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

Luchao Qi, Da Peng, Xi Wang

## Abstract:

Driver mutations are mutations within a gene that confers a selective growth advantages thus causing cancer. Knowing the co-occurence or mutually exclusive relationships between driver mutations can provide valuable information in studying the aberrant pathways in the formation of tumors. To test the mutually exclusive relationship, we utilized an existing R package tool called DISCOVER. However, currently there are no good tool to compute the co-occurrence between sample tests. As a result, we developed our own method of permutation test for finding significant co-occurring pairs. Our results have decent amount of overlap with existing literatures. At the meantime, we do identify some other key gene hubs in each cancer types that are not mentioned in those literatures. For example, HUWE1 in UCEC, APOB in cross cancer comparisons.

## 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[9]. 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 different cancers. However, due to time limitation, our invented correlation testing algorithm cannot produce fully confident result, and further verifications are encouraged before drawing conclusion.

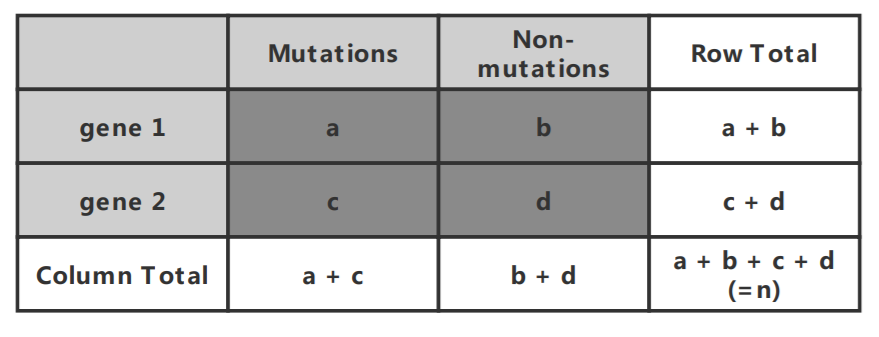
## Method:

#### ENCODING MUTATION MATRIX FROM TCGA DATABASE

We extracted mutation annotation format datasets from TCGA database using TCGAbiolinks R package. From the raw MAF file, we performed 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 encoded the MAF file into a mutation matrix with rows representing driver genes and column representing patient bar code. For non-mutation and silent mutation, we encode that as a 0, and for any other types of somatic mutation, we encode it as 1.

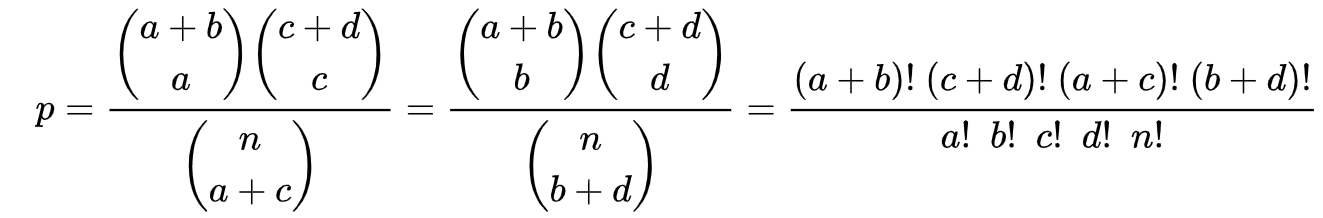
#### SELECT CO-OCCURRING DRIVER MUTATIONS: FISHER’S EXACT TEST

43956 combinations of gene pairs from 299 genes are selected to perform Fisher’s exact test. Before proceeding with the Fisher test, we first calculated the number of mutations and non-mutations in all gene pairs and denoted those numbers in a contingency matrix. We represented the cells by the letters a, b, c and d, called the totals across rows and columns marginal totals, and represented the grand total by n. So the contingency matrix looks like this:



*Figure 1. Contingency matrix of the encoded gene sequences. Letter a for the number of mutations in gene 1, b for number of non-mutations in gene 1, c for number of mutations in gene 2 and d for number of non-mutations in gene 2. The total number of mutations and non-mutations is n*

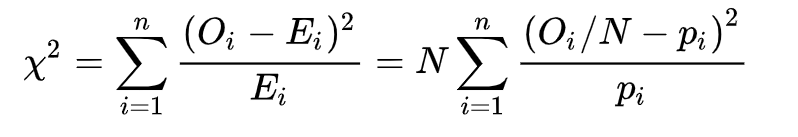
Fisher showed that the probability of obtaining any such set of values was given by the hypergeometric distribution:



We performed two-tailed test over all combinations from 299 genes and 4163 significant gene pairs were filtered out.

#### SELECT CO-OCCURRING DRIVER MUTATIONS: CHI-SQUARED TEST

Chi-squared test is based on the contingency matrix we have shown before in fisher’s exact test. When the sample size gets larger, chi-squared test performs better than fisher’s exact test. The value of the test-statistic is



where

*Oi* = Pearson's cumulative test statistic

*Ei* = the expected (theoretical) count of type i, asserted by the null hypothesis that the fraction of type i in the population is *pi* ( In contingency matrix *pi*  represents the proportion of mutation or non-mutation in each gene)

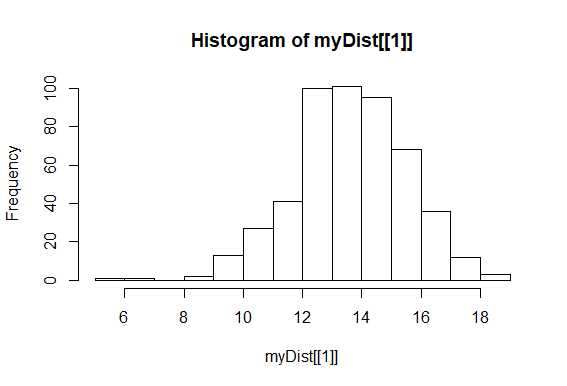
*n* = the number of cells in the table

*N* = Total number of mutations and non-mutations

Similar to the process in fisher’s exact test, two-tailed chi-squared test was performed over all combinations from 299 genes and 1568 gene pairs were filtered out.

#### SELECT CO-OCCURRING DRIVER MUTATIONS: PERMUTATION TEST

To select for co-occurring driver mutations, we designed our own permutation tests to test significance of all driver gene pairs. Our modified permutation test involves first selecting a pair of genes from the mutation matrix we encoded, randomly permute one of the gene row and count the number of co-occurrences in the permuted data. After 500 iterations, there will be a distribution in the number of co-occurrences of the permuted data in **Figure 2**. That distribution would be the null hypothesis.



*Figure 2. Sample distribution of permuted data using permutation tests*

After the distribution for the co-occurrences are calculated, we calculated the one tail p-value of our observed data. If the p-value of observed data is less than 0.05 then it is considered significant.

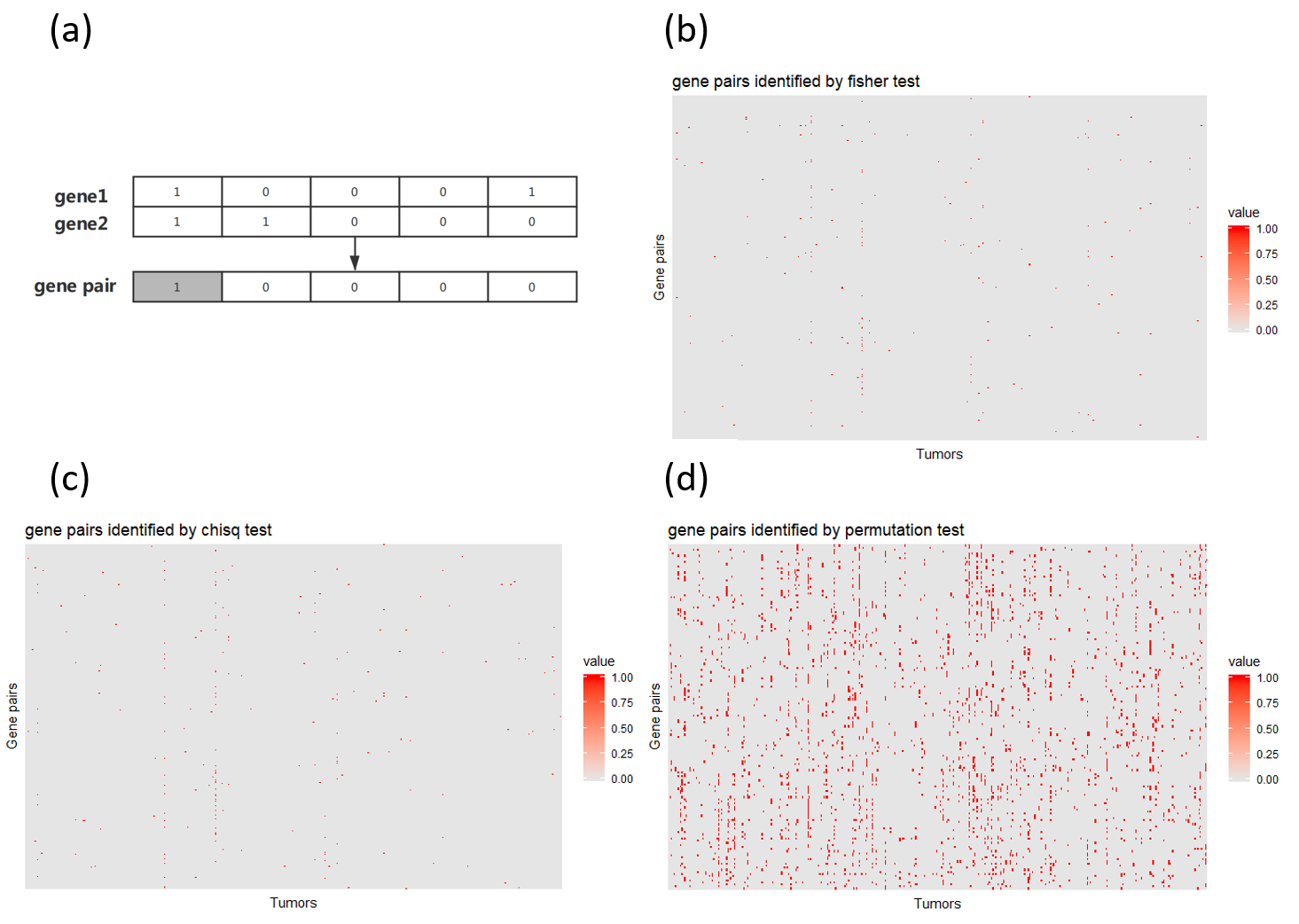
#### SELECT MUTUALLY EXCLUSIVE DRIVER MUTATIONS

We used an existing tool created called DISCOVER to find significant mutually exclusive gene pairs [1].

## RESULTS:

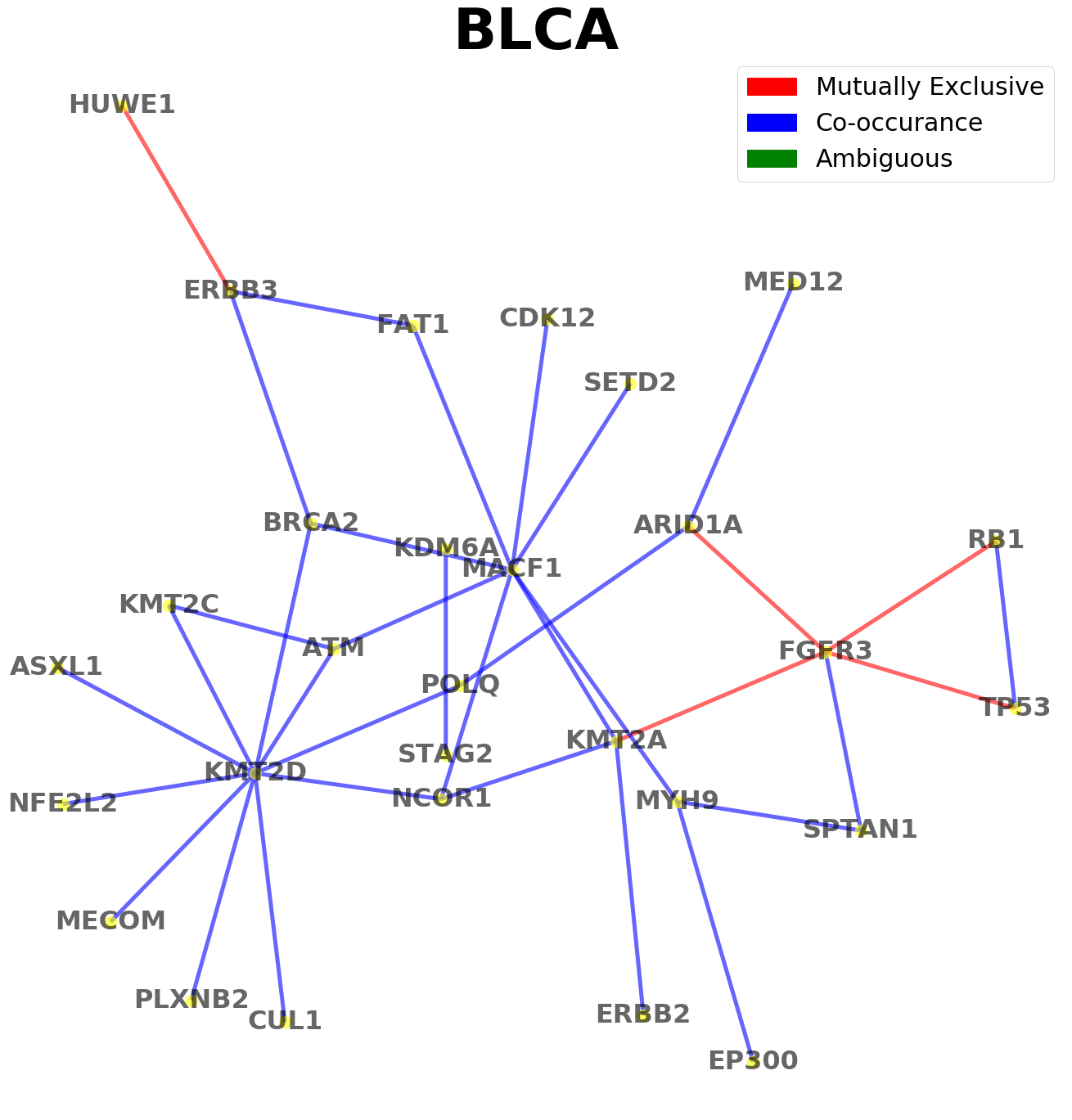
#### COMPARISON BETWEEN TRADITIONAL FISCHER AND OUR PERMUTATION TEST

After performing fisher’s exact test and chi-square test, significant gene pairs with co-occurring mutations are extracted for comparison. Genes with synchronous mutations in the same sample (i.e. gene 1 and gene 2 are both encoded as 1 in the same sample) are encoded as 1 for further visualization. We used heatmap to visualize the results after hypothesis testing. Gene pairs are colored in red for value 1 and non-colored for value 0 in heatmap. The density of red dots in permutation test is higher than that in fisher’s exact test and chi-squared test, which means the overlap of co-occurring mutations in gene pairs are more common in the results of permutation test. Overall, these results show that permutation test performs better than fisher’s exact test and chi-squared test with higher positive rate. Therefore we decide to use the custom permutation test for finding co-occurrences for this project.

**

*Figure 3. a. Gene pairs with synchronous mutations are encoded as 1 and 0 otherwise. b. After Fisher’s exact test, encoded gene pairs are visualized in heatmap. c. Heatmap of gene pairs after Chi-squared test. d. Compared to the results of fisher’s exact test and Chi-squared test, the heatmap of significant gene pairs after permutation test shows denser red dots which means higher true positive rate.*

#### BLCA CO-OCCURRENCE AND MUTUALLY EXCLUSIVE DRIVER GENES INTERACTION NETWORK



*Figure 4. The co-occurence and mutually exclusive driver gene interactions in BLCA*

Using the permutation method and DISCOVER tool, we selected the top significant driver gene pairs’ relationship pattern. Shown in **Figure 4**, we identified several important driver genes that act as hubs in the relationship network. FGFR3 signaling has been previously identified to be an important pathway in the formation of tumorigenesis in papillary bladder cancer [13]. Our model captured a significant co-occurrence relationship between FGFR3 and TP53 which has been reported in literature to have synergistic effect in cell proliferation and decrease the survival rate of patients [13]. In addition, RB1, as a tumor suppressor gene has been shown to have co-occurrence relationship with FGFR3 which suggests that the driver gene pairs have synergistic effect when both circuitries are aberrant. This relationship was also documented in previous literature indicating that a oncogene and tumor suppressor gene could provide evolutionary advantages in terms of tumorigenesis. [15]

Other hub driver genes that was uncovered through our model include KMF2D and MACF1 have also been discussed in literature to have mutually exclusive relationship with other driver genes suggesting that it is more evolutionarily more favorable to only have mutation in only one of those genes.

#### GLIOMA CO-OCCURRENCE AND MUTUALLY EXCLUSIVE DRIVER GENES INTERACTION NETWORK

|  |  |
| --- | --- |
|  |  |

*Figure 5. The co-occurence and mutually exclusive driver gene interactions in gliomas (lower grade glioma on the left, glioblastoma on the right)*

Another cancer type that we studied in depth is glioma (brain tumor). In comparison with current literature [11][12], we achieved high degree of overlap. For instance, the subnetwork in lower grade glioma, Figure 5 Left, involving TP53, ATRX, IDH1, CYC and EGFR was the result of a published studies and was accurately selected with the correct interaction relationship. However, an important hub gene that was discovered using our method is IDH2, which was not focused from previous studies. It has been previously shown that IDH2 is a critical biomarker for lower grade glioma hence known to have frequent interaction with other driver genes such as IDH1 and EGFR [12]. Although the relationship between IDH1 and APOB is unknown.

For glioblastoma (GBM), there is also a high amount of overlaps between the interaction networks found in literature and ours. The subnetworks involving EGFR, ATRX, PTEN, IDH1 and TP53 was also found in the relationship network that we constructed. In addition, we identified an important hub that further internet the established subnetwork, RB1. From previous studies, it has been identified that RB1 is an important tumor suppressor gene. Having any mutation that could disrupt its normal function can lead to uncontrolled cell proliferation and tumorigenesis. RB1 also appears to have mutually exclusive relationship between TP53 which makes sense biologically since TP53 is also a tumor suppressor gene. This suggests that genes or pathways with similar pathways might often have a mutually exclusive relationships.

#### OTHER CANCER TYPES

|  |  |
| --- | --- |
|  | b) |
| c) | d) |

*Figure 6. a) co-occurrence and mutually exclusive relationship in BRCA cohorts. b) co-occurrence and mutually exclusive relationship in UCEC cohorts. c) cross cancer type mutually exclusive network. d) cross cancer type co-occurrence network.*

We have performed many other cancer types testing and cross-cancer testing analysis. We are going to discuss BRCA and UCEC in details, and compare the results with a publication from Dao, et al. [14] that has performed systematic analysis on BRCA and UCEC as well.

Regarding BRCA, our co-occurrence analysis, in Figure 6(a), identifies several significant gene hubs, including TP53, CDH1, PIK3CA, MAP2K4, MAP3K1, AKT1, FOXA1. Interestingly, in Dao et al.’s study for mutual exclusivity between modules, different module-based network results were presented depending on functional module or co-occurrence module. The gene hubs in each modules are different depending on module type. For functional module study, most modules hubs of Dao’s network overlap with ours, with the exception of Module 5 (MUC4, MUC16, MUC12, MUC5B). However, Dao’s co-occurrence modules network is very different from its functional modules. As a result, only significant genes like TP53, MAP2K4, PIK3CA in our network are identified in Module 1 and 4 in Dao’s paper, whereas Dao’s Module 2, 3, 5 and 6 are missing in our network.

Regarding UCEC, Dao et al.’s study has presented a mutual exclusivity module network based on functionality. Our network, in Figure 6(b), has identified gene hubs like ARID1A, PIK3R1, TP53, CCND1, CTNNB1, PPP2R1A, PIK3CA. We found that our network overlapped half with the paper, with the exception of Module 1, 4 and 6. Since Dao’s paper studies both BRCA and UCEC cancer types, it is worth noticing that TTN has been listed in all Dao’s network, but our network failed to capture this important gene hub. Additionally, although the paper does not include a co-occurrence network result, our network found an interesting co-hub centering around HUWE1 and TP53. Specifically, our network proposes that HUWE1 and (ZFHX3, TAF1, SPTA1) have co-occurrence relationship, while TP53 and the same (ZFHX3, TAF1, SPTA1) have mutually exclusive relationship.

In addition, we have also performed cross-cancer testings. The DISCOVER paper [1] tested and analyzed on 12 Tumour Types, and we tested on 13 Tumour types, with 7 cancer types overlapping. Therefore it would be partially meaningful to compare the network results.

Regarding the conclusion on Co-occurrence, the DISCOVER paper believes that there is no evidence for widespread co-occurrence, and even with loosened threshold at FDR of 3%, the paper only stated one potential gene-pair: TP53 & MYC-amplified. According to Mitsui’s paper [10], MYC-amplified would usually occur together with EGFR and ERBB4, and we can see the relationship partially revealed by the hub APOB in Figure 6(c). In this aggregated network, we label the edge with number of repetitive occurrence across cancer types, In addition, we have identified other potential hubs colored, such as KMT2D & MACF1.

Our Mutually Exclusive Network, shown in Figure 6(d) is drawn in a similar manner. We have a great percent of overlapping with the paper’s network. Notethat our 13 cancer types and DISCOVER paper’s 12 cancer types only have 7 cancer overlapping. Achieving such similarity should be considered a solid result.

#### DISCUSSION

We have shown that using our method of permutation testing to identify significant yields much less false positive than the traditional methods of identifying co-occurring driver genes such as Fisher’s test or Chi-square test. Comparing with previous literature, we successfully reconstructed several relationship gene networks using our method of permutation test for finding co-occurrences and DISCOVER tool for finding mutually exclusive driver genes. In addition, we identified several important driver hub genes that are crucial for cancer development but was not mentioned in previous studies.

However, there are limitation in our method of permutation tests such as it is purely statistical without constraints of biological significance. In addition, the number of permutations that we performed is not sufficient enough to provide a very accurate estimation of the distribution. This problem can be alleviated with better computational power.

There are several driver gene hubs that was identified through this study. Well known cancer driving genes such as TP53 is the center hub gene for pan cancer mutually exclusive network. It suggests that TP53 is crucial in the development of many cancer and that many of the other genes in that network probably have similar functions. The center hub for co-occurence gene network is APOB.

Going forward, we hope to further validate our results that we found using our custom permutation test with current literatures or perform experimental testing for validation.

#### ACKNOWLEDGE

We would like to thank Dr. Rachel Karchin for her immense support.

#### REFERENCE

1. Canisius, S., Martens, J. W. M., & Wessels, L. F. A. (2016). A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence. *Genome Biology,* *17*(1), 261.
2. 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.
3. Shah, M. A., Denton, E. L., Arrowsmith, C. H., Lupien, M., & Schapira, M. (2014). A global assessment of cancer genomic alterations in epigenetic mechanisms. *Epigenetics Chromatin,* *7*(1), 29.
4. Cerchia, L., & Franciscis, V. D. (2007). Nucleic acid-based aptamers as promising therapeutics in neoplastic diseases. *Methods in Molecular Biology,* *361*, 187.
5. Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology,* *19*(1A), 68-77.
6. 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. *Sci Rep,* *6*(1), 38988.
7. Youn, A., & Simon, R. (2011). Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics,* *27*(2), 175-181.
8. Remy, E., Rebouissou, S., Chaouiya, C., Zinovyev, A., Radvanyi, F., & Calzone, L. (2016). A modeling approach to explain mutually exclusive and co-occurring genetic alterations in bladder tumorigenesis. *Cancer Research,* *76*(2), 4042-4052.
9. Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., & Weerasinghe, A., et al. (2018). Comprehensive characterization of cancer driver genes and mutations. *Cell,* *173*(2), 371.
10. Mitsui, Fumihiko, et al. “Non-Incidental Coamplification of Myc and ERBB2, and Myc and EGFR, in Gastric Adenocarcinomas.” Nature News, Nature Publishing Group, 13 Apr. 2007, [www.nature.com/articles/3800777](http://www.nature.com/articles/3800777).
11. A Shah, Muhammad & Denton, Emily & H Arrowsmith, Cheryl & Lupien, Mathieu & Schapira, Matthieu. (2014). A global assessment of cancer genomic alterations in epigenetic mechanisms. Epigenetics & chromatin. 7. 29. 10.1186/1756-8935-7-29.
12. Cohen AL, Holmen SL, Colman H. IDH1 and IDH2 mutations in gliomas. Curr Neurol Neurosci Rep. 2013;13(5):345. doi:10.1007/s11910-013-0345-4
13. Geelvink M, Babmorad A, Maurer A, et al. Diagnostic and Prognostic Implications of FGFR3high/Ki67high Papillary Bladder Cancers. Int J Mol Sci. 2018;19(9):2548. Published 2018 Aug 28. doi:10.3390/ijms19092548
14. Dao, P., Kim, Y. A., Wojtowicz, D., Madan, S., Sharan, R., & Przytycka, T. M. (2017). BeWith: A Between-Within method to discover relationships between cancer modules via integrated analysis of mutual exclusivity, co-occurrence and functional interactions. PLoS computational biology, 13(10), e1005695. doi:10.1371/journal.pcbi.1005695
15. Remy, Elisabeth, et al. “A Modeling Approach to Explain Mutually Exclusive and Co-Occurring Genetic Alterations in Bladder Tumorigenesis.” Cancer Research, American Association for Cancer Research, 1 Oct. 2015, cancerres.aacrjournals.org/content/75/19/4042.