Messe Wien Exhibition & Congress Center

International Myeloma Society

Vienna, Austria

8-11 September 2021

AI-based models for the identification of critical genetic biomarkers to distinguish MM from MGUS using the WES data

Vivek Ruhela[1], Akanksha Farswan[2], Ritu Gupta[3], Sriram K[1], Gurvinder Kaur[3], Anubha Gupta[2]

1 Department of Computational Biology & Centre for Computational Biology, IIIT-Delhi

2 SBILab, Department of Electronics and Communication Engineering, IIIT-Delhi

3 Laboratory Oncology Unit, Dr. B. R.A. IRCH, AIIMS, New Delhi

INTRODUCTION

- Multiple Myeloma (MM), a neoplasm of malignant plasma cells in the bone marrow (BM) , is preceded by the precancerous stage of Monoclonal Gammopathy of Undetermined Significance (MGUS).
- MGUS and MM both share several common genetic and genomic abnormalities that makes distinction between them challenging.

AIM

- The aim is (i) to identify genetic factors responsible for disease progression from MGUS to MM.
- (ii) Identify pivotal genetic biomarkers that distinguishes MGUS from MM.

METHODS

Materials:

Total tumor-normal matched pairs of MM samples: 1174 WES samples (1092 samples from MMRF CoMMpass Study (phs000748; phs000348) + 82 samples from AIIMS).

- **Total MGUS samples:** 61 WES samples (33 samples from EGA (EGAD00001001901) + 28 samples from AIIMS)

Method:

- **Variant Calling and annotation:** SNV is identified by 4 variant callers: MuSE [1], Mutect2 [2], Somatic-Sniper [3] and Varscan2 [4] and variant annotation using ANNOVAR [5].
- **Variant Filtration:** Filtered benign SNVs baesd on ClinVar database. In the analysis following 18 variant types are considered: Exonic, UTR3, UTR5, intergenic, synonymous, upstream, downstream, missense, inframe insertion, inframe deletion, frameshift, stop loss, stop gain, start lost, stop gained + frameshift, stop gained + inframe deletion, start lost + inframe deletion and protein altering variants.
- **Significantly mutated genes:** Identified significantly mutated genes from each variant caller individually for MGUS and MM using dndscv [6].
- Extracted top 250 mutated genes from each variant caller and took the union of top genes for MGUS and MM. This gave 1316 genes in total.
- **Feature matrix preparation:** Feature extraction for all 18 variant types for each gene gave a total of 162 features and the dimension of feature matrix is of 1316x162 for each patient. Dimensionality reduction was carried out for each gene using PCA that reduced the feature from 162 to 3. Finally, 3948 features for each patient was extracted.
- **AI-based workflow for classification:** Classification of MGUS vs. MM was carried out by taking five cost-sensitive models: (i) Random forest (ii) decision tree (iii) logistic regression (iv) SVM, and (v) XGBoost classifier. 5-fold cross validation was performed.
- **Extracting top genes:** The top ranking genes for MGUS and MM were obtained by mapping the corresponding top features from top-2 performing classifiers.

RESULTS

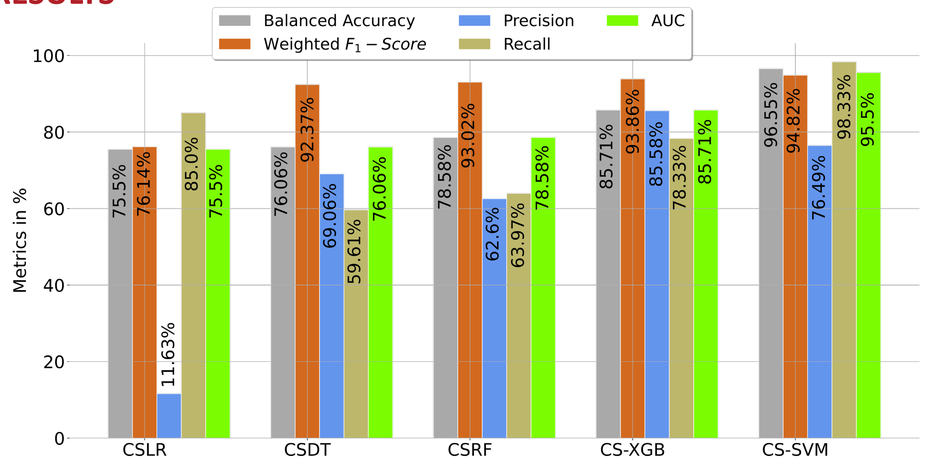

Figure 1: (A) Performance of five cost-sensitive ML models (CSLR: Cost-Sensitive Logistic Regression, CSDT: Cost-Sensitive Decision Tree, CDRF: Cost-Sensitive Random Forest, CSXGB: Cost-Sensitive and CS-SVM: Cost-Sensitive Support Vector Machine) on metrics of balanced accuracy, weighted F1-score, Precision, Recall, and AUC. (B) MCC scores of five cost-sensitive models.

Top mutated genes in MGUS and MM

Gene mutated in MM	Gene mutated in MGUS
HLA-DQB1 IRF1, MUC6, FGFR3, MUC4, HOXA1, ITPR3, HIST1H1E, MUC12, ITGA2, HLA-DQA2, HUWE1, IGLL5, HLA-DRB5, HLA-DQB2, ILK.	MUC3A, HLA-A, HLA-C, IRF4, JAK1, HDAC2, HLA-DQA1, FRG1, HS6ST1, H2AFV, and HLA-DRB1.

Table 1: List of top genes that were found mutated in MGUS (right)and MM (left)

DISCUSSION

- The cost-sensitive SVM (CS-SVM) classifier (Figure-1) outperformed all the other classifiers in terms of balanced accuracy (96.55%), weighted F1-score (0.9482), Matthew correlation coefficient (MCC, 0.8162,), precision (0.7649), recall (0.9833) and area under curve (AUC, 0.955).
- For MM, the top mutated genes got from top-2 performing classifiers are HLA-DQB1, IRF1, ITPR3, HOXA1, HIST1H1E, HUWE1, IGLL5, HIPK3, HLA-DQA2, HLA-DRB5, and ILK (Table-1).
- For MGUS, the top mutated genes got from top-2 performing classifiers are HLA-A, HLA-C, IRF4, JAK1, HDAC2, HLA-DQA1, HS6ST1, H2AFV, and HLA-DRB1 (Table-1).

CONCLUSION

- AI-based model is able distinguish MGUS and MM patients having several common genomic abnormalities in between them.
- Ai-based model is able to identify important mutated genes for MGUS (HLA-A, HLA-C, HDAC2, HLA-DQA1, etc.) and MM (i.e. HLA-DQB1, IRF1, HOXA1, HIST1H1E, etc.) that make both differentiable.

REFERENCES

1. Fan, Y., Xi, L., Hughes, D.S., Zhang, J., Zhang, J., Futreal, P.A., Wheeler, D.A. and Wang, W., 2016. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. Genome biology, e17(1), pp.1-11.
2. Benjamin, D., Sato, T., Cibulskis, K., Getz, G., Stewart, C. and Lichtenstein, L., 2019. Calling somatic SNVs and indels with Mutect2. Biorxiv, p.861054.
3. Larson, D.E., Harris, C.C., Chen, K., Koboldt, D.C., Abbott, T.E., Dooling, D.J., Ley, T.J., Mardis, E.R., Wilson, R.K. and Ding, L., 2012. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics, e28(3), pp.311-317.
4. Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L. and Wilson, R.K., 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome research, e22(3), pp.568-576.
5. Wang, K., Li, M. and Hakonarson, H., 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research, 38(16), pp.e164-e164.
6. Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K., Van Loo, P., Davies, H., Stratton, M.R. and Campbell, P.J., 2017. Universal patterns of selection in cancer and somatic tissues. Cell, 171(5), pp.1029-1041.

ACKNOWLEDGEMENTS

This work was supported by a grant from the 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. The authors would also like to thank the Centre of Excellence in Healthcare, IIIT-Delhi, for support in their research.

CONTACT INFORMATION

Vivek Ruhela (Email-id: vivekr@iiitd.ac.in)