100]. 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 [10].

Moreover, we meticulously curated a 282-gene panel by rigorous analysis of variant profiles (SNVs, CNVs, SVs, and LOF) from the top 500 genes. Our focus was solely on MM-relevant genes 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 [12, 83–87], and amp(1q) [12] and exclusive to MM, including del(13q), del(16q), del(17p), etc. [12] are shown in Table-4, Table-5.

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 the MAPK signaling, NF-kappaB, PI3K-AKT, and apoptosis pathways ranked prominently among the significantly altered pathways. The bubble plot in Figure-6(B) showcases the top 50 pathways linked with the proposed 282-genes panel, along with 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 classifier of the Bio-DGI (PPI9) model, we preserved the learned adjacency matrix. Employing the algorithm outlined in Section 2.2.3 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 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 both high node influence and GHIS scores. Notably, in the first gene community (Figure-3(A)), genes like UBC, USP6, PRIM2, and USP34; POTEM in third gene community (Figure-3(C)); LILRA1, LILRB1 in fourth gene community (Figure-3(D)) acted as central genes and may play a significant role in MM pathogenesis. We strongly recommend further analysis on these central genes to unveil their role in disease progression.

Furthermore, as we explored gene influences within each gene community (depicted in Figure-3(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 actively participating in these events. For instance, in the first gene community (Figure-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-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 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* 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-4 and Table-5. 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.

Conclusions

Distinguishing Multiple Myeloma (MM) from its precursor stage, Monoclonal Gammopathy of Undetermined Significance (MGUS) at the genomic level and identifying those who are at risk of progression to overt MM, presents a formidable challenge due to overlapping genomic characteristics. Unveiling the underlying pathogenesis of MM necessitates the identification of 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 282-genes panel design, we introduced the novel AI-based bio-inspired Bio-DGI (PPI9) model, encompassing gene interactions from nine Protein-Protein Interaction (PPI) databases and exonic mutational profiles from three global MM and MGUS repositories (AIIMS, EGA, and MMRF). The Bio-DGI (PPI9) model demonstrated superior performance both quantitatively and qualitatively, ensuring application-aware interpretability. Notably, the model identified the most number of previously reported genes, including Oncogenes (OGs), Tumor Suppressor Genes (TSGs), Oncogene Databases (ODGs), and Associated Genes (AGs), 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 gene 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 non-synonymous mutations, allele depth of synonymous SNVs and total number of other SNVs were found to be the most contributing genomic biomarker in distinguishing MM from MGUS. Subsequently, the significant alterations on chromosomes 1, 14, and 19 in MM cohort suggest major inclusion of the genes associated from these chromosomes. 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.

ACKNOWLEDGEMENTS

This work was supported by grant from Department of Biotechnology, Govt. of India [Grant: BT/MED/30/SP11006/2015] and Department of Science and Technology, Govt. of India [Grant: DST/ICPS/CPS-Individual/2018/279(G)]. Authors acknowledge dbGaP (Project#18964) for providing authorized access to the MM datasets (phs000748 and phs000348). Funding support for the phs000348 study was provided by the Multiple Myeloma Research Foundation in collaboration with the Multiple Myeloma Research Consortium. Assistance with data generation, processing and analysis was provided by the Broad Institute Genome Sequencing, Genetic Analysis, and Biological Samples Platforms. The datasets used for the analyses described in this work were obtained from dbGaP through dbGaP accession number phs000348.v1.p1. Data of study phs000748 were generated as part of the Multiple Myeloma Research Foundation CoMMpass [SM] (Relating Clinical Outcomes in MM to Personal Assessment of Genetic Profile) study (www.themmrf.org). We also acknowledge EGA (EGAD00001001901) for providing authorized access to the MGUS data. Authors would also like to thank the Centre of Excellence in Healthcare, IIIT-Delhi for their support in this research. Authors acknowledge the valuable insights provided by Dr. Satish Sankaran, Dr. Nandini Pal Basak, and Dr. Mohit Malhotra from Farcast Biosciences Pvt Ltd, India that greatly improved the gene panel designing workflow.

Bibliography

1. 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.
2. S Vincent Rajkumar. Mgus and smoldering multiple myeloma: update on pathogenesis, natural history, and management. *ASH Education Program Book*, 2005(1):340–345, 2005.
3. Alessandro Laganà, Itai Beno, David Melnekoff, Violetta Leshchenko, Deepu Madduri, Dennis Ramdas, Larysa Sanchez, Scot Niglio, Deepak Perumal, Brian A Kidd, et al. Precision medicine for relapsed multiple myeloma on the basis of an integrative multiomics approach. *JCO precision oncology*, 2:1–17, 2018.
4. 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.
5. Sarah A Holstein and Philip L McCarthy. Immunomodulatory drugs in multiple myeloma: mechanisms of action and clinical experience. *Drugs*, 77:505–520, 2017.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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):2997, 2014.
11. 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.
12. 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.
13. 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.
14. 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.
15. María-Victoria Mateos and Jesús F San Miguel. Management of multiple myeloma in the newly diagnosed patient. *Hematology 2014, the American Society of Hematology Education Program Book*, 2017(1):498–507, 2017.

16. 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.
17. 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.
18. 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.
19. 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.
20. Brian S White, Irena Lanc, Julie O'Neal, Harshath Gupta, Robert S Fulton, Heather Schmidt, Catrina Fronick, Edward A Belter Jr, 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):35, 2018.
21. Xiuquan Du, Shiwei Sun, Changlin Hu, Yu Yao, Yuanting Yan, and Yanping Zhang. Deepppi: boosting prediction of protein–protein interactions with deep neural networks. *Journal of chemical information and modeling*, 57(6):1499–1510, 2017.
22. Shu-Bo Zhang and Qiang-Rong Tang. Protein–protein interaction inference based on semantic similarity of gene ontology terms. *Journal of theoretical biology*, 401:30–37, 2016.
23. Stefan R Maetschke, Martin Simonsen, Melissa J Davis, and Mark A Ragan. Gene ontology-driven inference of protein–protein interactions using inducers. *Bioinformatics*, 28(1):69–75, 2012.
24. Ioan Ieremie, Rob M Ewing, and Mahesan Niranjan. Transformergo: predicting protein–protein interactions by modelling the attention between sets of gene ontology terms. *Bioinformatics*, 38(8): 2269–2277, 2022.
25. 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.
26. 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.
27. Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.
28. Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):5233, 2019.
29. Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.
30. 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.
31. 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.
32. 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.
33. David Benjamin, Takuto Sato, Kristian Cibulskis, Gad Getz, Chip Stewart, and Lee Lichtenstein. Calling somatic snvs and indels with mutect2. *BioRxiv*, page 861054, 2019.
34. 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.
35. 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.
36. 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.
37. 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.
38. 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.
39. 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.
40. 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.
41. 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.
42. 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.
43. 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.
44. 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.
45. 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.

46. 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.
47. 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.
48. 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.
49. 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.
50. 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.
51. 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.
52. Fabian Pedregosa, Gaël 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*, 12:2825–2830, 2011.
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. 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.
55. 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.
56. 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.
57. 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):3835, 2019.
58. 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.
59. 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.
60. 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.
61. Julia Steinberg, Frantisek Honti, Stephen Meader, and Caleb Webber. Haploinsufficiency predictions without study bias. *Nucleic acids research*, 43(15):e101–e101, 2015.
62. 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.
63. 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.
64. 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.
65. 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.
66. 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.
67. 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.
68. 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.
69. 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.
70. 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.
71. 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.
72. 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.
73. 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.

74. 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.
75. 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.
76. Jayakumar R Nair, Justin Caserta, Krista Belko, Tyger Howell, G Fetterly, Carmen Baldino, and Kelvin P Lee. Novel inhibition of pim2 kinase has significant anti-tumor efficacy in multiple myeloma. *Leukemia*, 31(8):1715–1726, 2017.
77. 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.
78. 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.
79. 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.
80. Jérôme Pagès. *Multiple factor analysis by example using R*. CRC Press, 2014.
81. 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.
82. Terry Therneau et al. A package for survival analysis in s. *R package version*, 2(7), 2015.
83. 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.
84. 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.
85. Joan Bladé. Monoclonal gammopathy of undetermined significance. *New England Journal of Medicine*, 355(26):2765–2770, 2006.
86. 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.
87. 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.
88. 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.
89. 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.
90. 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.
91. 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.
92. 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.
93. 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.
94. 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.
95. 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.
96. 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.
97. 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.
98. 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.
99. 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.
100. 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.

101. 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.

102. 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.

103. 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.

104. 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.

105. 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.

106. 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.

107. 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.

108. 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):2569, 2019.

109. 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.

110. 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.

111. 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.

112. Timothy M Schmidt, Rafael Fonseca, and Saad Z Usmani. Chromosome 1q21 abnormalities in multiple myeloma. *Blood cancer journal*, 11(4):83, 2021.