Cancer Genomics - Analysis Plan

Courtney Vaughn, Isaac Robson, Kentaro Hoffman, and John Sperger

March 26th, 2018

1 Introduction

Our group will analyze the publically available serous ovarian cancer data available through The Cancer Genome Atlas. The TCGA data includes progression-free survival time after treatment with Platinum-based. While there is no direct data on treatment regimen, ovarian cancer treatment plans are very standardized thus survival time correlates with treatment response allowing us to study the molecular basis for varying response to treatment. Given that Platinum-based chemotherapies act by damaging DNA, it is assumed that tumor cells that have better repair efficiencies. We will test this hypothesis by creating clusters of tumor samples based on DNA damage response gene expression and looking for differences in survival between the clusters.

2 Data & Pre-processing

There are RNA-Seq samples for 376 cases in the available data for the TCGA-OV project. We will restrict the sample to only those with clinical data available.

3 Analysis Plan

3.1 Subtype Replication

Our first planned analysis is a reproduction of the hierarchical clustering analysis[2] done on the mRNA data to replicate the four substypes identified in the initial report by the TCGA network [4]. This will show that the intial results are reproducible, and help ensure that any additional results we may discover are not due to differences in the pre-processing stage. This assumes the gene list is available somewhere. If we can't get the gene list we'll probably skip this step.

3.2 Identify gene list of DNA damage response genes

The DNA Damage Response (DDR) gene set was retrieved from the Gene Ontology Consortium. It contains 830 genes of which 808 appear in the TCGA sam-

ples. Gene list: `http://amigo.geneontology.org/amigo/search/bioentity?q=DNA%20damage%20response`

3.3 Clustering by DNA Damage Response

It is assumed that less efficient DNA damage repair will correlate with better outcomes for patients. We will test this hypothesis by exploring the difference in treatment response between tumors cluster by DDR gene expression patterns. We will use differential expression of the 830 DDR genes to create clusters. We will first create a hierarchical clustering as an exploratory analysis, and then will use consensus clustering to select the final number of clusters [3].

3.3.1 Cluster comparison

We will compare the DDR subtypes to the original unsupervised clusters by comparing an individual's unsupervised and DDR cluster membership and seeing the percentage of overlap across different clusters. We will particularly look to see if any DDR clusters overlap or provide additional sub-subtypes to the unsupervised subtypes.

3.4 Mining differential correlation

Differential Corrrelation Mining (DCM) is a method developed by Kelly Bodwin, Kai Zhang, and Andrew Nobel for identifying sets of variables where the average pairwise correlation between variables in a set is higher under one sample condition than the other [1] . We will split the mRNA-seq data into clusters based on the DDR subgroups identified in the previous stage of the analysis and run DCM to identify sets of differentially correlated genes across subtypes. We will then investigate the gene sets to identify potential underlying biological mechanisms. Existing analyses focus on gene counts, and it is possible for counts to remain similar while correlations change. This is potentially interesting because differential correlation could indicate that a group of genes which normally work together aren't functioning together in the same way in a particular subtypes. If a subtype has few members (say ¡10) we will exclude it from this analysis.

3.5 Survival Analysis

We will fit Kaplan-Meier Survival curves for each of the DNA damage response subtypes we identify to estimate the unadjusted difference in survival. Then we will fit a Cox proportional hazards model with the DNA damage response subtypes and clinical variables as predictors to estimate the effect of the DNA damage response subtypes on survival controlling for clinical factors like age. To better understand the predictive accuracy and the robustness of the model, we will split the data into training and test data sets using a 70/30 train/test

split and report the prediction error on the test set. If there are not enough events we will instead use 10-fold cross validation to assess performance.

3.6 Repair Pathway

If there are differences in survival between subgroups we will follow up to see which repair pathways are most important to drug response. This is an interesting question since people often assume repair is a mechanism of resistance to platinums but there really isn't any strong evidence for this.

References

[1] Kelly Bodwin, Kai Zhang, and Andrew Nobel. A testing-based approach to the discovery of differentially correlated variable sets, 2016.

[2] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, December 1998.

[3] Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning*, 52(1):91–118, July 2003.

[4] Cancer Genome Atlas Research Network et al. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609, 2011.