Am J Cancer Res 2023;13(4):1155-1187
www.ajcr.us /ISSN:2156-6976/ajcr0148190

Original Article

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

Vivek Ruhela[1,2], Lingaraja Jena[3], Gurvinder Kaur[3], Ritu Gupta[3], Anubha Gupta[2]

[1]Department of Computational Biology, Indraprastha Institute of Information Technology-Delhi (IIIT-D), New Delhi, India; [2]SBI Lab, Department of Electronics and Communication Engineering & Centre of Excellence in Healthcare, Indraprastha Institute of Information Technology-Delhi (IIIT-D), New Delhi, India; [3]Laboratory Oncology Unit, Dr. B.R.A. IRCH, All India Institute of Medical Sciences (AIIMS), New Delhi, India

Received December 2, 2022; Accepted March 23, 2023; Epub April 15, 2023; Published April 30, 2023

Abstract: Identification of the genomic features responsible for the progression of Multiple Myeloma (MM) cancer from its precancerous stage MGUS can improve the understanding of the disease pathogenesis and, in devising suitable preventive and treatment measures. We have designed an innovative AI-based model, namely, the Bio-inspired Deep Learning architecture for the identification of altered Signaling Pathways (BDL-SP) to discover pivotal genomic biomarkers that can potentially distinguish MM from MGUS. The proposed BDL-SP model comprehends gene-gene interactions using the PPI network and analyzes genomic features using a deep learning (DL) architecture to identify significantly altered genes and signaling pathways in MM and MGUS. For this, whole exome sequencing data of 1174 MM and 61 MGUS patients were analyzed. In the quantitative benchmarking with the other popular machine learning models, BDL-SP performed almost similar to the two other best performing predictive ML models of Random Forest and CatBoost. However, an extensive post-hoc explainability analysis, capturing the application specific nuances, clearly established the significance of the BDL-SP model. This analysis revealed that BDL-SP identified a maximum number of previously reported oncogenes, tumor-suppressor genes, and actionable genes of high relevance in MM as the top significantly altered genes. Further, the post-hoc analysis revealed a significant contribution of the total number of single nucleotide variants (SNVs) and genomic features associated with synonymous SNVs in disease stage classification. Finally, the pathway enrichment analysis of the top significantly altered genes showed that many cancer pathways are selectively and significantly dysregulated in MM compared to its precursor stage of MGUS, while a few that lost their significance with disease progression from MGUS to MM were actually related to the other disease types. These observations may pave the way for appropriate therapeutic interventions to halt the progression to overt MM in the future.

Keywords: AI in cancer, haematological malignancy, multiple myeloma, MGUS, genomic aberrations, ShAP

Introduction

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

Multiple studies involving exome data have been performed to understand the genomic abnormalities driving tumor progression in MM. Exome data analysis of MM patients has revealed that the primary events in MM are either hyperdiploidy, i.e., trisomy of chromosomes 3, 5, 7, 9, 11, 15, 17 and/or 21, or non-hyperdiploidy involving translocations affecting the genes encoding immunoglobulin (Ig) heavy chains (IGH)-mainly t(4;14), t(6;14), t(11;14), t(14;16), and t(14;20) [1]. Primary events are then followed by multiple secondary events that are secondary translocations: t(8;14) linked with *MYC* overexpression, loss of heterozygosity, copy number variations (CNV), acquired mutations, and epigenetic modifications [1], contributing to tumorigenesis. Initial deep sequencing studies on 38 whole-genome sequencing (WGS) and 23 whole-exome sequencing (WES) MM patients revealed frequent mutations in NF-kB signaling pathway and activating mutations in the oncogene *BRAF* [2]. In another study based on the WES data of 84 MM patients, *SP140, LTB, ROBO1*, and *EGR1* genes were identified as the novel drivers of MM [3]. Similarly, the analysis of 463 WES data of MM patients revealed 15 recurrently mutated genes: *IRF4, KRAS, NRAS, MAX, HIST1H1E, RB1, EGR1, TP53, TRAF3, FAM46C, DIS3, BRAF, LTB, CYLD, and FGFR3* [4]. Further, the analysis of same 463 MM samples reported *RAS* and *NF-Kappa-B* pathways as most altered signaling pathways. Furthermore, the same study reported that the mutations in *CCND1* and DNA repair pathway genes-*TP53, ATM*, and *ATR*, adversely impacted the overall survival, while the alterations in *IRF4* and *EGR1* were associated with a favorable overall survival.

Another study on the exome data analysis of 203 MM patients demonstrated tumor heterogeneity with subclonal pattern of mutations and multiple mutations within the same pathway in the same patient [5]. A recent study on 62 newly diagnosed MM (NDMM) patients reported the association of changes in the cellular prevalence of mutations with disease progression [6]. Another study explored oncogenic dependencies between mutations in driver genes, hyperdiploidy events, primary translocations, and copy number alterations in MM patients [7]. Associations were established between t(4;14) and mutations in *FGFR3*, *DIS3*, and *PRKD2*; t(11;14) and mutations in *CCND1* and *IRF4*; t(14;16) and mutations in *MAF, BRAF, DIS3*, and *ATM*; and hyperdiploidy with gain 11q and mutations in *FAM46C*, and *MYC* rearrangements [7]. A recent study demonstrated the co-occurrence of mutations within the same or a different clone and the clonal shifts in the co-occurring and mutually exclusive mutations with progression in MM [8]. Similar phenomena may be occurring from the stage of MGUS to overt MM and require to be evaluated. Analysis of WES data of unpaired samples of MGUS and MM has been carried out by several groups [9-12]. These studies have demonstrated a less complex genomic architecture in MGUS compared to MM with fewer mutations and lower TMB in MGUS. In a landmark study, the analysis of paired samples of MGUS and MM reaffirmed the clonal heterogeneity and presence of majority of genomic changes at MGUS stage [13]. The existence of the majority of genomic abnormalities seen in MM at the MGUS stage poses a challenge in distinguishing MM from MGUS based on the genomic signatures and in defining critical genomic events responsible for the progression of MGUS to MM [9-13].

The early diagnosis of MM and the identification of relevant differentiating genomic biomarkers between MGUS and MM present several challenges at the genomic-level and the subject-level. The unavailability of paired sequencing data (that is, sequencing data of MGUS and MM from the same sample), because all the MGUS subjects do not progress to MM, and the unavailability of reliable workflows for analyzing a pool of a large mutational information to decipher accurate and reliable genomic information, biomarkers, and significantly altered pathways pose key challenges at the genomic-level. Moreover, at the subject-level, limited information in the studies about the time intervals of a subject's treatment and death times pose key challenges in pursuing disease progression and a reliable identification of critical genes, genomic features, and signaling pathways for targeted therapeutics.

With advancements in bioinformatics and increasing inclination toward machine learning (ML) or deep learning (DL), newer methods are being developed for deducing salient information from the genomics data. For example, ML models have been developed to predict the survival outcome and treatment sensitivity in multiple myeloma [14, 15]. Similarly, AI-assisted risk stratification models for the prediction of survival and deciding the treatment regimen have been developed for the newly diagnosed multiple myeloma patients [16, 17]. Pathway enrichment analysis and classification has been shown to improve with the imputation of missing values in the microarray data of blood cancers via ML methods [18, 19]. ML/DL methods have also been proposed to detect somatic mutations from whole exome sequencing data [20, 21], prediction of copy number variants from whole exome data [22-24], driver genes in cancer [25-29] and, prediction of the survival-outcome and treatment-sensitivity in MM [14, 15].

However, the multi-dimensional analysis of exonic mutational profiles from exome sequencing data with gene-gene interaction has not yet been explored. This can be a promising direction for detecting key biomarkers in any cancer type. In recent years, geometric deep learning (GDL) has emerged to incorporate graph structures into a deep learning framework. Graph Convolutional Networks (GCNs) [30, 31], a type of GDL, can learn gene regulatory networks and do disease classification based on the network topology and disease-associated features, enabling an integration of graph-based data with genomic profiles [32]. The protein-protein interaction (PPI) network captures the physical interactions between proteins in an organism. Since the level of proteins and their interplay govern the molecular, cellular, and signaling controls which are the key to gene-level functionality and can help in capturing disease specific information, PPI networks can be immensely helpful if utilized alongside genomic information. A study on the exploration of the PPI network reported that the disease-related components in the PPI network are likely to be found in the network-based vicinity of disease components [33]. Similarly, another study on the PPI network revealed that the genes that contribute to a common disorder show an increased tendency of their protein-protein interactions [34]. These observations indicate that, due to the interconnected nature of a PPI network, genes belonging to similar diseases have a high predilection for interacting with other genes, forming a disease module. Therefore, identifying such genes or disease modules with the help of the PPI network can divulge the disease-related signaling pathways or other disease genes. These observations motivated us to incorporate the biological interactions in between genes as a key attribute of the bio-inspired BDL-SP model. Thus, we have incorporated the PPI information from the STRING database [35], which is the most comprehensive and global PPI network.

Motivated by the above discussion, this study addresses the problem of identifying significant biomarkers that differentiate MGUS from MM by incorporating a multidimensional analysis of exome profiles and their PPI network in a bio-inspired deep learning-based architecture from signaling pathways (BDL-SP) model. One of the challenges with this task is the ability to analyze a large amount of mutational information, a significant amount of which overlap in MGUS and MM samples. Since this mutational information is not easy to decipher for extracting differentiable patterns among MGUS and MM, the current literature shows this gap. To address the above gap, we have designed and implemented a GCN-based model, a *bio-inspired deep learning-based architecture from signaling pathways (BDL-SP)*, for extracting important genomic information to discern MGUS and MM. BDL-SP model uses single nucleotide variation (SNV) profiles of the significantly altered genes from the exome sequencing data along with the topological features of the PPI network, with an aim to identify pivotal biomarkers that can distinguish MGUS from MM. An in-depth analysis has been carried out for the identification of significantly altered genes and pathways that are specifically associated with MM and may be beneficial for the early identification of MGUS patients who are at a high risk of progression to the malignant MM stage. This work can further lead to the identification of novel therapeutic targets, thereby, preventing or delaying the malignant transformation of MGUS to MM.

For post-hoc model explainability, ShAP (SHapley Additive exPlanations) [36] algorithm is considered as one of the emerging and preferred approaches for decoding a DL model as well as for estimating feature importance based on their contribution to the model's predictions. The ShAP algorithm incorporates model-agnostic approximations and uniformly characterizes an approach for model explainability [37, 38]. We aimed to use ShAP for post-hoc explainability in order to extract the underlying cause of the model's predictions by analyzing the ShAP score of each individual gene and genomic feature. We ranked the significantly altered genes based on their contribution to disease classification using the ShAP score. Among all the ML models trained in this study, BDL-SP model reported the highest numbers of previously reported driver genes, oncogenes, TSGs, and actionable genes in the top-ranked significantly altered genes compared to the other models. BDL-SP model also shows novel genes in the top-ranking genes that are not reported in MM but found significantly altered and contributing significantly to the disease prediction. We performed pathway enrichment analysis for top-500 significantly mutated genes. We analyzed whether an altered signaling pathway becomes more or less significant with disease progression from MGUS to MM. We observed that several signaling pathways either become significant (from being insignificant at MGUS) or become more significant with disease progression from MGUS to MM.

We benchmarked the BDL-SP with several baseline ML models both quantitatively and qualitatively, and observed that BDL-SP outperformed the other models in both aspects. With the help of the BDL-SP model, we identified the genes and their corresponding enriched signaling pathways that significantly contributed to MM disease development. The BDL-SP model's findings helped us to improve the understanding of cell transformation from premalignant to malignant state and strategic diagnosis to support the early detection of transformation to MM.

Material and methods

Whole-exome sequencing datasets of MM and MGUS patients

In this work, we utilized two external whole-exome sequencing (WES) datasets available with controlled access and one in-house WES dataset of MM and MGUS patients. These datasets are: 1) Multiple Myeloma Research Foundation (MMRF) CoMMpass data (of American population), 2) EGA dataset (of European population), and 3) AIIMS WES dataset (of Indian population). The MMRF CoMMpass (https://research.themmrf.org) is an open-source, extensive clinical and molecular database of multiple myeloma. The majority of MM samples in MMRF CoMMpass dataset (>75%) were collected from the people of American ethnicity. The MMRF CoMMpass dataset is aimed to provide molecular characterization and to correlate clinical datasets of MM patients for finding new, actionable targets to facilitate future clinical trial designs [39]. In our study, we have included 1092 bone marrow (BM) samples of MM collected from the GDC portal via dbGaP authorized access (phs000748; phs000348). This is to note that the MMRF dataset also contained 20 peripheral blood (PB) samples that were not included in this study for the uniformity of the data. Similarly, the European Genome-phenome Archive (EGA) contains more than 700 studies of multiple diseases (including cancer and non-cancer) worldwide. EGA (http://www.ebi.ac.uk/ega/) was launched in 2008 by the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) to provide secure storage of biological data and distribution only to authorized users [40]. The whole exome sequencing data of 33 MGUS European patients were obtained from the EGA repository EGAD00001001901. Besides the above two external datasets, we also included the WES data collected in-house from patients of Indian origin registered at All India Institute of Medical Sciences (AIIMS), New Delhi, India. This dataset included 82 MM and 28 MGUS samples. We have used the tumor-normal matched pairs of all BM samples obtained from MMRF, EGA and AIIMS WES data repository. Thus, we have included MGUS and MM WES datasets from three different databases.

Methods

Data pre-processing: Four variant callers, namely, MuSE [41], Mutect2 [42], VarScan2 [43], and Somatic-Sniper [44], were used for finding the variants in MM patients from the MMRF CoMMpass study. Therefore, for each patient, four variant call format (VCF) files corresponding to each variant caller were downloaded from the GDC portal via dbGaP authorized access (phs000748; phs000348). Exome data obtained from EGA and AIIMS were processed with an exome sequencing pipeline [45] using BWA [46] and GATK [47], which is also considered a standard pipeline and mostly adopted to process the exome sequencing data. Similar to the MMRF data, the single nucleotide variants (SNVs) in EGA and AIIMS exome sequencing data were extracted using MuSE, Mutect2, VarScan2, and SomaticSniper variant callers. SNVs were annotated using ANNOVAR tool [48] that provides information about mutated genes, mutation type, the property of being deleterious or not, and clinical validation. In our study, we considered 23 types of functionally significant SNVs clustered into three groups based on their functional impact as follows: 1) *Non-Synonymous (NS) SNV Group*: This group consists of non-synonymous SNVs, exonic, ncRNA_exonic, stop gain, stop loss, start loss, exonic; splicing, splicing, frameshift insertion, and frameshift deletion type SNVs; 2) *Synonymous SNV Group*: This group consists of synonymous SNVs, UTR3 and UTR5 SNVs; and 3) *Other SNV Group*: This group consists of non-frameshift insertion/deletion/substitution, intronic, intergenic, ncRNA_intronic, upstream, downstream, unknown, and ncRNA_splicing SNVs. The benign SNVs were filtered out using the FATHMM-XF method [49]. Genomic annotations of SNVs (i.e., SNV type, mutated gene name, etc.) obtained from ANNOVAR were pooled and analyzed to identify the top significantly mutated genes using the 'dndscv' tool [50] based on the *q*-value (≤0.05) in both MM and MGUS individually. Union of significantly mutated genes from all four variant callers for MM (1174 patients) and MGUS (61 patients) groups led to 617 and 362 genes, respectively, and further union of the genes mentioned above yielded a total of 824 genes (Table S1 of supplementary material). For each gene, a total of 28 genomic features were created that includes total variant count and the distributive statistics (maximum, mean, median, and standard deviation) of variant allele frequency (VAF) and allele depth (AD) of each of the three groups of SNVs (NS SNV group, synonymous SNV group, and Other SNV group). A detailed description of the 28 genomic features is presented in **Figure 1**. The complete AI workflow is presented in **Figure 2**. For gene-gene interaction network information, we used the STRING database to get protein-protein interaction (PPI) of 824 significantly altered genes. The STRING database contains all the known and predicted associations of protein-protein interactions, including physical and functional associations for more than 14000 organisms.

Proposed shallow bio-inspired deep learning architecture from signaling pathways (BDL-SP): The conventional convolutional neural network (CNN) often fails to learn data of non-Euclidean space because non-Euclidean data cannot be modeled into *n*-dimensional linear space. The protein-protein interaction (PPI) network used in our model has a similar underlying non-Euclidean structure. Thus, a Graph Convolutional Network (GCN) could help us learn PPI data of non-Euclidean space. The proposed BDL-SP model carries out disease classification using a graph convolutional network that learns significant features from the exomic mutational profiles of genes interacting among each other according to the PPI network interactions. The mathematical description of GCN model is as follows:

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

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

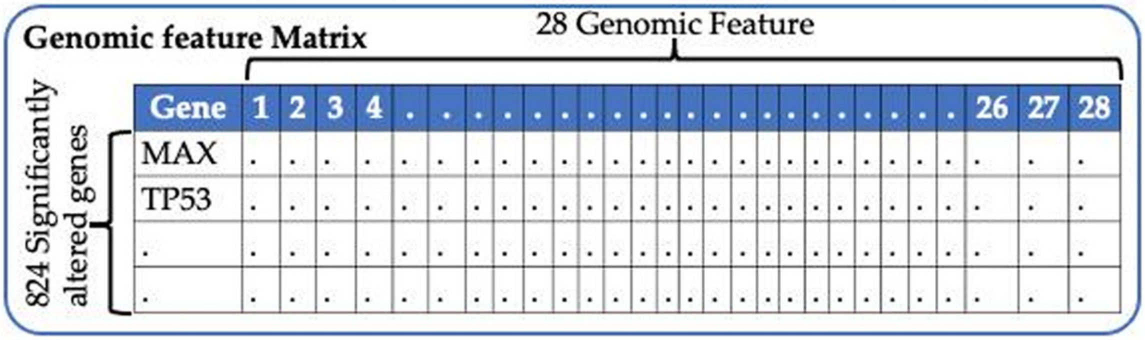
Where L denoted the normalized graph laplacian defined below.

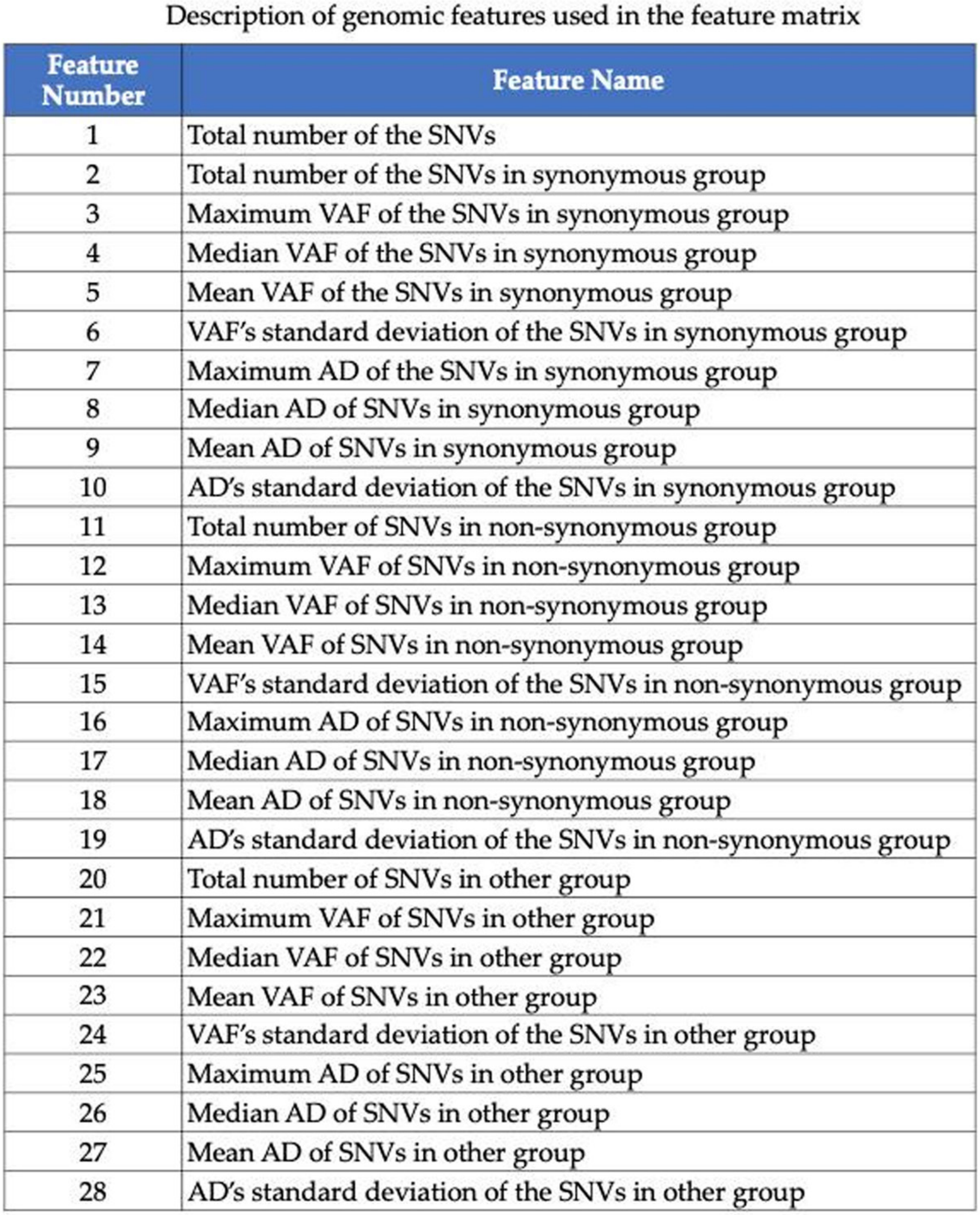
$$L = I - D^{-\frac{1}{2}} \tilde{A} D^{-\frac{1}{2}} = U \wedge U^{T}$$

Where $D_{i,j} = \sum_{j=1}^{n} A(i,j)$, degree matrix of the graph and $\tilde{A} = A+I$ where A is the adjacency matrix, U is the matrix of eigenvectors of graph, $\wedge$ denote the respective eigenvectors, and $W \in R^{p \times m}$ (where m corresponds to the number of filters in the graph convolution) denotes a learnable weight matrix. A GCN model transform a graph into the spectral domain by graph Fourier transformation [30] defined as below:

$$x * g = UgU^{T}x$$

The above Fourier transformation can be computed by approximating Chebyshev polynomi-




Description of genomic features used in the feature matrix

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

Figure 1. Schematic layout of genomic feature matrix used for the training of proposed BDL-SP model. The dimension of the genomic feature matrix is 824×28 with 824 significantly altered genes (See Table S1 of supplementary material) and 28 genomic features obtained from MMRF, EGA and AIIMS WES datasets using the AI-based workflow shown in **Figure 2**. The genomic features were extracted from three groups of SNVs, namely, 1. Non-synonymous SNV group, 2. Synonymous SNV group, and 3. Other SNV group. A total of nine features were extracted for each SNV group to learn the distributive statistics (maximum, mean, median, and standard deviation). The full form of abbreviations used in this figure are as follows: SNVs = Single Nucleotide Variations, VAF = Variant Allele Frequency, and AD = Allele Depth.

als and the renormalization trick mentioned in [30] as:

$$Z = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X\,W$$

The infographic representation of the architecture of BDL-SP with end-to-end pipeline is shown in **Figure 3** and is explained here. The BAM files from EGA and AIIMS datasets, and VCF files from the MMRF dataset are processed to extract 824 significantly altered genes using the dndscv tool (as shown in the WES Data pre-processing block) in **Figure 3**. The interaction among these 824 genes is extracted using the protein-protein interaction (PPI) network (from the STRING database). A network of nodes and edges is set-up using this information, where each node denotes one of these 824 genes and each link implies that the two nodes/genes of that link were connected as per the PPI network. Each node is set up with its 28-length feature-vector extracted earlier (as shown in **Figure 1**). Hence, the 28-length genomic feature-vectors of all 824 genes are added to the network established using the PPI network. This input layer is followed by two hidden layers of GCN, that are further followed by one fully connected layer of 824 neurons to 2 neurons giving output through log-softmax activation function. Since there were 95% samples of MM class and 5% samples of MGUS class, which made the data highly imbalanced (class imbalance ratio = 19.22), a cost-sensitive loss function was utilized to train the BDL-SP model in order to deal with the data imbalance problem. BDL-SP is trained in a supervised fashion, where the MM/MGUS target class label along with the feature matrix of 824×28 is provided as an input to the architecture. The network is trained until the loss reduces and saturates. Five-fold cross-validation is performed that led to the training of five BDL-SP classifiers, one for each fold of test data. Next the ShAP algorithm is used on these five trained BDL-SP classifiers to obtain the top genomic features and significantly altered signaling pathways as explained in the next subsection. The setting of layers of BDL-SP and the hyperparameters values are shown in **Table 1**.

Quantitative benchmarking of BDL-SP model with traditional machine learning classifiers: We have benchmarked the proposed BDL-SP model with six baseline ML models (random forest, decision tree, logistic regression, XGBoost, CatBoost, and SVM from scikit-sklearn [51]). The conventional cost-blind machine learning models do not account for the imbalanced classes in the data and tend to make decisions favoring the majority class resulting in misclassification. In the case of medical diagnosis, such misclassification can lead to erroneous direction of strategic treatment, causing patients to suffer. In our study, there were 95% samples of MM class and 5% samples of MGUS class, which made the data highly imbalanced (class imbalance ratio = 19.22). Therefore, we have used cost-sensitive ML models to account for the class imbalance in our data and model benchmarking. During training, the cost-sensitive loss function penalizes the mistake in identifying each MGUS sample (minority class) more compared to the mistake in identifying each MM sample (majority class). This ensures that the classifier is not biased to the majority class and learns to identify the samples of both the classes. These baseline models are trained with the traditional data pre-processing pipeline using principal component analysis (PCA). Each baseline ML model was trained exhaustively with five-fold cross-validation, where the confusion matrix of the hold-out set was kept separate for each fold. The final confusion matrix was obtained by adding the confusion matrices of all five hold-out sets and the performance metrics were calculated for each ML model.

Qualitative application-aware post-hoc benchmarking of BDL-SP model using ShAP: ShAP (SHapley Additive exPlanations) is an algorithm

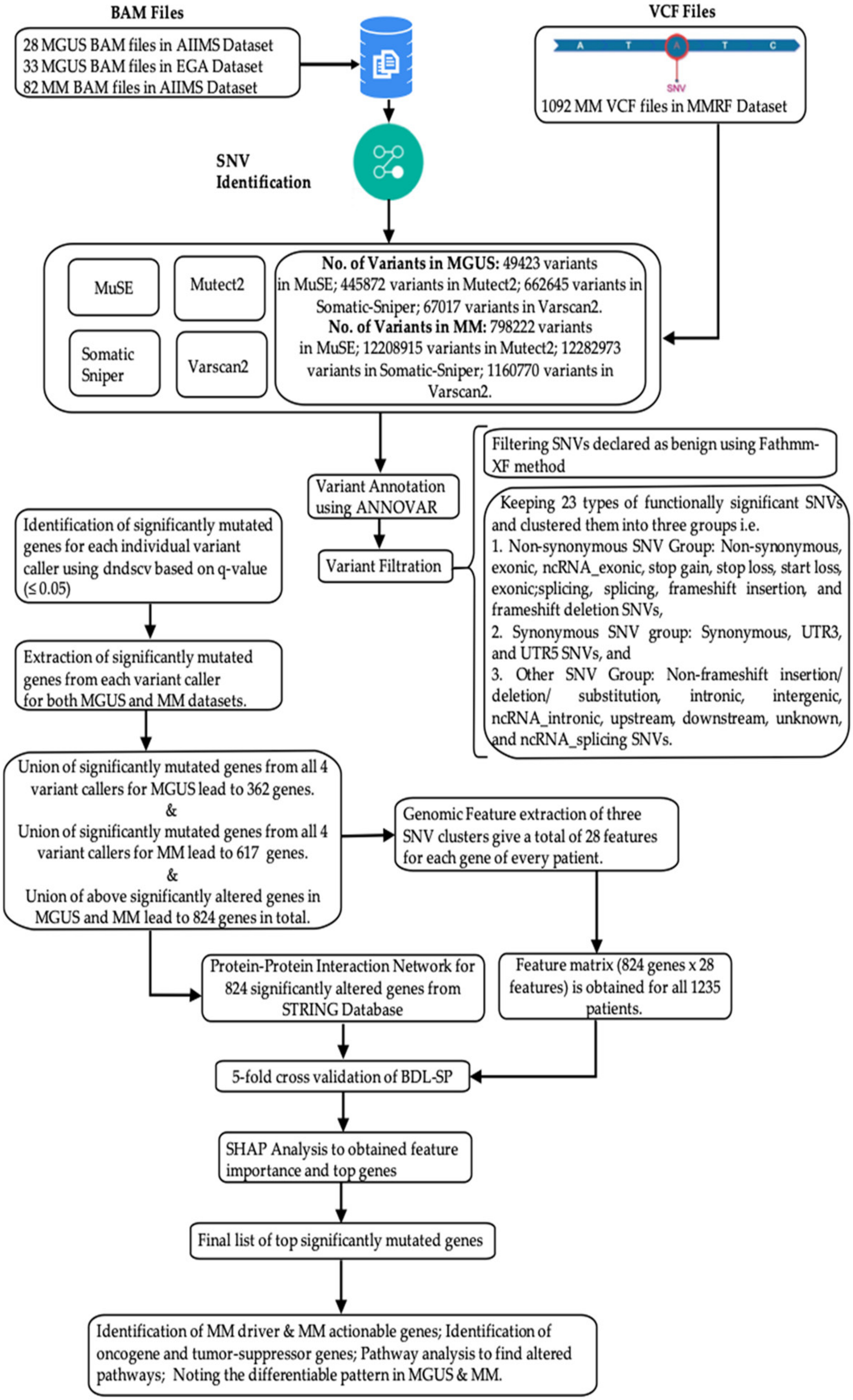

BAM Files
28 MGUS BAM files in AIIMS Dataset
33 MGUS BAM files in EGA Dataset
82 MM BAM files in AIIMS Dataset
VCF Files
A T A T C
SNV
1092 MM VCF files in MMRF Dataset
SNV Identification
MuSE
Mutect2
Somatic Sniper
Varscan2
No. of Variants in MGUS: 49423 variants in MuSE; 445872 variants in Mutect2;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

Figure 2. AI-based workflow to infer differentiable genomic biomarkers to identify MGUS and MM using the whole-exome sequencing (WES) data.

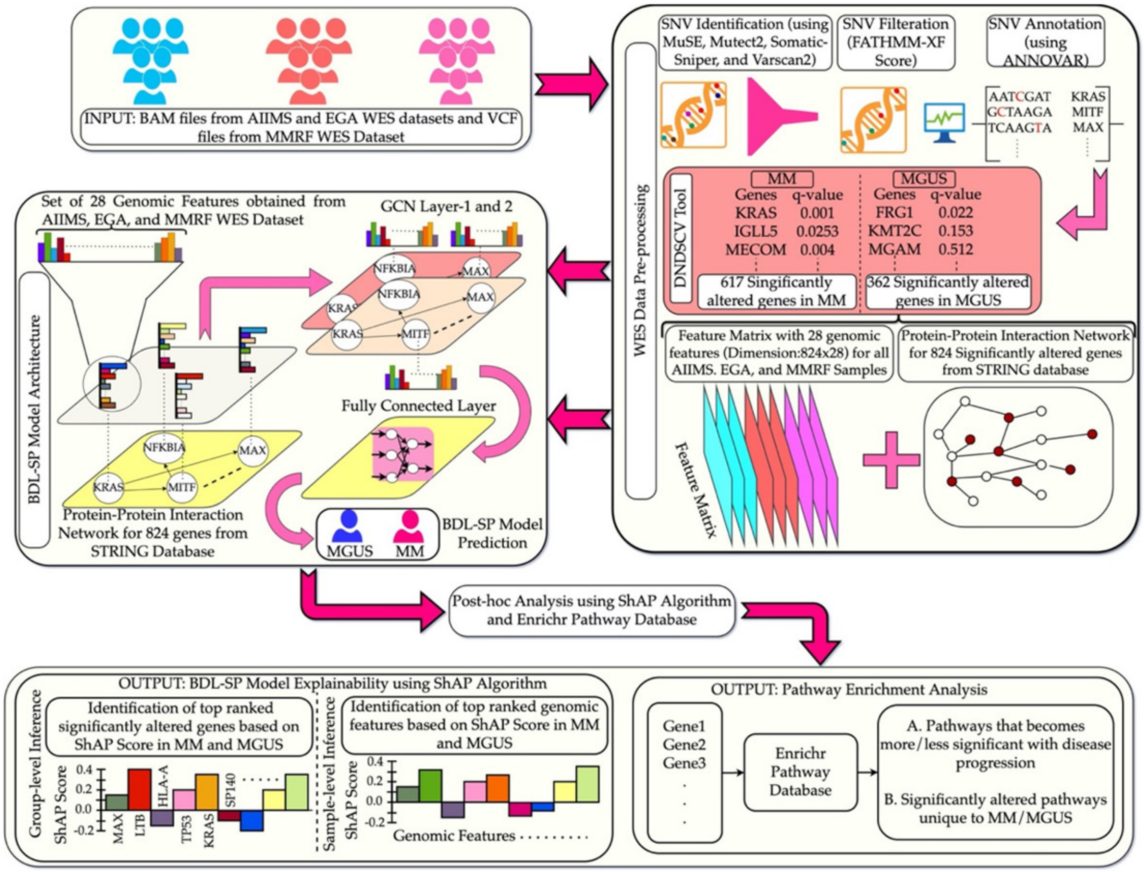

INPUT: BAM files from AIIMS and EGA WES datasets and VCF files from MMRF WES Dataset
WES Data Pre-processing
SNV Identification (using MuSE, Mutect2, Somatic-Sniper, and Varscan2)
SNV Filteration (FATHMM-XF Score)
SNV Annotation (using ANNOVAR)
AATCGAT
GCTAAGA
TCAAGTA
KRAS
MITF
MAX
DNDSCV Tool
MM
Genes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

Figure 3. Infographic representation of the proposed AI-based BDL-SP model architecture and the application-aware post-hoc analysis for the identification of pivotal genomic biomarkers that distinguish MGUS from MM. The BAM files from EGA and AIIMS datasets and VCF files from MMRF dataset are processed to extract 824 significantly altered genes using the dndscv tool (as shown in the WES Data pre-processing block). The interaction among these 824 genes is extracted using the protein-protein interaction (PPI) network (from STRING database). A network of nodes and edges is set-up using this information, where each node denotes one of these 824 genes and each link implies that the two nodes/genes of that link were connected as per the PPI network. Each node is set up with its 28 genomic features extracted for the corresponding gene as explained earlier. This input layer is followed by two hidden layers of GCN, one fully connected layer, and a softmax layer at the output. Thus, each subject's WES data is analyzed and the feature vectors of all 824 genes are extracted. These are given as input along with the subject's MM/MGUS target class label to train the GCN in a supervised mode. Once the BDL-SP model is learned to distinguish MGUS from MM, the top genomic features and significantly altered signaling pathways were obtained from ShAP algorithm and the Enrichr Pathway Database.

that measures the significance of an attribute in the prediction of a model, scoring each attribute proportional to its contribution. Therefore, it was utilized for the post-hoc explainability of the BDL-SP model. The most-contributing genomic features and significantly altered genes at the group (i.e., either MGUS or MM) as well as at the individual sample-level were identified. Since five-fold validation was carried out during training, the ShAP algorithm was applied on each trained classifier to obtain the significant genomic attributes (both genes and genomic features) for each sample. Note that the ShAP score can either be positive or negative. Here, the positive ShAP score for an attribute indicates its contribution to the model's prediction toward the MGUS class (positive class), while the negative score indicates its

Table 1. Hyperparameters values and layer dimensions of the BDL-SP model architecture

GCN Architecture Attribute/Hyperparameter	Hyperparameter Value
No. of GCN Layers	2
GCN layer dimensions	Input sample dimension: 824×28 1st Layer (For each node): 28×7 2nd Layer (For each node): 7×1 Output dimension: 824×1
Output Linear Layer dimension	824×2 (number of classes = 2)
Activation Function	LeakyReLU (0.1)
Dropout	0.75
Cost function and Adjusted Cost for class imbalance	Cost Function: Cross-Entropy Loss Cost Adjusted: 20.0
GCN Weight Initialization	Uniform Xavier

contribution to the model's prediction toward the MM class (negative class). Therefore, the higher the magnitude of the ShAP score, the higher its impact on the model's positive class outcome. Moreover, only those samples were considered for extracting ShAP interpretability that were correctly predicted by at least one of the five classifiers.

Next, we devised the algorithms for the estimation of the best ShAP score 1) for all 824 significantly altered genes (**Figure 4A**) and 2) for all 28 genomic features (**Figure 4B**) at a sample-level to understand their contribution to the BDL-SP model's prediction. The pseudo-codes with mathematical description for estimating the best ShAP scores for genes and genomic features are provided in **Table 2**, Algorithm-A, and Algorithm-B. The algorithms shown in **Figure 4A** and **4B** take the sample feature matrix as input and estimate the best ShAP scores for genes and genomic features at a sample-level. For each sample feature matrix, the corresponding sample class was predicted using all five trained classifiers of the BDL-SP model and the ShAP algorithm was applied only on those classifiers that predicted the sample's class correctly. Here, the ShAP score for all the genomic attributes were collected at the classifier-level and the sample-level. For each genomic attribute, the best ShAP score was first calculated at a classifier level and then the final best ShAP score was estimated among all classifiers at a sample-level. For each gene, we first collected the ShAP score of all 28 genomic features at a sample-level and then grouped them based on their positive and negative signs. Next, we compared the absolute value of the sum of ShAP scores of genomic features having the positive ShAP scores with the absolute sum of those having the negative ShAP scores. The ShAP score having the highest absolute value was considered as the best ShAP score for that gene and the classifier. This step was repeated for all those classifiers that predicted the sample's class correctly and the best ShAP score was saved for each of the classifiers. The ShAP score having the highest absolute value among all the classifiers was considered as the best ShAP score for a gene at a sample-level. For a better clarity of the steps employed in the estimation of the best ShAP scores of significantly altered genes and genomic features at a sample-level, one may refer to **Figure 4A** and Algorithm-A of **Table 2**, **Figure 4B** and Algorithm-B of **Table 2**, respectively.

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

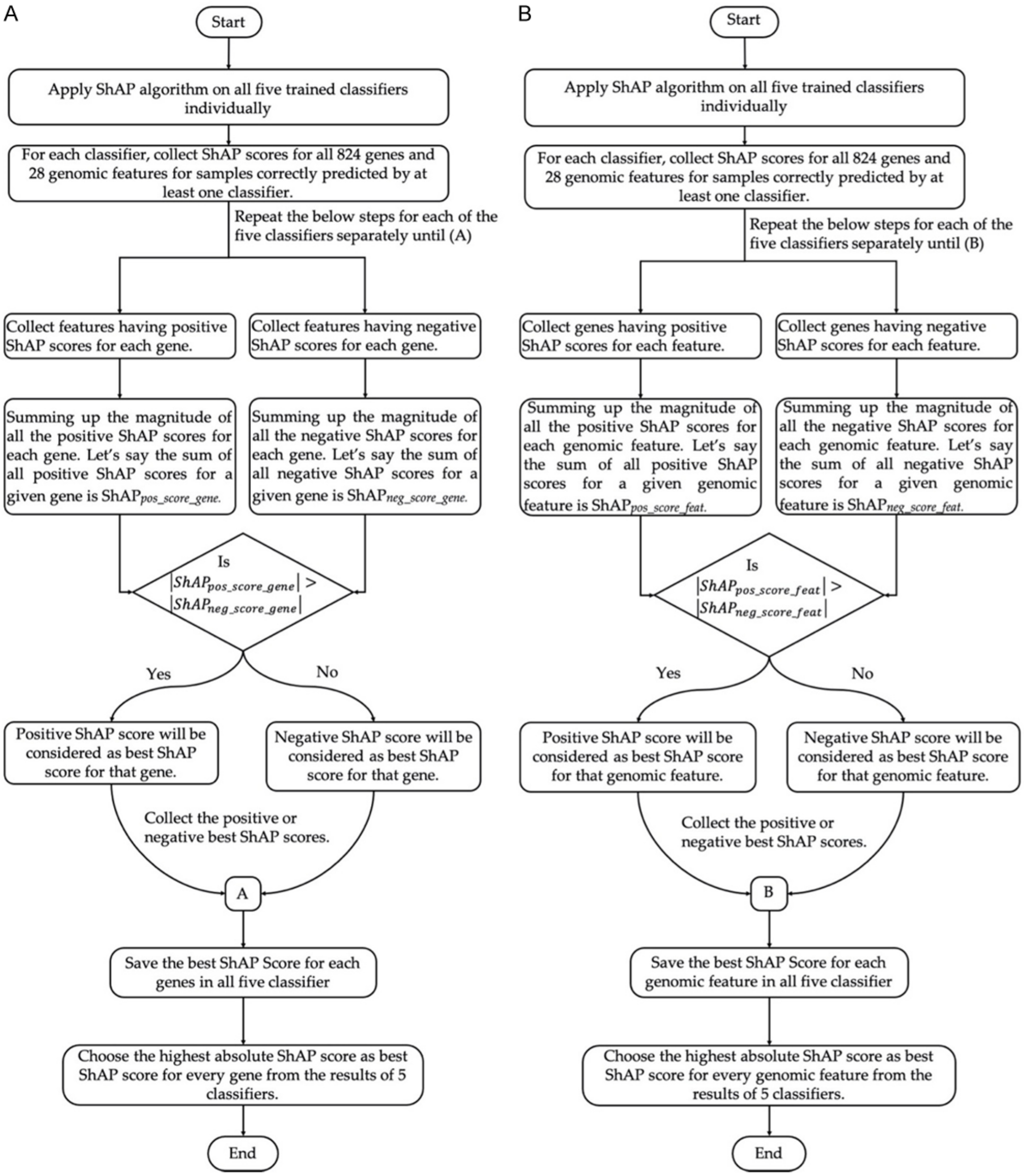

A
Start
Apply ShAP algorithm on all five trained classifiers individually
For each classifier, collect ShAP scores for all 824 genes and 28 genomic features for samples correctly predicted by at least one classifier.
Repeat the below steps for each of the five classifiers separately until (A)
Collect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

Figure 4. Flowchart showing steps for estimating the best ShAP score for (A) 824 significantly altered genes and (B) 28 genomic features at sample-level to reveal their contribution to the BDL-SP model prediction.

butes were identified at the group-level and at the sample-level.

Further, the top-ranked significantly altered genes revealed by BDL-SP were also compared with the multiple myeloma related studies to identify the previously reported significantly altered genes. We included information from multiple databases for model validation and post-hoc analysis at gene level analysis (OncoKB, COSMIC, IntoGen, and TargetDB databases). We downloaded a list of 1064 cancer genes from OncoKB [52] to deduce the oncogenes and tumor-suppressor genes in our top mutated genes. Further, 318 oncogenes and 320 tumor-suppressor genes obtained

Table 2. (A) Pseudo-codes of algorithm A for estimating the best ShAP score of 824 genes and (B) Algorithm B for estimating the best ShAP score of 28 genomic features at a sample level

Algorithm A: Estimate the Best ShAP Score (BSS) for each gene at a sample level

```
1. Fivefold classifiers = [List of five classifiers trained on each fold of test dataset]
2. CPC = [List of correct prediction classifiers i.e. classifiers that correctly predicted the sample's class]
3. SFM = Sample feature matrix
4. Genes = [List of 824 genes]
5. $GFPS_{g|c}$ = [List of genomic features having positive ShAP score for a gene "g" and classifier "c"]
6. $GFNS_{g|c}$ = [List of genomic features having negative ShAP score for a gene "g" and classifier "c"]
7. $CSG_{g|c}$ = Best ShAP score of gene "g" and classifier "c"
8. $ACGS[classifier]_{g}$ = [List of best ShAP scores of gene "g" for all the classifiers that correctly predicted the sample]
9. $BSG_{g}$ = Best ShAP score of gene "g" among all classifiers
10. $LBSG_{genes}$ = List of the best ShAP score of all the genes among all the classifiers
11. function BSS gene (SFM)
12. for classifier in (Fivefold classifiers) do
    A. Predict the sample's class with the help of a classifier
    B. if classifier predict the sample class correctly then
        a. Append classifier in CPC list
        b. Apply ShAP algorithm on the classifier
        c. Collect the ShAP score for all 824 genes on their respective 28 GF for that classifier
13. for gene in Genes do
    A. for classifier in CPC do
        a. $GFPS_{gene|classifier} \leftarrow$ Collect features having positive ShAP score
        b. $GFNS_{gene|classifier} \leftarrow$ Collect features having negative ShAP score
        c. If $|\sum GFPS_{gene|classifier}|>|\sum GFNS_{gene|classifier}|$ then
        d. $CSG_{gene|classifier} \leftarrow GFPS_{gene|classifier}$
        e. else
        f. $CSG_{gene|classifier} \leftarrow GFNS_{gene|classifier}$
        g. $ACGS[classifier]_{gene} \leftarrow CSG_{gene|classifier}$
    B. $BSG_{gene} \leftarrow ACGS[argmax[|CSG|\ for\ CSG\ in\ ACSG]]$
    C. $LBSG_{genes}[gene] \leftarrow BSG_{gene}$
14. Output: $LBSG_{genes}$
```

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

```
1. Fivefold classifiers = [List of five classifiers trained on each fold of test dataset]
2. CPC = [List of correct prediction classifiers i.e. classifiers that correctly predicted the sample's class]
3. SFM = Sample feature matrix
4. Genomic Features = [List of 28 GFs]
5. $GPS_{gf|c}$ = [List of genes having positive ShAP score for a genomic feature and classifier]
6. $GNS_{gf|c}$ = [List of genes having negative ShAP score for a genomic feature and classifier]
7. $CSGF_{gf|c}$ = Best ShAP score of GF "gf" and classifier "c"
8. $ACGFS[classifier]_{gf}$ = [List of best ShAP scores of gene "gf" for all the classifiers that correctly predicted the sample]
9. $BSGF_{gf}$ = Best ShAP score of GF "gf" among all classifiers
10. $LBSGF_{gfs}$ = List of the best ShAP score of all GF among all classifiers
11. function BSS genomic feature (SFM)
12. for classifier in (Fivefold classifiers) do
    A. Predict the sample's class with the help of a classifier
    B. if classifier predict the sample class correctly then
        a. Append classifier in CPC list
        b. Apply ShAP algorithm on the classifier
        c. Collect the ShAP score for all 824 genes on their respective 28 GF for that classifier
```

```
13. for feature in Genomic features do
    A. for classifier in CPC do
        a. $GPS_{feature|classifier}$ ← Collect genes having positive ShAP score
        b. $GNS_{feature|classifier}$ ← Collect genes having negative ShAP score
        c. if $|\sum GPS_{feature|classifier}| > |\sum GNS_{feature|classifier}|$ then
        d. $CSGF_{gf|classifier} \leftarrow GPS_{feature|classifier}$
        e. else
        f. $CSGF_{gf|classifier} \leftarrow GNS_{feature|classifier}$
        g. $ACGFS[classifier]_{gf} \leftarrow CSGF_{gf|classifier}$
    B. $BSGF_{gf} \leftarrow ACGFS[argmax[|CSGF|\ for\ CSGF\ in\ ACGFS]]$
    C. $LBSGF_{gfs}[feature] \leftarrow BSGF_{gf}$
14. Output: $LBSGF_{gfs}$
```

from COSMIC database [55] were also used to deduce oncogenes and tumor-suppressor genes in our top-mutated genes. Similarly, we created a list of MM driver genes reported by [7, 53]. MM Driver genes were also extracted from IntoGen database [54] (https://www.intogen.org/) to infer MM drivers genes present in our top mutated gene list. Finally, a list of 180 actionable genes from the COSMIC database [55] and 135 actionable genes from the TargetDB database [56] was used to infer the actionable genes present in our top mutated gene list. The top ranked significantly altered genes were grouped in four categories based on their functional significance as follows: 1. Oncogenes (OGs); 2. Tumor-Suppressor genes (TSGs); 3. Onco-driver genes (ODGs); 4. Actionable genes (AGs).

The top-ranked significantly altered genes in each of the above gene categories were then collected at the group-level (MM/MGUS) and the sample-level. We also checked the role of genomic features on the disease classification in post-hoc interpretability analysis of the BDL-SP model.

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

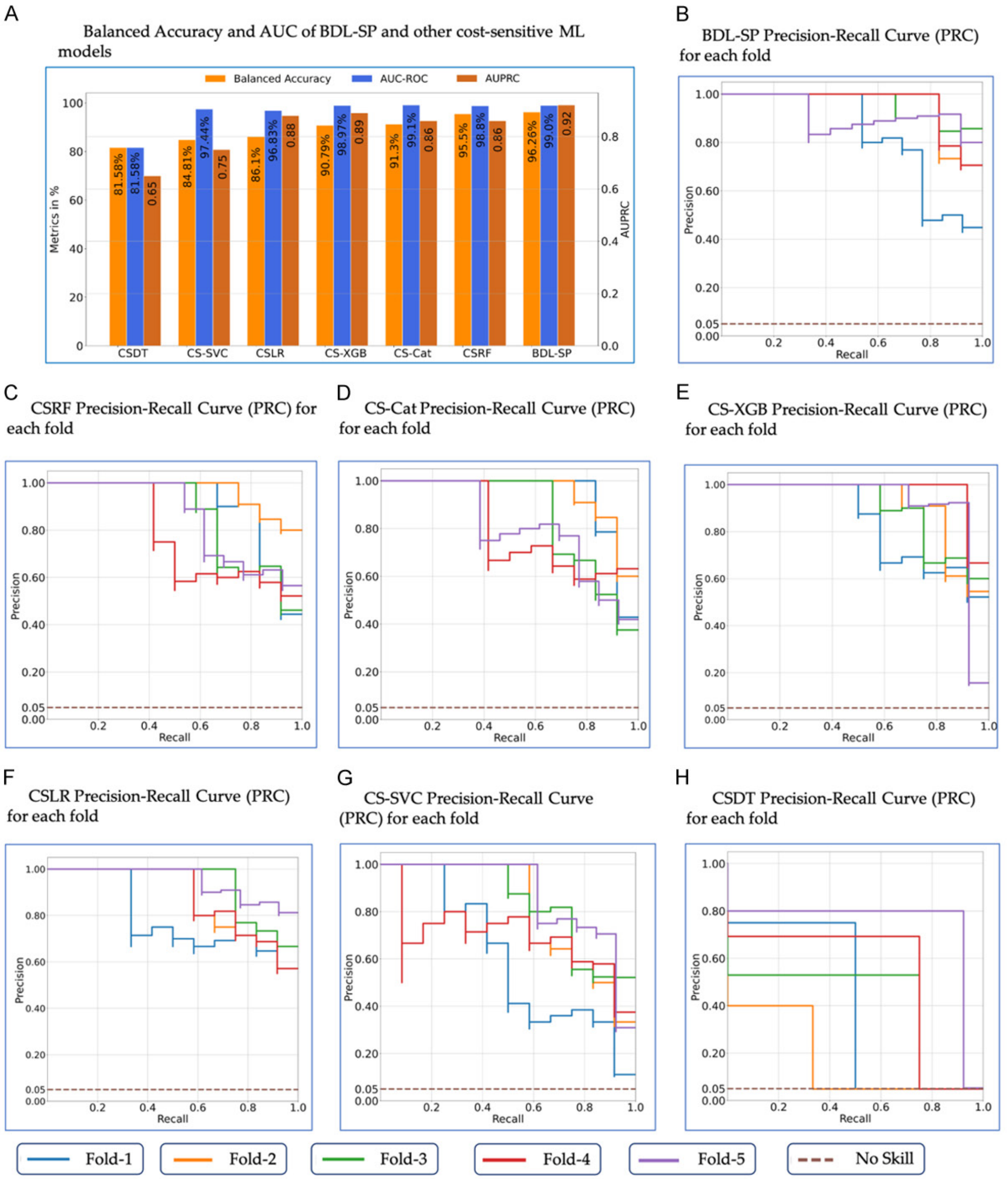
Gene pathway analysis: The significant genes identified by BDL-SP, which helped in differentiating MM from MGUS, were mapped back to the significant gene list obtained for MM and MGUS using the dndscv tool. Some genes were found to be common in both the groups, while some were found to be significantly mutated either in MGUS or in MM only. Pathway analysis was done on the top-500 genes obtained from the BDL-SP model. KEGG and Reactome pathways were found via Enrichr gene set enrichment analysis web server [57-59].

Results

Using the dndscv tool (as shown in **Figure 2**), 362 and 617 significantly altered genes were identified in MGUS and MM, respectively. Of these, 155 genes were common in MGUS and MM. The complete list of all 824 genes is shown in Table S1 of supplementary material. We then inferred the important genes that were accountable for distinguishing MGUS from MM as obtained through our graph-based BDL-SP model.

Comparative performance of BDL-SP and standard ML models

Using our AI-based workflow of BDL-SP (**Figures 2** and **3**), we trained the BDL-SP model with a 5-fold cross-validation and compared its performance with six standard cost-sensitive machine learning models. Results of the BDL-


A Balanced Accuracy and AUC of BDL-SP and other cost-sensitive ML models

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

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

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

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

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

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

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

Figure 5. (A) The benchmarking of the performance of BDL-SP with six cost-sensitive ML models on the metrics of balanced accuracy, AUC, and AUPRC (Area under Precision-Recall Curve). Precision-Recall Curves (PRC) for all five folds of (B) BDL-SP, (C) CSRF, (D) CS-Cat, (E) CS-XGB, (F) CSLR, (G) CS-SVC, and (H) CSDT. No skill line is also shown in each of the AUPRC plots that represent the inability of the classifier to correctly classify a sample. The full form of the abbreviation used in this figures are as follows: CSDT = Cost-Sensitive Decision Tree, CS-SVC = Cost-Sensitive Support Vector Machine, CSLR = Cost-Sensitive Logistic Regression, CS-XGB = Cost-Sensitive XGBoost, CS-Cat = Cost-Sensitive CatBoost, and CSRF = Cost-Sensitive Random Forest.

SP model and all the six cost-sensitive classifiers are presented in **Figure 5**. The proposed BDL-SP model outperformed the rest of the models in terms of the balanced accuracy and AUPRC (area under precision-recall curve), while the area under the curve (AUC) was highest (and equal) for the top three models. BDL-SP model performed best with a balanced

Table 3. Types of four different gene categories (OG, TSG, ODG, and AG) and their counts in 824 significantly altered genes

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

Table 4. 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 (BDL-SP, CS-RF, and CS-Cat)

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

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

accuracy of 96.26%. Cost-sensitive Random Forest (CS-RF) performed the next best with a balanced accuracy of 95.5%, and cost-sensitive Catboost (CS-Cat) performed the third best with a balanced accuracy of 91.3% (**Figure 5A**). All these three models reported an AUC value of 0.99. BDL-SP model also outperformed other models on AUPRC scoring the highest AUPRC of 0.92, while the AUPRC of both CSRF and CS-Cat model was 0.86 (**Figure 5B-D**). This is to note that AUPRC is one of the most important quantitative metrics and is more relevant compared to AUC on the unbalanced data. BDL-SP outperformed the other models on AUPRC with a great margin. This shows that, quantitatively, BDL-SP performed best, with the CS-RF model being the second best model.

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

Out of 824 significantly altered genes identified from the workflow shown in **Figure 2**, there were 31 oncogenes (OGs) (e.g. *KRAS, LTB, CARD11, NOTCH1*, etc.), 43 tumor-suppressor genes (TSGs) (e.g. *HLA-A/B/C, TRAF3, TP53, SDHA*, etc.), ten genes that were both oncogenes and driver genes (*KRAS, LTB, NRAS, FGFR3, BRAF*), and 19 actionable genes (e.g. *MITF, ARID1B, ARID2, RPTOR*, etc.) (**Table 3**). This full list of genes is provided in Table S1 of supplementary material. For each of the top-three models, we have considered only those genes in the top-250 or top-500 gene list that have a non-zero ShAP score in the post-hoc explainability analysis. The total counts of previously reported genes as found in the top-250 and top-500 genes of the top-three models is shown in **Table 4**.

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

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

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

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

1. *Category-1*: Pathways that become more significant with disease progression from MGUS to MM.

2. *Category-2*: Pathways that become less significant with disease progression from MGUS to MM.

3. *Category-3*: Significantly altered pathways observed only in MM and and not observed in MGUS.

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

The complete list of significantly altered pathways for the above mentioned four categories are provided in Tables S4 and S5 of supplementary material. In Category-1 of significantly altered pathways, 05 KEGG and 09 Reactome pathways became more significant as the disease progressed from MGUS to MM (**Figure 6**).

Table 5. List of 4 categories of previously reported genes as found in the post-hoc analysis based top-250 and top-500 genes of the top-3 models (BDL-SP, CS-RF, and CS-Cat)

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

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

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

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

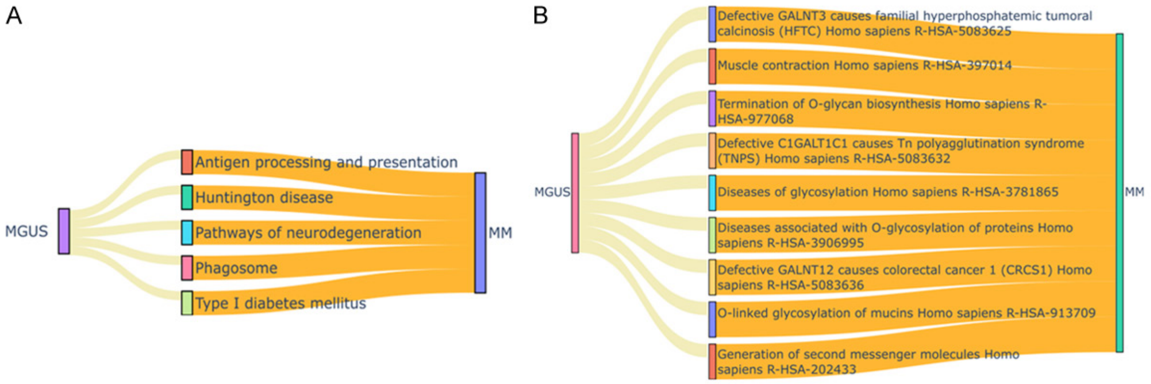

A
MGUS
Antigen processing and presentation
Huntington disease
Pathways of neurodegeneration
Phagosome
Type I diabetes mellitus
MM
B
MGUS
Defective GALNT3 causes familial hyperphosphatemic tumoral calcinosis (HFTC) Homo sapiens R-HSA-5083625
Muscle contraction Homo sapiens R-HSA-397014
Termination of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

Figure 6. Pathway enrichment analysis of the top-genes obtained from BDL-SP model. A. KEGG Pathways that gained more significance during progression from MGUS to MM. B. Reactome Pathways gained more significance during the progression from MGUS to MM. Here in both of the figures, pale golden and orange ribbon means significant p-adjusted value (<0.05); orange refers to more significant and pale golden color refers to less significant.

In Category-2, no pathway became less significant with disease progression in KEGG and in Reactome. In Category-3, 103 KEGG pathways and 125 Reactome pathways were observed as significantly altered only in MM and not in MGUS (**Figures 7** and **8**). We further observed that 14 out of 103 KEGG pathways and 14 out of 125 Reactome pathways had no overlapping genes with the set of 199 significantly altered genes in MGUS. Finally, in Category-4, no KEGG pathway, but one Reactome pathway was observed as significantly altered only in MGUS and not in MM (**Figure 9**). Further, we observed that several signaling pathways such as Calcium signaling, B-cell receptor signaling, MAPK signaling pathway, regulation of actin cytoskeleton, etc. were significantly altered only in MM (p-adjusted value >0.05) and were not observed to be significantly altered in MGUS. The KEGG pathways that were significantly involved in disease progression from MGUS to MM with highlighted top-ranking genes identified by BDL-SP are shown in **Figure 10**.

Explainability of the BDL-SP model using ShAP algorithm

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

Ranking of genes at the group-level from the explainability analysis of BDL-SP using ShAP: Based on the best ShAP score estimated for each genomic attribute using the algorithms shown in **Figure 4A** and **4B**, we ranked all the significantly altered genes at the group-level (MM/MGUS) to identify the top genes that significantly contributed to the model's prediction. The gene ranking of all 824 genes at group-level is shown in the beeswarm plot in Table S6 of supplementary material. In the beeswarm plot, each sample is represented as a dot, and the color of each dot corresponds to the best ShAP score of the gene. We have also highlighted all the previously reported genes of high relevance in MM in the beeswarm plot. In our analysis *KIR3DL2, IGLL5*, and *FCGR2A* are observed to be the top three genes based on their best ShAP scores in MGUS and MM samples from among the 824 significantly altered genes. Several previously reported driver genes in MM such as *IGLL5, HLA-A, KRAS, LTB, TP53, EGR1, FGFR3, NFKBIA, IRF1, NRAS*, etc. are observed in these top-ranked genes. Similarly, the previously reported oncogenes such as *CARD11, NOTCH1, VAV1, IRS1, MGAM, ABL2*, etc., and tumor-suppressor genes such as *HLA-B, HLA-C, SDHA*, etc. are observed in the top-ranked genes in our analysis. Also, many actionable genes are observed among the top genes, such as *KRAS, NOTCH1, TP53, FGFR3, ARID1B*, etc.

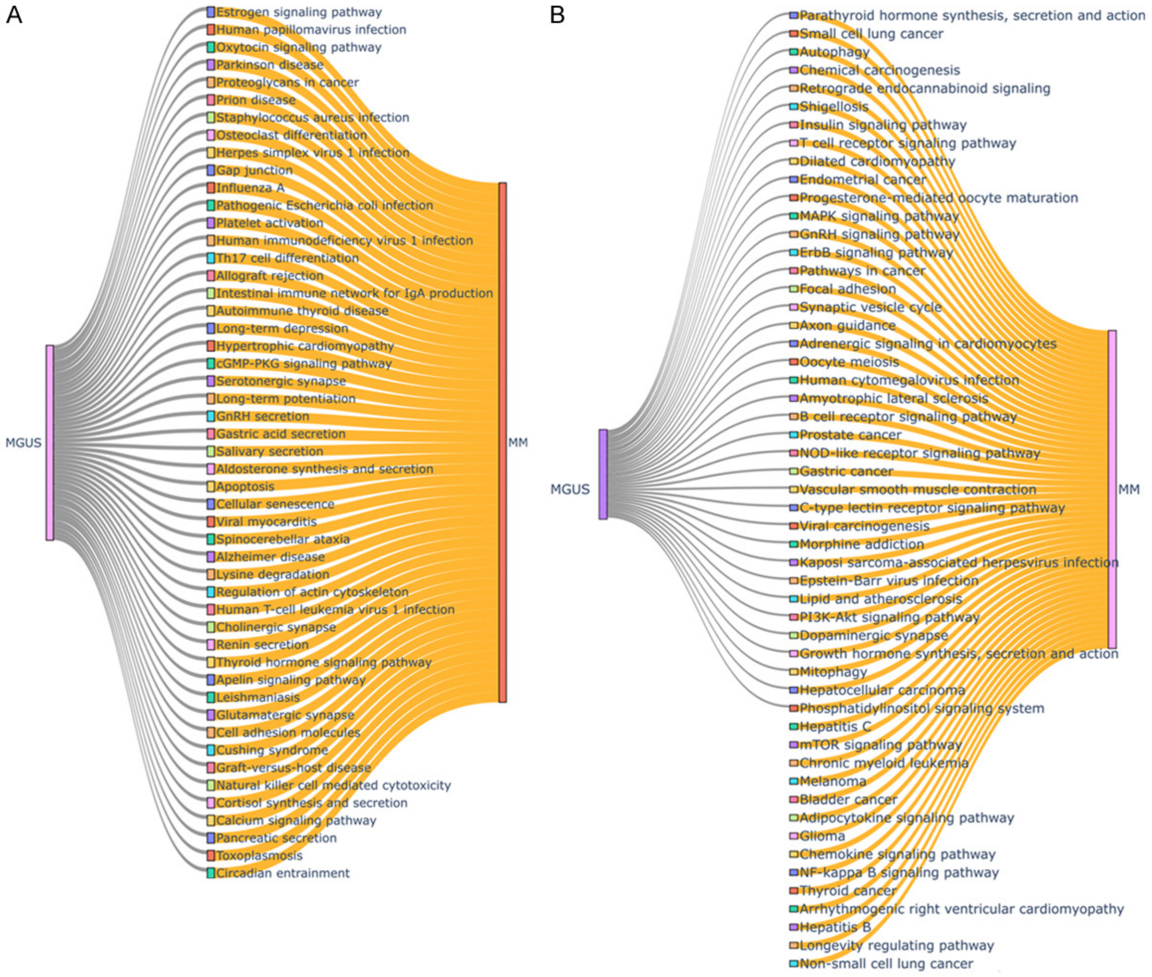


Figure 7. A, B. Pathway enrichment analysis of the top-genes obtained from BDL-SP model: KEGG Pathways that are uniquely significant in MM. In the above figure, orange ribbon means significant p-adjusted value (<0.05) and gray color refers to non-significant (p-adjusted value >0.05). There were a total of 108 KEGG pathways observed as significantly altered. Due to the large number of altered pathways, the above river plot was splitted into two parts to get more clarity.

Ranking of genes at the sample-level from the explainability analysis of BDL-SP using ShAP: In the sample-level analysis, we ranked genes found significantly altered in a sample according to their best ShAP scores estimated using the algorithm shown in **Figure 4A** and Algorithm A of **Table 2**. We observed that several previously reported OGs, TSGs, ODGs, and AGs were found in the top-ranked gene list of each sample. On assessing the ShAP scores of top significantly altered genes among all MM and MGUS samples, we observed that the mean ± standard deviation of the 100th ranked gene's ShAP score for all MM and MGUS samples is 0.0174±0.0037 and 0.0171±0.0040, respectively. Further, the ShAP score reduced to a considerably low value as we moved to a lower rank. Hence, we considered the top-100 significantly altered genes from all MM and MGUS samples based on their best ShAP scores for further analysis. The violin distribution plots for four gene groups of previously reported genes for all MM versus MGUS samples, only MGUS samples of EGA and AIIMS datasets, and only MM samples of MMRF and AIIMS datasets are shown in **Figure 11A-C**, respectively.

Analysis in MM & MGUS samples with ethnicity: We performed the statistical comparison of the disease stages (MM/MGUS) across American, European, and Indian populations (as mentioned in Section-2.1) on the basis of the number of previously reported genes in four gene groups using unpaired Wilcoxon ranksum

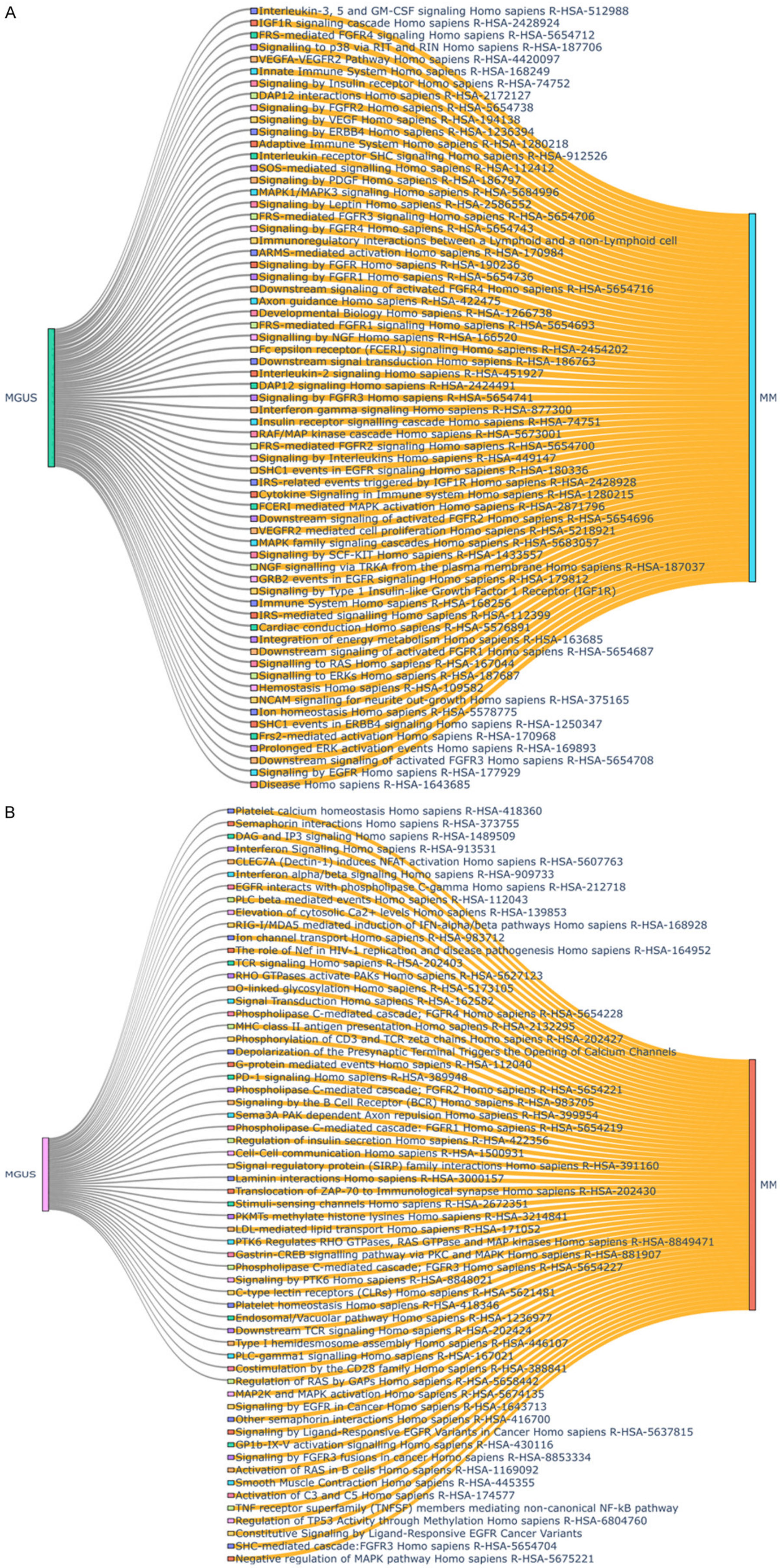

A
MGUS
MM
Interleukin-3, 5 and GM-CSF signaling Homo sapiens R-HSA-512988
IGF1R signaling cascade Homo sapiens R-HSA-2428924
FRS-mediated FGFR4 signaling Homo sapiens R-HSA-5654712
Signalling to p38 via RIT and RIN Homo sapiens R-HSA-187706
VEGFA-VEGFR2 Pathway Homo sapiens R-HSA-4420097
Innate Immune System Homo sapiens R-HSA-168249
Signaling by Insulin receptor Homo sapiens R-HSA-74752
DAP12 interactions Homo sapiens R-HSA-2172127
Signaling by FGFR2 Homo sapiens R-HSA-5654738
Signaling by VEGF Homo sapiens R-HSA-194138
Signaling by ERBB4 Homo sapiens R-HSA-1236394
Adaptive Immune System Homo sapiens R-HSA-1280218
Interleukin receptor SHC signaling Homo sapiens R-HSA-912526
SOS-mediated signalling Homo sapiens R-HSA-112412
Signaling by PDGF Homo sapiens R-HSA-186797
MAPK1/MAPK3 signaling Homo sapiens R-HSA-5684996
Signaling by Leptin Homo sapiens R-HSA-2586552
FRS-mediated FGFR3 signaling Homo sapiens R-HSA-5654706
Signaling by FGFR4 Homo sapiens R-HSA-5654743
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell
ARMS-mediated activation Homo sapiens R-HSA-170984
Signaling by FGFR Homo sapiens R-HSA-190236
Signaling by FGFR1 Homo sapiens R-HSA-5654736
Downstream signaling of activated FGFR4 Homo sapiens R-HSA-5654716
Axon guidance Homo sapiens R-HSA-422475
Developmental Biology Homo sapiens R-HSA-1266738
FRS-mediated FGFR1 signaling Homo sapiens R-HSA-5654693
Signalling by NGF Homo sapiens R-HSA-166520
Fc epsilon receptor (FCERI) signaling Homo sapiens R-HSA-2454202
Downstream signal transduction Homo sapiens R-HSA-186763
Interleukin-2 signaling Homo sapiens R-HSA-451927
DAP12 signaling Homo sapiens R-HSA-2424491
Signaling by FGFR3 Homo sapiens R-HSA-5654741
Interferon gamma signaling Homo sapiens R-HSA-877300
Insulin receptor signalling cascade Homo sapiens R-HSA-74751
RAF/MAP kinase cascade Homo sapiens R-HSA-5673001
FRS-mediated FGFR2 signaling Homo sapiens R-HSA-5654700
Signaling by Interleukins Homo sapiens R-HSA-449147
SHC1 events in EGFR signaling Homo sapiens R-HSA-180336
IRS-related events triggered by IGF1R Homo sapiens R-HSA-2428928
Cytokine Signaling in Immune system Homo sapiens R-HSA-1280215
FCERI mediated MAPK activation Homo sapiens R-HSA-2871796
Downstream signaling of activated FGFR2 Homo sapiens R-HSA-5654696
VEGFR2 mediated cell proliferation Homo sapiens R-HSA-5218921
MAPK family signaling cascades Homo sapiens R-HSA-5683057
Signaling by SCF-KIT Homo sapiens R-HSA-1433557
NGF signalling via TRKA from the plasma membrane Homo sapiens R-HSA-187037
GRB2 events in EGFR signaling Homo sapiens R-HSA-179812
Signaling by Type 1 Insulin-like Growth Factor 1 Receptor (IGF1R)
Immune System Homo sapiens R-HSA-168256
IRS-mediated signalling Homo sapiens R-HSA-112399
Cardiac conduction Homo sapiens R-HSA-5576891
Integration of energy metabolism Homo sapiens R-HSA-163685
Downstream signaling of activated FGFR1 Homo sapiens R-HSA-5654687
Signalling to RAS Homo sapiens R-HSA-167044
Signalling to ERKs Homo sapiens R-HSA-187687
Hemostasis Homo sapiens R-HSA-109582
NCAM signaling for neurite out-growth Homo sapiens R-HSA-375165
Ion homeostasis Homo sapiens R-HSA-5578775
SHC1 events in ERBB4 signaling Homo sapiens R-HSA-1250347
Frs2-mediated activation Homo sapiens R-HSA-170968
Prolonged ERK activation events Homo sapiens R-HSA-169893
Downstream signaling of activated FGFR3 Homo sapiens R-HSA-5654708
Signaling by EGFR Homo sapiens R-HSA-177929
Disease Homo sapiens R-HSA-1643685
B
MGUS
MM
Platelet calcium homeostasis Homo sapiens R-HSA-418360
Semaphorin interactions Homo sapiens R-HSA-373755
DAG and IP3 signaling Homo sapiens R-HSA-1489509
Interferon Signaling Homo sapiens R-HSA-913531
CLEC7A (Dectin-1) induces NFAT activation Homo sapiens R-HSA-5607763
Interferon alpha/beta signaling Homo sapiens R-HSA-909733
EGFR interacts with phospholipase C-gamma Homo sapiens R-HSA-212718
PLC beta mediated events Homo sapiens R-HSA-112043
Elevation of cytosolic Ca2+ levels Homo sapiens R-HSA-139853
RIG-I/MDA5 mediated induction of IFN-alpha/beta pathways Homo sapiens R-HSA-168928
Ion channel transport Homo sapiens R-HSA-983712
The role of Nef in HIV-1 replication and disease pathogenesis Homo sapiens R-HSA-164952
TCR signaling Homo sapiens R-HSA-202403
RHO GTPases activate PAKs Homo sapiens R-HSA-5627123
O-linked glycosylation Homo sapiens R-HSA-5173105
Signal Transduction Homo sapiens R-HSA-162582
Phospholipase C-mediated cascade; FGFR4 Homo sapiens R-HSA-5654228
MHC class II antigen presentation Homo sapiens R-HSA-2132295
Phosphorylation of CD3 and TCR zeta chains Homo sapiens R-HSA-202427
Depolarization of the Presynaptic Terminal Triggers the Opening of Calcium Channels
G-protein mediated events Homo sapiens R-HSA-112040
PD-1 signaling Homo sapiens R-HSA-389948
Phospholipase C-mediated cascade; FGFR2 Homo sapiens R-HSA-5654221
Signaling by the B Cell Receptor (BCR) Homo sapiens R-HSA-983705
Sema3A PAK dependent Axon repulsion Homo sapiens R-HSA-399954
Phospholipase C-mediated cascade: FGFR1 Homo sapiens R-HSA-5654219
Regulation of insulin secretion Homo sapiens R-HSA-422356
Cell-Cell communication Homo sapiens R-HSA-1500931
Signal regulatory protein (SIRP) family interactions Homo sapiens R-HSA-391160
Laminin interactions Homo sapiens R-HSA-3000157
Translocation of ZAP-70 to Immunological synapse Homo sapiens R-HSA-202430
Stimuli-sensing channels Homo sapiens R-HSA-2672351
PKMTs methylate histone lysines Homo sapiens R-HSA-3214841
LDL-mediated lipid transport Homo sapiens R-HSA-171052
PTK6 Regulates RHO GTPases, RAS GTPase and MAP kinases Homo sapiens R-HSA-8849471
Gastrin-CREB signalling pathway via PKC and MAPK Homo sapiens R-HSA-881907
Phospholipase C-mediated cascade; FGFR3 Homo sapiens R-HSA-5654227
Signaling by PTK6 Homo sapiens R-HSA-8848021
C-type lectin receptors (CLRs) Homo sapiens R-HSA-5621481
Platelet homeostasis Homo sapiens R-HSA-418346
Endosomal/Vacuolar pathway Homo sapiens R-HSA-1236977
Downstream TCR signaling Homo sapiens R-HSA-202424
Type I hemidesmosome assembly Homo sapiens R-HSA-446107
PLC-gamma1 signalling Homo sapiens R-HSA-167021
Costimulation by the CD28 family Homo sapiens R-HSA-388841
Regulation of RAS by GAPs Homo sapiens R-HSA-5658442
MAP2K and MAPK activation Homo sapiens R-HSA-5674135
Signaling by EGFR in Cancer Homo sapiens R-HSA-1643713
Other semaphorin interactions Homo sapiens R-HSA-416700
Signaling by Ligand-Responsive EGFR Variants in Cancer Homo sapiens R-HSA-5637815
GP1b-IX-V activation signalling Homo sapiens R-HSA-430116
Signaling by FGFR3 fusions in cancer Homo sapiens R-HSA-8853334
Activation of RAS in B cells Homo sapiens R-HSA-1169092
Smooth Muscle Contraction Homo sapiens R-HSA-445355
Activation of C3 and C5 Homo sapiens R-HSA-174577
TNF receptor superfamily (TNFSF) members mediating non-canonical NF-kB pathway
Regulation of TP53 Activity through Methylation Homo sapiens R-HSA-6804760
Constitutive Signaling by Ligand-Responsive EGFR Cancer Variants
SHC-mediated cascade:FGFR3 Homo sapiens R-HSA-5654704
Negative regulation of MAPK pathway Homo sapiens R-HSA-5675221

Figure 8. A. Pathway enrichment analysis of the top-genes obtained from BDL-SP model: Reactome Pathways that are uniquely significant in MM. In the above figure, orange ribbon means significant p-adjusted value (<0.05) and gray color refers to non-significant (p-adjusted value >0.05). There were a total of 134 Reactome pathways observed as significantly altered. Due to the large number of altered pathways, the above river plot was splitted into two parts to get more clarity. B. Pathway enrichment analysis of the top-genes obtained from BDL-SP model: Reactome Pathways that are uniquely significant in MM. In the above figure, orange ribbon means significant p-adjusted value (<0.05) and gray color refers to non-significant (p-adjusted value >0.05). There were a total of 134 Reactome pathways observed as significantly altered. Due to the large number of altered pathways, the above river plot was splitted into two parts to get more clarity.

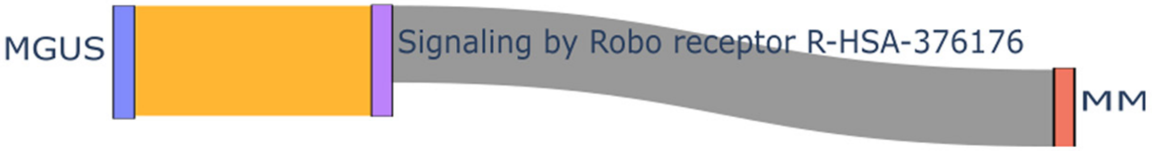


Figure 9. Pathway enrichment analysis of the top-genes obtained from BDL-SP model: Reactome Pathways that are uniquely significant in MGUS. In the above figure, orange ribbon means significant p-adjusted value (<0.05) and gray color refers to non-significant (p-adjusted value >0.05).

test. We observed that the number of genes in OG, ODG, and AG gene groups are significantly different between the disease stages (MM and MGUS) in the analysis of combined data of different geographic populations (**Figure 11A**). Further, the medians of the number of genes in the OG, TSG, and AG gene groups were observed to be higher than the respective medians in the precursor stage (MGUS) (**Figure 11A**). Similarly, comparing the number of genes in all four gene groups between the MGUS samples of Indian (AIIMS dataset) and European (EGA dataset) population, the number of genes in OG, ODG, and AG gene groups were observed to be significantly different (**Figure 11B**). On the contrary, the number of genes in all four gene groups were not found to be statistically and significantly different in MM samples of Indian (AIIMS dataset) and American population (MMRF dataset) (**Figure 11C**). These observations indicate that ethnicity might be playing a significant role in the disease development and thus, ethnicity-specific analysis can be helpful in further gaining in-depth insights into the disease biology of the premalignant stage of MM (MGUS).

Genomic feature ranking at a sample-level using ShAP analysis: Besides identifying the top-significantly altered genes in MM and MGUS, we also ranked the genomic features based on their contribution in the model prediction. A set of 28 genomic features (**Figure 1**) were used to train the BDL-SP model. These genes were ranked on the basis of their ShAP scores. The algorithm for estimating the best ShAP score for each genomic feature is shown in **Figure 4B**. We observed that the total number of SNVs, total number of SNVs in the *Other SNV group* (as shown in **Figure 1**), and VAF's standard deviation of SNVs in the *Other SNV group* were the top three genomic features, while VAF's standard deviation of SNVs in the *Non-synonymous SNV group*, VAF's standard deviation of SNVs in the *Synonymous SNV group*, and AD's standard deviation of SNVs in the *Non-synonymous SNV group* were the least contributing genomic features. The beeswarm plot for genomic feature ranking from BDL-SP model post-hoc analysis using ShAP is shown in **Figure 12**.

Discussion

It is well established that MM evolves through premalignant stages driven by the acquisition of multiple genomic aberrations [60]. Though a few studies have analyzed the progression from MGUS to MM [10, 13], a limited amount of information is available on the notable biomarkers responsible for this transformation. However, if known apriori, appropriate treatment at the MGUS stage can help control the progression of MGUS to MM, thereby preventing the complications associated with MM, reducing morbidity, and increasing the overall survival of these patients. Thus, it is crucial to unravel the genomic features responsible for the malignant transformation of MGUS to MM.

In this work, we addressed the challenge of extracting relevant MM and MGUS differentiat-

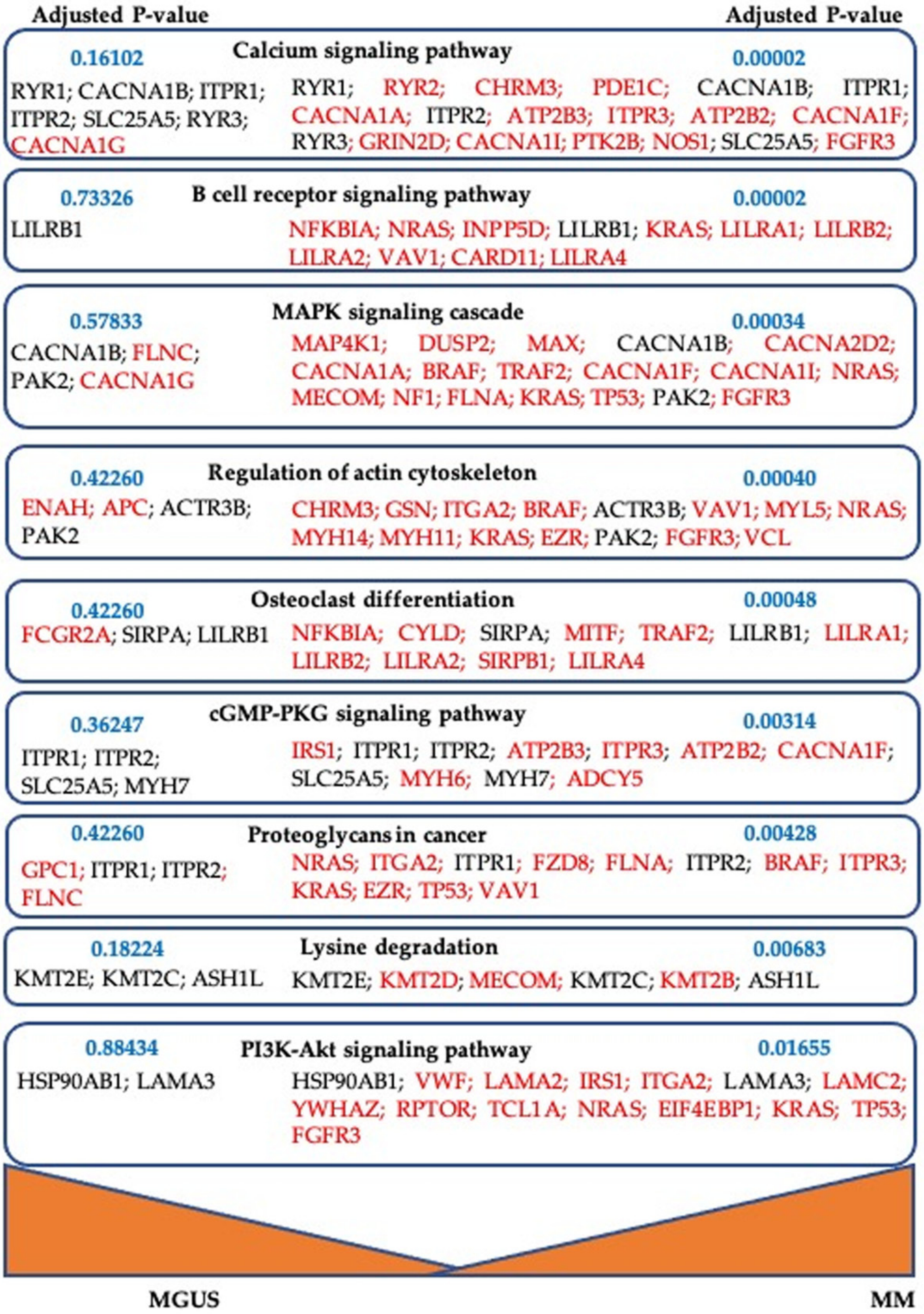
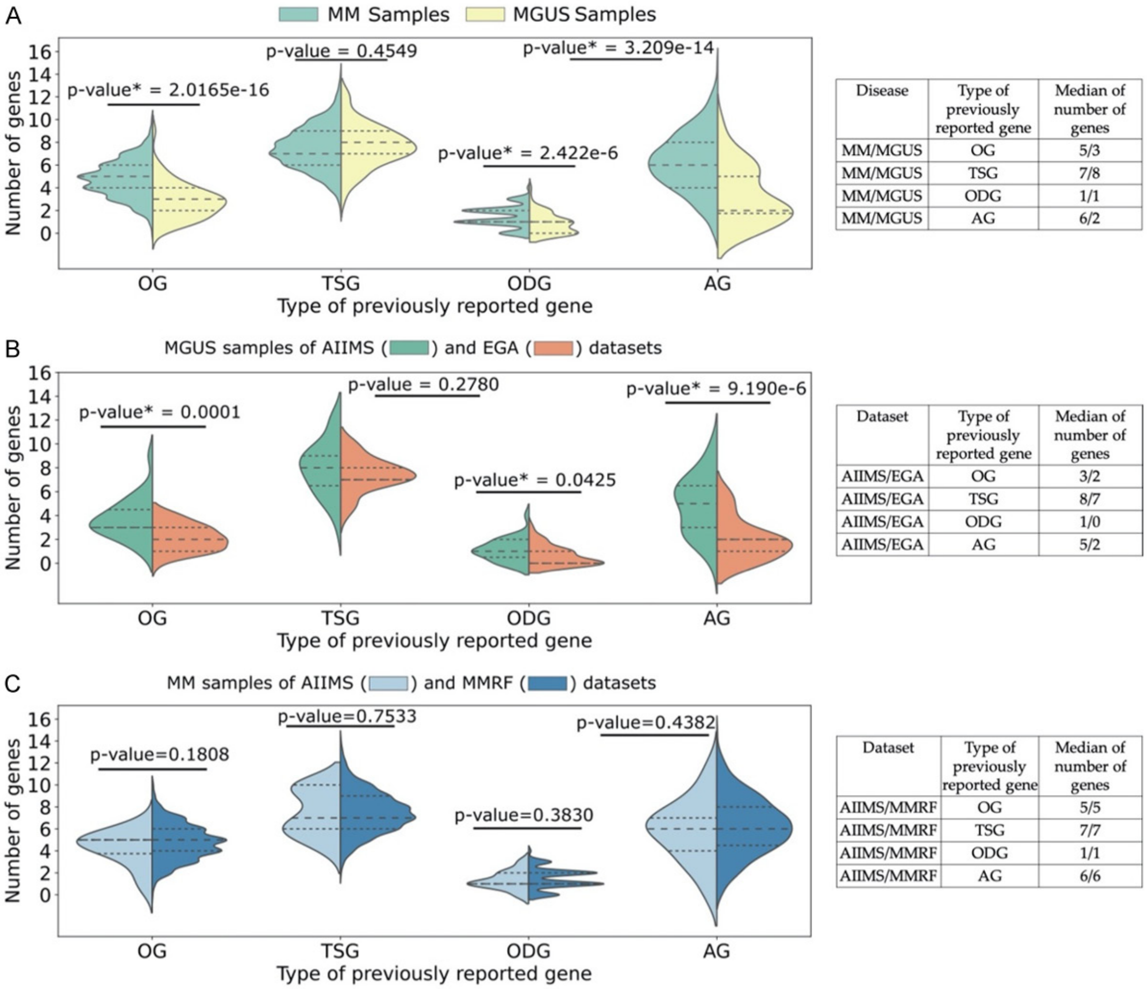

Adjusted P-value
Adjusted P-value
0.16102
Calcium signaling pathway
0.00002
RYR1; CACNA1B; ITPR1; ITPR2; SLC25A5; RYR3; CACNA1G
RYR1; RYR2; CHRM3; PDE1C; CACNA1B; ITPR1; CACNA1A; ITPR2; ATP2B3; ITPR3; ATP2B2; CACNA1F; RYR3; GRIN2D; CACNA1I; PTK2B; NOS1; SLC25A5; FGFR3
0.73326
B cell receptor signaling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

Figure 10. KEGG pathways found to be significantly involved in progression of MGUS to MM. Top genes that were identified by post-hoc analysis of BDL-SP using ShAP algorithm as significantly mutated either in MGUS only or in MM only (acting as differentiators of MGUS and MM) are shown in red colored font.



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

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

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

Figure 11. A. The distribution of the number of previously reported genes in four gene groups (OGs, TSGs, ODGs, and AGs) found significantly altered and ranked in top-100 across all MM and MGUS samples (combined dataset of MMRF, EGA, and AIIMS samples). B. The distribution of the number of previously reported genes found significantly altered and ranked in top-100 across all MGUS samples in EGA and AIIMS datasets. C. The distribution of the number of previously reported genes found significantly altered and ranked in top-100 across all MM samples in MMRF and AIIMS datasets (OG: Oncogenes, TSG: Tumor-Suppressor Genes, ODG: Both oncogenes and driver genes, AG: Actionable Genes). The *P*-value shown with each violin plot was estimated using unpaired Wilcoxon ranksum statistical test to check whether the number of genes in a particular gene group is significantly different from their respective counts in the other group. The gene group having *P*-value with superscript "*" (star) symbol represents that the number of genes in that gene group are significantly different compared to the other group. The table on the right of each figure shows the median of the number of genes in each gene group for disease stages (MM/MGUS) and datasets (MMRF, EGA, and AIIMS). Note: To have a better view of the violin plots, refer to the colored version of this figure.

ing genomic attributes from the pool of a large mutational information available for each patient. Our proposed BDL-SP based workflow has been successful in accomplishing this task. In the pre-processing of the data, we identified significantly mutated genes for each variant caller and then took the union of them so that we do not miss any important gene. Thus, a large cohort size and an ensemble of four variant callers enabled us to obtain generalizable mutational information, driver genes, and altered pathway information. Recently, graph-based learning has been extensively explored for deciphering crucial information such as disease progression, identification of novel biomarkers for targeted drug therapy, etc. in genomics. For example, the graph-based model was used to learn the temporal graphs of diagnosis (Dx), procedure (Px) and prescription (Rx) of multiple myeloma patients from the sequen-

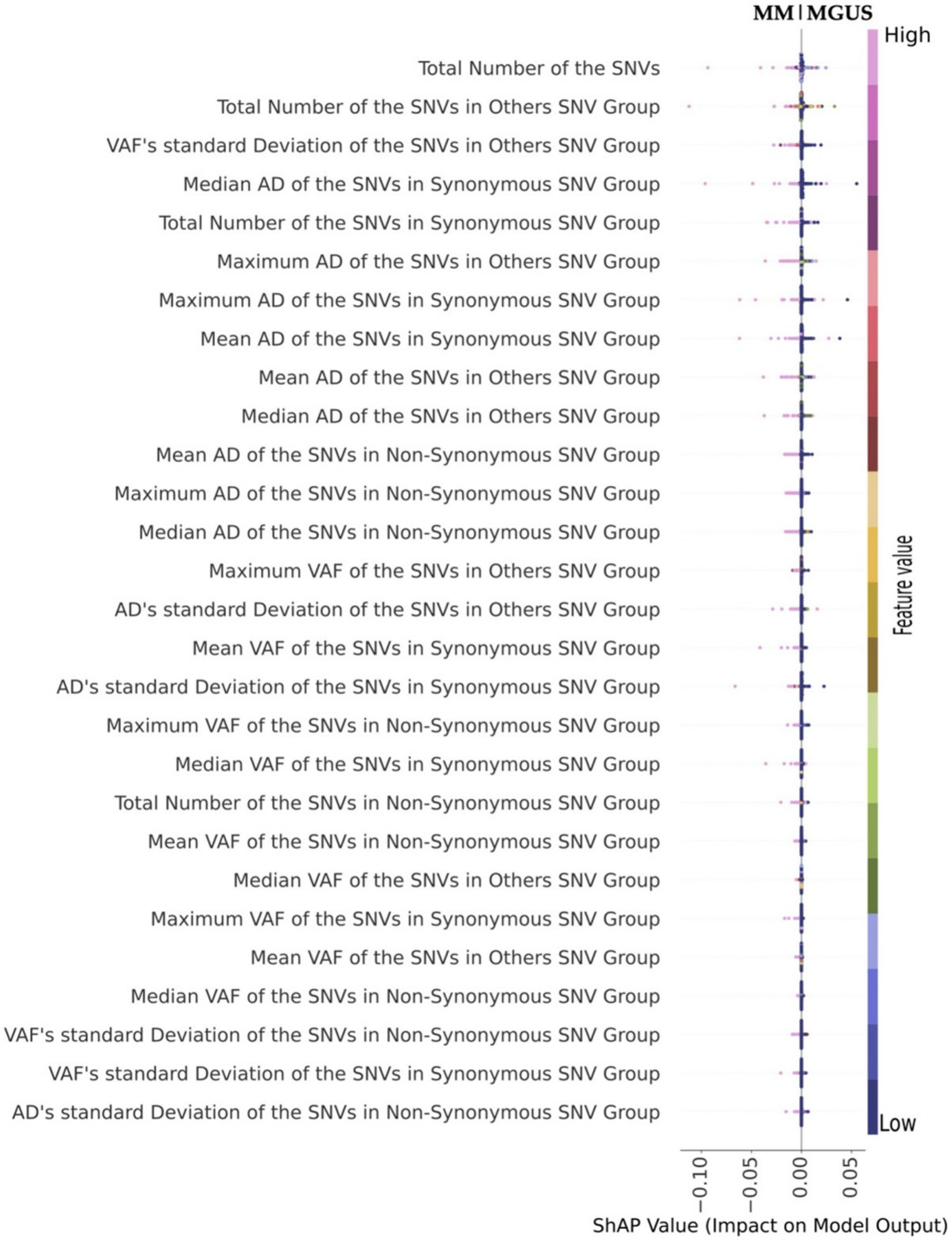


Figure 12. Genomic feature ranking based on the BDL-SP model's post-hoc explainability in MM and MGUS using ShAP algorithm. Each genomic feature is ranked according to their best ShAP score estimated using the algorithm shown in **Figure 4** and **Table 2**. The negative ShAP score represents the contribution of the genomic feature towards MM, while the positive ShAP score represents the contribution of the genomic feature towards MGUS. Further, each dot in the individual scatter plot of the genomic feature represents a sample and the color of dot represents the value of that genomic feature with the color-codes as follows: the dark blue color represents low and the pink color represents high value of the genomic feature. Note: Refer to the colored version of this figure for a clear view of the sample distribution for each genomic feature.

tial electronic health records (EHR) and predict a patient's response to treatment [61]. Till now, graph-based learning approaches have not yet been explored to identify the underlying difference between MM and its precursor stage (MGUS). In our BDL-SP model, we have used the connectionist model of graph-based learning to learn genomic mutational profiles (as node features) that were extracted from the WES datasets of AIIMS, EGA, and MMRF. We additionally utilized the gene-gene interaction information from PPI network to identify the pivotal biomarkers that can differentiate MM from MGUS.

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

1. The identification of pivotal biomarkers using WES datasets of MM of three populations (American, European, and Indian) increases the robustness of the workflow by enhancing its ability to assess the granular-level insights of mutational profiles from multiple datasets of different geographic locations/ethnicities.

2. Because of the pathogenic nature of deleterious SNVs, only deleterious SNVs were considered for identifying the significantly altered genes in the proposed workflow. We observed that the total number of SNVs were reduced considerably after variant filtration of benign SNVs using the FATHMM-XF [49] method.

3. An analysis of the genomic mutational profile along with the gene-gene interaction information enables this workflow to look at interdependencies between genes, making it a complete bio-inspired workflow.

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

The complete list of top significantly altered genes identified by the best three performing models (BDL-SP, CS-RF, and CS-Cat) is provided in **Table 5**. Of all, our proposed BDL-SP model identified the highest number of previously reported OGs, TSGs, ODGs, and AGs compared to the other standard ML methods. This shows that our GCN-based BDL-SP workflow is indeed capable of extracting the differentiating genomic features robustly that are otherwise difficult to obtain. Many of the top-ranking genes in the present study included known cancer drivers (*IGLL5, HLA-A, KRAS, LTB*, etc.), oncogenes (*KRAS, NRAS, FGFR3, BRAF*, etc.), tumor-suppressor genes (*HLA-A, LTB, TRAF3, EGR1, SAMHD1, DIS3, ARID2, CYLD, SP140*, etc.) and actionable genes (*KRAS, TP53, NF1, NFKBIA, ARID2*, etc.) having high relevance in MM. Interestingly, some TSGs (*HLA-B/C, NOTCH1, SDHA, MITF, ARID1B, FANCD2, KMT2D, APC, CMTR2* and *AMER1*) and oncogenes (*CARD11, NOTCH1, VAV1, IRS1, MGAM, ABL2, TCL1A, PGR, MITF, RPTOR, TERT, BRD4, MECOM*, and *TAL1*) that are so far not reported as drivers in MM, were also listed in the top ranking genes of BDL-SP. Further focused studies are required to validate the above finding and to check the functional status and other characteristics of these genes before classifying them as MM drivers.

Pathway analysis on MM and MGUS genes revealed that the MM related pathways such as MAPK, cGMP-PKG, B-cell receptor, etc. were not significantly altered in MGUS (adj *P*-value ≥0.05) and became significant in MM (adj *p*-value ≤0.05). We observed that several OGs, TSGs, ODGs, and AGs associated with the significantly altered pathways were found significantly altered only in the MM cohort and not in the MGUS cohort (See **Figure 13**). Here, the additional alterations in several previously reported genes such as *BRAF, FGFR3, IRS1, MAX, KRAS*, etc. assisted the malignant progression of MGUS to MM. Our pathway analysis also demonstrated that some pathways that lost their statistical significance from MGUS to MM were actually related to the other cancer types.

However, the results in our study are unique because we have demonstrated that these pathways are selectively and significantly dysregulated in MM compared to its precursor stage of MGUS due to a distinct set of genes that are differentially mutated in the two dis-

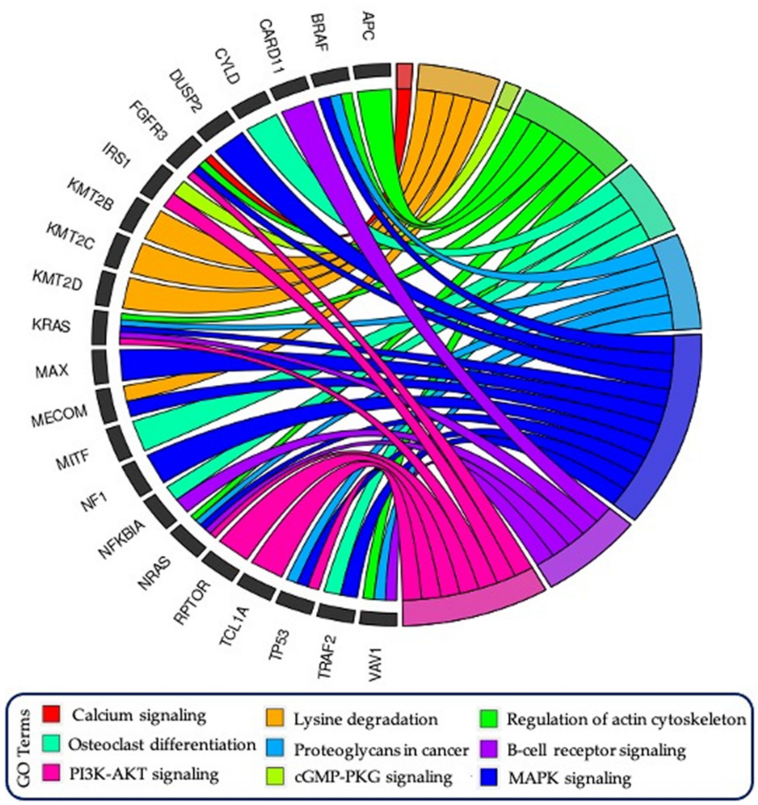


Figure 13. GOChord plot reveals the association of driver/TSG/Onco/Actionable genes associated with important pathways. The gene *KMT2C* was observed to be significantly mutated in MGUS and MM, while the gene *APC* was mutated only in MGUS. All other genes were observed to be significantly mutated in MM only.

eased stages. These observations warrant further investigations to decipher if any of these differentiating genes could become druggable targets especially during the early phase of MGUS. Some of the key genes and pathways that are selectively altered at the MM stage, such as *FGFR3, BRAF* and *MAP kinase* pathways, are actionable and hence, targeted therapy for them is under evaluation in clinical trials [62]. *FGFR3* is a partner gene in t(4;14) that has been observed as a significantly altered gene in all datasets of MM and MGUS. However, the poor prognostic impact of *FGFR3* has been linked to activating mutations in the *FGFR3* gene rather than the fusion event which exerts its influence via activation of *WHSC/MMSET* genes and is responsive to proteasome inhibitors [63, 64]. Besides, single case reports demonstrating efficacy of *BRAF* inhibitors in relapsed refractory MM with *BRAF* mutations, a recent report on NCI-Match trial shows promising results for *BRAF* inhibitors, Dabrafenib and Trametinib, in patients harboring tumors with *BRAF V600E* mutations including MM [65]. Many structural variations (SVs) observed in MM such as *IgH translocations, 1q gain, 1p del* are also observed at the MGUS stage. However, *C-MYC* alterations, which are predominantly structural variations, are secondary events and are seen in nearly 40% of MM patients [66]. Lack of analysis of structural variants is one of the limitations of this study. However, we did observe mutations in *MAX* at the MM stage, which is known to dimerize with *C-MYC* and influence the transcription of multiple genes and thus, the pathogenesis of MM [67].

The frequently observed complex genomic traits that can drive the disease progression from MGUS to MM can be 3'UTR/5'UTR mutations [12], copy number variants, structural variants [1, 60], and loss of the ability of the dysfunctional immune environment to control virulent cell clones [68]. Akin to higher levels of disease load in MM compared to its precursor states, measurable disease load, increased number of non-synonymous mutations in MM compared with MGUS [9-12] and increased levels of deregulated cytokines in relapsed refractory MM compared to treatment naïve MM has been reported [69].

In addition, Mikalasova et al. In MGUS, a positive correlation between the increasing chromosome changes and the somatic gene mutations, and absence of *MYC translocation* and *TP53 deletions* or mutations has been observed [11]. From the genomic profile analysis of paired MGUS-MM and SMM-MM samples, it has been observed that as the disease progressed, the number of NS mutations actually decreased in 70% samples. This observation is in contrast to reports on unpaired samples, where an increase in the NS mutations has been reported from MGUS to MM [9-12]. Further, the comparisons of unpaired MGUS/SMM and MM samples have shown the mutational similarity of MGUS/SMM with MM [13]. Based on this observation, we hypothesize that the progression is associated with an altered landscape of acquired mutations, rather than an increased total mutational burden.

The post-hoc explainability of the BDL-SP model using ShAP algorithm revealed the top genomic attributes (genomic features and significantly altered genes) at both the group- and sample-levels. At group-level, all the 824 significantly altered genes were ranked according to their ShAP score using the algorithm shown in **Figure 4** (Table S6 of supplementary material) and top-500 genes were further compared with the literature (Table S7 of supplementary material). Several significantly altered genes found in our analysis were previously reported as driver genes in [7, 53, 54], oncogenes and TSGs in [52], and actionable genes in [55, 56], while some genes such as *KIR3DL2, FCGR2A, LILRB1/2, KIR2DL1/4* etc. were novel that contributed significantly in disease classification (See Table S6 of supplementary material). The *KIR* framework genes (*KIR3DL2/2DL4*) were among the top significantly altered genes with highest ShAP scores. The KIR gene complex on chromosome 19 encodes a series of inhibitory or activatory *KIR* genes expressed on *NK cells* [70-72]. These receptors serve as *HLA* ligands and modulate *NK* cell immune function against tumors [70]. A few activating genes in the *KIR* family (*KIR2DS4* and *KIR2DS5*) have been shown to have higher prevalence in MM patients [70] than healthy people. The *KIR*s have also been reported to influence the efficacy of therapies such as that of isatuximab [71]. The second topmost gene with the highest ShAP score was *IGLL5*. Again, the *IGLL5* gene undergoes point mutations and *IGLL5/IGH* translocations in MM [73]. Point mutations are largely mutually exclusive of *RAS* mutations and associated with greater risk of disease progression. Similarly, other genes such as *HLA-A/B/C, FCGR2A* and *LILRB1/2* reported in previous studies are also shown to have a significant role in MM [74-77]. Given the crucial role of these top immune related genes highlighted by the ShAP ranking in our study suggests their potential role as drivers of progression and disease stratifying biomarkers.

We have also highlighted the impact of ethnicity (**Figure 11**) among three groups of American (MMRF), European (EGA), and Indian (AIIMS) population. The number of OGs, ODGs, and AGs were significantly different in the MM samples of American and Indian population, and MGUS samples of European and Indian population (**Figure 11A**). Also, the median of the number of OGs and AGs increased with the disease progression from MGUS to MM. This increase could be due to the increasingly active participation of OGs and AGs in disease progression from MGUS to MM. Similarly, the number of OGs, ODGs, and AGs are significantly different in the MGUS samples of Indian population and MGUS samples of European population (**Figure 11B**). Here, we also observed the increasingly active participation of OGs, ODGs, and AGs in MGUS samples of Indian and European population. On the other hand, the number of previously reported genes (OGs, ODGs, TSGs, and AGs) present in the MM samples of the American and Indian population were not found to be statically different (**Figure 11C**). These observations indicate that the impact of ethnicity on disease biology can not be overlooked and might be an important factor during the initial phase or development phase of MM. Further analysis of ethnicity-specific information to infer the responsible prognostic factor for disease development and progression is strongly suggested.

The sample-wise gene-ranking highlighted their contribution at the individual sample-level. The study in [11] showed that the transition from MGUS to MM is due to the acquisition of mutations in critical driver genes and oncogenes. Interestingly, we have observed that not only driver genes and oncogenes, but several TSGs and actionable genes were also altered significantly in MGUS (**Figure 11**). Further, the role of oncogenes increased as disease progressed from MGUS to MM. On comparing the top contributing genomic features in MGUS and MM samples, we observed that the genomic features related to the *Synonymous SNVs group* (a group of UTR3, synonymous, and UTR5 type SNVs) and the *Other SNVs group* (a group of Non-frameshift insertion/deletion/substitution, intronic, intergenic, ncRNA_intronic, upstream, downstream, unknown, and ncRNA_splicing SNVs) contributed largely in disease classification as compared to the genomic fea-

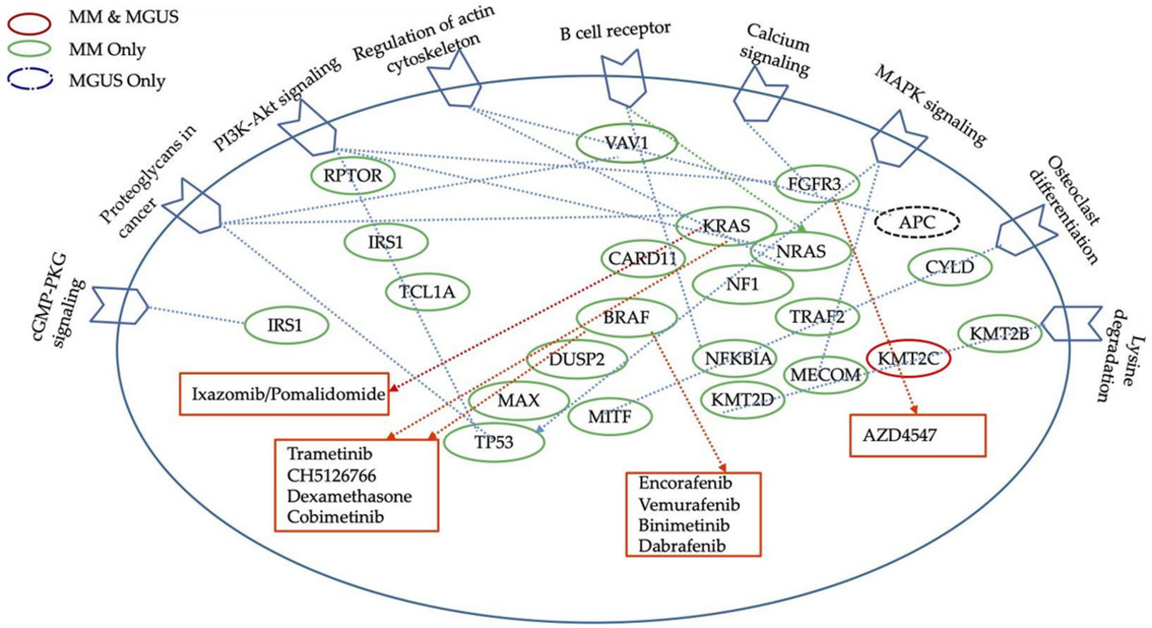


Figure 14. Important pathways significantly altered in MM. Drugs used for pathway-directed therapies associated with mutations in genes are also shown with red colored text-boxes and arrows.

tures of the *Non-synonymous SNVs group* (**Figure 12**).

Although the role of synonymous SNVs are unclear in MM, recent studies have observed these synonymous SNVs as significant contributors in multiple cancer types [78-82]. Further exploration of differentially affected biological pathways may provide the pathogenic link between MM, its precursor (MGUS or SMM), and overt disease stages so as to find appropriate targeted therapy to halt the progression from precursor to stage to MM (**Figure 14**). We have shown in the current study that the incremental accumulation of key mutations tilts the balance of biological pathways in favor of progression from the state of MGUS to MM in a large cohort of unpaired MGUS-MM samples. Some of these pathways are actionable and, targeting them may enable us to reverse the balance in favor of a controlled and relatively indolent clinical course. Further, AI-based workflow has assisted in successfully differentiating MGUS from MM. We have shown in our study that our trained machine learning (ML) classifiers are able to identify pivotal genomic biomarkers helpful in distinguishing MM and MGUS, thereby, leading to a better understanding of malignant transformation of MGUS to MM and prognostication.

Conclusion

MGUS and MM share many common features such as genomic biomarkers and structural variants, although MGUS has a relatively less complex genomic profile than MM. Therefore, it is a challenging task to distinguish MM from MGUS. In our proposed work, we have presented an innovative, bio-inspired AI-based workflow BDL-SP to identify pivotal genomic biomarkers to distinguish MGUS from MM. The proposed graph convolutional network based BDL-SP model is able to extract discriminative genomic biomarkers for identifying MM and MGUS samples. BDL-SP outperformed all the baseline ML-based models. Further, using the application-aware interpretability analysis of the trained AI model, we have demonstrated a way to identify the best AI model from among the multiple machine learning or deep learning models that may have performed similarly on the quantitative metrics on the available data. In the post-hoc interpretability benchmarking, BDL-SP outperformed all the baseline models by identifying the highest number of previously reported genes such as *KRAS, BRAF, LTB, NRAS, FGFR3, NF1, NFKBIA, ARID2, RB1, HLA-A, TP53, SP140, TRAF3, EGR1, IRF1, SAMHD1, DIS3, CYLD, KMT2B/C, MAX, ZFHX3 and NCOR1,* that are of high relevance in MM.

Further, some of the genes that acted as differentiable biomarkers included TSGs (*HLA-B/C, NOTCH1, SDHA, MITF, ARID1B, FANCD2, KMT2D, APC, CMTR2*, and *AMER1*) and oncogenes (*CARD11, NOTCH1, VAV1, IRS1, MGAM, ABL2, TCL1A, PGR, MITF, RPTOR, TERT, BRD4, MECOM*, and *TAL1*) that have not yet been identified as MM drivers. These require validation by future studies before being declared as MM drivers. We further validated our findings by performing pathway analysis on the top mutated genes. It was inferred from the pathway analysis that several signaling pathways such as Calcium signaling pathway, B-Cell receptor signaling pathway, PI3K-Akt signaling pathway, MAPK signaling pathway, etc. are selectively and more significantly dysregulated with disease progression. Additional mutations in driver genes, critical oncogenes, tumor-suppressor genes, and actionable genes fostered the transformation of benign MGUS to MM. Similarly, the genomic mutation associated with the Synonymous SNV group (synonymous SNVs, UTR3, and UTR5) were found to be the most significantly contributing biomarker differentiating MM from MGUS. These observations may hold great significance from a therapeutic point of view. We observed that the number of oncogenes, driver genes, and actionable genes in the MGUS samples of European and Indian populations were statistically different. Although no population specific differences were observed in our analysis of the MM data consisting of the American and Indian population, the results on MGUS data indicates that the impact of ethnicity on the disease biology of MM should be further explored.

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). 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. Sriram K. (Department of Computational Biology, IIIT-Delhi) and Akanksha Farswan (SBI Lab, Department of ECE, IIIT-Delhi) that greatly improved the AI-based workflow.

Voluntary written informed consent was obtained from all the study individuals.

Disclosure of conflict of interest

None.

Address correspondence to: Anubha Gupta, SBI Lab, Department of Electronics and Communication Engineering & Centre of Excellence in Healthcare, Indraprastha Institute of Information Technology-Delhi (IIIT-D), New Delhi 110020, India. ORCID: 0000-0002-7752-1926; E-mail: anubha@iiitd.ac.in; Ritu Gupta, Laboratory Oncology Unit, Dr. B.R.A. IRCH, All India Institute of Medical Sciences (AIIMS), Ansari Nagar, New Delhi 110029, India. ORCID: 0000000153644086; E-mail: drritugupta@gmail.com; drritu.laboncology@aiims.edu

References

[1] Manier S, Salem K, Glavey SV, Roccaro AM and Ghobrial IM. Genomic aberrations in multiple myeloma. Cancer Treat Res 2016; 169: 23-34.

[2] Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, Sougnez C, Schinzel AC, Harview CL, Brunet JP, Ahmann GJ, Adli M, Anderson KC, Ardlie KG, Auclair D, Baker A, Bergsagel PL, Bernstein BE, Drier Y, Fonseca R, Gabriel SB, Hofmeister CC, Jagannath S, Jakubowiak AJ, Krishnan A, Levy J, Liefeld T, Lonial S, Mahan S, Mfuko B, Monti S, Perkins LM, Onofrio R, Pugh TJ, Rajkumar SV, Ramos AH, Siegel DS, Sivachenko A, Stewart AK, Trudel S, Vij R, Voet D, Winckler W, Zimmerman T, Carpten J, Trent J, Hahn WC, Garraway LA, Meyerson M, Lander ES, Getz G and Golub TR. Initial genome sequencing and analysis of multiple myeloma. Nature 2011; 471: 467-472.

[3] Bolli N, Avet-Loiseau H, Wedge DC, Van Loo P, Alexandrov LB, Martincorena I, Dawson KJ, Iorio F, Nik-Zainal S, Bignell GR, Hinton JW, Li Y, Tubio JM, McLaren S, O'Meara S, Butler AP, Teague JW, Mudie L, Anderson E, Rashid N, Tai YT, Shammas MA, Sperling AS, Fulciniti M, Richardson PG, Parmigiani G, Magrangeas F, Minvielle S, Moreau P, Attal M, Facon T, Futreal PA, Anderson KC, Campbell PJ and Munshi NC. Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. Nat Commun 2014; 5: 2997.

[4] Walker BA, Boyle EM, Wardell CP, Murison A, Begum DB, Dahir NM, Proszek PZ, Johnson DC, Kaiser MF, Melchor L, Aronson LI, Scales M, Pawlyn C, Mirabella F, Jones JR, Brioli A, Mikulasova A, Cairns DA, Gregory WM, Quartilho A, Drayson MT, Russell N, Cook G, Jackson GH, Leleu X, Davies FE and Morgan GJ. Mutational spectrum, copy number changes, and outcome: results of a sequencing study of patients with newly diagnosed myeloma. J Clin Oncol 2015; 33: 3911-20.

[5] Lohr JG, Stojanov P, Carter SL, Cruz-Gordillo P, Lawrence MS, Auclair D, Sougnez C, Knoechel B, Gould J, Saksena G, Cibulskis K, McKenna A, Chapman MA, Straussman R, Levy J, Perkins LM, Keats JJ, Schumacher SE, Rosenberg M; Multiple Myeloma Research Consortium, Getz G and Golub TR. Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. Cancer Cell 2014; 25: 91-101.

[6] Farswan A, Jena L, Kaur G, Gupta A, Gupta R, Rani L, Sharma A and Kumar L. Branching clonal evolution patterns predominate mutational landscape in multiple myeloma. Am J Cancer Res 2021; 11: 5659-5679.

[7] Walker BA, Mavrommatis K, Wardell CP, Ashby TC, Bauer M, Davies FE, Rosenthal A, Wang H, Qu P, Hoering A, Samur M, Towfic F, Ortiz M, Flynt E, Yu Z, Yang Z, Rozelle D, Obenauer J, Trotter M, Auclair D, Keats J, Bolli N, Fulciniti M, Szalat R, Moreau P, Durie B, Stewart AK, Goldschmidt H, Raab MS, Einsele H, Sonneveld P, San Miguel J, Lonial S, Jackson GH, Anderson KC, Avet-Loiseau H, Munshi N, Thakurta A and Morgan GJ. Identification of novel mutational drivers reveals oncogene dependencies in multiple myeloma. Blood 2018; 132: 587-597.

[8] Kaur G, Jena L, Gupta R, Farswan A, Gupta A and Sriram K. Correlation of changes in subclonal architecture with progression in the MMRF CoMMpass study. Transl Oncol 2022; 23: 101472.

[9] Mikulasova A, Smetana J, Wayhelova M, Janyskova H, Sandecka V, Kufova Z, Almasi M, Jarkovsky J, Gregora E, Kessler P, Wrobel M, Walker BA, Wardell CP, Morgan GJ, Hajek R and Kuglik P. Genomewide profiling of copy-number alteration in monoclonal gammopathy of undetermined significance. Eur J Haematol 2016; 97: 568-575.

[10] Walker BA, Wardell CP, Melchor L, Brioli A, Johnson DC, Kaiser MF, Mirabella F, Lopez-Corral L, Humphray S, Murray L, Ross M, Bentley D, Gutiérrez NC, Garcia-Sanz R, San Miguel J, Davies FE, Gonzalez D and Morgan GJ. Intraclonal heterogeneity is a critical early event in the development of myeloma and precedes the development of clinical symptoms. Leukemia 2014; 28: 384-390.

[11] Mikulasova A, Wardell CP, Murison A, Boyle EM, Jackson GH, Smetana J, Kufova Z, Pour L, Sandecka V, Almasi M, Vsianska P, Gregora E, Kuglik P, Hajek R, Davies FE, Morgan GJ and Walker BA. The spectrum of somatic mutations in monoclonal gammopathy of undetermined significance indicates a less complex genomic landscape than that in multiple myeloma. Haematologica 2017; 102: 1617-1625.

[12] Farswan A, Gupta A, Jena L, Ruhela V, Kaur G and Gupta R. Characterizing the mutational landscape of MM and its precursor MGUS. Am J Cancer Res 2022; 12: 1919-1933.

[13] Dutta AK, Fink JL, Grady JP, Morgan GJ, Mullighan CG, To LB, Hewett DR and Zannettino ACW. Subclonal evolution in disease progression from MGUS/SMM to multiple myeloma is characterised by clonal stability. Leukemia 2019; 33: 457-468.

[14] Mosquera Orgueira A, González Pérez MS, Díaz Arias JÁ, Antelo Rodríguez B, Alonso Vence N, Bendaña López Á, Abuín Blanco A, Bao Pérez L, Peleteiro Raíndo A, Cid López M, Pérez Encinas MM, Bello López JL and Mateos Manteca MV. Survival prediction and treatment optimization of multiple myeloma patients using machine-learning models based on clinical and gene expression data. Leukemia 2021; 35: 2924-2935.

[15] Venezian Povoa L, Ribeiro CHC and Silva ITD. Machine learning predicts treatment sensitivity in multiple myeloma based on molecular and clinical information coupled with drug response. PLoS One 2021; 16: e0254596.

[16] Farswan A, Gupta A, Gupta R, Hazra S, Khan S, Kumar L and Sharma A. AI-supported modified risk staging for multiple myeloma cancer useful in real-world scenario. Transl Oncol 2021; 14: 101157.

[17] Farswan A, Gupta A, Sriram K, Sharma A, Kumar L and Gupta R. Does ethnicity matter in multiple myeloma risk prediction in the era of genomics and novel agents? Evidence from real-world data. Front Oncol 2021; 11: 720932.

[18] Farswan A and Gupta A. TV-DCT: method to impute gene expression data using DCT based sparsity and total variation denoising. ICASSP 2019; 1244-1248.

[19] Farswan A, Gupta A, Gupta R and Kaur G. Imputation of gene expression data in blood cancer and its significance in inferring biological pathways. Front Oncol 2020; 9: 1442.

[20] Anzar I, Sverchkova A, Stratford R and Clancy T. NeoMutate: an ensemble machine learning framework for the prediction of somatic mutations in cancer. BMC Med Genomics 2019; 12: 63.

[21] Hsu YC, Hsiao YT, Kao TY, Chang JG and Shieh GS. Detection of somatic mutations in exome

sequencing of tumor-only samples. Sci Rep 2017; 7: 15959.

[22] Pounraja VK, Jayakar G, Jensen M, Kelkar N and Girirajan S. A machine-learning approach for accurate detection of copy number variants from exome sequencing. Genome Res 2019; 29: 1134-1143.

[23] Huang T, Li J, Jia B and Sang H. CNV-MEANN: a neural network and mind evolutionary algorithm-based detection of copy number variations from next-generation sequencing data. Front Genet 2021; 12: 700874.

[24] Hill T and Unckless RL. A deep learning approach for detecting copy number variation in next-generation sequencing data. G3 (Bethesda) 2019; 9: 3575-3582.

[25] Li Y and Luo Y. Performance-weighted-voting model: an ensemble machine learning method for cancer type classification using whole-exome sequencing mutation. Quant Biol 2020; 8: 347-358.

[26] Collier O, Stoven V and Vert JP. LOTUS: a single- and multitask machine learning algorithm for the prediction of cancer driver genes. PLoS Comput Biol 2019; 15: e1007381.

[27] Han Y, Yang J, Qian X, Cheng WC, Liu SH, Hua X, Zhou L, Yang Y, Wu Q, Liu P and Lu Y. DriverML: a machine learning algorithm for identifying driver genes in cancer sequencing studies. Nucleic Acids Res 2019; 47: e45.

[28] Zeng Z, Mao C, Vo A, Li X, Nugent JO, Khan SA, Clare SE and Luo Y. Deep learning for cancer type classification and driver gene identification. BMC Bioinformatics 2021; 22 Suppl 4: 491.

[29] Luo P, Ding Y, Lei X and Wu FX. deepDriver: predicting cancer driver genes based on somatic mutations using deep convolutional neural networks. Front Genet 2019; 10: 13.

[30] Kipf TN and Welling M. Semi-supervised classification with graph convolutional networks. ArXiv Preprint 2016; ArXiv: 1609.02907.

[31] Bruna J, Zaremba W, Szlam A and LeCun Y. Spectral networks and locally connected networks on graphs. ArXiv Preprint 2013; ArXiv: 1312.6203.

[32] Schulte-Sasse R, Budach S, Hnisz D and Marsico A. Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. Nat Mach Intell 2021; 3: 513-526.

[33] Barabási AL, Gulbahce N and Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet 2011; 12: 56-68.

[34] Goh KI, Cusick ME, Valle D, Childs B, Vidal M and Barabási AL. The human disease network. Proc Natl Acad Sci U S A 2007; 104: 8685-8690.

[35] Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, Jensen LJ and von Mering C. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res 2021; 49: D605-D612.

[36] Lundberg SM and Lee SI. A unified approach to interpreting model predictions. NuraIPS 2017; 30.

[37] Ahmad MA, Eckert C and Teredesai A. Interpretable machine learning in healthcare. ACM-BCB 2018; 559-560.

[38] Du M, Liu N and Hu X. Techniques for interpretable machine learning. CACM 2019; 63: 68-77.

[39] Keats JJ, Craig DW, Liang W, Venkata Y, Kurdoglu A, Aldrich J, Auclair D, Allen K, Harrison B, Jewell S, Kidd PG, Correll M, Jagannath S, Siegel DS, Vij R, Orloff G, Zimmerman TM; Mmrf CoMMpass Network, Capone W, Carpten J and Lonial S. Interim analysis of the MMRF CoMMpass Trial, a longitudinal study in multiple myeloma relating clinical outcomes to genomic and immunophenotypic profiles. Blood 2013; 122: 532.

[40] Lappalainen I, Almeida-King J, Kumanduri V, Senf A, Spalding JD, Ur-Rehman S, Saunders G, Kandasamy J, Caccamo M, Leinonen R, Vaughan B, Laurent T, Rowland F, Marin-Garcia P, Barker J, Jokinen P, Torres AC, de Argila JR, Llobet OM, Medina I, Puy MS, Alberich M, de la Torre S, Navarro A, Paschall J and Flicek P. The European genome-phenome archive of human data consented for biomedical research. Nat Genet 2015; 47: 692-695.

[41] Fan Y, Xi L, Hughes DS, Zhang J, Zhang J, Futreal PA, Wheeler DA and Wang W. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. Genome Biol 2016; 17: 178.

[42] Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C and Lichtenstein L. Calling somatic SNVs and indels with Mutect2. BioRxiv 2019; 861054.

[43] Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L and Wilson RK. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 2012; 22: 568-576.

[44] Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, Ley TJ, Mardis ER, Wilson RK and Ding L. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics 2012; 28: 311-317.

[45] Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S and DePristo MA. From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics 2013; 43: 11.10.1-11.10.33.

[46] Li H and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009; 25: 1754-1760.

[47] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M and DePristo MA. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010; 20: 1297-1303.

[48] Wang K, Li M and Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010; 38: e164.

[49] Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR and Campbell C. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics 2018; 34: 511-513.

[50] Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, Davies H, Stratton MR and Campbell PJ. Universal patterns of selection in cancer and somatic tissues. Cell 2017; 171: 1029-1041.

[51] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V and Vanderplas J. Scikit-learn: machine learning in python. J Mach Learn Res 2011; 12: 2825-2830.

[52] Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, Rudolph JE, Yaeger R, Soumerai T, Nissan MH, Chang MT, Chandarlapaty S, Traina TA, Paik PK, Ho AL, Hantash FM, Grupe A, Baxi SS, Callahan MK, Snyder A, Chi P, Danila D, Gounder M, Harding JJ, Hellmann MD, Iyer G, Janjigian Y, Kaley T, Levine DA, Lowery M, Omuro A, Postow MA, Rathkopf D, Shoushtari AN, Shukla N, Voss M, Paraiso E, Zehir A, Berger MF, Taylor BS, Saltz LB, Riely GJ, Ladanyi M, Hyman DM, Baselga J, Sabbatini P, Solit DB and Schultz N. OncoKB: a precision oncology knowledge base. JCO Precis Oncol 2017; 2017: PO.17.00011.

[53] Maura F, Bolli N, Angelopoulos N, Dawson KJ, Leongamornlert D, Martincorena I, Mitchell TJ, Fullam A, Gonzalez S, Szalat R, Abascal F, Rodriguez-Martin B, Samur MK, Glodzik D, Roncador M, Fulciniti M, Tai YT, Minvielle S, Magrangeas F, Moreau P, Corradini P, Anderson KC, Tubio JMC, Wedge DC, Gerstung M, Avet-Loiseau H, Munshi N and Campbell PJ. Genomic landscape and chronological reconstruction of driver events in multiple myeloma. Nat Commun 2019; 10: 3835.

[54] Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, Mularoni L, Pich O, Bonet J, Kranas H, Gonzalez-Perez A and Lopez-Bigas N. A compendium of mutational cancer driver genes. Nat Rev Cancer 2020; 20: 555-572.

[55] Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, Fish P, Harsha B, Hathaway C, Jupe SC, Kok CY, Noble K, Ponting L, Ramshaw CC, Rye CE, Speedy HE, Stefancsik R, Thompson SL, Wang S, Ward S, Campbell PJ and Forbes SA. COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res 2019; 47: D941-D947.

[56] De Cesco S, Davis JB and Brennan PE. TargetDB: a target information aggregation tool and tractability predictor. PLoS One 2020; 15: e0232644.

[57] Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW and Ma'ayan A. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 2016; 44: W90-W97.

[58] Xie Z, Bailey A, Kuleshov MV, Clarke DJB, Evangelista JE, Jenkins SL, Lachmann A, Wojciechowicz ML, Kropiwnicki E, Jagodnik KM, Jeon M and Ma'ayan A. Gene set knowledge discovery with enrichr. Curr Protoc 2021; 1: e90.

[59] Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR and Ma'ayan A. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics 2013; 14: 128.

[60] Lopez-Corral L, Sarasquete ME, Beà S, García-Sanz R, Mateos MV, Corchete LA, Sayagués JM, García EM, Bladé J, Oriol A, Hernández-García MT, Giraldo P, Hernández J, González M, Hernández-Rivas JM, San Miguel JF and Gutiérrez NC. SNP-based mapping arrays reveal high genomic complexity in monoclonal gammopathies, from MGUS to myeloma status. Leukemia 2012; 26: 2521-2529.

[61] Wei G, Zhou L, Lu LL and Romano M. Sequential EHR-based dynamic graph network for multiple myeloma detection and feature interaction investigation. J Clin Oncol 2022; 40: e13591.

[62] National Cancer Institute (NCI). Targeted therapy directed by genetic testing in treating patients with advanced refractory solid tumors, lymphomas, or multiple myeloma (The MATCH Screening Trial). NLM Identifier: NCT0246-5060; 2020.

[63] Gupta R, Kaur G, Kumar L, Rani L, Mathur N, Sharma A, Dahiya M, Shekhar V, Khan S, Mookerjee A and Sharma OD. Nucleic acid based risk assessment and staging for clinical practice in multiple myeloma. Ann Hematol 2018; 97: 2447-2454.

[64] Benard B, Christofferson A, Legendre C, Aldrich J, Nasser S, Yesil J, Auclair D, Liang W, Lonial S and Keats JJ. FGFR3 mutations are an adverse prognostic factor in patients with t(4;14)(p16;q32) multiple myeloma: an mmrf commpass analysis. Blood 2017; 130: 3027.

[65] Salama AKS, Li S, Macrae ER, Park JI, Mitchell EP, Zwiebel JA, Chen HX, Gray RJ, McShane LM, Rubinstein LV, Patton D, Williams PM, Hamilton SR, Armstrong DK, Conley BA, Arteaga CL, Harris LN, O'Dwyer PJ, Chen AP and Flaherty KT. Dabrafenib and trametinib in patients with tumors with BRAFV600E mutations: results of the NCI-MATCH trial subprotocol H. J Clin Oncol 2020; 38: 3895-3904.

[66] Misund K, Keane N, Stein CK, Asmann YW, Day G, Welsh S, Van Wier SA, Riggs DL, Ahmann G, Chesi M, Viswanatha DS, Kumar SK, Dispenzieri A, Gonzalez-Calle V, Kyle RA, O'Dwyer M, Rajkumar SV, Kortüm KM, Keats JJ; MMRF CoMMpass Network; Fonseca R, Stewart AK, Kuehl WM, Braggio E and Bergsagel PL. MYC dysregulation in the progression of multiple myeloma. Leukemia 2020; 34: 322-326.

[67] Garcia SB, Ruiz-Heredia Y, Da Via M, Gallardo M, Garitano-Trojaola A, Zovko J, Raab MS, Sonneveld P, Braggio E, Stewart AK and Einsele H. Role of MAX as a tumor suppressor driver gene in multiple myeloma. Blood 2017; 130: 4347.

[68] Ohmine K and Uchibori R. Novel immunotherapies in multiple myeloma. Int J Hematol 2022; 115: 799-810.

[69] Jasrotia S, Gupta R, Sharma A, Halder A and Kumar L. Cytokine profile in multiple myeloma. Cytokine 2020; 136; 155271.

[70] Hoteit R, Bazarbachi A, Antar A, Salem Z, Shammaa D and Mahfouz R. KIR genotype distribution among patients with multiple myeloma: higher prevalence of KIR 2DS4 and KIR 2DS5 genes. Meta Gene 2014; 2: 730-736.

[71] Sun H, Martin TG, Marra J, Kong D, Keats J, Macé S, Chiron M, Wolf JL, Venstrom JM and Rajalingam R. Individualized genetic makeup that controls natural killer cell function influences the efficacy of isatuximab immunotherapy in patients with multiple myeloma. J Immunother Cancer 2021; 9: e002958.

[72] Mahaweni NM, Ehlers FAI, Bos GMJ and Wieten L. Tuning natural killer cell anti-multiple myeloma reactivity by targeting inhibitory signaling via KIR and NKG2A. Front Immunol 2018; 9: 2848.

[73] White BS, Lanc I, O'Neal J, Gupta H, Fulton RS, Schmidt H, Fronick C, Belter EA Jr, Fiala M, King J, Ahmann GJ, DeRome M, Mardis ER, Vij R, DiPersio JF, Levy J, Auclair D and Tomasson MH. A multiple myeloma-specific capture sequencing platform discovers novel translocations and frequent, risk-associated point mutations in IGLL5. Blood Cancer J 2018; 8: 35.

[74] Lozano E, Díaz T, Mena MP, Suñe G, Calvo X, Calderón M, Pérez-Amill L, Rodríguez V, Pérez-Galán P, Roué G, Cibeira MT, Rosiñol L, Isola I, Rodríguez-Lobato LG, Martin-Antonio B, Bladé J and Fernández de Larrea C. Loss of the immune checkpoint CD85j/LILRB1 on malignant plasma cells contributes to immune escape in multiple myeloma. J Immunol 2018; 200: 2581-2591.

[75] Kang X, Kim J, Deng M, John S, Chen H, Wu G, Phan H and Zhang CC. Inhibitory leukocyte immunoglobulin-like receptors: immune checkpoint proteins and tumor sustaining factors. Cell Cycle 2016; 15: 25-40.

[76] Beksac M, Gragert L, Fingerson S, Maiers M, Zhang MJ, Albrecht M, Zhong X, Cozen W, Dispenzieri A, Lonial S and Hari P. HLA polymorphism and risk of multiple myeloma. Leukemia 2016; 30: 2260-2264.

[77] Kassem S, Diallo BK, El-Murr N, Carrié N, Tang A, Fournier A, Bonnevaux H, Nicolazzi C, Cuisinier M, Arnould I, Sidhu SS, Corre J, Avet-Loiseau H, Teillaud JL, van de Velde H, Wiederschain D, Chiron M, Martinet L and Virone-Oddos A. SAR442085, a novel anti-CD38 antibody with enhanced antitumor activity against multiple myeloma. Blood 2022; 139: 1160-1176.

[78] Chu D and Wei L. Nonsynonymous, synonymous and nonsense mutations in human cancer-related genes undergo stronger purifying selections than expectation. BMC Cancer 2019; 19: 359.

[79] Sharma Y, Miladi M, Dukare S, Boulay K, Caudron-Herger M, Groß M, Backofen R and Diederichs S. A pan-cancer analysis of synonymous mutations. Nat Commun 2019; 10: 2569.

[80] Soussi T, Taschner PE and Samuels Y. Synonymous somatic variants in human cancer are not infamous: a plea for full disclosure in databases and publications. Hum Mutat 2017; 38: 339-342.

[81] Teng H, Wei W, Li Q, Xue M, Shi X, Li X, Mao F and Sun Z. Prevalence and architecture of post transcriptionally impaired synonymous mutations in 8,320 genomes across 22 cancer types. Nucleic Acids Res 2020; 48: 1192-1205.

[82] Supek F, Miñana B, Valcárcel J, Gabaldón T and Lehner B. Synonymous mutations frequently act as driver mutations in human cancers. Cell 2014; 156: 1324-1335.