

A Comprehensive Targeted Sequencing Panel for Genomic Aberration Profiling using Directed Gene-Gene Interactions in Multiple Myeloma

Vivek Ruhela



INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI



Outline

- Background
- Challenges and Motivation
- Description of Global WES data repositories
- Infographic abstract of proposed Bio-DGI model
- AI-driven workflow for Genomic feature extraction of significantly altered genes
- Post-hoc analysis of Bio-DGI model for identification of top genomic attributes
 - Algorithm for estimation of best ShAP score for genomic attributes
 - Quantitative and qualitative benchmarking of proposed Bio-DGI model
 - Ranking of Genomic features
- Pathway Enrichment Analysis using Enrichr and ranking of significantly altered pathways
- Identification of Gene Communities and significantly contributing genes in communities
- Identification of other genomic variant profiles (CNVs, SVs, LOF, Haploinsufficiency)
- Geo2R validation of top-ranked significantly altered genes
- Workflow for identifying candidate driver gene panel
- Workflow for Survival Analysis
- Key features of 282 gene panel
- Key findings from post-hoc analysis of proposed Bio-DGI model
- Conclusions
- References

Background

- Multiple Myeloma (MM) and Monoclonal Gammopathy of Undetermined Significance (MGUS) are both plasma cell disorders, but they represent distinct stages along the continuum of disease progression.
- MM is characterized by malignant plasma cell proliferation, organ damage, and clinical symptoms, while MGUS is a precursor condition with no clinical manifestations. Discriminating between MM and MGUS is critical for appropriate patient management, as the prognosis and treatment strategies differ significantly.
- Clinical differentiation between MM and MGUS is challenging due to overlapping genomic and clinical features. Both conditions may present with monoclonal proteinemia and bone marrow plasma cell infiltration, making it essential to investigate genetic markers and gene-gene interactions to improve diagnostic accuracy [1].
- Further, the genomic complexity and instability in MM represents a unique model for studying gene-gene interactions due to its genetic heterogeneity and the role of aberrant plasma cell interactions with the bone marrow microenvironment [2].
- Advances in high-throughput genomic profiling techniques, such as whole exome sequencing (WES), whole-genome sequencing (WGS), single-cell sequencing etc., have allowed for comprehensive analysis of genetic alterations in MM and MGUS. These approaches enable the identification of gene-gene interactions that underlie the transition from MGUS to MM [3].

References:

1. Rajkumar, S. V. (2016). MGUS and Smoldering Multiple Myeloma: Update on Pathogenesis, Natural History, and Management. Hematology, American Society of Hematology Education Program, 2016(1), 426-433.
2. Anderson, K. C., & Carrasco, R. D. (2011). : Mechanisms of Disease, 6, 249-274.
3. Bolli, N., et al. (2014). Heterogeneity of genomic evolution aPathogenesis of myeloma. Annual Review of Pathologynd mutational profiles in multiple myeloma. Nature Communications, 5, 2997

Background

- Several genomic studies have identified several key genetic alterations associated with MM, including chromosomal translocations, mutations in tumor suppressor genes (e.g., TP53), and dysregulation of oncogenes (e.g., MYC) [4]. These genetic alterations often act in concert through gene-gene interactions to drive disease progression.
- The advent of next-generation sequencing (NGS) technology has extended the cancer biology knowledge through detection of genomic abnormalities such as deleterious mutations, chromosomal translocations and copy number abnormalities (CNA) [5].
- In the quest to differentiate MM from MGUS, identifying driver mutations that initiate and sustain malignant plasma cell proliferation is crucial. Mutations in genes like BRAF and DIS3 have been associated with MM progression, while MGUS is characterized by fewer and less impactful mutations [6].
- Several targeted sequencing panels have been designed and proposed for comprehensive profiling of genomic aberrations in MM including SNVs, CNVs, and SVs [6-8]. However, none of the previous recommended panels have prioritized genes based on gene-gene interactions and considered the most influencing genes that can also act as a differentiating biomarker in between MM and MGUS.
- We are motivated to design a targeted sequencing panel for comprehensive profiling of genomic aberrations to capture uniquely defined characteristic of MM and MGUS for identifying MGUS samples likely to progress MM and high risk MM samples.

References:

- 4.Walker, B. A., & Morgan, G. J. (2012). Genetic and genomic insights into myeloma. *FEBS Letters*, 586(17), 2766-2773.
- 5.Chapman, M. A., et al. (2011). Initial genome sequencing and analysis of multiple myeloma. *Nature*, 471(7339), 467-472.
- 6.Yellapantula, V., Hultcrantz, M., Rustad, E.H., Wasserman, E., Londono, D., Cimera, R., Ciardiello, A., Landau, H., Akhlaghi, T., Mailankody, S. and Patel, M., 2019. Comprehensive detection of recurring genomic abnormalities: a targeted sequencing approach for multiple myeloma. *Blood Cancer Journal*, 9(12), p.101.
- 7.Cutler, S.D., Knopf, P., Campbell, C.J., Thoni, A., Abou El Hassan, M., Forward, N., White, D., Wagner, J., Goudie, M., Boudreau, J.E. and Kennedy, B.E., 2021. DMG26: A Targeted Sequencing Panel for Mutation Profiling to Address Gaps in the Prognostication of Multiple Myeloma. *The Journal of Molecular Diagnostics*, 23(12), pp.1699-1714.
- 8.Sudha, P., Ahsan, A., Ashby, C., Kausar, T., Khera, A., Kazeroun, M.H., Hsu, C.C., Wang, L., Fitzsimons, E., Salminen, O. and Blaney, P., 2022. Myeloma Genome Project Panel is a comprehensive targeted genomics panel for molecular profiling of patients with multiple myeloma. *Clinical Cancer Research*, 28(13), pp.2854-2864.

Challenges and Motivation

Challenges:

- MGUS patients share many genetic aberrations of MM without showing any overt clinical symptoms of MM that makes it challenging to differentiate MGUS from MM.
- The early diagnosis of MM and the identification of relevant differentiating genetic and genomic biomarkers between MGUS and MM present several challenges at the genomic-level and the subject-level [9].
- Genomic risk stratification in the asymptomatic disease states of MM (MGUS) requires to include the genomic abnormalities that can define all risk-group stages of patients and need to include recurrently mutated genes. The identification of such recurrently mutated genes is challenging due to similar genomic complexity of MM and MGUS [8].
- Genomic data encompassing single nucleotide variations (SNVs), copy number variations (CNVs), structural variations (SVs), and loss-of-function (LOF) events can be vast and high-dimensional. Analyzing such data efficiently and extracting meaningful insights is computationally intensive.
- MM and MGUS exhibit substantial heterogeneity in terms of clinical, genetic, and molecular characteristics. Deciphering the intricate interplay of various genetic alterations and their diverse effects on disease progression poses a significant challenge.

References:

9. Ruhela, V., Jena, L., Kaur, G., Gupta, R. and Gupta, A., 2023. BDL-SP: A Bio-inspired DL model for the identification of altered Signaling Pathways in Multiple Myeloma using WES data. American Journal of Cancer Research, 13(4), p.1155.

Challenges and Motivation

Motivations:

- AI-based techniques have emerged as valuable tools for integrating diverse genomic data and identifying key gene-gene interactions. These approaches have the potential to uncover complex networks that differentiate MM from MGUS [10].
- In recent years, the deep learning-based methods, particularly Graph Convolutional Networks (GCNs), have gained prominence in the analysis of biological networks, including Protein-Protein Interaction (PPI) networks. GCNs are adept at capturing intricate relationships and interactions within these networks, making them an ideal tool for studying gene-gene interactions.
- Identifying key differentiating gene-gene interactions between MM and MGUS is paramount for early diagnosis, prognosis, and the development of targeted therapies. A model that achieves this can significantly impact patient outcomes and healthcare strategies.
- A robust targeted genomics panel can guide the design and optimization of clinical trials by identifying potential targets for drug development and predicting treatment response based on patients' genomic profiles.

References:

10. Tovar, N., et al. (2018). Integrated Single-Cell Transcriptomics and Chromatin Accessibility Analysis Reveals Regulators of Differentiation in Distinct B-Cell Subpopulations. *Bioinformatics*, 34(24), 4142-4150.

Description of Global WES data repositories

- We have considered three datasets in this work (shown Table-1) that consists of whole exome sequencing (WES) samples from three different ethnicities.
- The overall datasets contains 1154 MM samples and 61 MGUS samples that makes it an imbalanced datasets (95% MM and 5% MGUS cases).

Table-1: Three global WES repositories were used in this work for identification of pivotal biomarkers to differentiate MM and MGUS.

| Data Repository | # of MM/# MGUS Patients | Ethnicity | Sources | Available data type |
|---------------------------|-------------------------|----------------------|------------------------------------|---------------------|
| phs0000748 | MM patients: 1092 | American Population | MMRF [11] | Processed VCF files |
| EGAD00001001901 | MGUS: 33 | Caucasian Population | European Genome Archive (EGA) [12] | Processed BAM files |
| PRJNA685283, PRJNA 694218 | MM: 82, MGUS: 28 | Asian Population | AIIMS [9] | Raw FASTQ files |

- We have also used the clinical information of MM samples of MMRF and AIIMS database. The clinical information for MMRF and AIIMS database is obtained from MMRF CoMMpass [11] and Laboratory Oncology, Dr. Ritu Gupta's Lab, AIIMS [9].

References:

11. Keats JJ, Craig DW, Liang W, Venkata Y, Kurdo- glu A, Aldrich J, Auclair D, Allen K, Harrison B, Jewell S, Kidd PG, Correll M, Jagannath S, Sie- gel DS, Vij R, Orloff G, Zimmerman TM; Mmrfr 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 genom- ic and immunophenotypic profiles. Blood 2013; 122: 532.
12. 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.

Infographic abstract of proposed Bio-DGI Model

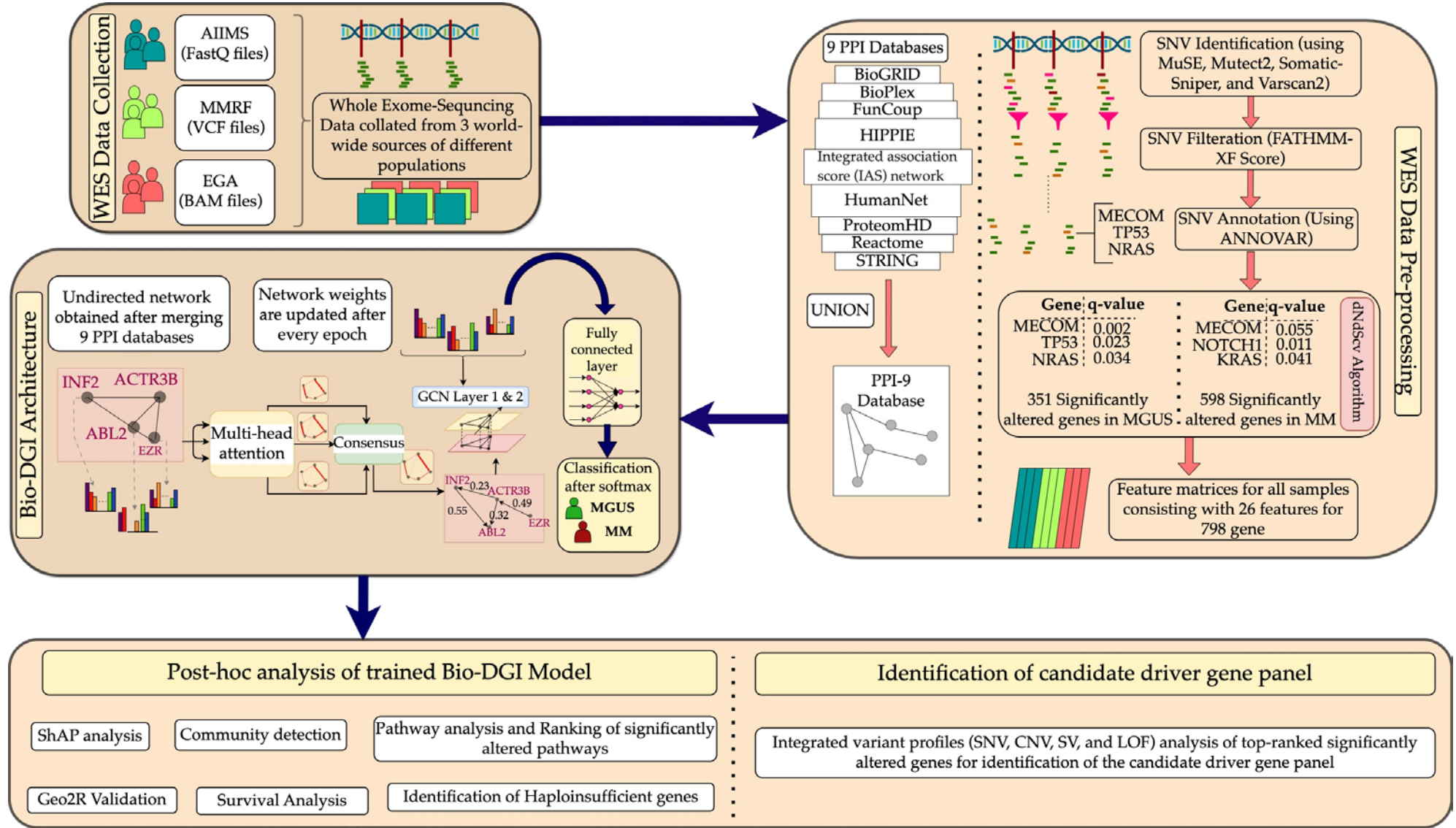


Figure-1: Infographic representation of proposed Bio-DGI model architecture and post-hoc analysis for the identification of pivotal genomic biomarkers, gene communities and key gene-gene interaction that can distinguish MGUS from MM.

Algorithm for estimation of best ShAP score for genomic attributes (genes)

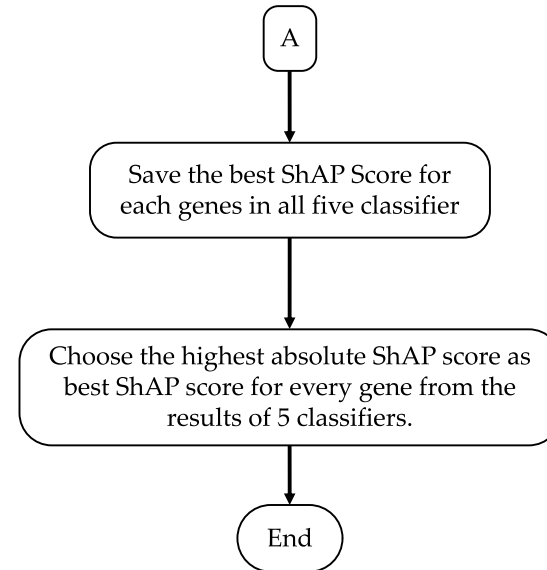
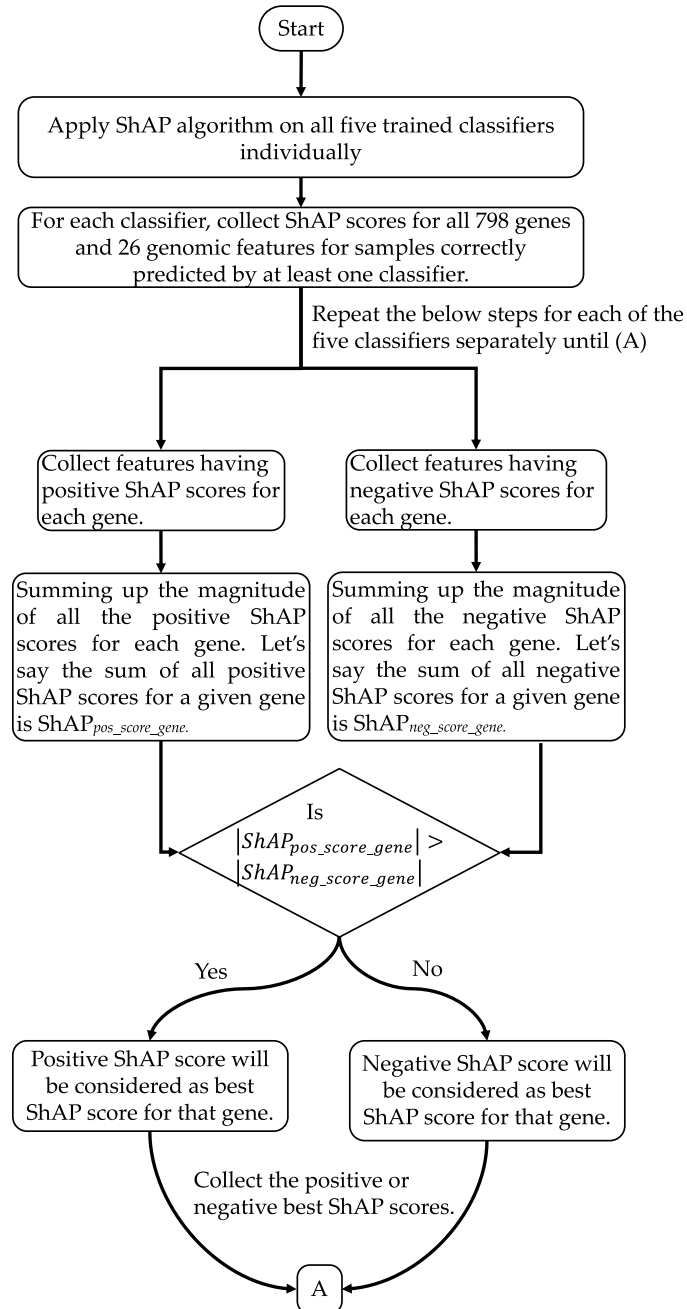


Figure-2 (A): Algorithm for the estimation of the best ShAP score of significantly altered genes at a sample level (adopted from our previous related study [6]).

Algorithm for estimation of best ShAP score for genomic attributes (genomic feature)

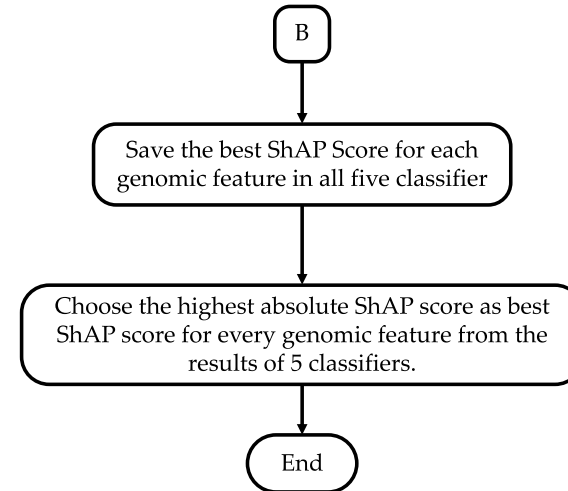
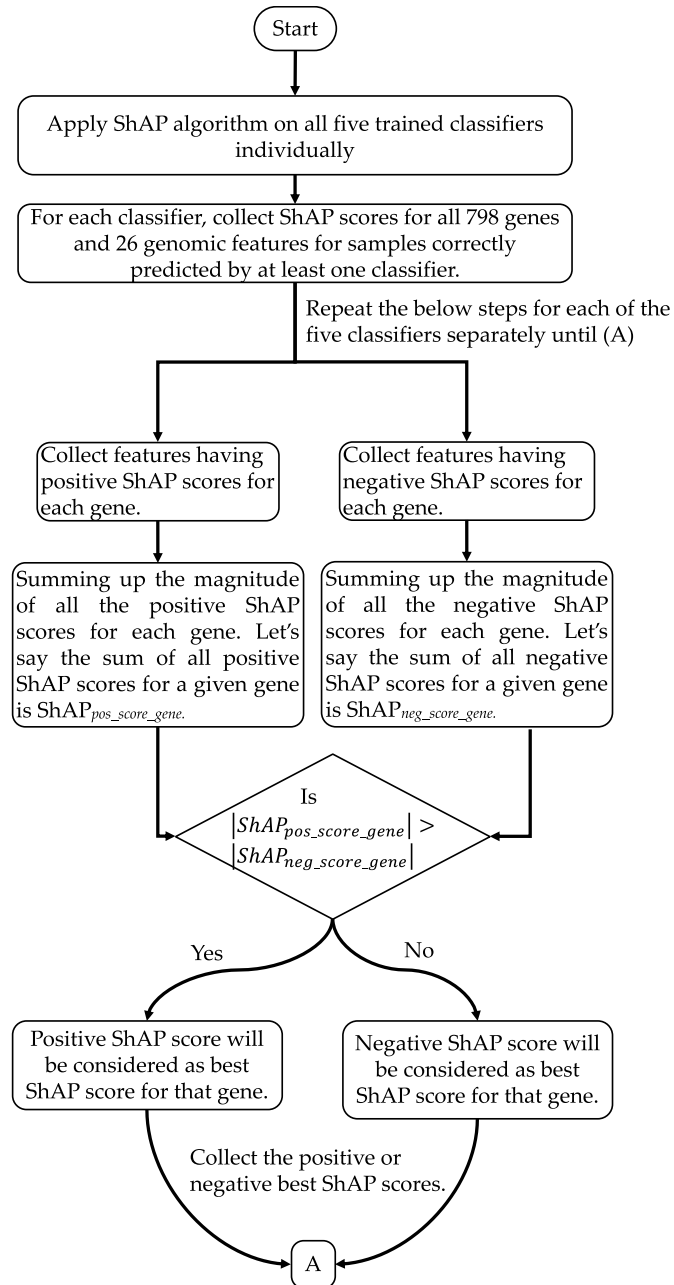
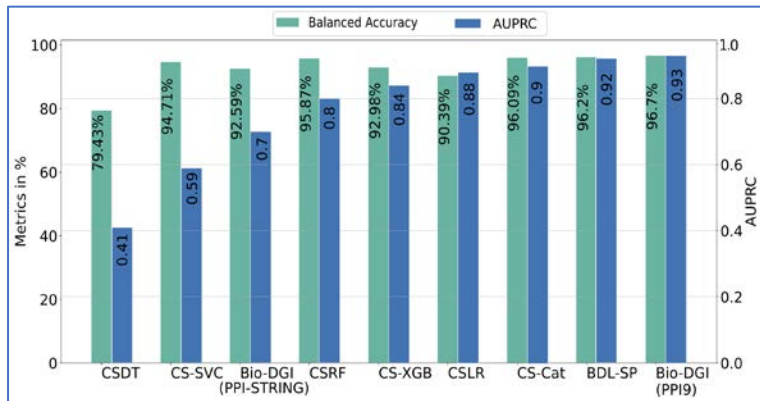


Figure-2 (B): Algorithm for the estimation of best ShAP score of genomic features at a sample level (adopted from our previous related study [9]).

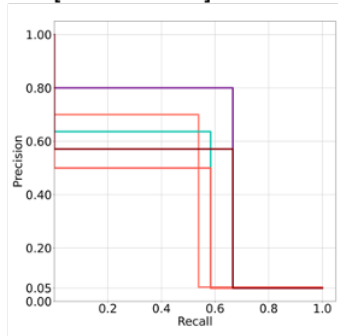
Quantitative benchmarking of Bio-DGI model

(A)

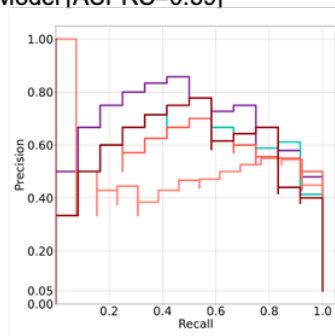


(B)

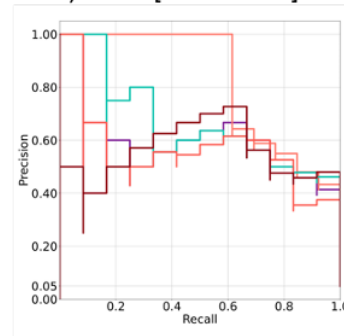
(i) PRC Curves for CSDT Model [AUPRC=0.41]



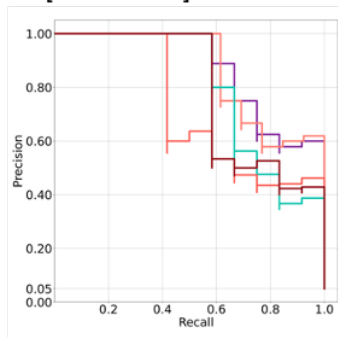
(ii) PRC Curves for CS-SVC Model [AUPRC=0.59]



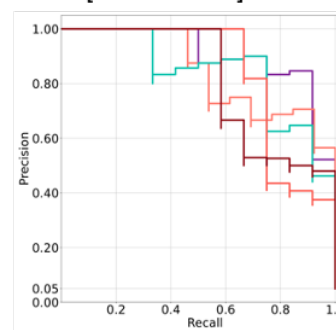
(iii) PRC Curves for Bio-DGI (PPI-STRING) Model [AUPRC: 0.7]



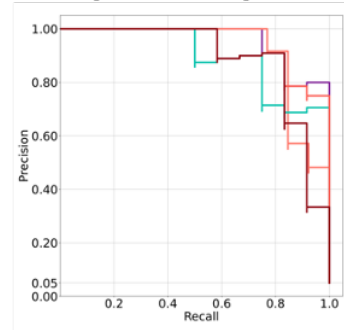
(iv) PRC Curves for CSRF Model [AUPRC: 0.8]



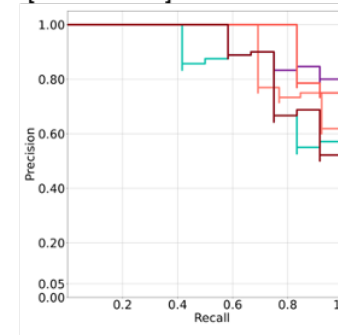
(v) PRC Curves for CS-XGBoost Model [AUPRC: 0.84]



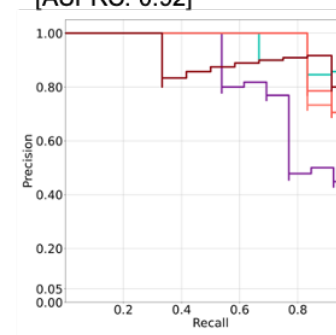
(vi) PRC Curves for CSLR Model [AUPRC: 0.88]



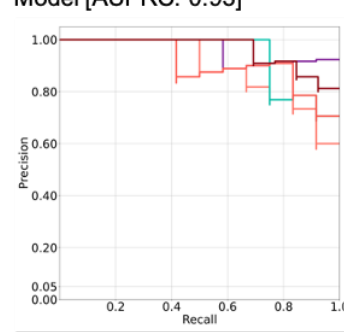
(vii) PRC Curves for CS-Cat Model [AUPRC: 0.9]



(viii) PRC Curves for BDL-SP Model [AUPRC: 0.92]



(ix) PRC Curves for Bio-DGI (PPI9) Model [AUPRC: 0.93]



Legend: Fold-1 (Purple), Fold-2 (Orange), Fold-3 (Teal), Fold-4 (Red), Fold-5 (Dark Red)

Table-2: Confusion matrix of top-3 performing models

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

Figure-3: (A) Benchmarking of proposed Bio-DGI model with baseline ML and DL models, (B) Precision-Recall curves (PRC) for all five folds of (i) CSDT, (ii) CS-SVC, (iii) Bio-DGI (PPI-STRING), (iv) CS-RF, (v) CS-XGB, (vi) CSLR, (vii) CS-Cat, (viii) BDL-SP, and (ix) Bio-DGI (PPI9).

Qualitative benchmarking of proposed Bio-DGI model

Post-hoc Analysis of top-performing models

Table-2: Types of four different gene categories (OG, TSG, ODG, and AG) and their counts in 798 significantly altered genes

| Gene Type | Total number of previously reported genes present in the list of 798 significantly altered genes |
|---------------------------------------|--|
| Oncogenes (OG) | 31 |
| Tumor Suppressor genes (TSG) | 43 |
| Both Oncogenes and driver genes (ODG) | 10 |
| Actionable genes (AG) | 19 |

Table-3: Counts of four gene categories in top-250 and top-500 genes obtained from the post-hoc analysis of top performing models (CS-Cat, Bio-DGI (PPI STRING), BDL-SP, and Bio-DGI (PPI9)).

| Genes | CS-Cat Model | | | | Bio-DGI (PPI-STRING) Model | | | | BDL-SP Model | | | | Bio-DGI (PPI9) Model | | | |
|---------|--------------|-----|-----|----|----------------------------|-----|-----|----|--------------|-----|-----|----|----------------------|-----------|----------|-----------|
| | OG | TSG | ODG | AG | OG | TSG | ODG | AG | OG | TSG | ODG | AG | OG | TSG | ODG | AG |
| Top 250 | 0 | 0 | 0 | 0 | 18 | 24 | 7 | 13 | 20 | 21 | 7 | 11 | 23 | 26 | 8 | 14 |
| Top 500 | 0 | 0 | 0 | 0 | 28 | 41 | 9 | 19 | 27 | 37 | 8 | 17 | 28 | 41 | 9 | 19 |

Qualitative benchmarking of Bio-DGI model

Table-4(A): Oncogenes (OG) and actionable genes (AG) reported by Bio-DGI (PPI STRING), BDL-SP and Bio-DGI (PPI9) model.

| Genes | Bio-DGI (PPI STRING) Model | | BDL-SP Model | | Bio-DGI (PPI9) Model | |
|---------|--|--|---|---|--|--|
| | OG | AG | OG | AG | OG | AG |
| Top 250 | BIRC6, BRAF, BRD4, FGFR3, IRS1, KMT2D, KRAS, MUC16, MUC4, NOTCH1, NRAS, PABPC1, RPTOR, SETD1A, TERT, TP53, TRRAP, USP6 | ARID1B, ARID2, BRAF, BRD4, FANCD2, FGFR3, KRAS, NF1, NOTCH1, NRAS, RPTOR, TP53, TYRO3 | ABL2, BIRC6, BRAF, CARD11, FGFR3, IRS1, KRAS, LTB, MGAM, MITF, MUC16, MUC4, NOTCH1, NRAS, PABPC1, PGR, RPTOR, TCL1A, TP53, VAV1 | ARID2, BRAF, FGFR3, KRAS, MITF, NF1, NOTCH1, NRAS, RPTOR, TP53, TYRO3 | ABL2, BIRC6, BRAF, BRD4, CARD11, FGFR3, FUBP1, IRS1, KMT2D, KRAS, MUC16, MUC4, NOTCH1, NRAS, PABPC1, PGR, RPTOR, SETD1A, TERT, TP53, TRRAP, USP6, VAV1 | ARID1B, ARID2, BRAF, BRD4, FANCD2, FGFR3, KRAS, NF1, NFKBIA, NOTCH1, NRAS, RPTOR, TP53, TYRO3 |
| Top 500 | ABL2, BIRC6, BRAF, BRD4, CARD11, FGFR3, FUBP1, IRS1, KMT2D, KRAS, LTB, MECOM, MGAM, MITF, MUC16, MUC4, NOTCH1, NRAS, PABPC1, PGR, RPTOR, SETD1A, TAL1, TERT, TP53, TRRAP, USP6, VAV1 | APC, ARID1B, ARID2, BRAF, BRD4, FANCD2, FGFR3, KRAS, MITF, NF1, NFKBIA, NOTCH1, NRAS, RAD54B, RB1, RPTOR, RTEL1, TP53, TYRO3 | ABL2, BIRC6, BRAF, BRD4, CARD11, FGFR3, IRS1, KMT2D, KRAS, LTB, MACC1, MECOM, MGAM, MITF, MUC16, MUC4, NOTCH1, NRAS, PABPC1, PGR, RPTOR, TAL1, TCL1A, TERT, TP53, TRRAP, VAV1 | APC, ARID1B, ARID2, BRAF, BRD4, FANCD2, FGFR3, KRAS, MITF, NF1, NFKBIA, NOTCH1, NRAS, RB1, RPTOR, TP53, TYRO3 | ABL2, BIRC6, BRAF, BRD4, CARD11, FGFR3, FUBP1, IRS1, KMT2D, KRAS, LTB, MECOM, MGAM, MITF, MUC16, MUC4, NOTCH1, NRAS, PABPC1, PGR, RPTOR, SETD1A, TAL1, TERT, TP53, TRRAP, USP6, VAV1 | APC, ARID1B, ARID2, BRAF, BRD4, FANCD2, FGFR3, KRAS, MITF, NF1, NFKBIA, NOTCH1, NRAS, RAD54B, RB1, RPTOR, RTEL1, TP53, TYRO3 |

Qualitative benchmarking of Bio-DGI model

Table-4(B): Tumor suppressor genes (TSG) and Both oncogenes and driver genes (ODG) reported by Bio-DGI (PPI STRING), BDL-SP and Bio-DGI (PPI9) model.

| Genes | Bio-DGI (PPI STRING) Model | | BDL-SP Model | | Bio-DGI (PPI9) Model | |
|---------|--|--|--|---|--|--|
| | TSG | ODG | TSG | ODG | TSG | ODG |
| Top 250 | ARID1B, ARID2, DIS3, EGR1, EP400, FANCD2, HLA-A, HLA-B, HLA-C, KMT2B, KMT2C, KMT2D, MAX, MYH11, NCOR1, NF1, NOTCH1, PABPC1, RPL10, SDHA, SIRPA, TERT, TP53, TRAF3 | BRAF, FGFR3, KRAS, NRAS, PABPC1, TP53, TRRAP | ARID2, ATP2B3, CYLD, DIS3, EGR1, HLA-A, HLA-B, HLA-C, IRF1, KMT2C, LTB, MITE, NF1, NOTCH1, PABPC1, SAMHD1, SDHA, SIRPA, SP140, TP53, TRAF3 | BRAF, FGFR3, KRAS, LTB, NRAS, PABPC1, TP53 | ARID1B, ARID2, DIS3, EGR1, EP400, FANCD2, FUBP1, HLA-A, HLA-B, HLA-C, IRF1, KMT2B, KMT2C, KMT2D, MAX, MYH11, NCOR1, NF1, NFKBIA, NOTCH1, PABPC1, SDHA, SIRPA, TERT, TP53, TRAF3 | BRAF, FGFR3, FUBP1, KRAS, NRAS, PABPC1, TP53, TRRAP |
| Top 500 | ACVR1B, AMER1, APC, ARID1B, ARID2, ATP2B3, CYLD, DDX41, DIS3, EGR1, EP400, FANCD2, FUBP1, HLA-A, HLA-B, HLA-C, IRF1, KMT2B, KMT2C, KMT2D, LTB, MAX, MITE, MYH11, NCOR1, NF1, NFKBIA, NOTCH1, PABPC1, RB1, RPL10, RTEL1, SAMHD1, SDHA, SIRPA, SP140, TERT, TP53, TRAF3, WNK2, ZFHX3 | BRAF, FGFR3, FUBP1, KRAS, LTB, NRAS, PABPC1, TP53, TRRAP | ACVR1B, AMER1, APC, ARID1B, ARID2, ATP2B3, CMTR2, CYLD, DIS3, EGR1, FANCD2, HLA-A, HLA-B, HLA-C, IRF1, KMT2B, KMT2C, KMT2D, LTB, MAX, MITE, MYH11, NCOR1, NF1, NFKBIA, NOTCH1, PABPC1, RB1, SAMHD1, SDHA, SIRPA, SP140, TERT, TP53, TRAF3, WNK2, ZFHX3 | BRAF, FGFR3, KRAS, LTB, NRAS, PABPC1, TP53, TRRAP | ACVR1B, AMER1, APC, ARID1B, ARID2, ATP2B3, CYLD, DDX41, DIS3, EGR1, EP400, FANCD2, FUBP1, HLA-A, HLA-B, HLA-C, IRF1, KMT2B, KMT2C, KMT2D, LTB, MAX, MITE, MYH11, NCOR1, NF1, NFKBIA, NOTCH1, PABPC1, RB1, RPL10, RTEL1, SAMHD1, SDHA, SIRPA, SP140, TERT, TP53, TRAF3, WNK2, ZFHX3 | BRAF, FGFR3, FUBP1, KRAS, LTB, NRAS, PABPC1, TP53, TRRAP |

Genomic Feature Ranking from Bio-DGI (PPI9) post-hoc analysis

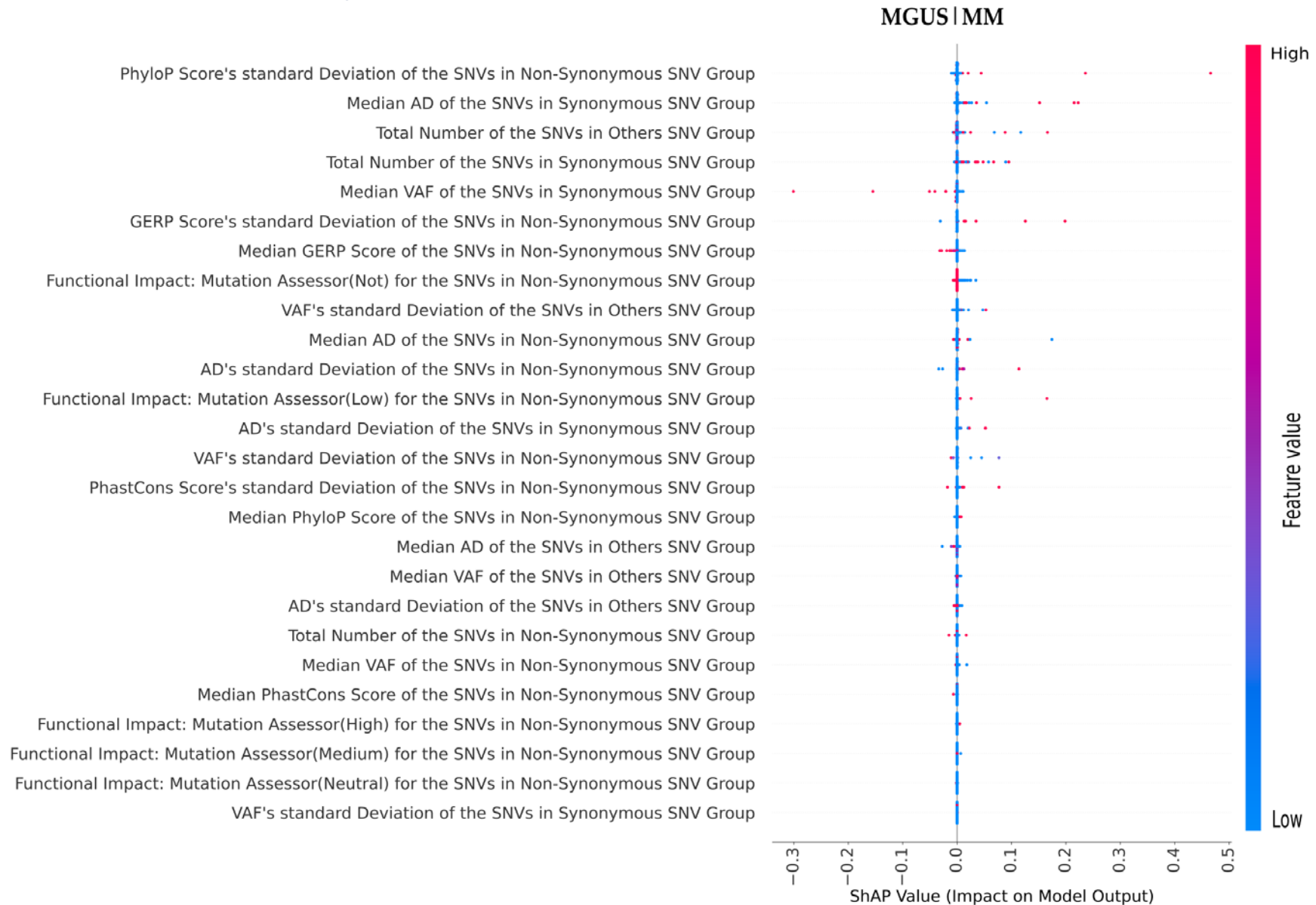


Figure-4: Genomic Feature ranking based on the Bio-DGI (PPI9) post-hoc explainability in MM and MGUS using ShAP algorithm

Pathway Enrichment Analysis using Enrichr and ranking of significantly altered pathways

- We performed the pathway enrichment analysis using Enrichr pathway database [13-15] for identifying the significantly altered pathways associated with top-ranked exclusively and significantly altered genes in MM and MGUS (obtained from post-hoc analysis of Bio-DGI (PPI9) model).
- In MGUS, a total of 7 KEGG and 10 Reactome pathways were found significantly altered. On the other hand, in MM, a total of 105 KEGG and 84 Reactome pathways were found as significantly altered.
- We categorized the significantly altered pathways into 4 categories based on their significance with disease progression from MGUS to MM shown in Table-5.

Table-5: Categories of significantly altered pathways based on their significance with disease progression and number of pathways present in each category

| S. No. | Pathway Category | Number of significantly altered KEGG Pathways | Number of significantly altered Reactome Pathways |
|--------|---|---|---|
| 1 | Pathways observed significantly altered only in MM and not in MGUS | 98 | 75 |
| 2 | Pathways that become more significant with disease progression from MGUS to MM | 7 | 8 |
| 3 | Pathways observed significantly altered only in MGUS and not in MM | 0 | 1 |
| 4 | Pathways that become less significant with disease progression from MGUS to MM. | 0 | 1 |

References:

13. M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, et al., Enrichr: a comprehensive gene set enrichment analysis web server 2016 update, *Nucleic acids research* 44 (W1) (2016) W90–W97.
14. Z. Xie, A. Bailey, M. V. Kuleshov, D. J. Clarke, J. E. Evangelista, S. L. Jenkins, A. Lachmann, M. L. Wojciechowicz, E. Kropiwnicki, K. M. Jagodnik, et al., Gene set knowledge discovery with enrichr, *Current protocols* 1 (3) (2021) e90.
15. E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. V. Meirelles, N. R. Clark, A. Ma'ayan, Enrichr: interactive and collaborative html5 gene list enrichment analysis tool, *BMC bioinformatics* 14 (1) (2013) 1–14.

Pathway Enrichment Analysis using Enrichr and ranking of significantly altered pathways

- We also ranked KEGG and Reactome pathways based on their adjusted p-value (Figure-5 in the next slide). The pathways related to cancer such as Antigen processing, Calcium signaling pathway, Adaptive immune system, Human T-cell leukemia virus 1 infection were among the top ranked significantly altered pathways.
- We have observed that several MM related pathways such as NF-kB signaling pathway, PI3K signaling pathway, Calcium Signaling pathway, immune system pathway were among top-ranked pathways.

Identification of Gene Communities and significantly contributing genes in communities

Table-6: Steps for gene community identification using learned adjacency matrices obtained from Bio-DGI (PPI9) model

| | | | |
|--------|---|---|--|
| Step-1 | Fetch learned adjacency matrices for each fold | Fetch the learned adjacency matrix from each of the five Bio-DGI (PPI9) classifiers. | All five learned adjacency matrices are of dimension 798x798. |
| Step-2 | Community detection for each fold | <ul style="list-style-type: none"> For each of the five learned adjacency matrix, apply Leiden algorithm (LA) [16] for community detection. Within each fold, rank the communities based on the number of previously reported genes (OG, TSG, ODG, and AG) they contain. | In each of the five-fold learned adjacency matrices, we identified 5 communities in the first fold, 5 communities in the second, 6 communities in the third, 5 communities in the fourth, and 6 communities in the fifth fold learned adjacency matrix. |
| Step-3 | Gene selection from top-communities for each fold | <ul style="list-style-type: none"> For each fold, select the top three communities based on the number of previously reported genes (OG, TSG, ODG, and AG) present in the community. For each fold, generate a new learned adjacency matrix by taking union of the genes present in top three community. | The union of genes derived from the top three communities in each of the fold-1, fold-2, fold-3, fold-4, and fold-5 learned adjacency matrices resulted in gene sets comprising 500, 500, 539, 500, and 422 genes, respectively. Consequently, using the same gene sets in each of the five folds, we generated new learned adjacency matrices with dimensions of 500x500, 500x500, 539x539, 500x500, and 422x422, respectively. |
| Step-4 | Generation of a consensus of learned adjacency matrix | <ul style="list-style-type: none"> Merge all five of the new adjacency matrices obtained in the previous step (achieved by combining genes from the top three communities in each fold) by computing the mean of gene-gene interaction weights across these learned adjacency matrices. In case, if a particular gene-gene interaction is absent in any fold, assume a weight of zero for the corresponding interaction in that fold. | Upon merging the five new learned adjacency matrices, we obtained a consolidated adjacency matrix with dimensions of 690x690. |
| Step-5 | Community detection on merged adjacency matrix | On the merged adjacency matrix obtained from Step-4, apply LA for community detection using the new gene-gene interaction weights. | A total of five communities were identified with community size 202, 125, 122, 104, and 21. |

References: 16. Traag, V.A., Waltman, L. and Van Eck, N.J., 2019. From Louvain to Leiden: guaranteeing well-connected communities. Scientific reports, 9(1), p.5233.

Details of communities identified using the learned adjacency matrix

Table-7: Number of previously reported genes (OG, TSG, ODG, and AG) in the identified gene communities

| Community Number | 1 | 2 | 3 | 4 | 5 |
|---------------------|---|---|-------|---|-------------|
| Number of Oncogenes | 11 | 14 | 0 | 4 | 2 |
| Oncogenes Names | BIRC6, BRD4, FUBP1, KMT2D, MECOM, PABPC1, SETD1A, TAL1, TERT, TRRAP, USP6 | ABL2, BRAF, FGFR3, IRS1, KRAS, MACC1, MAML2, MGAM, MITE, NOTCH1, NRAS, PGR, RPTOR, TP53 | - | CARD11, LTB, TCL1A, VAV1 | MUC16, MUC4 |
| Number of TSGs | 21 | 8 | 1 | 11 | 0 |
| TSG Names | AMER1, ARID1B, ARID2, CMTR2, DDX41, DIS3, EP400, FANCD2, FUBP1, KMT2B, KMT2C, KMT2D, MAX, NCOR1, NF1, PABPC1, RB1, RPL10, RTEL1, TERT, ZFH3 | ACVR1B, APC, BTG1, EGR1, MITE, NOTCH1, SDHA, TP53 | MYH11 | CYLD, HLA-A, HLA-B, HLA-C, IRF1, LTB, NFKBIA, SAMHD1, SIRPA, SP140, TRAF3 | - |
| Number of ODGs | 3 | 6 | 0 | 1 | 0 |
| ODG Names | FUBP1, PABPC1, TRRAP | BRAF, FGFR3, KRAS, MAML2, NRAS, TP53 | | LTB | |
| Number of AG | 8 | 10 | 0 | 1 | 0 |
| AG Names | ARID1B, ARID2, BRD4, FANCD2, NF1, RAD54B, RB1, RTEL1 | APC, BRAF, FGFR3, KRAS, MITE, NOTCH1, NRAS, RPTOR, TP53, TYRO3 | - | NFKBIA | - |

Color codes for community visualization

Table-8: Color coding for gene community visualization

| S. No. | Gene Category | Color Notation |
|--------|---|----------------|
| 1 | Actionable Genes (AG) | Blue |
| 2 | Tumor Suppressor Genes (TSG) | Cyan |
| 3 | Oncogenes (OG) | Red |
| 4 | Both Oncogenes and Driver Genes (ODG) | Magenta |
| 5 | Other Genes (Not reported in any gene categories) | Gray |

Criteria Used for Nodes Ranking:

- We employed Katz Centrality scores [17] to assess the influence of nodes within the community, considering both their global and local impact.
- In the community context, a larger the node size signifies a greater degree of influence within that community.

References:

17. Katz, L., 1953. A new status index derived from sociometric analysis. *Psychometrika*, 18(1), pp.39-43.

Top Genes in 1st gene community

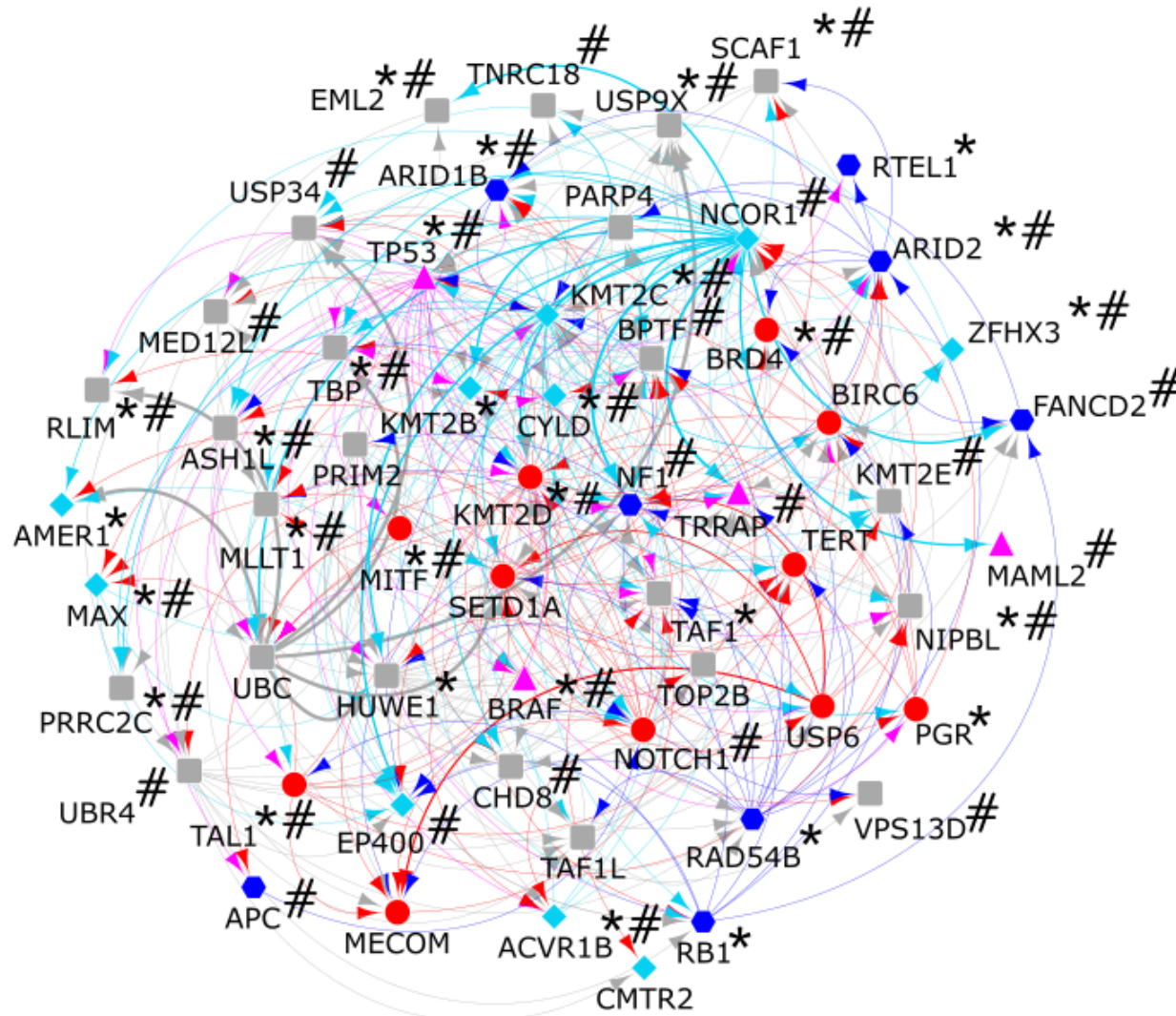


Figure-6: First gene community displayed with top-250 genes present within the community. This figure showcase the previously reported genes (OG, TSG, ODG, AG) regardless of their rank, alongside other non-reported genes (in magenta color) within the top-250 ranks respectively. Genes marked with "*" are included in the 282-genes panel. Additionally, genes marked with "#" possess a high likelihood of being haploinsufficient, with a GHIS score > 0.52. Lastly, "O" and "T" symbols denote genes involved in oncogenic and transformative genomic events, respectively.

Colour Notations:

AG – Blue hexagon (●)

OG – Red circle (●)

TSG – Cyan diamond (◆)

ODG – Magenta triangle (▲)

Genes that are part of communities and not OG/TSG/ODG/AG – Gray square (■)

Top Genes in 2nd gene community

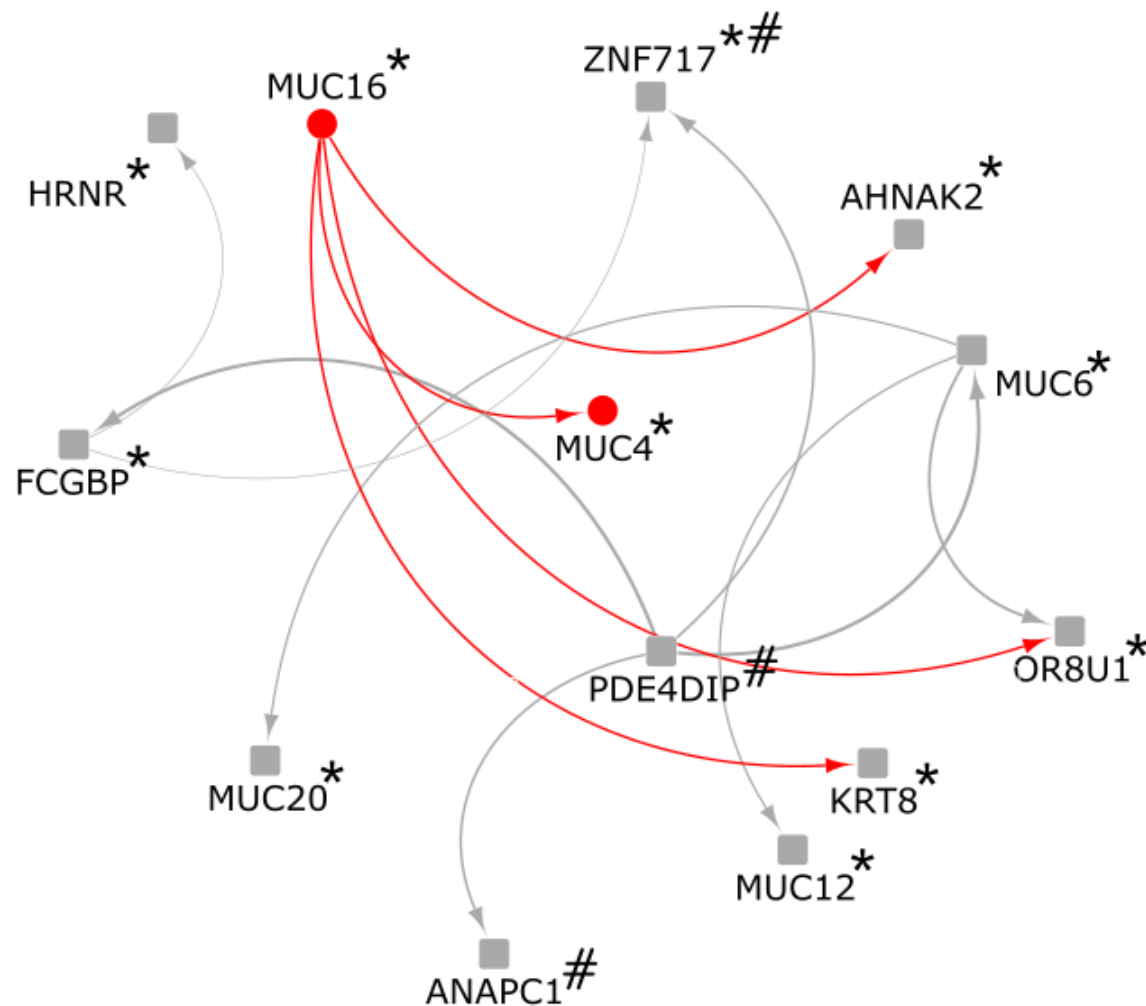


Figure-7: Second gene community displayed with top-250 genes present within the community. This figure showcase the previously reported genes (OG, TSG, ODG, AG) regardless of their rank, alongside other non-reported genes (in magenta color) within the top-250 ranks respectively. Genes marked with "*" are included in the 282-genes panel. Additionally, genes marked with "#" possess a high likelihood of being haploinsufficient, with a GHIS score > 0.52. Lastly, "O" and "T" symbols denote genes involved in oncogenic and transformative genomic events, respectively.

Colour Notations:

AG – Blue hexagon (■)

OG – Red circle (●)

TSG – Cyan diamond (◆)

ODG – Magenta triangle (▲)

Genes that are part of communities and not OG/TSG/ODG/AG – Gray square (■)

Top Genes in 3rd gene community

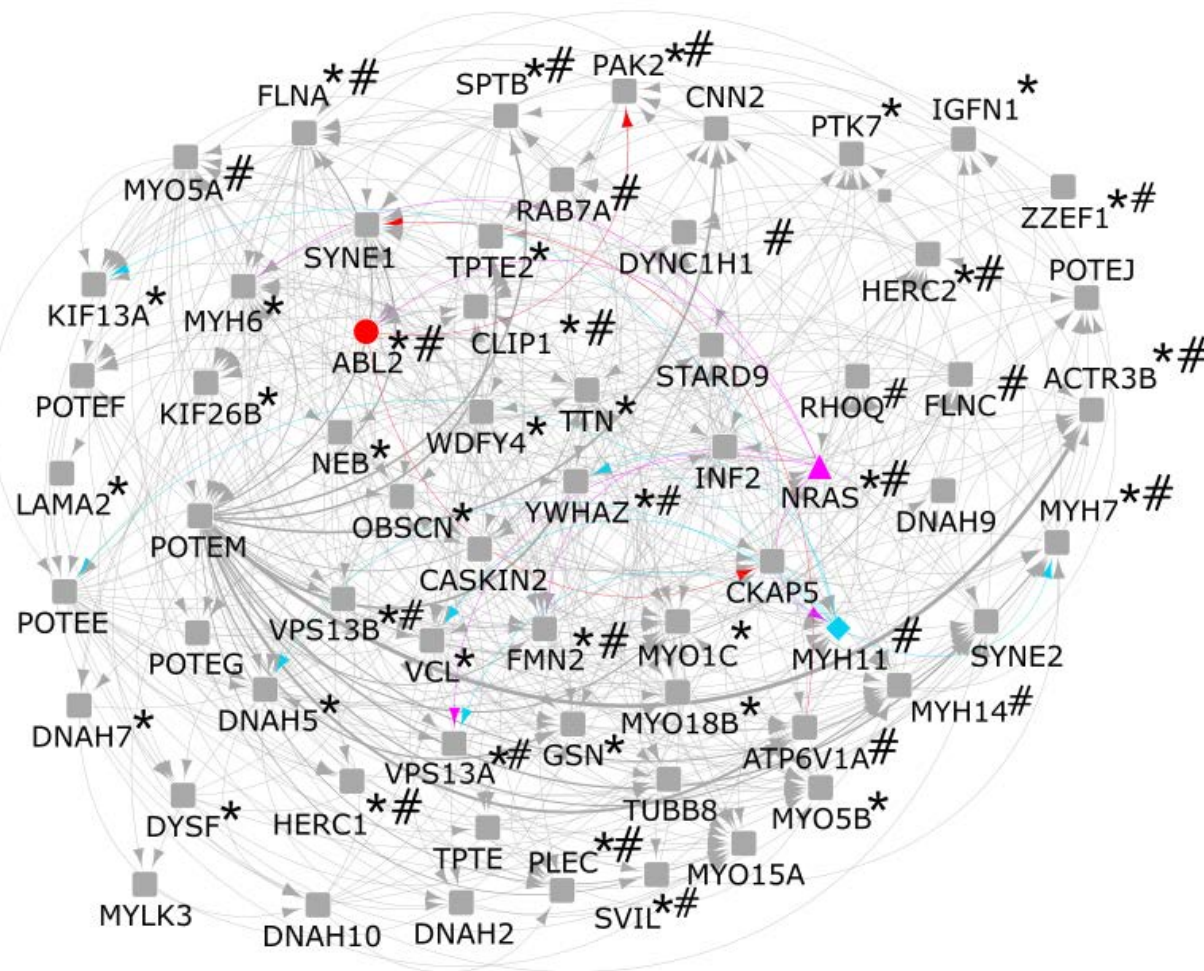


Figure-8: Third gene community displayed with top-250 genes present within the community. This figure showcase the previously reported genes (OG, TSG, ODG, AG) regardless of their rank, alongside other non-reported genes (in magenta color) within the top-250 ranks respectively. Genes marked with "*" are included in the 282-genes panel. Additionally, genes marked with "#" possess a high likelihood of being haploinsufficient, with a GHIS score > 0.52. Lastly, "O" and "T" symbols denote genes involved in oncogenic and transformative genomic events, respectively.

Colour Notations:

AG – Blue hexagon (■)

OG – Red circle (●)

TSG – Cyan diamond (◆)

ODG – Magenta triangle (▲)

Genes that are part of communities and not OG/TSG/ODG/AG – Gray square (■)

Top Genes in 4th gene community

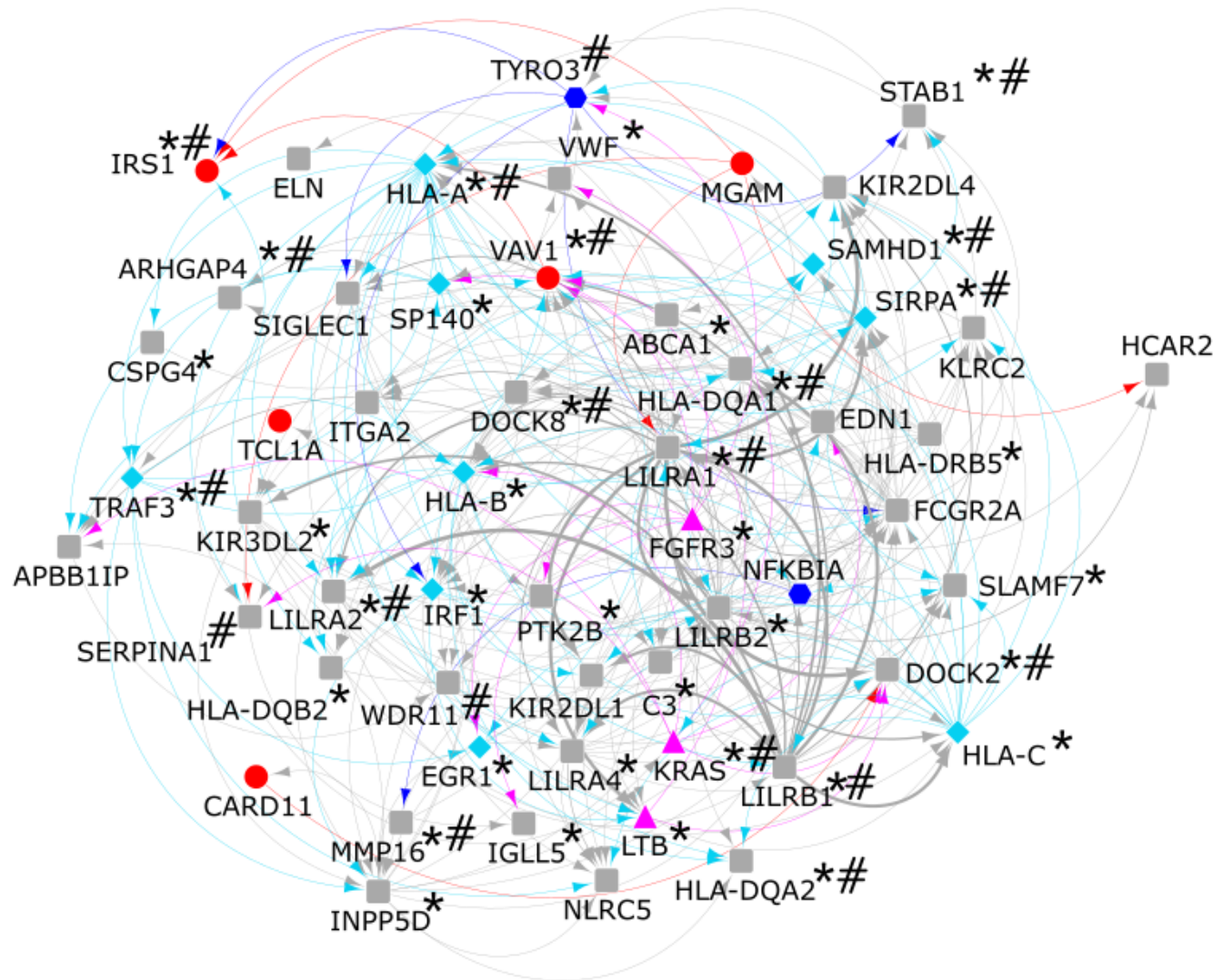


Figure-9: Fourth gene community displayed with top-250 genes present within the community. This figure showcase the previously reported genes (OG, TSG, ODG, AG) regardless of their rank, alongside other non-reported genes (in magenta color) within the top-250 ranks respectively. Genes marked with "*" are included in the 282-genes panel. Additionally, genes marked with "#" possess a high likelihood of being haploinsufficient, with a GHIS score > 0.52. Lastly, "O" and "T" symbols denote genes involved in oncogenic and transformative genomic events, respectively.

Colour Notations:

AG – Blue hexagon (■)

OG – Red circle (●)

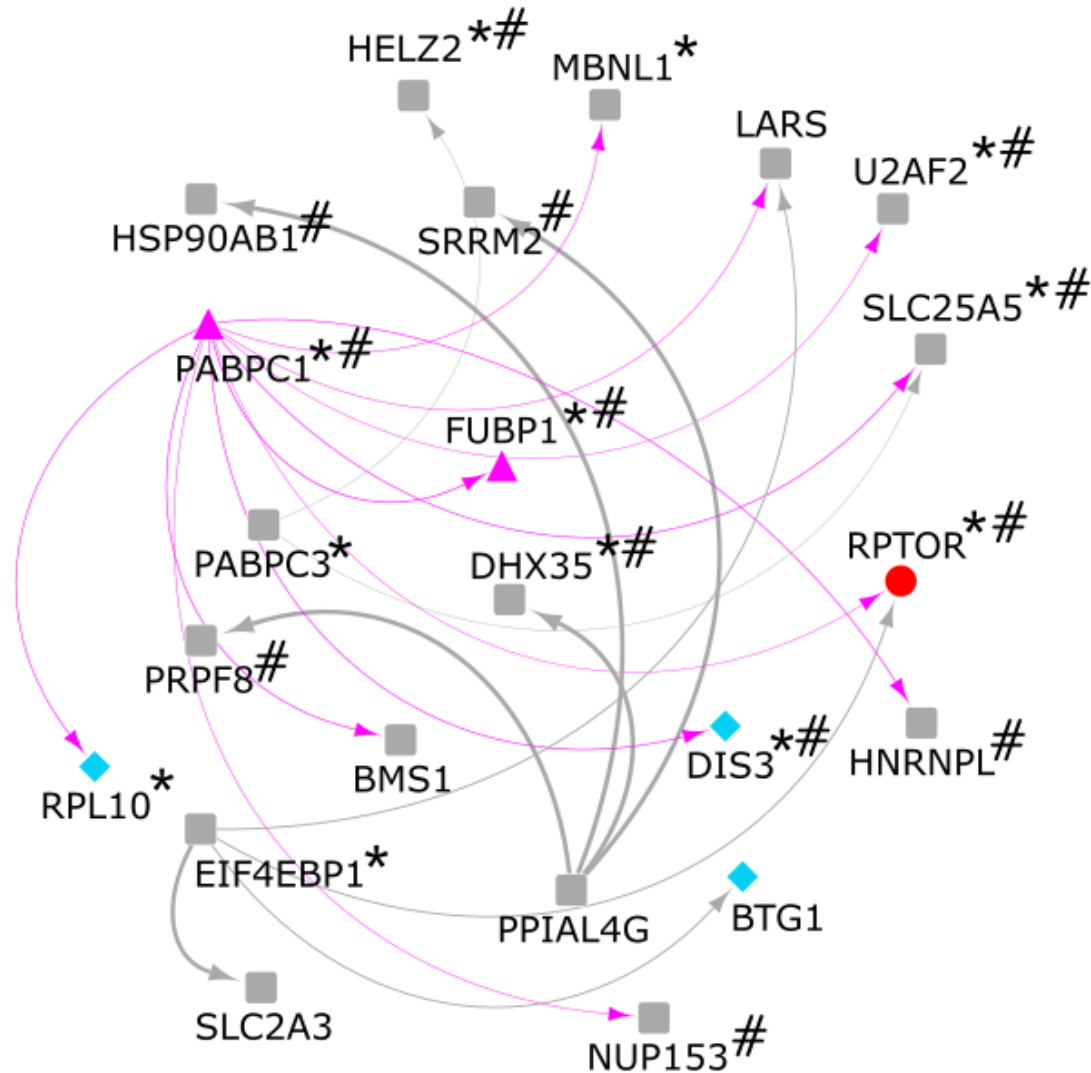
TSG – Cyan diamond (◆)

ODG – Magenta triangle (▲)

Genes that are part of communities and not OG/TSG/ODG/AG – Gray square (■)

Top Genes in 5th gene community

Figure-10: Fourth gene community displayed with top-250 genes present within the community. This figure showcase the previously reported genes (OG, TSG, ODG, AG) regardless of their rank, alongside other non-reported genes (in magenta color) within the top-250 ranks respectively. Genes marked with "*" are included in the 282-genes panel. Additionally, genes marked with "#" possess a high likelihood of being haploinsufficient, with a GHIS score > 0.52. Lastly, "O" and "T" symbols denote genes involved in oncogenic and transformative genomic events, respectively.



Colour Notations:

AG – Blue hexagon (■)

OG – Red circle (●)

TSG – Cyan diamond (◆)

ODG – Magenta triangle (▲)

Genes that are part of communities and not OG/TSG/ODG/AG – Gray square (■)

Identification of other genomic variant profiles (CNVs, SVs, LOF, Haploinsufficiency)



Identification of CNVs in top-ranked genes:

- We employed CNVKit [18] tool for identifying copy number variations (CNVs) in top-ranked significantly altered genes across MM and MGUS cohort of AIIMS and EGA database.
- For MMRF dataset, we have downloaded the segmentation data using MMRF CoMMpass (version 12) [11] that were further used to estimate the copy number using CNVkit tool.
- Further, we checked for number of deletion, loss, gain, and amplification for each gene across entire MM and MGUS cohort to account the impact of copy number variations on each disease stage (MM and MGUS).

Identification of SVs in top-ranked genes:

- We downloaded the structural variants identified using whole genome data of MMRF from MMRF CoMMpass and, for each of top-500 gene, we extracted the required information such as total number of samples having SVs in the gene, number of samples having inversion in the gene, etc.

References:

18. Talevich, E., Shain, A.H., Botton, T. and Bastian, B.C., 2016. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. PLoS computational biology, 12(4), p.e1004873.

Identification of other genomic variant profiles (CNVs, SVs, LOF, Haploinsufficiency)



Identification of Loss-of-Function in top-ranked genes:

- To identify the genes having loss-of-function, we assessed all transcripts against specific criteria. If a transcript satisfied any of the following conditions: deletion of over half of the coding sequence, deletion of the start codon, deletion of the first exon, deletion of a splice signal, or deletion causing a frameshift, it was considered to exhibit loss-of-function [19].
- We processed MM and MGUS cohort for identifying the number of samples having LOF present in the gene based on above-mention criteria.

Identification of Haplo-insufficiency in top-ranked genes:

- For haplo-insufficiency in the top-ranked genes, we downloaded the haplo-insufficiency score from two previously published studies, namely GHIS score [20], and DECIPHER score [19].
- The high value of GHIS score and the low value of DECIPHER score for a gene signifies the higher degree of haplo-insufficiency in the gene.

References:

19. Huang, N., Lee, I., Marcotte, E.M. and Hurles, M.E., 2010. Characterising and predicting haploinsufficiency in the human genome. PLoS genetics, 6(10), p.e1001154.
20. Steinberg, J., Honti, F., Meader, S. and Webber, C., 2015. Haploinsufficiency predictions without study bias. Nucleic acids research, 43(15), pp.e101-e101.

Inference from CNV, SV, and LOF variant profiles

- Analyzing the distribution of primary genomic events (such as CNVs, SVs, and LOF) chromosome-wise, we observed that chromosome 1 and chromosome 19 exhibited the highest impact among the 22 pairs of chromosomes, along with chromosome X and chromosome Y (Figure-11 (A)-(C)).
- Among the four types of CNVs, gene amplifications were the most prevalent CNVs in the top 500 genes across the MM cohort, accounting for 58.3% of occurrences (see Figure-11 (D)).
- Our analysis revealed that among the five types of structural variations, gene inversions were the most frequent SVs across the MM cohort in the top 500 genes, constituting 65.1% of the occurrences (see Figure-5 (E)).
- We have observed that chromosome 1, 2, 3, 6, and 7 were most affected by inversion structural variant. Furthermore, chromosome 1, 4, 7, 14, and 21 were most affected by translocation structural variant in MM (Figure-11 (F) and 11 (G)).
- Genes with structural variants, particularly translocations, were among previously reported OGs (e.g., FGFR3, BIRC6, PGR, RPTOR, BRAF), TSGs (e.g., KMT2C, TRAF3, HLA-A, ARID2, NF1), and AGs (e.g., ARID1B, RPTOR, FANCD2).
- Similarly, genes with CNVs were among previously reported OGs (e.g., MUC16, TP53, KMT2D, USP6), TSGs (e.g., HLA-A/B/C, KMT2B, VAV1, NF1, TRRAP), and AGs (e.g., HLA-C, BRD4, NOTCH1).

Inference from CNV, SV, and LOF variant profiles

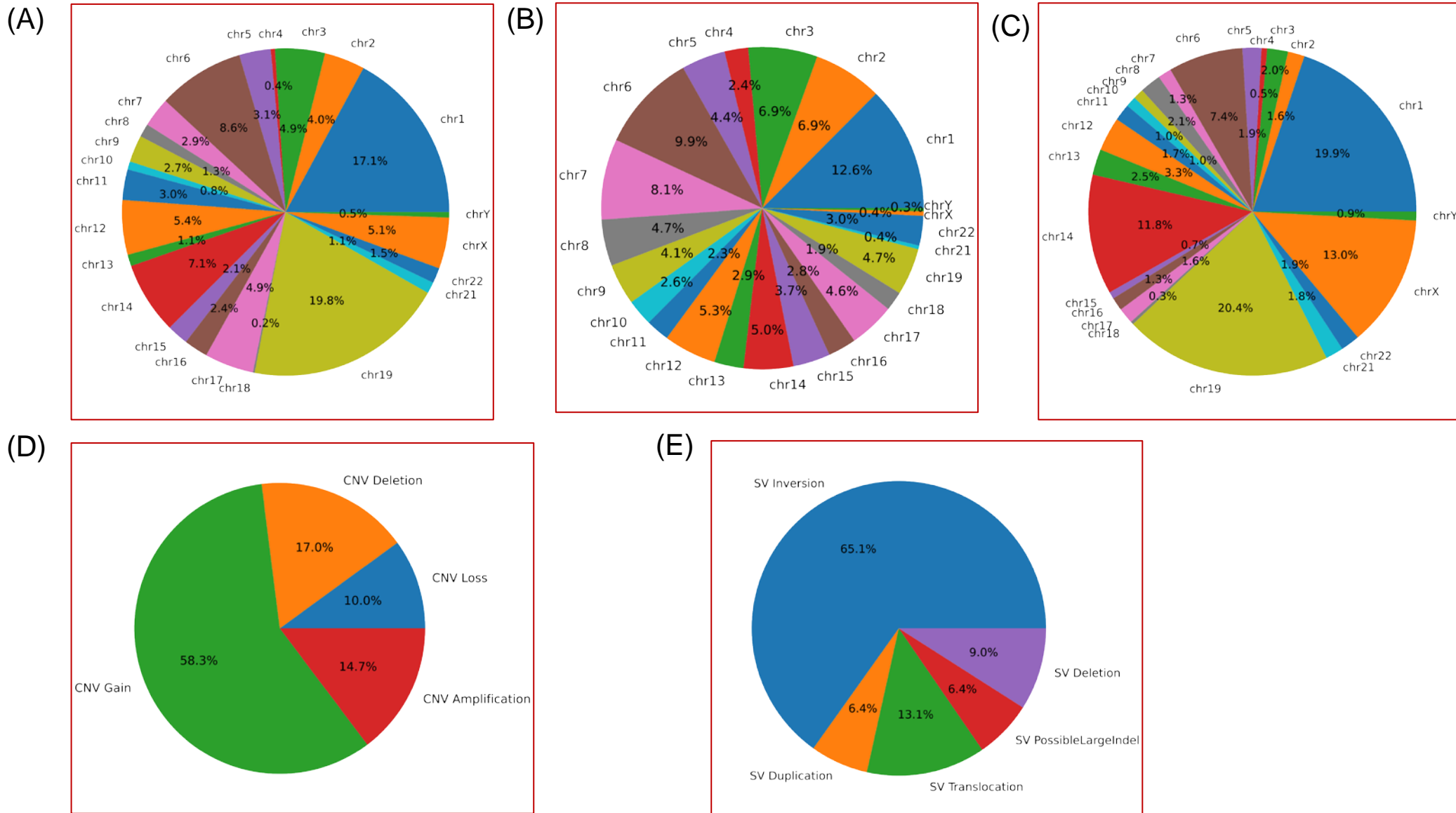


Figure-11: Chromosome-wise distribution of (A) CNVs, (B) SVs, and (C) LOF, (D) Distribution of CNV types identified in MM samples from AIIMS and MMRF dataset, and (E) Distribution of SV types identified in MM samples from MMRF dataset.

Inference from CNV, SV, and LOF variant profiles

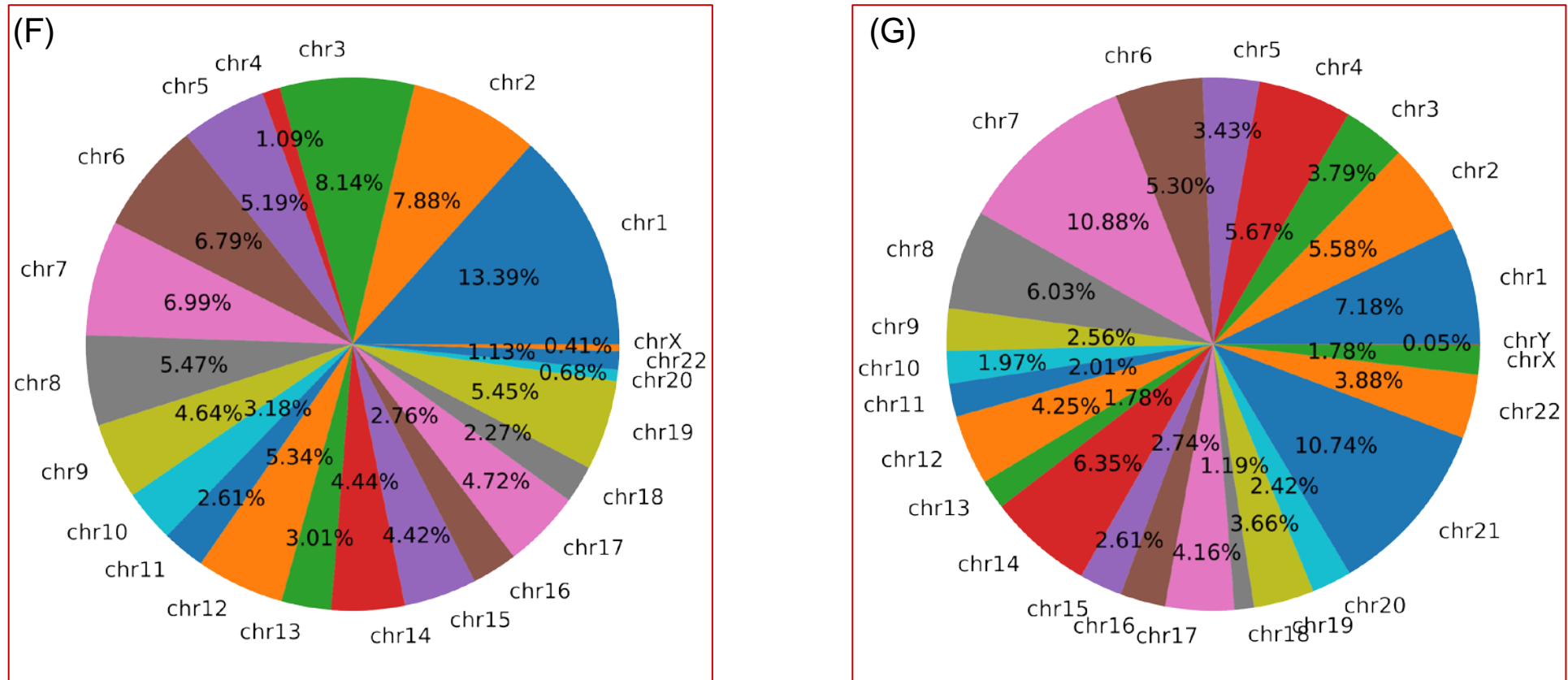
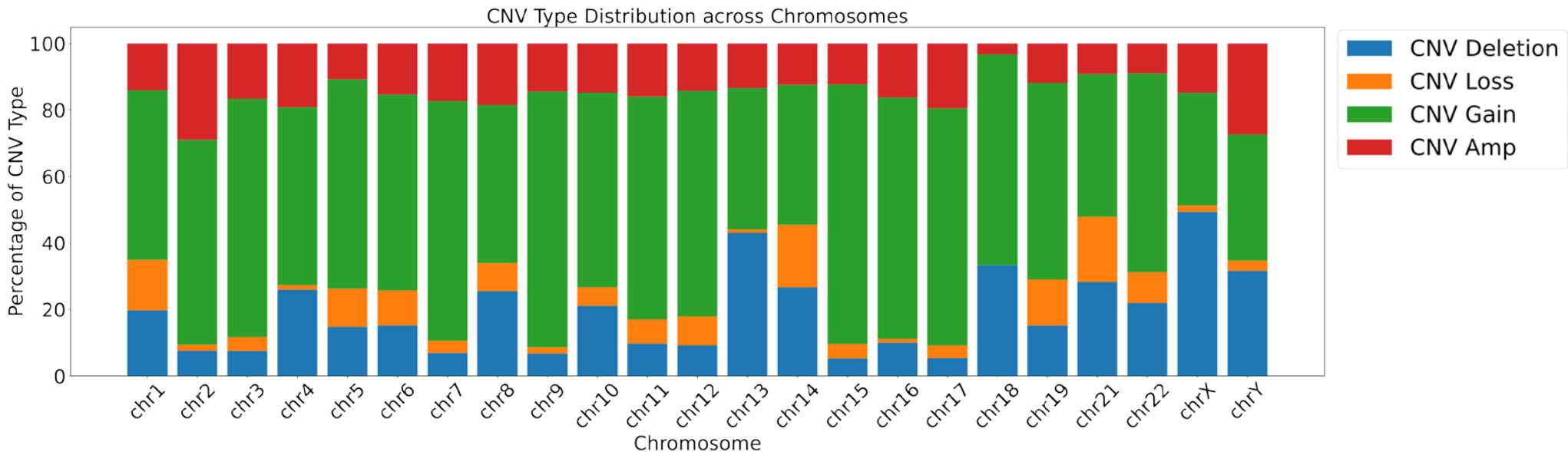


Figure-11 (continue): Chromosome-wise distribution of (F) inversion and (G) Translocation structural variant

(A)



(B)

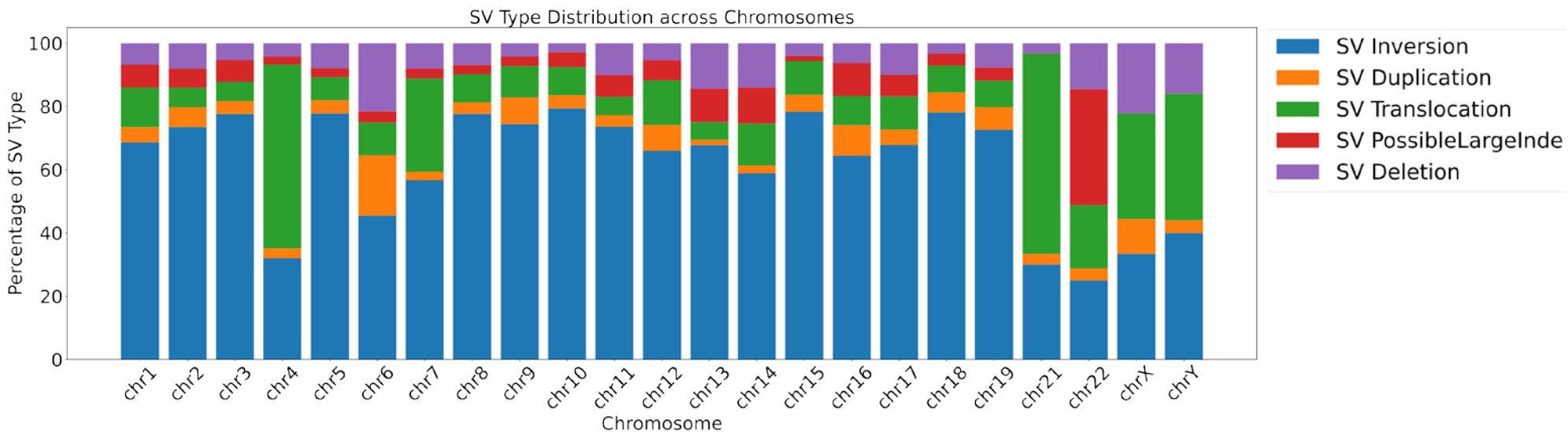


Figure-12: Distribution of (A) CNVs and (B) SVs types at chromosome level.

Geo2R validation of top-ranked significantly altered genes



- We Validated the significance of the top-ranked genes obtained from post-hoc analysis of Bio-DGI model using datasets related to MM-related studies with the help of Geo2R based analysis.
- In this approach, we considered 11 MM-related studies [21-33] and identified the significantly deregulated genes (adj p-value ≤ 0.05) using Geo2R tool and compared them with our top-ranked genes.
- Out of top ranked 500 genes, 488 genes were validated in at least one MM-related studies. Further, out of top-500 genes, 127 (25.4%) and 111 (22.2%) genes were found significantly deregulated in MM in datasets related to four and five MM-related studies (Figure-13).

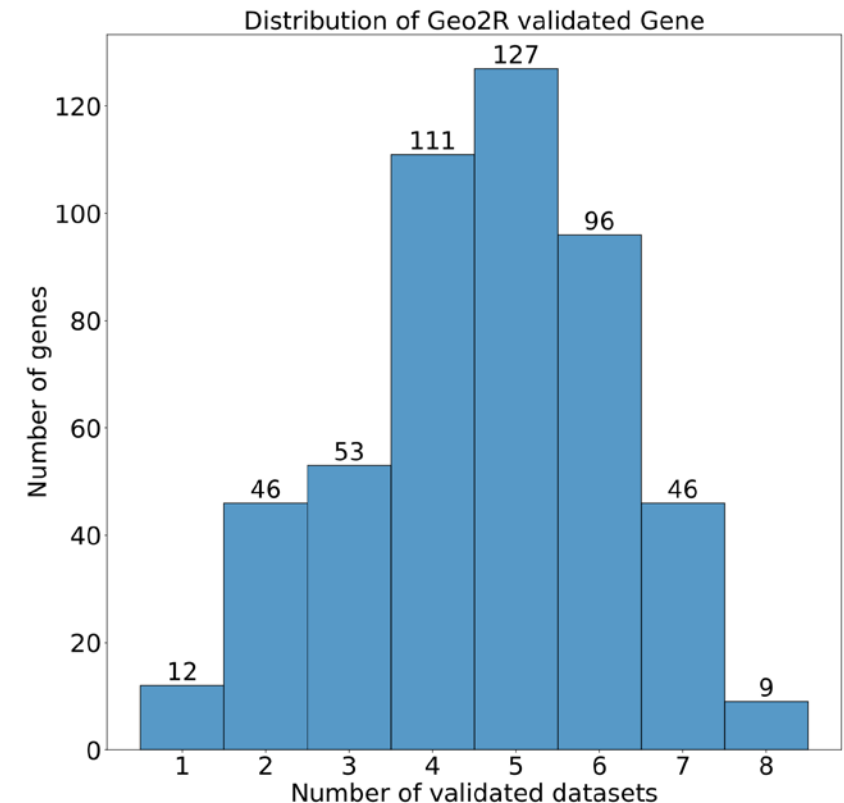


Figure-13: Distribution of genes validated with MM-related studies using Geo2R. The x-axis represents the number of MM-related studies in which the gene has been validated and y-axis represent the number of genes.

Geo2R validation of top-ranked significantly altered genes



References:

21. Farswan A, Gupta A, Gupta R, Kaur G. Imputation of Gene Expression Data in Blood Cancer and Its Significance in Inferring Biological Pathways. *Front Oncol* 2019;9:1442. PMID: 31970084
22. Garcia-Gomez A, Li T, de la Calle-Fabregat C, Rodríguez-Ubreva J et al. Targeting aberrant DNA methylation in mesenchymal stromal cells as a treatment for myeloma bone disease. *Nat Commun* 2021 Jan 18;12(1):421. PMID: 33462210
23. Gutiérrez NC, Sarasquete ME, Misiewicz-Krzeminska I, Delgado M et al. Dereglulation of microRNA expression in the different genetic subtypes of multiple myeloma and correlation with gene expression profiling. *Leukemia* 2010 Mar;24(3):629-37. PMID: 20054351
24. Zhou Y, Chen L, Barlogie B, Stephens O et al. High-risk myeloma is associated with global elevation of miRNAs and overexpression of EIF2C2/AGO2. *Proc Natl Acad Sci U S A* 2010 Apr 27;107(17):7904-9. PMID: 20385818
25. Lionetti M, Biasiolo M, Agnelli L, Todoerti K et al. Identification of microRNA expression patterns and definition of a microRNA/mRNA regulatory network in distinct molecular groups of multiple myeloma. *Blood* 2009 Dec 10;114(25):e20-6. PMID: 19846888
26. Biasiolo M, Sales G, Lionetti M, Agnelli L et al. Impact of host genes and strand selection on miRNA and miRNA* expression. *PLoS One* 2011;6(8):e23854. PMID: 21909367
27. Amodio N, Di Martino MT, Foresta U, Leone E et al. miR-29b sensitizes multiple myeloma cells to bortezomib-induced apoptosis through the activation of a feedback loop with the transcription factor Sp1. *Cell Death Dis* 2012 Nov 29;3:e436. PMID: 23190608
28. Bolzoni M, Storti P, Bonomini S, Todoerti K et al. Immunomodulatory drugs lenalidomide and pomalidomide inhibit multiple myeloma-induced osteoclast formation and the RANKL/OPG ratio in the myeloma microenvironment targeting the expression of adhesion molecules. *Exp Hematol* 2013 Apr;41(4):387-97.e1. PMID: 23178378
29. Ronchetti D, Todoerti K, Tuana G, Agnelli L et al. The expression pattern of small nucleolar and small Cajal body-specific RNAs characterizes distinct molecular subtypes of multiple myeloma. *Blood Cancer J* 2012 Nov 23;2:e96. PMID: 23178508
30. Bong IPN, Ng CC, Othman N, Esa E. Gene expression profiling and in vitro functional studies reveal RAD54L as a potential therapeutic target in multiple myeloma. *Genes Genomics* 2022 Aug;44(8):957-966. PMID: 35689754
31. Nair JR, Caserta J, Belko K, Howell T et al. Novel inhibition of PIM2 kinase has significant anti-tumor efficacy in multiple myeloma. *Leukemia* 2017 Aug;31(8):1715-1726. PMID: 28008178
32. Sacco A, Federico C, Todoerti K, Ziccheddu B et al. Specific targeting of the KRAS mutational landscape in myeloma as a tool to unveil the elicited antitumor activity. *Blood* 2021 Nov 4;138(18):1705-1720. PMID: 34077955
33. Soncini D, Martinuzzi C, Becherini P, Gelli E et al. Apoptosis reprogramming triggered by splicing inhibitors sensitizes multiple myeloma cells to Venetoclax treatment. *Haematologica* 2022 Jun 1;107(6):1410-1426. PMID: 34670358

Workflow for identifying candidate driver gene panel

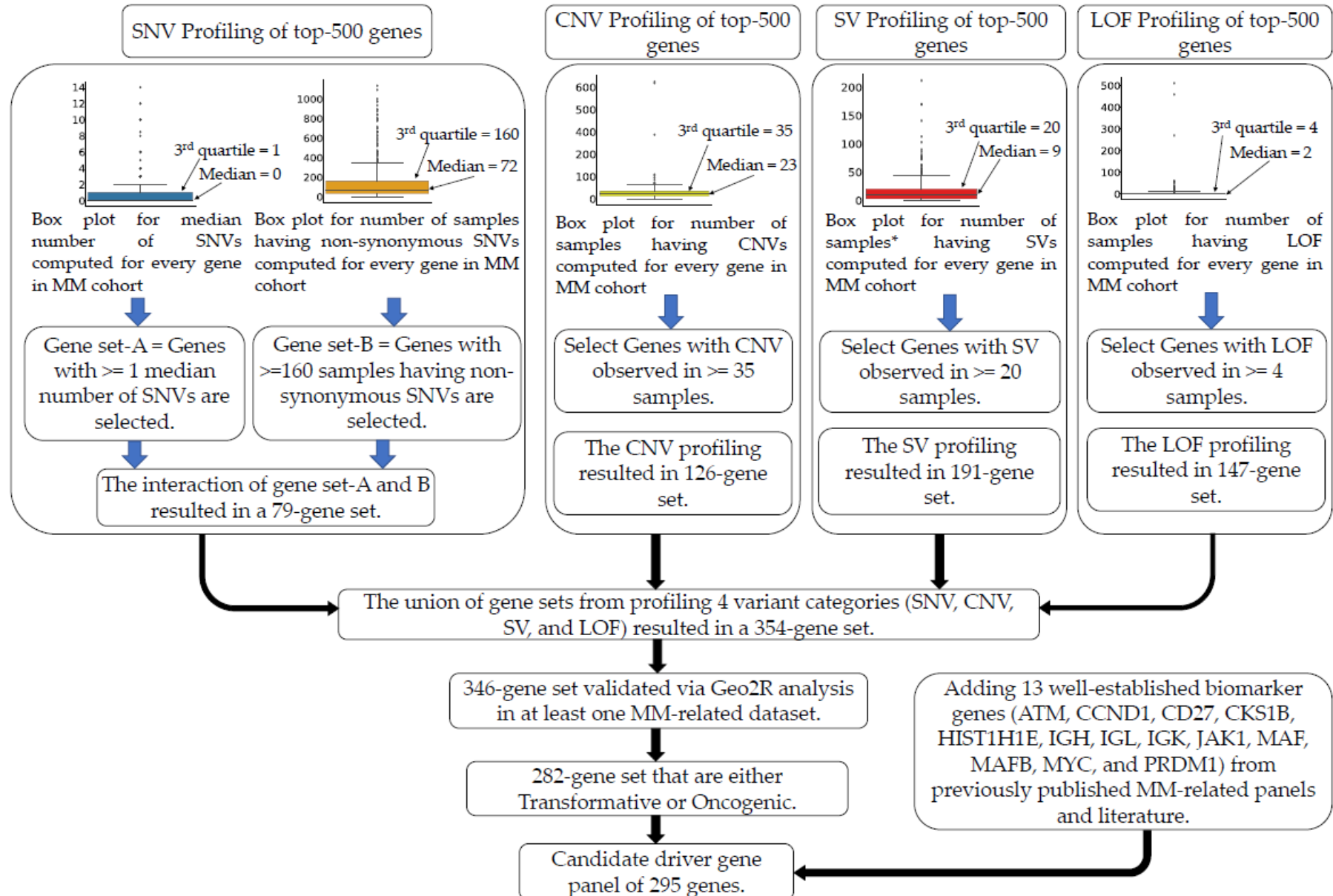


Figure-14: Workflow for filtration of top-500 genes for identifying the candidate driver gene panel.

Workflow for identifying candidate driver gene panel

- We further filtered top ranked genes based on their significant alterations and joint impact of their variant profiles in MM and MGUS.
- We considered the following five profiles for identifying the candidate driver gene panel i.e. 1. SNV profile, 2. CNV profile, 3. SV profile, 4. LOF profile, and 5. Geo2R Validation profile (workflow shown in Figure-14 in the previous slide).
- In addition to variant Geo2R validation profile, we categorized the top-ranked genes based on whether each gene exhibited significant alterations in MM and MGUS.
- Genes exclusively displaying significant alterations in MM were designated as "Transformative" (or disease progressing genes), while those significantly altered in both MM and MGUS were labeled as "Oncogenic" (or disease initiating genes).
- SNV profiling of the top 500 significantly altered genes involved filtering based on median SNV count and the number of samples with non-synonymous SNVs, resulting in a set of 79 genes.
- Furthermore, variant profiling for CNV, SV, and LOF involved filtering genes based on the number of samples exhibiting that particular variant type, yielding sets of 126, 191, and 147 genes, respectively.
- Combining all genesets derived from SNV, CNV, SV, and LOF variant profiles resulted in a comprehensive set of 354 genes.
- Only genes validated in at least one MM-related study using Geo2R validation were retained. Out of the 354 genes, 346 were validated through Geo2R validation analysis.
- In the final selection, 282 genes out of the 346 validated genes were chosen based on their significant alterations in MM and MGUS. Among these, 212 genes were observed to be transformative, and 70 genes were identified as oncogenic.

Workflow for identifying candidate driver gene panel

- We meticulously analyze the variant profiles (SNV, CNV, SV and LOF) for each gene and observed the following:
- A total of 4 genes, namely, RYR3, HLA-A, HLA-B, and HLA-DRB5, were found heavily mutated in all four variant profiles (shown in Figure-15).
- Further, A total of 32 genes were found heavily mutated in at least three variant profiles.
- For each gene, we have checked the most occurring molecular aberration such as CNV gain, CNV loss, SV translocation, LOF, etc. We observed that CNV gain is the most occurring molecular aberration found in 188 out of 282 genes, while LOF is the least occurring molecular aberration found in 12 out of 282 genes.
- Further, by combining CNVs, SVs, and LOFs, we have noted down the exon number and their genomic regions recommended for sequencing and can be helpful for further investigation from clinical perspectives.

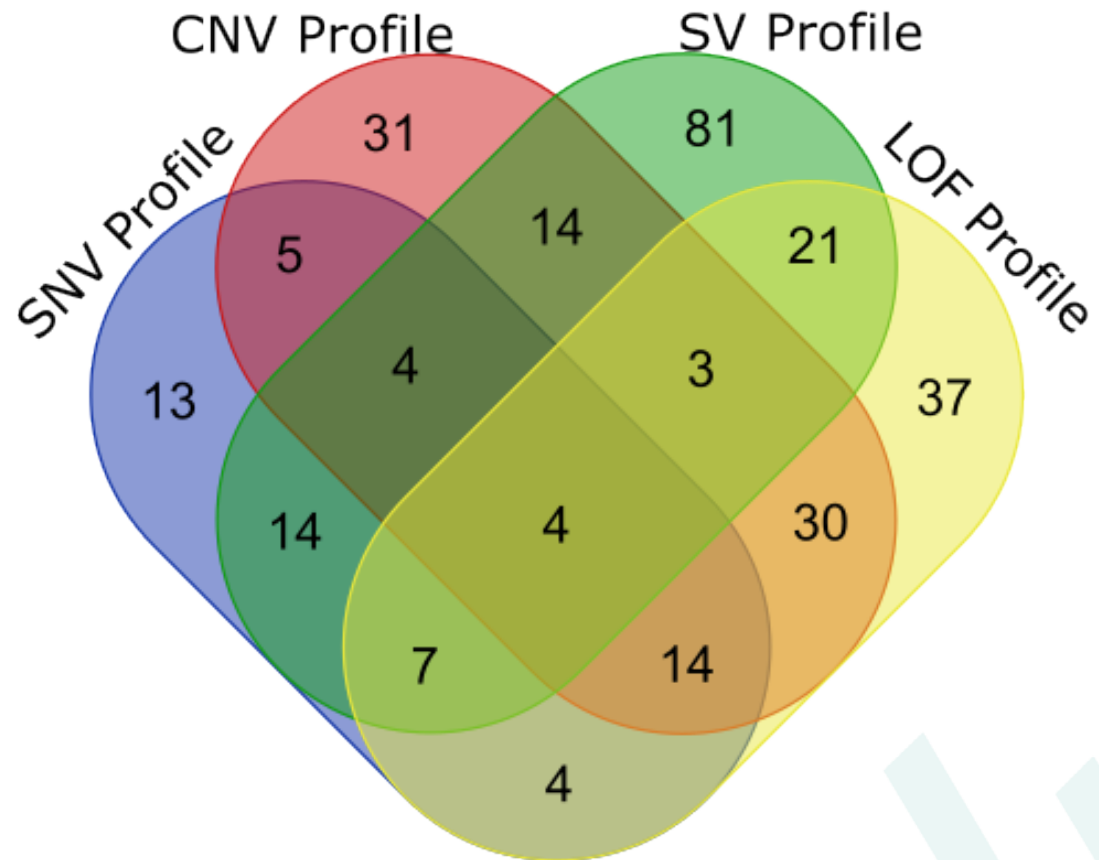


Figure-15: Venn diagram showing the comparison of filtered gene set between four variant category (SNV, CNV, SV, and LOF).

Workflow for Survival Analysis

- We conducted univariate survival analysis on a panel of 282 candidate driver genes to comprehend the influence of gene variant profiles on clinical outcomes in MM patients.
- To gauge the impact of gene variant profiles, we utilized two distinct approaches:
 1. We individually assessed the impact of each variant profile (SNV, CNV, SV, and LOF) on clinical outcomes. Univariate survival analysis was performed for each variant profile, providing insights into their respective impact (refer to Figure-16, Algorithm-A).
 2. Additionally, we amalgamated the four variant profiles (SNV, CNV, SV, and LOF) for each gene using Factor Analysis for Mixed Data (FAMD), enabling us to estimate a joint feature. Subsequently, we performed univariate survival analysis using the FAMD 1st component as a prognostic factor (refer to Figure-16, Algorithm-B).
- In total, 193 genes out of the 282 gene were found to significantly influence clinical outcomes in univariate survival analysis based on at least one prognostic factor through workflow-A (shown in the Figure-16 in the next slide).
- Moreover, 185 out of 282 genes demonstrated significance in univariate survival analysis based on the FAMD 1st component (a combined feature generated by integrating the four variant profiles for each gene) through workflow-B (shown in the next slide). Intriguingly, 139 genes out of these 185 were also identified as significant in univariate survival analysis using workflow-A.

Workflow for Survival Analysis

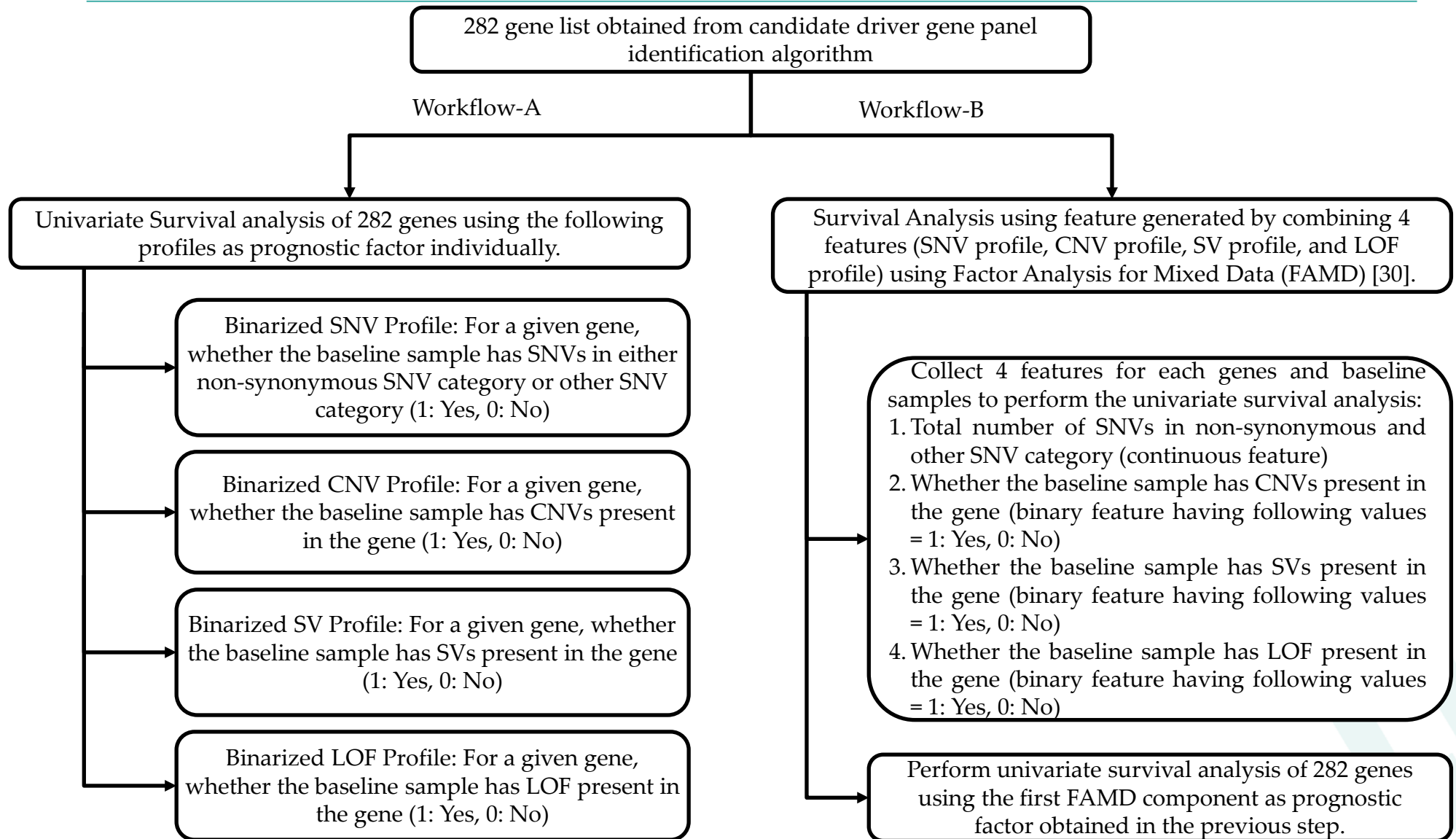


Figure-16: Survival analysis workflow for identifying the impact of gene variant profiles on MM patients clinical outcomes.

Haploinsufficient genes in 282-genes panel

1. Haploinsufficiency and Disease Phenotype:

- Haploinsufficiency arises when a single functional gene copy fails to sustain a normal phenotype, leading to an abnormal or disease phenotype. This inadequacy disrupts essential cellular processes critical for maintaining normalcy.

2. Disruption of Cellular Processes:

- A loss of function in haploinsufficient genes disrupts pivotal cellular processes such as growth regulation, DNA repair, and signaling pathways. This disruption profoundly impacts cell behavior and functions.

3. Haploinsufficiency in Tumor Suppressor Genes:

- The loss of one allele in a tumor suppressor gene due to structural variations (SVs) like chromosomal deletion, copy number variations (CNVs), or single nucleotide variations (SNVs) can lead to haploinsufficiency.
- Tumor suppressor genes are crucial in controlling cell growth and impeding tumor formation. The haploinsufficiency of these genes can significantly contribute to the development and progression of cancer, including MM.

Haploinsufficient genes in 282-genes panel

4. Highly Haploinsufficient Genes in the Proposed 282-Gene Panel:

- In our recommended 282-gene sequencing panel, 61 genes exhibit notably high haploinsufficiency with GHIS Scores [34] equal to or greater than 0.6 (the third quartile of GHIS scores for all 282 genes). Notable genes in this category include NRAS, MAX, RB1, ARID1B, ARID2, SLC25A5, CYLD, TP53, ZNF717, PABPC1, showing a high prevalence of SNVs, CNVs, SVs, and LOF.
- Further, a total of 127 genes out of the 282-gene panel were identified to have high haploinsufficiency (GHIS Score greater than or equal to 0.54, the median GHIS score of all 282 genes). Noteworthy genes within this group include BRAF, DOCK2/8, MITF, RLIIM, and RPTOR, further indicating the significant role of haploinsufficiency in MM progression.
- Influence of Haploinsufficient Genes in Gene Communities: Several MM-related genes like RPTOR, KRAS, HLA-A, HLA-DQA1/2, HLA-DQB2, ZNF717 were observed to actively participate in gene communities, exhibiting higher influence based on their Katz centrality scores. This high influence underscores the crucial role of haploinsufficiency in shaping gene networks and potentially driving MM disease progression.
- By strengthening these points, we emphasize the fundamental impact of haploinsufficiency on MM disease progression, shedding light on its intricate role within the genomic landscape of MM.

References:

34. Steinberg, J., Honti, F., Meader, S. and Webber, C., 2015. Haploinsufficiency predictions without study bias. *Nucleic acids research*, 43(15), pp.e101-e101.

Key features of 282-genes panel

- Several previously published MM and MGUS-related studies have identified the key genomic event and their impact on the specific disease stage. We categorized these genomic events based on their impact on particular disease stage i.e. oncogenic events and transformative events.
- The oncogenic genomic events are disease initiating events and observed in the early stage of the disease, while the transformative genomic events are disease progressing events that help the disease to progress further.
- Our recommended targeted sequencing panel of 282-genes contains genes having both oncogenic and transformative events. The presence of both category of genomic events enhances the prognostic capability for both MM and MGUS.
- We have listed the previously published MM and MGUS-related oncogenic and transformative genomic events in Table-9 and Table-10 (on subsequent slides). Furthermore, we have highlighted the key genomic events that can be detected using our recommended targeted sequencing panel.
- The presence of oncogenic genomic events in our recommended sequencing 282-genes panel facilitate the earliest stage detection or inception as the oncogenic genomic events support the accumulation of plasma cells in MGUS.
- Further, presence of transformative genomic events help find out the patients within MGUS that are more likely to progress to MM and high risk MM patients.
- In total, 239 genes out of the 282 gene were found to significantly influence clinical outcomes in univariate survival analysis. The remaining set of 43 genes which were not found significant in survival analysis contains several important genes such as CYLD, LILRB1, NFKBIA, NRAS, KMT2B, PABPC3, PGR were considered based on their key oncogenic and transformative genomic events.
- The total number of exon regions covered in our targeted sequencing panel are 9272 and total length of targeted sequencing coding region in the human genome is 2.557 Mb

List of previously published oncogenic genomic events

Table 4: List of key genomic events in MM and MGUS and overlapping of their associated genes with 295 genes panel.

(A) List of disease-initiating genomic events reported previously in both MM and MGUS and the overlapping of their associated genes with our proposed 295-genes panel

| S.No. | A. Genomic Events in MM and MGUS | B. Genes associated with the event (Column-A) | C. References for the genes shown in column-B | D. Whether the genes of Column-B are present in 295-genes panel | E. If yes, list of genes common with Column-B | F. Associated gene-gene interactions found by BIO-DGI (PPI9) in the 295-gene panel. |
|-------|----------------------------------|---|---|---|---|---|
| 1 | t(11;14) | <i>CCND1, BCL-2</i> | [65, 81] [82, 83] | Yes | <i>CCND1</i> | <i>BRD4, IRS1, KRAS, NRAS, TP53</i> |
| 2 | t(4;14) | <i>FGFR3</i> | [65, 80] [82, 83, 84] | Yes | <i>FGFR3</i> | <i>CYLD, DIS3, KRAS, NRAS</i> |
| 3 | t(14;16) | <i>MAF</i> | [82, 83] [90, 65] | Yes | <i>MAF</i> | <i>FLNA</i> |
| 4 | t(14;20) | <i>MAFB</i> | [65, 83] [90] | Yes | <i>MAFB</i> | <i>HUWE1, USP9X</i> |
| 5 | Amp(1q21) | <i>MCL1, CKS1B, ANP32E or BCL9</i> | [80] | Yes | <i>CKS1B</i> | <i>KPRP, BRD4, PLEC, USP9X</i> |
| 6 | Del(17p13) | <i>TP53</i> | [91] | Yes | <i>TP53</i> | <i>IRF4, KMT2B/C/D, KRAS</i> |
| 7 | <i>KRAS</i> mutations | <i>KRAS</i> | [65] | Yes | <i>KRAS</i> | <i>MAX, BRAF, EGR1, NF1</i> |
| 8 | <i>NRAS</i> Mutations | <i>NRAS</i> | [65] | Yes | <i>NRAS</i> | <i>ARID2, FGFR3, SP140</i> |
| 9 | <i>LTB</i> Mutations | <i>LTB</i> | [65] | Yes | <i>LTB</i> | <i>IRF1, NFKBIA, SP140</i> |
| 10 | <i>DIS3</i> mutations | <i>DIS3</i> | [65] | Yes | <i>DIS3</i> | <i>FGFR3, KRAS, NRAS</i> |
| 11 | <i>EGR1</i> mutations | <i>EGR1</i> | [65] | Yes | <i>EGR1</i> | <i>CYLD, FGFR3, KRAS, NRAS</i> |
| 12 | <i>MYC</i> Rearrangement | <i>IGH, IGL, IGK, NSMCE2, TXNDC5, FAM46C, FOXO3, IGJ, PRDM1</i> | [10] | Yes | <i>IGH, IGL, IGK, FAM46C</i> | <i>CYLD, DIS3, FGFR3, KRAS</i> |

(B) List of disease-transformative genomic events reported previously in MM but not in MGUS and the overlapping of their associated genes with our proposed 295 genes panel.

| S.No. | A. Disease-Transformative Genomic Events | B. Genes associated with the event (Column-A) | C. References for the genes shown in column-B | D. Whether the genes of Column-B are present in 295-genes panel | E. If yes, genes present in the 295-genes panel | F. Associated gene-gene alterations found by BIO-DGI (PPI9) in the 295-genes panel |
|-------|--|---|---|---|---|--|
| 1 | Del(13q14) | <i>RB1</i> | [92, 80] | Yes | <i>RB1, DIS3</i> | <i>BRAF, FGFR3, KRAS</i> |
| 3 | Del(16q23) | <i>CYLD</i> | [81] | Yes | <i>CYLD</i> | <i>DIS3, KRAS, NRAS</i> |
| 4 | Del(1p21) | <i>CDC14A</i> | [80, 86] | Yes | <i>FAM46C</i> | <i>CYLD, DIS3, FGFR3, KRAS</i> |
| 5 | Del(12p13) | <i>CD27</i> | [93] | Yes | <i>CD27</i> | <i>TRAF2, TRAF3, ATP2B3</i> |
| 6 | <i>TP53</i> Mutations | <i>TP53</i> | [94, 95] | Yes | <i>TP53</i> | <i>IRF4, KMT2B/C/D, KRAS</i> |
| 7 | <i>BRAF</i> Mutations | <i>BRAF</i> | [96, 97] | Yes | <i>BRAF</i> | <i>FGFR3, KRAS, NRAS</i> |
| 8 | Gain(9q) | <i>ABCA1, KCNT1, TRAF2, VPS13A</i> | [98, 99] | Yes | <i>ABCA1, KCNT1, TRAF2, VPS13A</i> | <i>BRAF, CYLD, IRF1</i> |
| 9 | del(14q) | <i>TRAF3</i> | [80] | Yes | <i>TRAF3</i> | <i>CYLD, DIS3, KMT2D</i> |
| 10 | del(17p) | <i>TP53</i> | [80] | Yes | <i>TP53</i> | <i>IRF4, KMT2B/C/D, KRAS</i> |
| 11 | del(8p) | <i>PTK2B, TP53</i> | [100] [101] | Yes | <i>PTK2B</i> | <i>EGR1, FGFR3, KRAS</i> |

Comparison with previously reported panels

Table-9: Comparison of previously published targeted NGS-based panels in Multiple Myeloma

| S. No. | Panel Reference, Publication year | Total number of genes in the proposed gene panel | Number of samples used for panel validation | Data Type | Detected variant profiles | Overlapping with 295-genes panel |
|--------|-----------------------------------|--|--|-----------------------|---|----------------------------------|
| 1 | Kortum et al [14], 2015 | 47 | 22 NDMM, 3 pretreated MM samples | WES | SNVs, clonal evolution analysis | 19 |
| 2 | Bolli et al [15], 2016 | 182 | 5 MM samples | WGS | SNVs, CNVs, SVs* | 26 |
| 3 | White et al [16], 2018 | 465 | 110 MM samples | WGS | SNVs, CNVs, SVs* | 47 |
| 4 | Cutler et al [12], 2021 | 26 | 76 (20 MGUS, 3 SMM, 52 MM, and 1 PCL) samples | WGS | SNVs, CNVs, Clinical validation using survival analysis | 18 |
| 5 | Sudha et al [13], 2022 | 228 | 185 MM samples | WGS | SNVs, CNVs, SVs* | 42 |
| 6 | Vivek et al (Current study) | 295 | 1215 (1,154 MM and 61 MGUS) samples + 11 MM-datasets | WES, microarray, mRNA | SNVs, CNVs, SVs, clinical validation using two-fold survival analysis | - |

“*”: SVs* represents the translocation structural variation involving IgH.

Key pathways associated to 295-genes panel

- Our comprehensive analysis identified a total of 189 significantly altered pathways, encompassing 105 KEGG [50] and 84 Reactome [51] pathways within the 282-gene panel.
- Among these pathways, 15 signaling pathways exhibited heightened significance as the disease progressed to MM. Additionally, 149 pathways displayed significant alterations exclusive to MM.
- Noteworthy pathways impacted by these alterations include the MAPK signaling pathway, Immune System, and the calcium signaling pathway.
- The MAPK signaling pathway, notably associated with KRAS gene mutation [53] and the t(4;14) translocation [54], underlines its critical role, particularly in the progression from MGUS to MM.
- Disruption of the calcium signaling pathway triggers intracellular calcium signaling, resulting in heightened intracellular calcium levels that activate various other signaling pathways, including NF-κB, PI3K-AKT, and MAPK signaling pathways [55].
- Recent findings have shed light on an association between T-cell dysfunction and immune system suppression in MM patients, emphasizing the vital role of the immune system pathway in MM progression [56].
- Figure-5 (next slide) presents a bubble plot for the top 50 significantly altered pathways associated with the 282-genes panel. This figure showcases pathway ranks, the number of significantly altered genes (out of 282) within each pathway, and $-\log_{10}(\text{adjusted } p\text{-value})$.

References:

50. Kanehisa, M. and Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1), pp.27-30.
51. Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senf-Ribeiro, A., Griss, J., Sevilla, C., Matthews, L., Gong, C. and Deng, C., 2022. The reactome pathway knowledgebase 2022. *Nucleic acids research*, 50(D1), pp.D687-D692.
52. Hideshima, T., Chauhan, D., Richardson, P., Mitsiades, C., Mitsiades, N., Hayashi, T., Munshi, N., Dang, L., Castro, A., Palombella, V. and Adams, J., 2002. NF-κB as a therapeutic target in multiple myeloma. *Journal of Biological Chemistry*, 277(19), pp.16639-16647.
53. van Nieuwenhuijzen, N., Spaan, I., Raymakers, R. and Peperzak, V., 2018. From MGUS to multiple myeloma, a paradigm for clonal evolution of premalignant cells. *Cancer research*, 78(10), pp.2449-2456.
54. Chesi, M., Brents, L.A., Ely, S.A., Bais, C., Robbiani, D.F., Mesri, E.A., Kuehl, W.M. and Bergsagel, P.L., 2001. Activated fibroblast growth factor receptor 3 is an oncogene that contributes to tumor progression in multiple myeloma. *Blood, The Journal of the American Society of Hematology*, 97(3), pp.729-736.
55. Cui, C., Merritt, R., Fu, L. and Pan, Z., 2017. Targeting calcium signaling in cancer therapy. *Acta pharmaceutica sinica B*, 7(1), pp.3-17.
56. Dhodapkar, M.V., 2023. The immune system in multiple myeloma and precursor states: lessons and implications for immunotherapy and interception. *American Journal of Hematology*, 98, pp.S4-S12.

Top-50 Significantly altered signaling Pathways

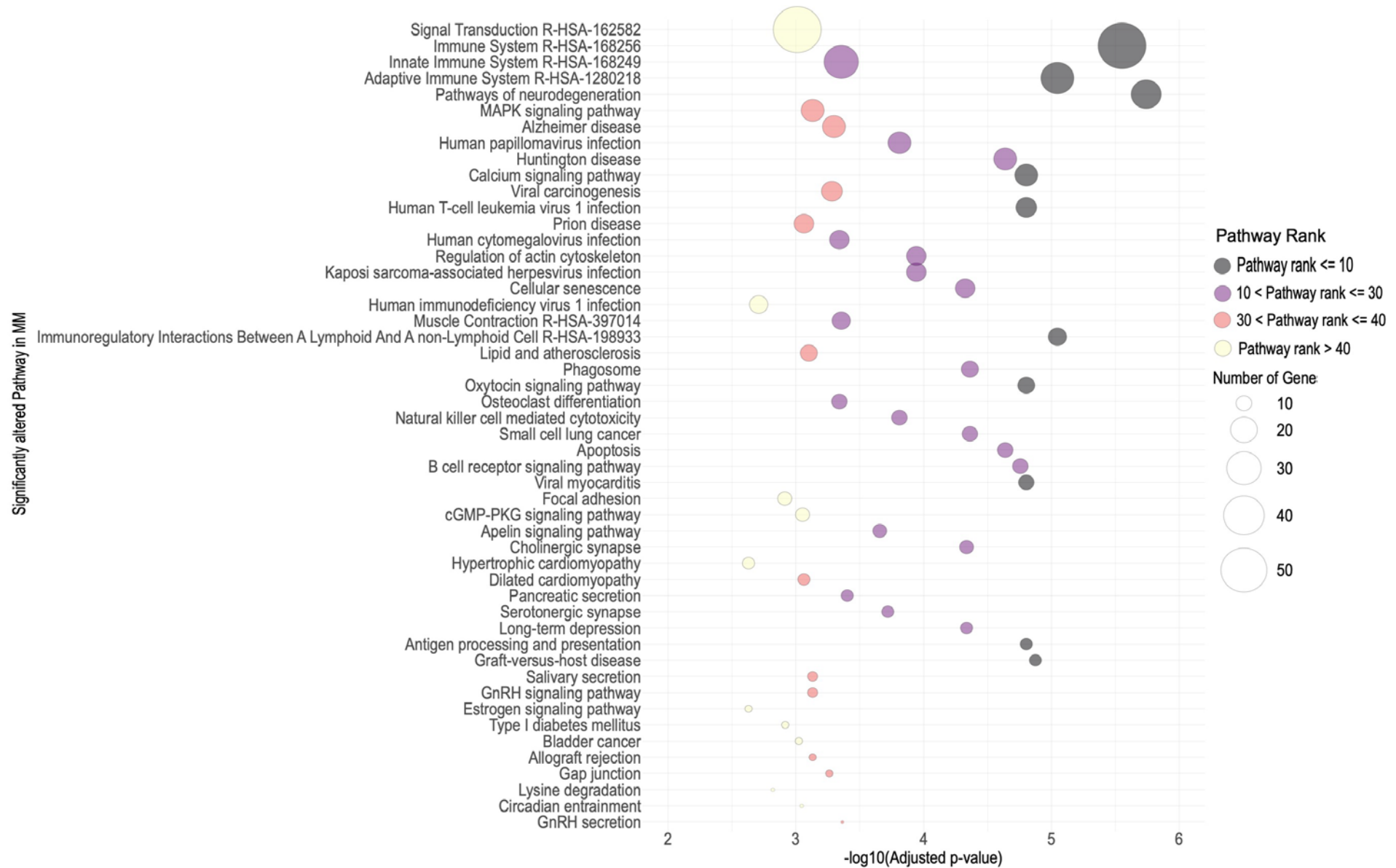


Figure-5: Bubble plot for Top-50 significantly altered pathways associated with 282-genes panel in MM. The x-axis in the bubble plot represents $-\log_{10}(\text{adjusted p-value})$, y-axis represent the pathway name, bubble size represents the number of significantly altered genes and bubble colour represents the pathway ranking.

Key findings from post-hoc analysis of proposed Bio-DGI model

- Our proposed Bio-DGI (PPI9) model was designed and trained in a bio-inspired manner that enables model to learn simultaneously from gene-gene interactions from nine PPI databases and genomic features for identifying the key differentiating gene-gene interactions from MM and MGUS genomic data.
- The integrated learning using topological information obtained from nine PPI network and exomic mutational profiles enables the Bio-DGI model to rank the genes and genomic features based on their role in disease progression more efficiently with few layers of graph convolutions network, and a multi-head attention modules than the traditional machine learning models that were trained only on the exomic mutational profiles.
- When two ML/DL model perform similar quantitatively, it would be better if the model is chosen on the basis of the interpretability with reference to the application domain. Here, The quantitative performance of top-3 models (Bio-DGI (PPI9), BDL-SP, and CS-Cat) were similar and on further post-hoc benchmarking revealed the Bio-DGI model has identified most number of previously reported genes.
- The genomic features associated to functional significance to non-synonymous SNVs and median AD of synonymous SNVs were observed to be the most contributing genomic features in differentiating MM from MGUS. Thus, further analysis on the clinical impact of non-synonymous SNVs is strongly suggested.
- Using this workflow, we have identified the the significantly altered pathways using Enrichr with the help of top-500 significantly altered genes obtained from Bio-DGI that become more/less significant with disease progression from MGUS to MM. We also ranked pathways based on their adjusted p-value. We observed that several top-ranked pathways are selectively dysregulated in MM and also, the pathways who lost their significance from MGUS to MM were actually related to other cancer types.
- We identified the five gene communities using the learned adjacency matrix obtained from trained Bio-DGI (PPI9) model. Each gene community contains the previously reported genes of multiple categories such as OG, TSG, ODG, and AG. Further, we have also identified the key member in each gene community using Katz centrality criterion that help to identify the key gene-gene interactions in the PPO network.

Key findings from post-hoc analysis of proposed Bio-DGI model

- We have observed that chromosome-1 and chromosome 19 were the most affected chromosomes by genomic alterations such as CNVs, SVs and LOFs. The genomic events in chromosome-1, such as 1q gain (associated with disease aggressiveness [54]) and 1p deletion (frequently observed in MGUS [55]); and chromosome-19 such as 19q deletion (common in MM and associated with disease progression [56]) and 19p gain (observed in MM [57]) has been published in many MM-related studies. Lastly, the interplay between alterations in these chromosomes and other genetic events can lead to increased genomic instability, facilitating the acquisition of additional mutations that promote MM aggressiveness [58].
- We validated the top-ranked genes with previously published MM-related studies using Geo2R analysis. A total of 488 genes, out of 500 were validated in Geo2R validation of at least one MM-related study.
- We proposed a recommended targeted sequencing panel using the post-hoc analysis of proposed Bio-DGI (PPI9) model and designed a robust 282 gene panel for improved prognostication and comprehensive detection of recurring genomic aberrations.
- The 282 gene panel contains genes of both oncogenic and transformative nature enabling earlier detection of disease, and identifying the high risk MM patients.
- Through a meticulous analysis of significantly altered pathways, we observed a substantial overlap with MM-related pathways, reinforcing the importance of the genes within the 282-gene panel. This overlap enhances the potential to identify MGUS samples likely to progress to MM.
- Further, a total of 127 out of 282 genes were found to be highly haploinsufficient and actively participating in the gene communities with high node influence.

References:

54. Schmidt, T.M., Fonseca, R. and Usmani, S.Z., 2021. Chromosome 1q21 abnormalities in multiple myeloma. *Blood cancer journal*, 11(4), p.83.
55. Chang, H., et al. (2006). The t (1; 14)(p22; q32) is nonrandom and specific for high-risk myeloma. *Genes, Chromosomes and Cancer*, 45(3), 213-220.
56. Nakamura, T., et al. (2018). BACH2 is a target for mutated IGHV3-7*04 in chronic lymphocytic leukemia. *Leukemia*, 32(6), 1378-1388.
57. Schop, R. F., et al. (1999). Karyotypic diversity in the progression of monoclonal gammopathy of undetermined significance to multiple myeloma. *Clinical Cancer Research*, 5(10), 2414-2422.
58. Bolli, N., et al. (2014). Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nature Communications*, 5, 2997.

Conclusion

- We designed and implemented an AI-driven work with novel “Graph network learning based directed gene-gene interactions in MM (Bio-DGI)” to identify the differentiating genomic biomarkers (such as gene-gene interactions, genomic features, etc.) in MM and MGUS.
- The PPI network can be helpful to infer the genes playing significant role in MM disease pathogenesis. We incorporated nine PPI databases to get a merged network of 798 significantly altered genes obtained using three global WES data repositories.
- On benchmarking the proposed Bio-DGI (PPI9) model with several baseline machine learning models, the Bio-DGI model performed similar in terms of balanced accuracy with top-3 models (BDL-SP and CS-Cat) and outperformed all baseline ML (CS-DT, CS-LR, CS-RF, CS-XGB, CS-Cat, CS-SVC) and DL (BIO-DGI (PPI-STRING) and BDL-SP) model in terms of AUPRC.
- When two model performed similar quantitatively, the model should be chosen on the basis of interpretability with reference to the application domain using proper post-hoc analysis/benchmarking.
- On post-hoc benchmarking of Bio-DGI (PPI9) model using ShAP algorithm, the Bio-DGI (PPI9) model reported the highest number of previously reported genes (OG, ODG, and AG) as comparison to the traditional ML models.
- We identified five gene communities using the learned adjacency matrices obtained from trained Bio-DGI (PPI9) model and significant community members using Katz centrality criterion. The interactions associated with genes having high Katz centrality were further proven to be crucial for MM disease progression.
- We further validated our findings by performing pathway analysis on the top mutated genes and observed that several signaling pathways such as Calcium signaling path- way, B-Cell receptor signaling pathway, PI3K-Akt signaling pathway, MAPK signaling pathway, etc. are selectively and more significantly deregulated with disease progression.
- We validated the top-ranked 500 genes with MM-related studies using Geo2R tool. A total of 488 genes⁴⁸ out of 500 genes were validated in at least one MM-related studies in Geo2R validation.

Conclusion

- We further filtered out the top-ranked 500 genes for identifying the candidate driver gene panel and designed a 282-genes panel by integrated analysis of their four variant profiles (SNV, CNV, SV, and LOF), significantly altered mutations, and Geo2R validated studies.
- Our proposed gene panel contains genes having both oncogenic and transformative genomic events and hence, can be helpful in earlier detection of MM, identifying high risk MM samples.
- We have also identified the impact of gene variant profiles on MM baseline samples clinical outcomes by considering individual variant profile as prognostic factor and by combining all variant profiles using FAMMD approach. A total of 193 and 185 genes were found affecting the MM patients clinical outcomes significantly using individual variant profiles and FAMMD component, respectively.
- Several MM-related genes exhibiting high haploinsufficiency were observed to actively participate in gene communities with high node influence.