Immunogenomics Report

Jing Wang

June 15, 2019

1 Motivation

This study aims to find mutated genes that are responsible for immune microenvironment and immunotherapy response. The problem is great importance. As we know, lots of patients with cancers have seen pronounced clinical response to immunotherapy, many more patients have shown minimal benefits to the same therapies [2]. The goal of our project is to find the real reason and personalize cancer immunotherapy.

2 Introduction of Data sets

TCGA cancer data set [5, 1] has 8,594 patients, each of which is described by 304 features including 299 gene symbols, cancer type, Gender, Race, Age, tumor mutational burden and Leukocyte fraction. There are totally 30 unique types of cancers in our data set, including OV, GBM, LUAD and so on. The distribution of cancer types are shown in Figure 2. It is shown that the number of patients for each cancer type is in the range of $[34, 911]$. There are an average of 286 patients for each caner type. We use a 30-dimensional zero-one vector to represent the cancer type feature where the position of one indicates the cancer type. The same procession is performed on the other clinical variables.

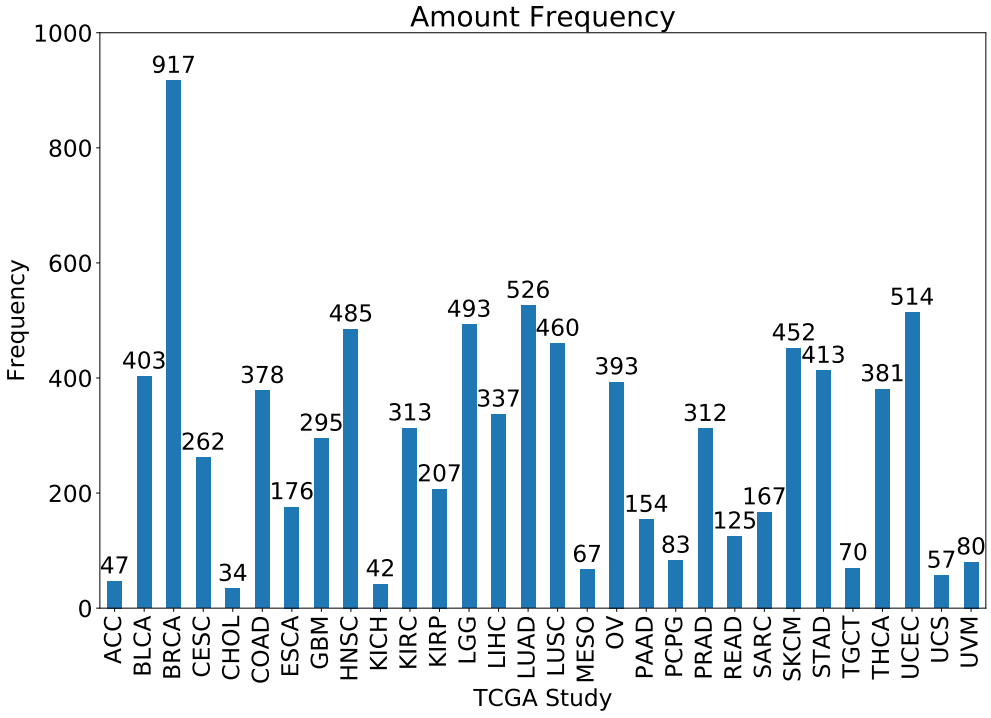

Figure 1: The distribution of patients with different cancer types.

MSKCC data set [6] has 1,662 advanced cancer patients treated with Immune checkpoint inhibitor (ICI) treatments. The patients miss genomic mutational information are removed. Each patient is represented by 474-dimensional genomic features and clinical variables including Age group (5), Sample Type (2), Cancer type (10), Drug type (3), Tumor Mutational Burden (1), Gender (2), Survival Status and Survival Time.

Goal. We select genes associated with Leukocyte fraction and survival time based on the TCGA data set and MSKCC data set independently, in the hope to see if there is an overlap between the two set.

3 Method

Notations. Let us instate several pieces of notation that are involved throughout the paper. We write bold capital letters such as $\mathbf{X}$ for matrices and its i-th column vector is denoted as $\mathbf{x}_i$. We use bold lowercase letters, such as $\mathbf{w}$, to denote a vector. The ℓ_1-norm and ℓ_2-norm of a vector $\mathbf{w}$ are denoted by $||\mathbf{w}||_1$ and $||\mathbf{w}||_2$ respectively. Given a data matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ (where n is the number of data points, and d is the feature dimension) and a target vector $\mathbf{y} \in \mathbb{R}^n$, we aim to learn a coefficient vector $\mathbf{w} \in \mathbb{R}^d$ which leads to an approximation $\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$ as close to the groundtruth $\mathbf{y}$ as possible.

We utilize elastic net and survival analysis methods including Kaplan-Meier and Cox for our data analysis. Here is a brief review of these methods.

- Elastic Net is a linear regression method that linearly combines the ℓ_1-norm and ℓ_2-norm penalties of the Lasso and ridge regression. It overcomes the limitations of Lasso when dealing with high-dimensional data with few data points. The objective function is given by:

$$\min \frac{1}{2n}||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2 + \alpha_1||\mathbf{w}||_2 + \alpha_2||\mathbf{w}||_2^2, \tag{1}$$

  where the parameters $\alpha_1 \geq 0$, $\alpha_2 \geq 0$ are tuned by cross-validation.

- Kaplan–Meier estimator [3] is a non-parametric statistic used to estimate the survival function from lifetime data. The estimator of the survival function $S(t)$ (the probability that life is longer than t) is given by:

$$S(t) = \prod_{i:t_i\leq t}(1 - \frac{d_i}{n_i}), \tag{2}$$

  with t_i a time when at least one event happened, d_i the the number of events (i.e., deaths) that happened at t_i and n_i the individuals known to have survived (have not yet had an event or been censored) up to t_i.

- Cox proportional hazards model, regularized by convex combinations of ℓ_1 and ℓ_2 penalties (elastic net)[4, 7]. The Cox model assumes a semi-parametric form for the hazard

$$h_i(t) = h_0(t)\exp(x_i^T\beta), \tag{3}$$

  where $h_i(t)$ is the hazard for patient i at time t, $h_0(t)$ is a shared baseline hazard, and $\beta \in \mathbb{R}^p$ is a vector. Inference is then made via the partial likelihood

$$L(\beta) = \prod_{i=1}^{m}\frac{\exp(x_{j(i)}^T\beta)}{\sum_{j\in R_i}\exp(x_j^T\beta)}, \tag{4}$$

where R_i is the set of indices, j, with those at risk at time t_i. [7] proposed an algorithm to find β which maximizes Equation (4) and subject to the constraint:

$$\alpha \sum |\beta_i| + (1-\alpha) \sum \beta_i^2 \leq c. \tag{5}$$

4 Result

TCGA dataset We plot the figures with coefficients learned by Elastic Net on samples with the same cancer subtype. We annotate the features with top 20 largest absolute coefficient values. Here are the examples on HNSC and BLCA cancer types.

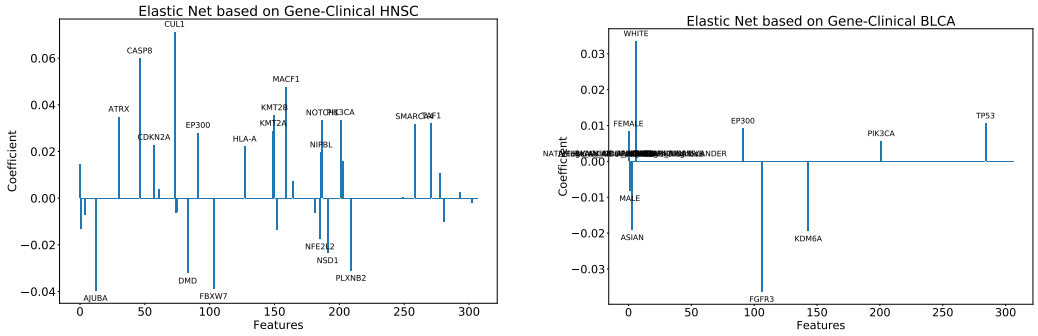

Figure 2: Coefficients of Elastic Net on HNSC and BLCA cancer subtypes.

MSKCC We plot the distribution of the data set and Kaplan-Meier curves with response to various variables in Figure 4, more results are shown in Figures 5.

We learn the coefficients of features learned by Cox model with ℓ_1 and ℓ_2 constraints. Figures shows the result on the samples with Renal Cell Carcinoma with ℓ_1 ratio 0.05.

5 Implementation

For the Elastic net, we utilize the scikit-learn `https://scikit-learn.org/stable/`. For the survival analysis, we use scikit-survival `https://github.com/sebp/scikit-survival`. Please refer to the websites for parameters tuning.

Here is the brief introduction of the codes:

- TCGA data set
  - clinical_dataprocessing_tcga
  - convert_txt_dataframe
  - elastic_gene: samples with genetic features
  - elastic_net_geneticlinic: samples with genetic and clinical features
  - elastic_net_geneclinicMutationBurden: samples with genetic, clinical and tumor mutational burden features.

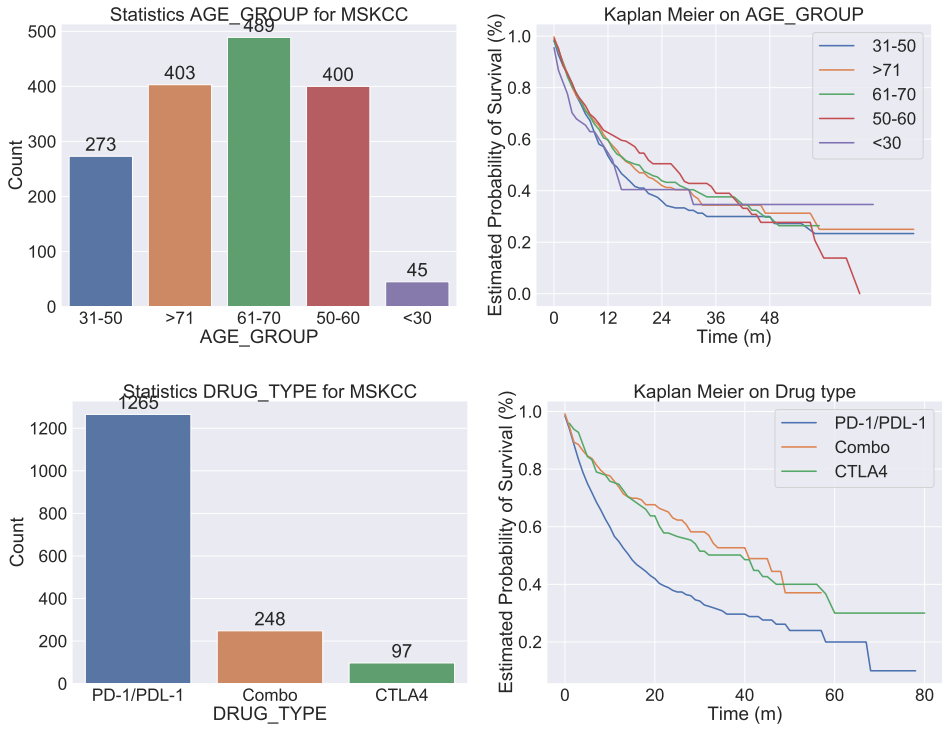

Figure 3: Effect of various variables on overall survival after ICI treatment-1

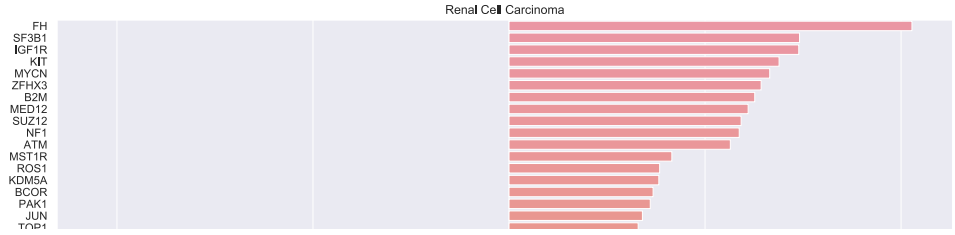

Figure 4: Part of coefficients of features learned by Cox model.

  - elastic_tcga
  - regression_data

- MSKCC data set
  - merge_two_files
  - plot_scripts
  - survival_analysis_cox: output log file "log_coxnet_0.5.txt" and "result_coxnet_Bladder Cancer_L1ratio_0.05.txt"
  - survival_analysis_kp
  - data_statistics_msk

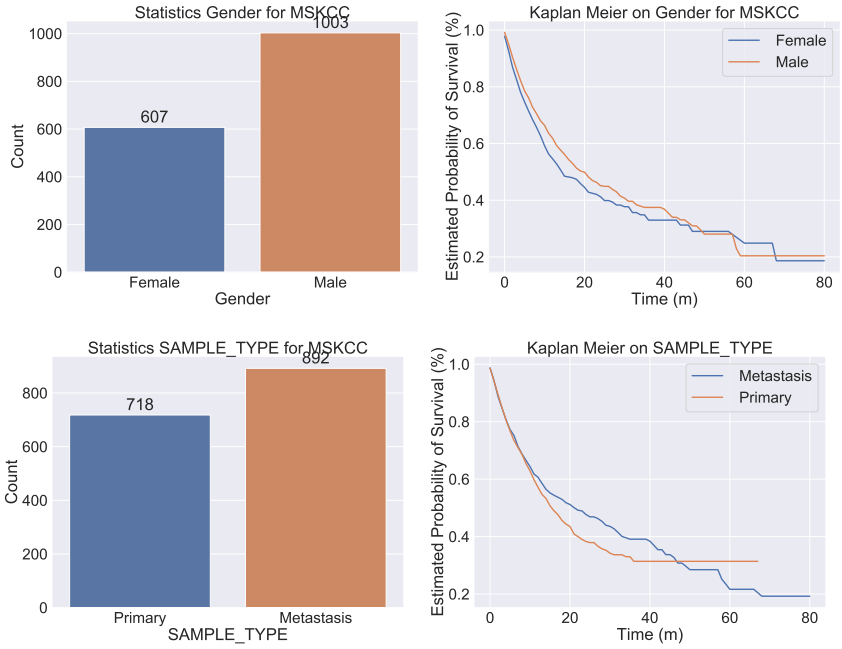

Figure 5: Effect of various variables on overall survival after ICI treatment-2

References

[1] Matthew H Bailey, Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, Michael C Wendl, Jaegil Kim, Brendan Reardon, et al. Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173(2):371–385, 2018.

[2] Mikhail Binnewies, Edward W Roberts, Kelly Kersten, Vincent Chan, Douglas F Fearon, Miriam Merad, Lisa M Coussens, Dmitry I Gabrilovich, Suzanne Ostrand-Rosenberg, Catherine C Hedrick, et al. Understanding the tumor immune microenvironment (time) for effective therapy. *Nature medicine*, 24(5):541, 2018.

[3] J Martin Bland and Douglas G Altman. Survival probabilities (the kaplan-meier method). *Bmj*, 317(7172):1572–1580, 1998.

[4] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[5] Kyle Ellrott, Matthew H Bailey, Gordon Saksena, Kyle R Covington, Cyriac Kandoth, Chip Stewart, Julian Hess, Singer Ma, Kami E Chiotti, Michael McLellan, et al. Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell systems*, 6(3):271–281, 2018.

[6] Robert M Samstein, Chung-Han Lee, Alexander N Shoushtari, Matthew D Hellmann, Ronglai Shen, Yelena Y Janjigian, David A Barron, Ahmet Zehir, Emmet J Jordan, Antonio Omuro, et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nature genetics*, 2019.

[7] Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software*, 39(5):1, 2011.