An Ensemble Machine Learning and Graph Networks Approach to Identify Biomarkers Responsible for Immune-Related Adverse Events

Kamil Can **Kural**[1], Emma C. **Scott**[2], Sean **Smith**[1], Luis **Santana-Quintero**[1], Ilya **Mazo**[1], Dickran **Kazandjian**[2], Tigran **Ghazanchyan**[2], Svetlana **Petrovskaya**[2], Yong **Zhang**[2], Amy **Rosenberg**[2], V. Ashutosh **Rao**[2], Jennifer L. **Marté**[3], Marc R. **Theoret**[4], Richard **Pazdur**[4], James L. **Gulley**[3], Julia A. **Beaver**[2] and Konstantinos **Karagiannis**[1]

1 – CBER Office of Biostatistics and Epidemiology, CBER HIVE
2 – CDER OOD/OBP Translational Research Laboratory
3 – National Cancer Institute, National Institutes of Health
4 – FDA Oncology Center of Excellence

Introduction

Detection of immune-related adverse events (irAEs) is critical to the treatment of cancer patients receiving immune checkpoint inhibitor (ICI) treatment. Understanding biomarkers that predict irAEs can significantly improve prognosis for patients receiving ICI therapies. The overall objective of this study is to validate the use of next-generation sequencing technology and bioinformatics to inform the use of immunotherapy for cancer.

Methods

Our dataset consisted of whole exome sequencing of DNA samples with germline mutations from 101 patients treated with ICIs as part of clinical trials (NCT01772004, NCT02517398, NCT02155647) performed at NCI. We used HIVE[1] pipelines and GATK HaplotypeCaller with GRCh37 as the reference genome to identify the germline mutations. The initial stage of machine learning model development consisted of a thorough feature selection scheme to reduce dimensionality and select the mutated genes that contribute most to the variance in the dataset. We utilized five statistical tests and machine learning models for feature selection. Selected features (genes) were based on the degree of consensus of the tests. Using data from BIOGRID, a protein-protein network map was constructed with Neo4j graph database, built on top of Hetionet [2] and pathway analysis was conducted that takes account of protein-protein interactions.

References

1. High-performance integrated virtual environment (HIVE): a robust infrastructure for next-generation sequence data analysis. Simonyan V et al. Database Oxford. (2016)
2. Himmelstein, Daniel Scott, et al. "Systematic Integration of Biomedical Knowledge Prioritizes Drugs for Repurposing." Elife (2017).
3. LIME: Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?: Explaining the predictions of any classifier."2016.
4. Shapley sampling values: Strumbelj, Erik, and Igor Kononenko. "Explaining prediction models and individual predictions with feature contributions." Knowledge and information systems (2014)

Results

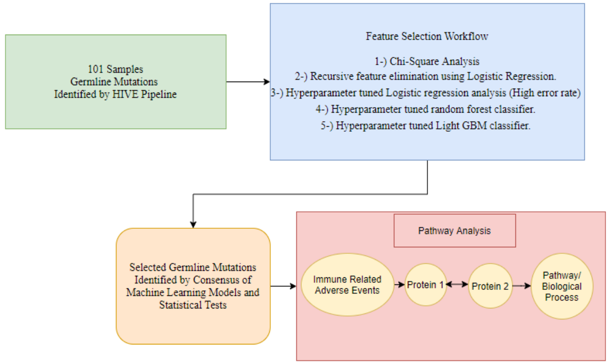
101 Samples
Germline Mutations
Identified by HIVE Pipeline
Feature Selection Workflow
1-) Chi-Square Analysis
2-) Recursive feature elimination using Logistic Regression.
3-) Hyperparameter tuned Logistic regression analysis (High error rate)
4-) Hyperparameter tuned random forest classifier.
5-) Hyperparameter tuned Light GBM classifier.
Selected Germline Mutations Identified by Consensus of Machine Learning Models and Statistical Tests
Pathway Analysis
Immune Related Adverse Events
Protein 1
Protein 2
Pathway/ Biological Process

Figure 1. irAE study analytical workflow . A ten-fold, stratified cross-validation LightGBM model trained by different thresholds identified 52 genes as the optimal number of features for the final model, predicting development of irAEs. The model is 100% accurate when it predicts on the randomly split, held out portion of the data with 21 samples.

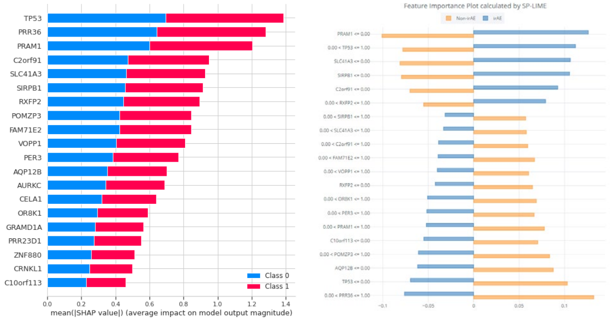

Figure 2.a. Shapley Feature Importance Plot

Figure 2.b. SP-LIME Feature Importance Plot

Figure 2.a and 2.b. Feature Importance for the final model, calculated by Shapley Additive Values and SP-LIME. The most important contributors to the classification performance are plotted using the LIME[3] and SHAP[4] packages.

Tests Conducted	Mean AUC for the features selected after conducted tests	Standard Deviation	Number of Total Genes
0	0.657	0.136	2996
1	0.641	0.166	457
2	0.622	0.139	244
3	0.714	0.173	131
4	0.742	0.225	52
5	0.842	0.150	7

Table 1: Feature selection workflow results and model performances with the criteria satisfying genes

go_id	go_name	PC	DWPC	n_genes
"GO:0050776"	"regulation of immune response"	138	7221100653677664.0	991
"GO:0060548"	"negative regulation of cell death"	173	7122618337074084.0	923
"GO:0001934"	"positive regulation of protein phosphorylation"	125	7052742894762540.0	999
"GO:1901701"	"cellular response to oxygen-containing compound"	128	7052306738753250.0	975
"GO:1902533"	"positive regulation of intracellular signal transduction"	128	6736515242973984.0	988
"GO:0010243"	"response to organonitrogen compound"	120	6521500788245640.0	940
"GO:0051276"	"chromosome organization"	193	6498705967736004.0	957
"GO:0043069"	"negative regulation of programmed cell death"	165	6350965673437296.0	859
"GO:0006468"	"protein phosphorylation"	147	6328224127037520.0	934
"GO:0043549"	"regulation of kinase activity"	137	6308078146792452.0	918
"GO:0043066"	"negative regulation of apoptotic process"	164	6284422222956000.0	850
"GO:0002684"	"positive regulation of immune system process"	116	6065185360956480.0	880
"GO:0045859"	"regulation of protein kinase activity"	134	5956800783885540.0	870

Table 2: Pathway analysis results for 52 identified genes

Tables 1 and 2 show the results of the pathway analysis and feature selection workflows. The pathway analysis was done for 52 genes that were picked up by at least 4 machine learning models and statistical tests.

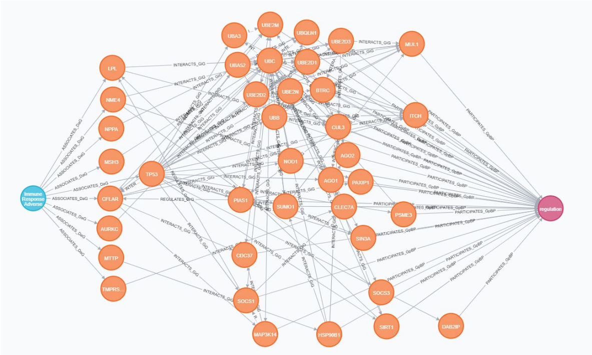

Figure 3. Regulation of Immune response. Pathway analysis conducted with the protein-protein Interactions network of 52 identified important genes elucidate the possible mechanisms behind immune related adverse events. (Not all genes in the graph are genes identified with the pipeline.)

Discussion/Conclusion

The pipeline support mechanism in HIVE enables rapid development and configuration of genomics pipelines for both scientific and regulatory applications. The Germline identification/NGS workflow as a part of HIVE is designed to facilitate and standardize the NGS analysis and biomarker identification.

Our results show the benefits of utilizing multiple machine learning models/statistical tests, which could be valuable in successfully identifying biomarkers responsible for irAEs. Conducting pathway analysis with the identified genes did not generate clear causal pathways likely due to inclusion of genes with unknown/distinct functions. However, by investigating protein-protein interactions of the candidate genes with germline mutations, we were able to identify potential contributing networks such as glucocorticoid signaling, B cell receptor signaling, CD40 signaling, role of PKR in interferon induction and antiviral response, macrophage activation, and neutrophil activation involved in immune response. Our plans include validating the results in a separate cohort in the future.