Informatics framework for clustering and deriving gene signatures for prognostic stratification of cancer patients

Segun Jung, Yingtao Bi, and Ramana V. Davuluri

Division of Health and Biomedical Informatics, Department of Preventive Medicine, Northwestern University Feinberg School of Medicine

Research Objective

- **Stratification of cancer patients into different molecular groups is essential for developing targeted therapies. High-throughput technologies have been extensively used for generating multi-omics data. Indeed, The Cancer Genome Atlas (TCGA; https://tcga-data.nci.nih.gov/tcga/) consortium has been accumulating large volumes of invaluable data. Despite the technological advances, analyzing and integrating the -omics data from different platforms, however, remains a challenge. We developed an informatics framework that integrates genomic and clinical data to stratify cancer patients into different molecular subgroups and predict clinically applicable phenotypes, such as survival.**

Methods

- **Dataset: 488 lung adenocarcinoma (LUAD) and 489 lung squamous cell carcinoma (LUSC)) were downloaded from TCGA data portal, from which 419 LUAD and 383 LUSC samples have relevant clinical information.**
- **To collect most variable and biologically relevant genes/isoforms, we used both mean absolute deviation (MAD) and Cox Proportional Hazards regression model with false discovery rate (FDR) < 0.05.**
- **To assign class labels, we used semi-supervised learning algorithm (k-means based consensus clustering) for 1000 runs with 80% sample-subsampling.**

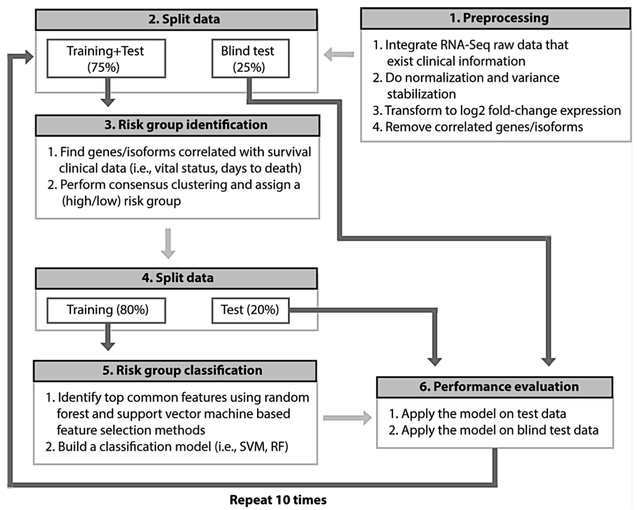

Figure 1. Overview of the computational pipeline.

- **For variable selection, we combined two recursive feature elimination algorithms based on support vector machine (SVM) and random forest (RF) to find clinically testable gene/isoform candidates.**
- **SVM and RF are the base algorithms to evaluate the risk prediction performance on two test dataset.**
- **The pipeline in Figure 1 is implemented in R language, incorporating with the following R packages: DESeq, Hmisc, caret, e1071, ConsensusClusterPlus, varSelRF, and msvmRFE**

Results

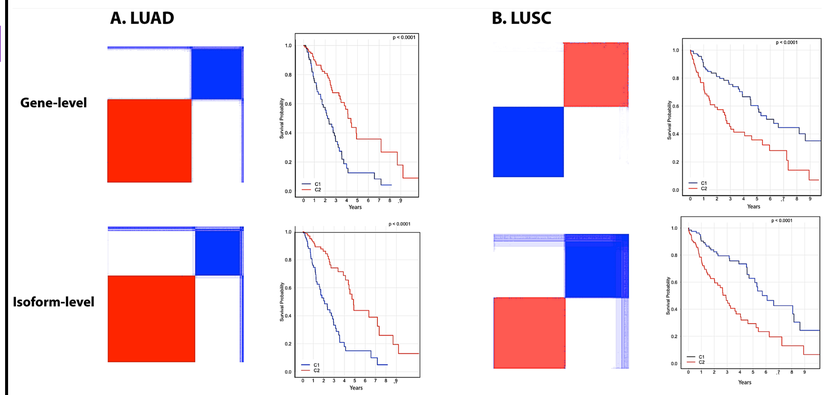

Figure 2. Consensus heatmaps show crisp clusters, associated with high- and low-risk groups, using 419 samples (1835 genes and 1092 isoforms) for LUAD and 383 samples (1064 genes and 626 isoforms) for LUSC, respectively.

Table 1. Average classification accuracy for 10 independent iterations. The number of features range from 9 – 75 (gene) and 9 – 120 (isoform) for LUAD, and 11 – 60 (gene) and 11 – 67 (isoform) for LUSC. We abbreviate Sensitivity, Specificity, and Positive Prediction Value as Sen, Spe, and PPV in the table, respectively.

Cancer Type		LUAD						LUSC					
		Gene (%)			Isoform (%)			Gene (%)			Isoform (%)		
		Sen	Spe	PPV	Sen	Spe	PPV	Sen	Spe	PPV	Sen	Spe	PPV
Test data	SVM	0.89	0.95	0.95	0.85	0.91	0.92	0.94	0.94	0.95	0.88	0.96	0.96
	RF	0.88	0.91	0.91	0.78	0.87	0.91	0.95	0.94	0.95	0.86	0.95	0.93

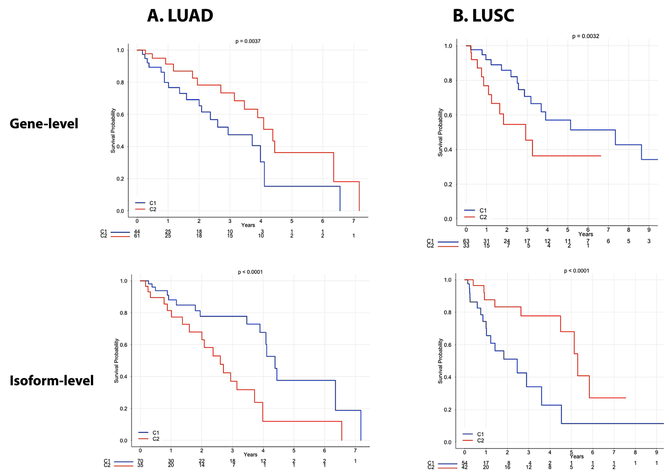

Figure 3. Classification models tested on blind test dataset (LUAD: 105 samples; LUSC: 96 samples) shows distinct risk difference with statistical significance using **45 genes and 24 isoforms for LUAD, and 60 genes and 14 isoforms for LUSC**.

Limitations

- Although we experimented multiple time the overall procedure to avoid any potential bias from data split (distribution) and clustering (randomness), we acknowledge it can only be overcome by experimental validation.
- Our pipeline is easy to maintain and update; however, current version is specifically designed for TCGA RNA-seq data where the raw data are processed by the rsem program.

Conclusions

- We implemented a pipeline (in R language) that automatically processes TCGA RNA-Seq raw data for high- and low-risk prediction (Figure 3).
- We tested our classification model on TCGA lung cancer data and achieved distinct risk separation with clinically testable gene/isoform dataset.
- We found top genes (LUAD: 19, LUSC: 16) and isoforms (LUAD: 22, LUSC: 20) for further experimental investigation.

NORTHWESTERN UNIVERSITY FEINBERG SCHOOL OF MEDICINE