Identify significant co-occurring or mutually exclusive mutated driver genes across cancer types

Luchao (Leon) Qi, Dan Peng, Xi (Stanley) Wang **Introduction:**

Cancer, being one of the deadliest diseases throughout the world, is a topic of interest for many biomedical researchers around the world. One difficult aspect in cancer treatment is the dynamic nature of cancer genomes.[1] With the rapid development and decrease in cost of sequencing technology, researchers have been able to collect and sequence tens of thousands patient tumors. The Cancer Genome Atlas is a public funded database that contains genomic profiles of more than 30 human tumors types[2] . By harvesting the large publicly available tumor genomic data in TCGA, we can begin to elucidate relationship between mutations in driver genes.

Major tumor sequencing projects have been conducted in the past few years to identify genes that contain ‘driver’ somatic mutations in tumor samples. These genes have been defined as those for which the non-silent mutation rate is significantly greater than a background mutation rate estimated from silent mutations[3]. Compiling a comprehensive list of trusted cancer driver genes is imperative for oncology diagnostics and drug development. While driver genes are typically discovered by analysis of tumor genomes, infrequently mutated driver genes often evade detection due to limited sample sizes. Here, we identify driver genes by using TCGA cohorts, integrating tumor genomics data with a wide spectrum of gene-specific properties to search for trusted drivers, functionally classify them, and detect if there are significant co-occurring or mutually exclusive mutated driver genes across cancer types[4].

Further investigation in the relationship between driver genes may provide new insights into etiology and clinical management [5]. Specifically, we would like to discover co-occurring and mutually exclusive driver gene-pair within the same cancer type, and ideally find correlation between sets of such significant pairs among different cancers. One of the hypothesis we propose is that certain correlated driver gene-pair pattern might exist between cancers in similar tissues ? physical locations.

**Method:** ENCODING MUTATION MATRIX FROM TCGA DATABASE: To initiate the project, we will be extracting mutation annotation format datasets from TCGA database using TCGAbiolinks R package. From the raw MAF file, we will perform selection of non-hypermutated samples and reliable cancer driver genes, as suggested in previous studies[6, 7]. After the filtration of samples driver genes, we then encode the MAF file into a mutation matrix with rows representing driver genes and column representing patient bar code. Given the

scope of this class project and constraint of time, we will consider a gene mutated in a patient sample if it is is either missense, nonsense, SNP, INDEL) and not mutated if there is no mutation called, encoded as 1 or 0 in the mutation matrix. In addition, we will curate a meta-data table that will map the patient’s barcode to a specific cancer type to elucidate the effect of driver gene in a cancer type population.

SELECTION OF DIFFERENTIALLY MUTATED DRIVER GENES: We will also perform a selection of driver genes that are significantly mutated or not mutated in a cancer type population to filter out driver genes that are not uniform across the population. There are two methods that we proposed of doing, one is to use template matching method demonstrated previously in research[9] and PCA. Our modified way of template matching is to plot 2 driver genes and their encoded mutation status (0, 1) from two cancer types, and perform linear regression on the data points. We then would select the genes that produce the low residual sum of squares through linear regression to indicate that the mutation status of the drive gene is fairly uniform across population.

PCA ANALYSIS: Other method of reducing the number of genes to study is PCA. Once expression matrix is found, in order to reduce dimension, principal component analysis would be performed. Redundant dimensionality (i.e. genes) can be eliminated and thus improve convergence time and the quality of results[10].

ENCODING GENE-PAIR WITHIN SAME CANCER: In order to understand the driver genes between cancer type, we should first understand the driver genes relationship within a cancer category. We could utilize the gene pair transformation to better study the relationships between driver genes and cancer types[8]. We will encode each patient’s significant pairs, i.e. co-occurring and mutually exclusive pairs, and statistically summarize the pattern among all patients having such cancer. The proposed encoding rule is the following: encode driver gene-pair as 1 if both genes are mutated, 0 if one is mutated and the other is not, -1 if neither is mutated. The hypothesis testing would be done to determine whether if the result is significant. This would provide information on whether a gene pair is co-occurring or mutually exclusive within a cancer type. We then will compare the gene pair status across cancer types to have a more comprehensive understanding of driver gene pairs across cancer types.

**Reference:**

1. Hanahan, D. , & Weinberg, R. A. . (2000). The hallmarks of cancer. , 100(1), 0-70.

2. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an

immeasurable source of knowledge. Contemp Oncol (Pozn). 2015;19(1A):A68-77.

3. Waks, Z. , Weissbrod, O. , Carmeli, B. , Norel, R. , Utro, F. , & Goldschmidt, Y. . (2016).

Driver gene classification reveals a substantial overrepresentation of tumor suppressors among very large chromatin-regulating proteins. *Scientific Reports,6*(1), 38988.

4. Youn, A. , & Simon, R. . (2011). Identifying cancer driver genes in tumor genome

sequencing studies. *Bioinformatics, 27*(2), 175-181.

5. Remy, E. , Rebouissou, S. , Chaouiya, C. , Zinovyev, A. , & Calzone, L. . (2015). A

modeling approach to explain mutually exclusive and co-occurring genetic alterations in bladder tumorigenesis. *Cancer Research, 75*(19), 4042-4052.

6. Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe,

A., ... Ding, L. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*, *173*(2), 371–385.e18. https://doi.org/10.1016/J.CELL.2018.02.060 7. K. Ellrott, M.H. Bailey, G. Saksena, K.R. Covington, C. Kandoth, C. Stewart, J. Hess, S.

Ma, M. McLellan, H.J. Sofia, *et al.*Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines Cell Syst., 6 (2018), 10.1016/j.cels.2018.03.002 8. Chopra P, Lee J, Kang J, Lee S. Improving cancer classification accuracy using gene

pairs. *PLoS One*. 2010;5(12):e14305. Published 2010 Dec 21. doi:10.1371/journal.pone.0014305 9. Pavlidis P, Noble WS. Analysis of strain and regional variation in gene expression in

mouse brain. *Genome Biol*. 2001;2(10):RESEARCH0042. 10. Yeung, K. Y. , & Ruzzo, W. L. . (2001). Principal component analysis for clustering gene

expression data. *Bioinformatics,17*(9), 763-774.