**T.R.**

**ANTALYA BILIM UNIVERSITY**

**INSTITUTE OF POSTGRADUATE EDUCATION**

**DISSERTATION MASTER'S PROGRAM OF ELECTRICAL AND COMPUTER ENGINEERING**

# UTILIZING MUTUAL EXCLUSIVITY FOR THE IDENTIFICATION OF CANCER DRIVER GENE MODULES

**DISSERTATION**

**Prepared By**

**Rafsan Ahmed**

**Student ID: 181212003**

**ANTALYA–2020**

**T.R.**

**ANTALYA BILIM UNIVERSITY**

**INSTITUTE OF POSTGRADUATE EDUCATION**

**DISSERTATION MASTER'S PROGRAM OF ELECTRICAL AND COMPUTER ENGINEERING**

# UTILIZING MUTUAL EXCLUSIVITY FOR THE IDENTIFICATION OF CANCER DRIVER GENE MODULES

**DISSERTATION**

**Prepared By**

**Rafsan Ahmed**

**Student ID: 181212003**

**Dissertation Advisors**

**Assoc. Prof. Dr. Hilal Kazan**

**Prof. Dr. Cesim Erten**

**ANTALYA–2020**

# APPROVAL/NOTIFICATION FORM

# ANTALYA BILIM UNIVERSITY
# INSTITUTE OF POST-GRADUATE EDUCATION

Rafsan Ahmed, a M.Sc. student of Antalya Bilim University, Institute of Post Graduate Education, Electrical and Computer Engineering owning student ID 181212003, successfully defended the thesis/dissertation entitled "Utilizing Mutual Exclusivity for the Identification of Cancer Driver Gene Modules", which he prepared after fulfilling the requirements specified in the associated legislation, before the jury whose signatures are below.

**Academic Tittle, Name-Surname, Signature**

**Thesis Advisor:** Assoc. Prof. Dr. Hilal Kazan ,.....

**Thesis Co-Advisor:** Prof. Dr. Cesim Erten ,.....

**Jury Member:** Assoc. Prof. Dr. Nurcan Tunçbağ ,.....

**Jury Member:** Asst. Prof. Dr. Shahram Taheri ,.....

**Jury Member:** Asst. Prof. Dr. Deniz Gençağa ,.....

**Director of The Institute:** Prof. Dr. Ibrahim Sani Mert ,.....

Date of Submission : 28 / 08 / 2020
Date of Defence : 10 / 09 / 2020

# ÖZET

## KARŞILIKLI DIŞLAMA KULLANILARAK KANSER SÜRÜCÜ GEN MODÜLLERININ BULUNMASI

Büyük kanser kohortlarından alınan genomik analizler, sadece mutasyon profillerine dayalı olarak sürücü genlerin tanımlanmasını engelleyen mutasyonel heterojenlik problemini ortaya çıkarmıştır. Bu sorunu çozmenin bir yolu, genlerin fonksiyonel modüllerde birlikte hareket ettiği bilgisini kullanmaktır. Mevcut protein-protein etkileşim ağlarında bulunan bağlantı bilgisi, genlerin mutasyon frekansları ve kanser mutasyonlarının mutual exclusivity ozelliği ile birlikte el alınarak kanser sürücü modüllerinin etkin bir şekilde bulunması için kullanılabilir.

Bu tezde, protein-protein etkileşim ağlarında mevcut olan bağlantıları, karşılıklı dışlama ve kapsam bilgileriyle birleştirerek ayrıt ağırlıkları tanımlayan ve daha sonra bu ağırlıkları kullanan MEXCOwalk adında bir rastgele yürüyüş algoritması onerilmektedir. MEXCOwalk TCGA pan-kanser verileri üzerinde bilinen kanser genlerini bulma, normal ve tümor orneklerini sınıflandırma ve belirli kanser türleri için zenginleşmiş mutasyon profillerine sahip gen setlerini modül olarak tanımlama açıarıindan var olan yontemlerden daha iyi performans gostermektedir. Ayrıca, MEXCOwalk pan-kanser verilerinde nadiren mutasyona uğramış kanser genlerini tespit etmede de başarılı olmuştur. Tezin, ikinci kısmında, mevcut karşılıklı dışlama bulma algoritmalarını değerlendirme amacıyla ağ merkezli yenilikçi bir yontem geliştirilmiş; iyi çalışan karşılıklı dışlama bulma algoritmalarının çıkıları MEXCOwalk ayrıt ağırlıklarında kullanıdığında performansın daha da iyileştiği gozlemlenmiştir.

*Anahtar sözcükler*: Karşılıklı Dışlama, Epistasis, kanser sürücü genleri, rastgele yürüyüş.

# ABSTRACT

## UTILIZING MUTUAL EXCLUSIVITY FOR THE IDENTIFICATION OF CANCER DRIVER GENE MODULES

Genomic analyses from large cancer cohorts have revealed the mutational heterogeneity problem which hinders the identification of driver genes based only on mutation profiles. One way to tackle this problem is to incorporate the fact that genes act together in functional modules. The connectivity knowledge present in existing protein-protein interaction networks together with mutation frequencies of genes and the mutual exclusivity of cancer mutations can be utilized to increase the accuracy of identifying cancer driver modules.

We present a novel edge-weighted random walk-based approach that incorporates connectivity information in the form of protein-protein interactions, mutual exclusion, and coverage to identify cancer driver modules. MEXCOwalk outperforms several state-of-the-art computational methods on TCGA pan-cancer data in terms of recovering known cancer genes, providing modules that are capable of classifying normal and tumor samples, and that are enriched for mutations in specific cancer types. MEXCOwalk identifies modules containing both well-known cancer genes and putative cancer genes that are rarely mutated in the pan-cancer data. We then take this approach one step further by devising a network-centric epistasis framework to evaluate the estimated values from existing mutual exclusivity finding algorithms and applying these values to MEXCOwalk. We observe a significant improvement in the recovery of known driver genes.

*Keywords:* Mutual Exclusivity, Epistasis, Cancer Driver Genes, Random Walk.

# DEDICATION AND ACKNOWLEDGMENT

I would like to thank my graduate studies supervisors Assoc. Prof. Dr. Hilal Kazan and Prof. Dr. Cesim Erten for their expert guidance and feedback during the course of this project. The motivation and support they provided through their mentorship not only steered me through this research but also helped pave my career.

I would also like to share my gratitutde towards my fellow M.Sc. researchers Aissa Houdjedj, Ilyes Baali, Ahmed Amine Taleb Bahmed and Yacine Marouf for their feedback, criticism, cooperation and for creating a wonderful research environment in the lab. I would like to thank Evis Hoxha for developing the base of this research. I would also like to thank Cansu Yalçin for her encouragement.

Lastly, I would like to thank my family and friends for supporting me spiritually throughout my research. I would not be here without them.

This work and relevant publications have been supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) - project 117E879 to H.K and C.E.)

# Contents

# INSTITUTE OF POSTGRADUATE EDUCATION ELECTRICAL AND COMPUTER ENGINEERING MASTERS PROGRAM WITH THESIS

## ACADEMIC DECLARATION

I hereby declare that this master's thesis titled "Utilizing Mutual Exclusivity for the Identification of Cancer Driver Gene Modules" has been written by myself under the academic rules and ethical conduct of the Antalya Bilim University. I also declare that the work attached to this declaration complies with the university requirements and is my work. I also declare that all materials used in this thesis consist of the mentioned resources in the reference list. I verify all these with my honor.

28/08/2020

Rafsan Ahmed

# List of Figures

# List of Tables

# ABBREVIATIONS

# CHAPTER 1

# 1. Introduction

Cancer is defined as the uncontrolled growth of abnormal cells in the body. It is a complex disease that affects a large number of people in the world. Second only to cardiovascular disease, cancer is one of the leading causes of death in the world, killing approximately 10 million people annually. The reason behind the complex nature of cancer is mutational heterogeneity, which suggests that different genes mutate in different tumors, making it difficult for a commmon targeted therapy.

Recent advances in high-throughput DNA sequencing technology have opened up a Pandora's box which allowed several projects such as the TCGA [1] to construct and release genomic data from thousands of tumors. This further gave rise to the design of several computational approaches for the systematic detection of cancer-related somatic mutations - genetic mutations that are not inherited but acquired after an individual is born. These computational approaches focus on prioritizing independent genes to provide a genomic landscape that facilitates the identification of hypothesized cancer driver genes, referring to genes that are causally linked to oncogenesis [2, 3, 4, 5]. However, mutational heterogeneity in cancer causes some genes to mutate in a large number of samples and the rest to mutate in fewer samples. This leads to something called a long tail phenomenon. If a certain computational approach for driver gene discovery is applied to such data without additional information, the majority of genes discoverd will be from the set of highly mutated genes and may have a bias against the genes in the long tail region.

Many recent computational methods integrate somatic mutation data with additional information in the form of interaction networks (e.g. protein-protein interaction network orPPI) or gene expression data. Since genes do not function alone but work interactively, focusing on funcional modules instead of individual genes allow us to get additional information. Although such gene rankings provide valuable insight regarding potential genes of interest, in many cases mutations at different loci could lead to the same disease [6]. This genetic heterogeneity may reflect an underlying molecular mechanism in which the cancer-causing genes form some kind of functional pathways or candidate driver modules. Genomic data has shown that, many cancer driver genes are mutated in tumors with a small number of mutations [7]. In other words, certain genes belonging to the same pathway do not mutate together in the same tumor, since either of them can lead to oncogenesis. This phenomenon is called Mutual Exclusivity. This mainly occurs due to the selection pressure and fitness advantage on genes. Since it is established that driver genes tend to be mutually exclusive in the same tumor, mutual exclusivity can be concurrently used as additional information to identify driver gene modules from genomic data.

In this thesis we aim to first identify cancer driver modules from genomic data by utilizing mutual exclusivity and interaction data. We develop a novel cancer driver module finding algorithm, MEXCOwalk, that utilizes this information. Secondly, we develop a network-centric framework to evaluate existing algorithms for finding mutual exclusivity. We then apply the mutual exlusivity values estimated by these algorithm to MEXCOwalk to assess whether they improve over original MEXCOwalk.

## 1.1. Contribution

We provide an overview of the research contributions in this thesis towards the problem of driver module identification in cancer as well as comparing different mutual exclusivity finding algorithms and finding the best outcome. We provide a novel combinatorial optimization problem definition to detect driver modules in cancer. We then develop an algorithm, MEXCOwalk, that identifies such driver modules in cancer from genomic data, by utilizing mutual exclusivity and coverage information. Next, we look at existing mutual exclusivity finding algorithms and evaluate the methods based on a novel network-centric framework. Finally, we conduct a thorough evaluation of the results, first by comparing the results from MEXCOwalk against other driver gene module finding algorithms,

followed by applying the estimated mutual exclusivity values from the existing mutual exclusivity finding algorithms to MEXCOwalk and evaluating the results.

## 1.2. Thesis Organization

The thesis is organized as follows:

- In Chapter 2, we provide a review of cancer biology including mutual exclusivity and epistasis. We also review driver gene module finding algorithms and mutual exclusivity algorithms.
- In Chapter 3, we establish a formal problem definition in identifying driver gene modules and discuss the proposed algorithm, MEXCOwalk. We then present the results of MEXCOwalk in comparison to state-of-the-art driver gene module finding algorithms and discuss the key biological insights of the results.
- In Chapter 4, we establish a formal problem definition for evaluation of statistical mutual exclusivity finding algorithms. We then present the results from our proposed network centric evaluations of mutual exclusivity algorithms.
- Finally, in Chapter 5, we discuss the results and summarize the thesis.

## 1.3. Publications

- R. Ahmed, I. Baali, C. Erten, E. Hoxha, and H. Kazan, "MEXCOwalk: mutual exclusion and coverage based random walk to identify cancer modules," *Bioinformatics*, vol. 36, pp. 872–879, 08 2019.

# CHAPTER 2

# 2. Background

In this chapter, we aim to cover the fundamental computational and biological aspects behind the thesis.

## 2.1. Mutational Heterogeneity in Cancer

Cancer is a disease caused by the uncontrollable cell division of abnormal cells. This occurs due to the accumulation of somatic mutations in different genes that control cell proliferation. Mutational heterogeneity in cancer allows different tumors to have mutations in different sets of genes. Even different cancerous tumors in the same patient can have mutations in significantly different genes. This creates the afforementioned *long tail phenomenon* which is when we observe some genes having mutations in a large number of samples and the rest having mutations in fewer samples. Eventually when we attempt to discover cancer driver genes, a selection advantage is present for genes with large mutation frequency. Since the mutational heterogeneity problem hinders the identification of cancer driver genes with low mutation frequency, additional information is required to discover such genes.

## 2.2. Mutual Exclusivity

Mutual exclusivity refers to the phenomenon that for a group of genes which exhibit evidence of shared functional pathway, simultaneous mutations in the same tumor are less frequent than is expected by chance [8]. When there is a mutation in a specific cancer driver gene, it is a causality of the presence of complimentary mutations in the

same pathway that act synergistically with that driver gene. Alternatively, multiple driver gene mutations in the same pathway is less likely to occur because they lead to the same functional effect [7]. Mutual exclusivity among genes generally arise from epistatic interaction, where the selection pressue is strong for a particular gene. However, mutual exclusivity can also arise from underlying biological properties, such as tumor subtypes.

## 2.3. A review of driver gene finding algorithms

Several computational methods have been suggested for the identification of candidate driver modules [9, 10, 11]. The module identification approaches as applied to cancer can be viewed in two broad categories based on the types of input data they employ: the de novo methods and knowledge based methods. Following sections contain brief descriptions of each method.

### 2.3.1. De novo methods

The de novo methods rely only on genetic data to discover novel genetic interactions, as well as cancer-related functional modules [12, 13, 14, 15]. Due to the large solution space such methods usually apply a prefiltering based on alteration frequency to reduce the inherent computational complexity which may reduce sensitivity by overlooking modules involving rare alterations [9].

### 2.3.2. Knowledge-based methods

Knowledge-based methods, in addition to genomic data, incorporate prior knowledge in the form of pathways, networks and functional phenotypes to identify driver modules. Such methods can be subcategorized based on the optimization goals set within the computational problem formulations they employ in defining the biologically motivated cancer driver module identification problem. The first subcategory consists of methods including Hotnet [16], Hotnet2 [17], Hierarchical Hotnet [18] which utilize the fact that a driver pathway tends to be perturbed in a relatively large number of patients. These methods informally optimize the coverage of the modules as identified by the mutation frequencies of the comprising genes over a cohort of samples constitutes an informal optimization goal. Heat-diffusion over an interaction network that diffuses the mutation frequencies throughout the network is a common attribute in these methods. The resulting

diffusion values are then employed to extract modules exhibiting a large degree of connectedness as formulated with an appropriate graph-theoretical connectivity definition, usually the strong connectivity.

The second subcategory of knowledge-based module identification methods incorporate an appropriate definition mutual exclusion in their computational problem formulations. Several cancer module identification methods incorporate this observation in the employed combinatorial optimization problem definitions. In MEMo, maximal clique extraction in a similarity graph derived from an interaction network or functional relation graph is used and the maximal cliques are postprocessed taking into account the mutual exclusivity results [19]. In Babur et al. [20] a method based on seed-and-growth on a network, where the growth strategy is determined with respect to a suitably defined mutual exclusion score is proposed to identify pan-cancer modules using TCGA data. BeWith proposes an ILP formulation that combines interaction density within a module and several mutual exclusion definitions as a maximization goal [21]. The ILP formulation incorporates constraints in the form of desired number of modules and maximum number of genes per module. MEMCover combines pairwise mutual exclusion scorses with confidence values of interactions in the network [22]. To maximize high-confidence interactions, mutual exclusion, and coverage simultaneously; heavy subnetworks covering every disease case at least k times are found following a greedy iterative seed-and-growth heuristic.

## 2.4. A review of statistical mutual exclusivity finding algorithms

As discussed previously, many computational methods aim to utilize mutual exclusivity for precision medicine and targeted therapy. Some statistical mutual exclusivity tests are based on the assumpton that genes' alterations across tumors are identically distributed. We choose five popular statistical mutual exclusivity methods: DISCOVER [23], Fisher's Exact Test, WExT [24], MEMO [19] and MEGSA [25]. Each of these methods outputs a p-value for each gene pair, where a small p-value indicates that the mutation profiles of the corresponding gene pair exert a much smaller overlap and thus larger mutual exclusivity than it would be expected by chance.

DISCOVER is a statistical independence test that utilizes tumor specific alteration probabilities, which is in contrast with many existing tests that consider identically distributed events. The method is able to detect significant mutual exclusivities without increasing the false positive rate. WExT is a weighted exact test that aims to find highly significant mutual exclusivities by considering per-event and per-sample mutation probabilities. Unlilke the other mutual exclusity finding algorithms, WExT employs a highly accurate saddlepoint approximation was used to calculate the p-values for pairwise mutual exclusivity without going through computationally expensive permutations tests. The MEMO algorithm uses mutual exclusivity to identify candidate drivers networks. For pairwise mutual exclusivity, a bipartite graph of patients and genes is constructed from the mutation data. An edge swapping method is applied and the results are compared with the original graph to derive p-values for pairwise mutual exclusivity. The algorithm was reimplemented from scratch to get pairwise mutual exclusivity and fit our data. Following MEMO [19], the edges from the binary matrix were swapped 100*edges times and this permutation was done 10000 times for each cancer type. MEGSA is a framework that extends pairwise analysis and searches for MEGS (Mutually Exclusive Gene Sets). The pairwise mutual exclusivity p-values for MEGSA are calculated applying chi-square cumulative probability less than or equal to the value of the log likelyhood calculated by the *funestimate* function.

Some statistical mutual exclusivity tests are based on the assumption that genes' alterations across tumors are identically distributed. Among the approaches considered in this study Fisher's Exact Test and MEGSA [25] belong to this category. However it has been observed that the number of alterations per tumor can vary quite considerably, even in tumors of the same type; colorectal tumors with microsatellite stability have a median of 66 non-synonymous mutations, but colorectal tumors with microsatellite instability have a median of 777 mutations [24]. It has been shown that under such settings the mutual exclusivity tests relying on identical alteration probabilities across tumors may lead to reduced sensitivity for mutual exclusivity analysis [23]. The effects of varying alteration probabilities on pairwise mutual exclusivity calculations have been formalized within the context of the so-called mutation load confounding (MLC) in a recent study by van de

Haar et al. [7]. MLC is a correlation between the number of statistically significant mutual exclusivity findings and the mutation load association (MLA) of a gene, where logistic regression is used to compute MLA as a standardized score of association between the mutation likelihood of each gene and the mutation load, that is the genome-wide number of somatic mutations observed in a tumor. Note that negative MLA values correspond to higher mutation frequencies in tumors with low mutation loads, whereas positive values correspond to higher mutation frequencies in tumors with high mutation loads. Strong negative correlations between the MLA of a gene and the number of statistically significant pairwise mutual exclusivities have been observed, implicating the finding that the more negative a gene's MLA, the higher the number of other genes that show mutual exclusivity with that particular gene [7].

However, such a negative correlation does not always imply true epistasis since a gene highly mutated in tumors wih low mutation loads, naturally has a better chance of forming mutually exclusive pairs with other genes. Thus extra sources of information are necessary to filter out the pairs with true epistasis relations among a set of statistically significant pairwise mutual exclusivities postulated by some exclusivity test. van de Haar et al. [7] make use of the subtype information for such a purpose and show that the mutation load confounding can be reduced by correcting via tumor subtype stratification. Such a correcton greatly reduces the number of gene pairs reported to show mutual exclusivity, especially for pairs that include genes with low MLA. A major drawback is the absence of subtype information for many tumors.

# CHAPTER 3

# 3. MEXCOwalk

In this chapter, we provide a formal definition of the cancer driver gene module discovery problem and discuss our proposed algorithm MEXCOwalk. We then compare the results of MEXCOwalk on pan-cancer data against existing state of the art driver gene module finding algorithms.

## 3.1. Methods

The following sections include a detailed description of the methods for MEXCOwalk.

### 3.1.1. Problem Definition

We provide a novel combinatorial optimization problem definition to detect driver modules in cancer. Such a definition is not only important for algorithmic purposes but also to serve as a measure of performance for alternative methods suggested for the problem.

Let $S_i$ denote the set of samples for which gene $g_i$ is mutated. Let $G = (V, E)$ represent the PPI network where each vertex $u_i \in V$ denotes a gene $g_i$ whose expression gives rise to the corresponding protein in the network and each undirected edge $(u_i, u_j) \in E$ denotes the interaction among the proteins corresponding to genes $g_i, g_j$. Henceforth we assume that $g_i$ denotes both the gene and the corresponding vertex in $G$.

Let $M \subseteq V$ be a set of genes denoting a *module*. We define the mutual exclusion of $M$

as,

$$MEX(M) = \frac{|\bigcup_{\forall g_i \in M} S_i|}{\sum_{\forall g_i \in M} |S_i|}$$

and the coverage of $M$ as,

$$CO(M) = \frac{|\bigcup_{\forall g_i \in M} S_i|}{|\bigcup_{\forall g_i \in V} S_i|}.$$

Let $P = \{M_1, M_2, \ldots M_r\}$ be a set of modules. Let $RS(M_q)$ denote the relative size of a module $M_q$ with respect to the total size, that is $RS(M_q) = \frac{|M_q|}{|\cup_{\forall M_t \in P} M_t|}$. We define the mutual exclusion score and the coverage score of a set of modules, so that each module $M_q$ contributes its share proportional to its relative size $RS(M_q)$ for the former, whereas for the latter the contribution of $M_q$ is proportional to the normalized value of $1 - RS(M_q)$. Intuitively, a large module with high mutual exclusion score should be rewarded, since as the size of the module increases the chances of achieving better mutual exclusion decrease. Analogously, a small module with high coverage score should be rewarded. Thus we define the mutual exclusion score of $P$ as,

$$MS(P) = \sum_{\forall M_q \in P} RS(M_q) \times MEX(M_q).$$

The coverage score of $P$ is defined as

$$CS(P) = \sum_{\forall M_q \in P} \frac{1 - RS(M_q)}{\sum_{\forall M_t \in P} 1 - RS(M_t)} \times CO(M_q),$$

if $|P| > 1$ and $CS(P) = CO(M_1)$, if $|P| = 1$.

For a graph $H$ and a set $M_q$ of genes, let $H(M_q)$ denote the subgraph of $H$ induced by the vertices corresponding to genes in $M_q$.

*Cancer driver module identification problem:* Given as input a PPI network $G$, $S_i$ for each gene $g_i$, integers $total\_genes$, and $min\_module\_size$, find a disjoint set of modules $P$ that maximizes the *driver module set score* defined as,

$$DMSS(P) = MS(P) \times CS(P) \tag{3.1}$$

and that satisfies the following:

1. For each $M_q \in P$, $G(M_q)$ is connected.
2. $|\bigcup_{\forall M_q \in P} M_q| = total\_genes$.

3. $min_{\forall M_q \in P} |M_q| = min\_module\_size$.

**Theorem 1.** *Cancer driver module identification problem is NP-hard.*

*Proof.* The transformation is from *Set Packing* which is shown to be NP-complete. In the Set Packing problem, given a collection $C$ of finite sets and a positive integer $K \leq |C|$, the problem is to find out whether $C$ contains at least $K$ mutually disjoint sets. The problem is NP-hard even when the size of each set is at most 3, which can easily be extended to the setting where the size of each set is exactly 3. Given an input to the Set Packing problem within this setting in the form of $K$ and $C$ such that for each $S \in C$, $|S| = 3$, we generate $G$ as a complete graph on $|C|$ vertices, corresponding to the set of genes, such that each finite set in $C$ corresponds to a set of samples $S_i$ for which gene $g_i$ is mutated. We set both $total\_genes$ and $min\_module\_size$ to $K$. The answer to the Set Packing problem is Yes, if and only if the maximized score of the cancer module identification problem is exactly $\frac{3 \times K}{|\bigcup_{\forall g_i \in V} S_i|}$. □

### 3.1.2. MEXCOWalk Algorithm

Since the problem is computationally intractable we provide a polynomial-time heuristic approach based on vertex and edge-weighted random walks on special graphs that incorporate mutual exclusion and coverage information as vertex and edge weights in the H.Sapiens PPI network. The pseudocode of our method is provided in Algorithm 1. There are three main steps of the algorithm, each of which is described in detail in the following subsections.

#### 3.1.2.1. Weight Assignment with MEX and CO

Given a PPI network $G = (V, E)$, we first construct a weighted graph $G_w$ that contains properly defined weights for vertices and edges. For each $g_i \in V$ we assign a weight, $w(g_i) = CO(\{g_i\})$, thus the weight corresponds to the mutation frequency of a gene. It represents the heat to be diffused from that vertex during the random walk procedure.

The weight of an edge incident on a $g_i$ should on the other hand reflect the ratio of heat transferred to $g_i$'s neighbors at each step of the random-walk. We

**Algorithm 1** *MEXCOwalk*

**Input:** PPI network $G = (V,E)$, $S_i$ for each gene $g_i$, integers $total\_genes, min\_module\_size$ and threshold $\theta$ with $0 < \theta \leq 1$.
**Output:** Set of driver modules $P$.

```
//1. Weight Assignment with MEX and CO
```
construct $G_w$ by assigning a weight to each $g_i \in V, e \in E$
```
//2. Edge-Weighted Random Walk
```
construct $G_d$ by applying weighted-random walk on $G_w$
```
//3. Constructing Set of Driver Modules
//Initial Candidate Modules
```
**repeat**
    $P = SCC(G_d)$
    remove $M_q \in P$ with $|M_q| < min\_module\_size$
    remove min-weight edge from $G_d$
**until** $|\bigcup_{\forall M_q \in P} M_q| == total\_genes$
```
//Split-and-extend on Large Modules
```
$split\_size = \max_{\forall G_d(M_q)} outdeg(G_d(M_q))$
**for** each $M_q \in P$ with $|M_q| > split\_size$ **do**
    remove $M_q$ from $P$ and let $L = \{G_d(M_q)\}$
```
    //Split
```
    **while** $L$ not empty **do**
        remove $G_c$ from $L$ and let $v'$ be max outdegree vertex in $G_c$
        remove $IN(v')$ from $G_c$ and insert it into $leaf_q$ or $seed_q$
        **for** each $M_j \in SCC(G_c)$ **do**
            insert $M_j$ into one of $L$, $leaf_q$, or $seed_q$
```
    //Extend
```
    **for** each $M_i$ in $leaf_q$ **do**
        merge $M_i$ with appropriate $M_j \in seed_q$
    insert modules in $seed_q$ into output set of modules $P$

formulate it so as to mimic the optimization goal defined in the problem definition. One option could be to define the weight solely in terms of the gene pair $g_i, g_j$. However such a simple weighting scheme may not be sufficient in practice, since the co-occurrence of a pair in a module increases the chances of the genes in their neighborhoods to coexist in the same module as well. This is especially important for the contribution of mutual exclusion in the edge-weight, as pairwise mutual exclusion values are almost always close to 1. In order to reflect these observations we consider an edge-weighting scheme where contribution of mutual exclusion is computed within the vertex neighborhoods. More specifically, let $N_e(g_i)$ denote the *extended neighborhood* of $g_i$, that is $N_e(g_i) = \bigcup_{\forall (g_i,g_j)\in E} g_j \cup \{g_i\}$. The contribution of mutual exclusion to the edge weight, $MEX_n(g_i, g_j)$ is the average of $MEX(N_e(g_i))$ and $MEX(N_e(g_j))$. Thus the weight of an edge $(g_i, g_j)$ is defined as,

$$w(g_i, g_j) = MEX_n(g_i, g_j) \times CO(\{g_i\}) \times CO(\{g_j\}).$$

The contribution of coverage is computed as a product so as to reduce the chances of a single gene with large coverage dominating the weights of incident edges. Furthermore, it allows the algorithm to favor more balanced coverages among equal-sized modules; coverage of 100 patients with a module containing a pair of genes, one covering 99 and the other only 1, is less preferable than a module with a pair where each gene covers 50 patients. Finally we note that, to further strengthen the impact of mutual exclusion on edge-weights, we introduce a threshold $\theta$, so that for pairs with $MEX_n$ score less than $\theta$, edge weights are assigned to 0.

#### 3.1.2.2. Edge-Weighted Random Walk

Once $G_w$ is constructed after vertex and edge weight assignments, we apply an insulated heat diffusion process on $G_w$ that can also be described as a random walk with restart on the graph. The random walk starts from a gene $g_s$. At each time step, with probability $1-\beta$, the random surfer follows one of the edges incident on the current node with probability proportional to the edge weights. Otherwise, with probability $\beta$, the walker restarts the walk from $g_s$. Here $\beta$

is called the restart probability. The transition matrix that corresponds to this process is defined as follows:

$$T_{ij} = \begin{cases} \frac{w(g_i,g_j)}{\sum_k w(g_k,g_j)}, & \text{if } (g_i,g_j) \in E \\ 0, & \text{otherwise} \end{cases}$$

$T_{ij}$ can be interpreted as the probability that a simple random walk will transition from $g_j$ to $g_i$. The random walk process can then be described as a network propagation process by the equation, $F_{t+1} = (1-\beta)TF_t + \beta F_0$, where $F_t$ is the distribution of walkers after $t$ steps and $F_0$ is the diagonal matrix with initial heat values, that is $F_0[i,i] = CO(g_i)$. One strategy to compute the final distribution of the walk is to run the propagation function iteratively for increasing $t$ values until $F_{t+1}$ converges [26]. Another strategy, which we chose to employ in our implementation, is to solve this system numerically using the equation, $F = \beta(I-(1-\beta)T)^{-1}F_0$ [17]. The edge-weighted directed graph $G_d$ is constructed by creating directed edge $[g_i,g_j]$ with weight $F[i,j]$, for every pair $i \neq j$.

We note that the idea of random walks with restart has been employed in the context of cancer module identification in previous work [16, 17, 27, 28, 18]. However as the concept of edge weights is absent, the transition probabilities in those studies are only based on the degrees of the vertices. In our case, the transition probabilities reflect the edge weights which in turn model the contribution of a pair of genes to the maximization score, when placed in the same module. Similar to the previous methods employing heat diffusion we assign $\beta = 0.4$.

#### 3.1.2.3. Constructing Set of Driver Modules

We have two main steps. We employ strongly connected components (SCC) as a primitive in both of the steps, similar to Hotnet2 and Hierarchical Hotnet. We first create an initial set of candidate modules. For this, we iteratively remove the smallest weight edge from $G_d$, add the strongly connected components (SCC) of $G_d$ into initial module set $P$, and remove all modules of size less

than $min\_module\_size$ from $P$, until the total number of genes in $P$ decreases to $total\_genes$.

Next we process large candidate modules via a *split-and-extend* procedure, where large modules to be thus processed are determined by a network property of the initial modules. We define the $split\_size$ to be the outdegree of any vertex in any of the subgraphs induced by the modules. Any initial candidate module $M_q$ of size greater than the $split\_size$ goes through the split-and-extend procedure.

The idea is to first extract *seed* modules that satisfy certain size and connectivity criteria, and extend them with small *leaf* modules. Given a directed graph $G_c$, let $IN(v')$ denote the *isolated neighborhood* of $v'$ in $G_c$, that is $w \in IN(v')$, if and only if $w \in N_e(v')$ and for any directed edge $[w,x]$ or $[x,w]$, $x \in N_e(v')$. The split phase of a module $M_q$ consists of removing $IN(v')$ from $G_d(M_q)$, where $v'$ is the vertex with largest degree in $G_d(M_q)$. Assuming its size is not less than $min\_module\_size$, $IN(v')$ is a seed module to be extended in the next phase, otherwise it is a leaf module that is to be attached to an appropriate seed module. The remainder of $G_d(M_q)$ goes through a *SCC* partitioning. Any resulting component of size larger than the $split\_size$ goes through the same split process, any component of size less than $min\_module\_size$ becomes a leaf module, and any other component in between these two sizes becomes a seed module. In the extend phase, each leaf module is merged with the seed module with which it has maximum number of connections in $G_d(M_q)$.

We note that the graph $G_d$ and the original PPI network $G$ may differ, even when we consider the undirected version of $G_d$ for comparison. Since the decisions made by the module construction step described in this subsection are based on the connections in $G_d$, it should be noted that the extracted modules may not exactly satisfy Constraint 1, that is the connectivity constraint of our main problem. However this should be considered a benefit rather than a drawback. It is well-known that PPI networks are incomplete and contain many false positives and negatives. Network prediction/smoothing methods based

solely on network topology have been suggested previously. Many such methods are unsurprisingly based on random walks and have been shown to perform quite well on PPI networks including those of the H.Sapiens [29, 30]. Thus the directed graph $G_d$, especially after dilutions via minimum weight edge removals, not only embeds mutual exclusion and coverage information as relevant to cancer module identification, but also is indirectly a smoothed version of the original PPI network $G$.

## 3.2. Results

We implemented the MEXCOWalk algorithm in Python. We compare MEXCOWalk results against those of three other existing knowledge-based cancer driver module identification methods: Hotnet2 [17], MEMCover [22], and Hierarchical Hotnet [18]. The first two benchmark algorithms are chosen as representatives of their respective subcategories; Hotnet2 is one of the most popular benchmark methods based on optimizing coverage via a heat-diffusion heuristic and MEMCover is a popular algorithm among those optimizing mutual exclusion as well as coverage via a greedy seed-and-growth heuristic. Hierarchical Hotnet is chosen as a third benchmark method, as it is one of the most recent cancer driver module identification methods.

### 3.2.1. Input Data

All four methods, including MEXCOWalk, assume same type of input data in the form of mutation data of available samples and a H.Sapiens PPI network. We employ somatic aberration data from TCGA, preprocessed and provided by [17]. The preprocessing step includes the removal of hypermutated samples and genes with low expression in all tumor types. After the filtering, the dataset contains somatic aberrations for 11,565 genes in 3,110 samples. The mutation frequency of a gene $g_i$ is calculated as the number of samples with at least one single nucleotide variation (SNV) or copy number alteration (CNA) in $g_i$ divided by the number of all samples. As for the H.Sapiens PPI network, we used the HINT+HI2012 network [17]. This is a combination of two interactome databases: HI-2012 prepublication data in human HI2 Interactome database2 (HI2012) [31] and high quality interactomes database (HINT) [32]. We execute each of the four algorithms on the largest connected component of this combined network that consists of 40,704 interactions among 9,858 proteins.

### 3.2.2. Parameter Settings

Regarding MEXCOWalk, we have settings for three parameters: the mutual exclusion threshold $\theta$, the $total\_genes$, and the $min\_module\_size$. We examine the distribution of $MEX_n$ scores of all edges to determine a meaningful range of values for the mutual exclusion score threshold $\theta$ and conclude that only settings of $\theta \geq 0.5$ are in that range. We present results for $\theta = 0.7$, as this setting provides the best results in terms of our main optimization goal defined in Equation 3.1. The results with other threshold values of $0.5, 0.6, 0.8, 0.9$ are available in the Supplementary. Regarding the $total\_genes$ parameter, within the scope of the evaluations discussed in the main document, we view it as the main independent variable; we obtain the results of each evaluation under the settings $total\_genes = 100, 200, \ldots, 2500$. Finally, we set $min\_module\_size$ to 3 for the results discussed in the main document, as this constitutes a nontrivial mimimum module size compatible with the problem definition. Further results applying to the settings of $min\_module\_size = 6, 9, 12$ are presented in the Supplementary Document.

For Hotnet2, we obtain results for varying values of $total\_genes = 100, 200, \ldots, 2500$, with the default value of $min\_module\_size = 3$. We present results of Hierarchical Hotnet with the default setting of its *clustering parameter* $\delta$, which outputs modules of size greater than one, with a total of 806 genes. Since some of these modules may contain modules with two genes, we generate a filtered version as well, where all such modules are removed, resulting in modules with a total of 554 genes. In what follows, we refer to the former version as *HierHotnet_v1* and the latter version as *HierHotnet_v2*.

For MEMCover, mutual exclusion scores are obtained from type-restricted permutation test with all pan-cancer samples, that is the TR_test. Because confidence scores are not available for HINT + HI2012 network, we set the confidence score of all edges to 1 when calculating the edge weights for the MEMCover model. We set the coverage parameter $k$ to its default value of 15. MEMCover introduces a parameter, $f(\theta)$, that is used to control the trade-off between the output number of modules and the average weights within each module. It indirectly controls the module sizes; the smaller $f(\theta)$, the larger the modules output by MEMCover in general. We consider three settings for the MEMCover algorithm, referred to as *MEMCover_v1*, *MEMCover_v2*, and *MEMCover_v3*, respectively.

For the first one, we assign $f(\theta) = 0.584$, which is achieved by setting $\theta$ parameter (not to be confused with the $\theta$ we employ in MEXCOWalk) to 40%, as recommended in the original paper. For the second one, we assign $f(\theta) = 0.03$, which is the setting that minimizes the percentage of size one and size two modules. Finally, the last one corresponds to the setting where $f(\theta) = 0.03$ and all modules of size $< 3$ are removed. To obtain results with varying *total_genes* from 100 to 2500 we consider the modules formed by the first *total_genes* many genes output by each version, since the order MEMCover outputs the modules reflects the algorithm's quality preferences. Values of *total_genes* larger than 1600 are not available for *MEMCover_v3* as it outputs 1684 genes in total.

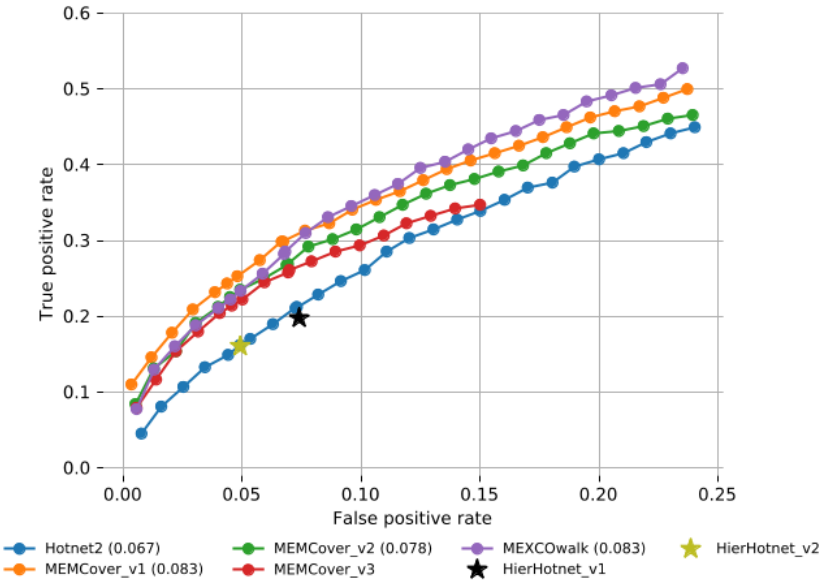


Figure 3.1: The fraction of recovered CGC genes for each *total_genes* value is shown with a ROC plot. AUROC values are written in parentheses for each applicable method.

### 3.2.3. Static Evaluations

Most of the existing driver module identification methods employ *static evaluations*, where the union of the genes in all the modules are compared against a reference set of cancer genes. Such evaluations measure the capability of an algorithm in dissecting cancer-related genes throughout the modules it provides, without regard for the specific modularity of the output. COSMIC Cancer Gene Census (CGC) database [33] is one such reference containing 616 genes with mutations that have been causally implicated in cancer. For consistency with previous work, our first evaluation compares the algorithms

based on their ability to recover these known cancer genes. Figure 3.1 plots the Receiver Operating Characteristic (ROC) curves of the set of genes in the union of modules each algorithm provides with respect to the CGC genes. *MEMCover_v1* and our model has the same Area Under the ROC (AUROC) value of 0.083. *MEMCover_v2* ranks the third. Although the areas are not comparable, *MEMCover_v3* outputs provide worse TP rates than those of *MEMCover_v2*. Finally, Hotnet2 and Hierarchical Hotnet recover fewer known cancer genes than the rest of the methods.

### 3.2.4. Modular Evaluations

Most cancer driver module identification studies evaluate their output modules via analysis methods including functional enrichment, survival analysis, or literature verification. However in general these evaluations are not systematic due to lack of evaluation metrics that quantify these types of analysis and as such are not amenable for comparisons among alternatives. Without such metrics it is not possible to provide a fair judgment of different methods as far as the quality of the modules is of concern, and not just the quality of the static gene set. We provide three modularity-based metrics and evaluate the output module sets of alternative methods based on these metrics.

#### 3.2.4.1. Driver Module Set Score

Our first evaluation metric is the main optimization goal of the cancer driver module identification problem, that is the driver module set scores (*DMSS*) defined in Equation 3.1. Fig. 3.2-A shows that MEXCoWalk discovers modules that have better *DMSS* values than the module sets of all the other methods. The difference is much more dramatic for smaller $total\_genes$ values such as 100 and 200. Those of Hierarchical Hotnet and Hotnet2 are among the worst, especially for settings of $total\_genes > 500$. *MEMCover_v1* performs worse than the two other MEMCover versions, as it provides many size 1 and size 2 modules. This finding demonstrates another merit of the *DMSS* definition; if there are many small modules, assuming the mutual exclusion does not decrease substantially by enlarging the modules, then our optimization score function prefers outputs with larger modules. Consider for instance, the following special case where we have 10 genes under consideration, each covering $x$ out

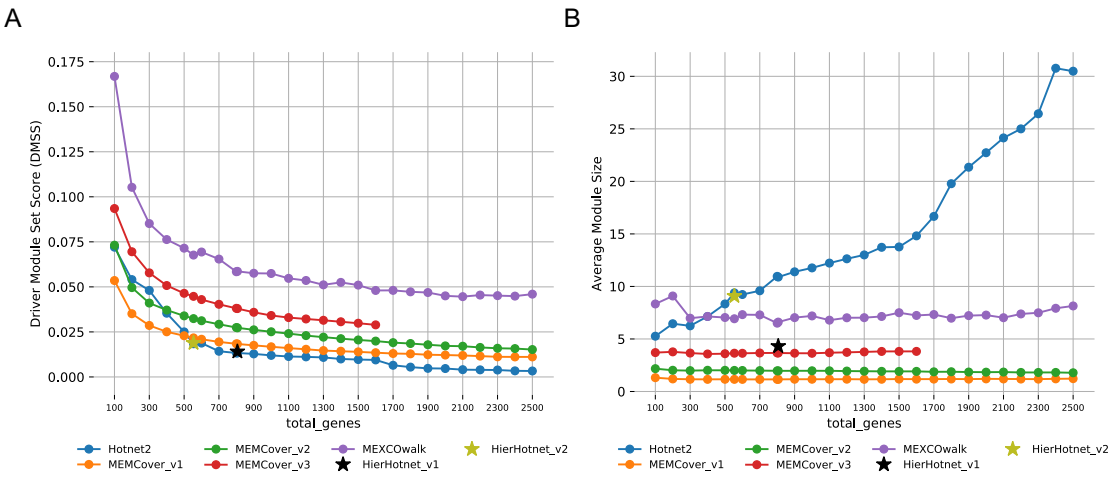


Figure 3.2: A) DMSS evaluations of output modules of MEXCOwalk, MEMCover, Hotnet2, and Hierarchical Hotnet for increasing values of $total\_genes$. B) Average modules sizes in the outputs of MEXCOwalk, MEMCover, Hotnet2, and Hierarchical Hotnet for increasing values of $total\_genes$.

of a total of $y$ samples. The output consisting of a set of modules each containing one gene has a $DMSS$ of $x/y$. On the other hand, assuming a $MEX$ score of $m$ for every pair of genes, the output with any pair of genes per module has a $DMSS$ of $2m^2x/y$. This implies that the latter is a more preferable module set than the former, as long as $m > \sqrt{1/2}$. It corresponds to the case where upto almost 58% of samples covered by a gene to be in the intersection of samples covered by another gene.

Analyzing the outputs of the methods with respect to the average module sizes in Fig. 3.2-B, we can observe that all MEMCover versions, especially *v1* and *v2*, provide module sets with very small average module sizes; for almost all values of $total\_genes$, MEMCover_v1 provides an average size of almost 1.2 genes per module, MEMCover_v2 provides an average of almost 1.9 genes per module, and MEMCover_v3 provides an average of almost 3.7 per module. MEXCOWalk provides module sets with average sizes ranging between 6.5 and 9, whereas Hotnet2 module sizes significantly increase proportional to $total\_genes$. The range of average module size values for Hotnet2 is approximately between 5 and 30.

#### 3.2.4.2. Cancer Type Specificity Score

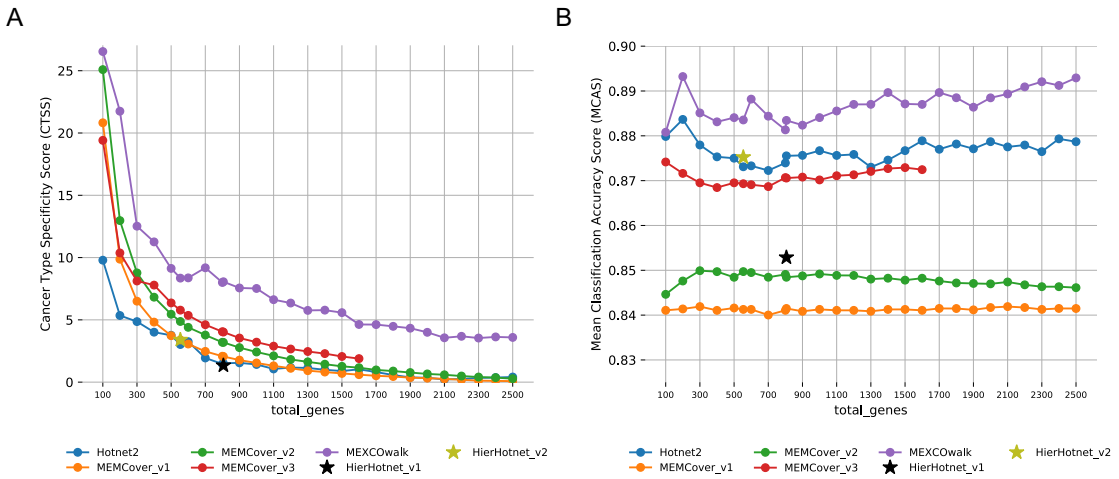


Figure 3.3: A) CTSS evaluations of output modules of MEXCOwalk, MEMCover, Hotnet2, and Hierarchical Hotnet for increasing values of $total\_genes$. B) MCAS evaluations of output modules of MEXCOwalk, MEMCover, Hotnet2, and Hierarchical Hotnet for increasing values of $total\_genes$.

Our second modularity-based evaluation metric is defined with respect to cancer type specificity. We test an output module set in terms of enrichment for mutations in a specific cancer type using Fisher's exact test. For a module $M$, let $S_M$ denote the set of patients where at least one of the genes in $M$ is mutated. For a cancer type $t$, let ${S_M}^t$ denote the subset of patients in $S_M$ diagnosed with cancer type $t$. Assuming $n_t$ denotes the number of patients of cancer type $t$ in the whole dataset, we calculate the Fisher's exact test with the following entries in the contingency table in row-major order: $|{S_M}^t|$, $n_t - |{S_M}^t|$, $\sum_{t' \neq t} |{S_M}^{t'}|$, $\sum_{t' \neq t} n_{t'} - |{S_M}^{t'}|$. We use the Bonferroni method for multiple testing correction.

Let $P = \{M_1, M_2, \ldots M_r\}$ be a set of modules. For each module $M_q \in P$, the described process results in a p-value for every cancer type $t$, denoted with ${p_q}^t$. We define the *cancer type specificity score* of $P$ as the average $-\log$ of best p-value per module. More formally,

$$CTSS(P) = \frac{\sum_{M_q \in P} -\log(\min_{\forall t}({p_q}^t))}{r} \tag{3.2}$$

Fig. 3.3-A shows the CTSS scores of the module sets provided by the methods under consideration. Compared to the other methods, MEXCOWalk provides a larger CTSS value for every setting of *total_genes*, indicating that the output modules are strongly enriched for particular cancer types. We also observe that module sets of MEMCover versions perform better than those of Hotnet2 and Hierarchical Hotnet.

#### 3.2.4.3. Mean Classification Accuracy Score

We examine the predictive value of an output set of modules in classifying tumor and normal samples of TCGA pan-cancer data consisting of 12 cancer types. For this, we employ k-nearest-neighbor classifier using Euclidean distance with $k = 1$, where the features are the expression values of the set of genes in a module. To evaluate the predictive performance of a module $M_q$, we use 10-fold stratified cross-validation accuracy, denoting it with $Acc(M_q, c)$ for a fold $c$. We can then define the Mean Classification Accuracy Score of a set of modules $P$ as,

$$MCAS(P) = \frac{\sum_{M_q \in P} \sum_{c=1}^{10} Acc(M_q, c)}{10 \times r} \tag{3.3}$$

The plots of the MCAS scores of the module sets of all four methods for varying *total_genes* are provided in t Fig. 3.3-B. MEXCOwalk consistently achieves the top accuracy for all settings of *total_genes*, implying that MEXCOwalk modules can more accurately perform tumor/normal classification than the other methods. Interestingly, Hierarchical Hotnet performs worse than Hotnet2. Among MEMCover models, MEMCover_v3 shows a better performance than MEMCover_v1 and MEMCover_v2, in contrast to their relative performances in recovering known cancer genes.

## 3.3. Analysis of MEXCOWalk Modules

Fig. 3.4-A shows the 12 modules that MEXCOWalk identifies when *total_genes* is set to 100. The sizes of the modules range between 3 and 31, and their coverage values range between 5% to 50%. Note that the edges correspond to the PPI network edges, whereas

the weight of an edge is the smaller of the weights of the corresponding directed edges from $G_d$ as computed through edge-weighted random walk and thus represents the degree of mutual exclusivity and coverage assigned by MEXCOWalk.

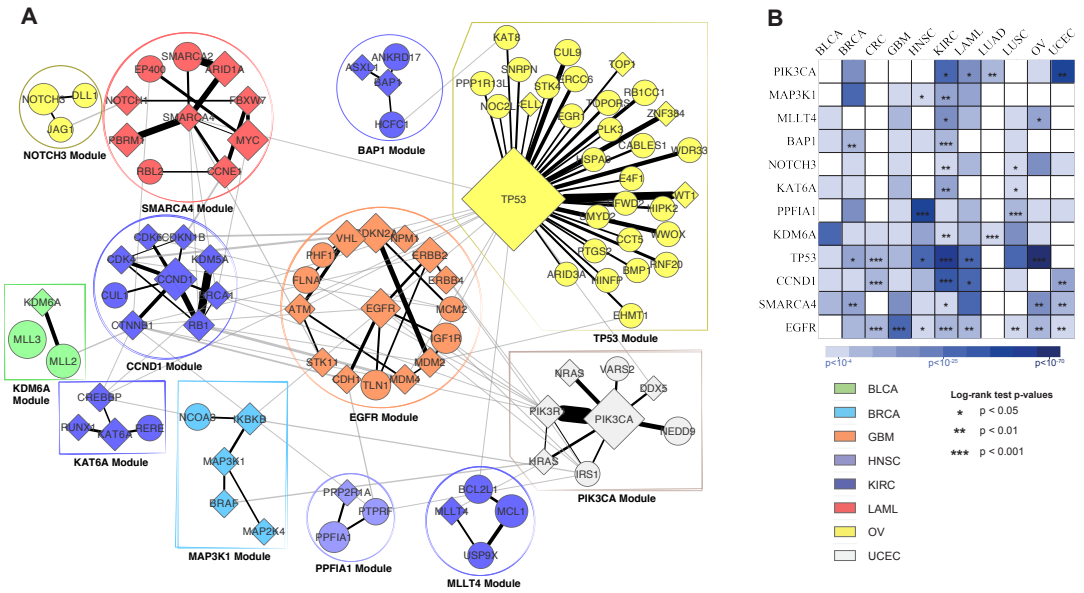


Figure 3.4: A) MEXCOWalk output modules when $total\_genes = 100$. Diamond shaped nodes correspond to CGC genes. Sizes of the nodes are proportional with mutation frequencies of corresponding genes. Edges within the module are colored black, whereas the edges between the modules are colored. Edge weights are reflected in the thicknesses of the line segments. Color of a module denotes the cancer type with the strongest enrichment for mutations in genes of that module. The legend for the color codes are shown on the right. Each module is named with the largest degree gene in the module. B) Results of cancer type specificity and survival analyses. Rows correspond to modules and columns correspond to cancer types. Colors of the matrix entries indicate the significance of enrichment for cancer types in terms of Fisher's exact test p-values. Stars indicate the significance of log-rank test p-values in survival analyses.

Many of these modules are part of well known cancer-related pathways such as those centered at EGFR, TP53, PIK3CA, and CCND1. Analyzing the interactions between the modules, EGFR module can be seen as an important hub module between many important modules such as the TP53 module, CCND1 module, and the PIK3CA module; without the EGFR module these three modules would almost be isolated in the induced subgraph. The EGFR module contains several known cancer genes many of which are related to cell cycle control: VHL, CDKN2A, NPM1, ERBB2, ERBB4, MDM2, MDM4, STK11, CDH1, ATM. Seven cancer types are enriched for mutations in this module with GBM being the most significant enrichment; Fisher's exact test p-value is $= 4.9e - 20$. Indeed, EGFR gene is mutated in more than half of all GBM patients and anti-EGFR agents are

already used for GBM treatment [34]. However, resistance to these agents is a major problem suggesting that treatment strategies might benefit from targeting multiple genes in this module. This module also contains TLN1, which is not one of the known cancer genes listed in CGC. However it is mutated in 104 patients across 10 cancer types and it has previously been associated with tumorigenecity and chemosensitivity [35, 36]. We investigate whether the genes in this module are predictive of patient survival profiles by calculating a risk score for each patient as in [37] and [38]. When we divide the GBM patients into two as training and test sets, the low-risk and high-risk thresholds that we identify from the training set are successful in stratifying the patients into low-risk and high-risk groups in the test set; the log-rank test p-value$= 0.0004$ (Fig. 3.4-B).

Our TP53 module includes 30 interactors out of 213 available in the HINT+HI2012 PPI network. TP53 shares the highest edge weight with WT1, which is a transcription factor that has roles in cellular development and cell survival. Another gene which has a large edge weight is CUL9. It is altered in 48 patients out of a total of 3110 samples, which would possibly make it easy to miss through single-gene tests.

The PIK3CA module identifies several genes in the PI3K pathway whose deregulation is critical in cancer development and progression [39]. The module provides a chance to observe the importance of incorporating mutual exclusion in MEXCOWalk. Among all the interactions presented in the induced subgraph of 100 genes in all 12 modules, the one between PIK3CA and PIK3R has the largest weight. These genes are mutated in 602 and 155 patients respectively, although the overlap between the two patient sets is only 18 indicating the high mutual exclusivity between the pair of genes.

The CCND1 module is yet another fairly known cancer driver module [40, 41]. Other than EGFR, it is the module that contains the most reference genes; all 9 genes in the module, except CUL1, are listed in the reference CGC database. It is shown that the mutations, amplification, and expression changes of these genes, which alter cell cycle progression, are frequently observed in a variety of tumors and contribute to tumorigenesis [41, 40]. Indeed, we find significant association of this module with patients' survival outcome in CRC, KIRC, LAML and UCEC types (Fig. 3.4-B).

One of the less known modules, the KDM6A module illustrates the advantage of using

random walk to compensate for the incomplete edges in the PPI network data. MLL3 gene does not have interactions with MLL2 and KDM6A in the original PPI network. Nevertheless, the heat diffussion process of MEXCOWalk diffuses a large enough heat due mainly to the network topology, the coverage, and the mutual exclusion values of the pairs. As a result the three genes are connected with large edge weights in $G_d$, demonstrating the quality of MEXCOWalk modules even with missing interaction data, as verified by the the predicted interaction between MLL3 and KDM6A [42].

# CHAPTER 4

# 4. Network Centric Epistasis Evaluation

We provide a description of the overall framework for the network-centric epistasis evaluations and briefly summarize the statistical mutual exclusivity tests under study.

## 4.1. Methods

The following sections include a detailed description of the methods for network-centric epistatic evaluation.

### 4.1.1. Problem definition

Statistical mutual exclusivity algorithms employ many different approaches to compute pairwise mutual exclusivity. However, a standard approach for evaluation of such methods do not exist. The results from these methods may falsely display a strong mutual exclusivity among pairs. To overcome this, we develop a network-centric epistasis evaluation framework based on the pairwise interactions among driver genes.

### 4.1.2. Network-centric Epistasis Evaluation Framework

The overall framework has two main components. The first one consists of definitions of statistics and metrics employed in the network-centric epistasis evaluations. Such metrics should properly quantify the gains attained by a network-centric approach of discerning pairwise epistasis relations from among those provided as a result of employing any appropriate pairwise mutual exclusivity test on a given cohort. The second component is inspired by the recent work of [7], where the effects of mutation load associations on the set of statisticaly significant findings of pairwise mutual exclusivities are

depicted through subtype-stratified analyses. We follow a similar methodology with our network-centric approach to verify whether the use of interactome information provides similar advantages in epistasis corrections of pairwise mutual exclusivity findings as the subtype-stratification idea suggested by [7] .

#### 4.1.3. Metrics for Network-centric Epistasis Evaluations

Assuming that cancer driver genes in the same pathway are more likely to show mutually exclusive mutation profiles, we utilize the interactome to devise a new strategy for evaluating the mutual exclusivity methods and the effects of the interactome information on quantifying true epistasis. The steps of the evaluation framework are described in pseudocode as shown in Algorithm 2. Let $\mathscr{G}, \mathscr{C}, \mathscr{T}, c$ denote respectively the input PPI network, the employed cohort, the statistical mutual exclusivity test undergoing the network-centric epistasis evaluations, and the type of the control group to be employed. Let $N_{CGC}(g_i)$ denote the set of *CGC* genes that are in the neighborhood of the node corresponding to gene $g_i$ in the PPI network $\mathscr{G}$. Corresponding to each *CGC* neighbor $g_j$ of the gene $g_i$, we randomly select a gene $g_r$ from a control group $\mathscr{X}_c(g_i)$, and compute 12 statistics, $Stat_1 \dots Stat_{12}$, based on the $-\log$-transformed p-values $p_{i,j}, p_{i,r}$ as computed by the mutual exclusivity test $\mathscr{T}$. To obtain robust results, the selection of the random genes from the control group is repeated 100 times and the median values of these 100 instances are taken into account.

With regard to the computed statistics, $Stat_1 \dots Stat_6$ compare the significance of *CGC* gene $g_i$'s mutual exclusivity with a neighbor *CGC* gene against its mutual exclusivity with a random gene from the control group. For instance, for each reference *CGC* gene $g_i$, $Stat_1$ counts the number of instances where a *CGC* neighbor gene shows significant mutual exclusivity with $g_i$, whereas a random gene from the control group does not. $Stat_7$ simply sums up the $-\log$-transformed p-values of all *CGC* gene pairs $g_i, g_j$ that are interacting in the PPI network. $Stat_8$ calculates the same sum but the p-values are computed with respect to the random genes from the control group, rather than the interacting *CGC* genes. $Stat_9$ computes the number of significantly mutually exclusive *CGC* gene pairs that have an interaction in the PPI, whereas $Stat_{10}$ counts the number of pairs found not to be significantly mutually exclusive. $Stat_{11}$ computes the number of significantly mutually

**Algorithm 2** Network-centric Epistasis Evaluation

**Input:** $\prec \mathscr{G}, \mathscr{C}, \mathscr{T}, c \succ$
**Output:** Relevant Statistics of the mutual exclusivity test $\mathscr{T}$

Execute $\mathscr{T}$ on $\mathscr{C}$
$p_{x,y}$: $-\log$ transformed p-value of $g_x, g_y$ as output by $\mathscr{T}$
**for** each gene $g_i$ in $CGC$ **do**
  **for** t = 1 to 100 **do**
    **for** k=1 to 12 **do**
      $C_k = 0$
    **for** each gene $g_j$ in $N_{CGC}(g_i)$ **do**
      $g_r$: randomly sampled gene from $\mathscr{X}_c(g_i)$
      **if** $p_{i,j} > -\log(0.05) > p_{i,r}$ **then** $C_1 \mathrel{+}= 1$
      **else if** $p_{i,j} > p_{i,r} > -\log(0.05)$ **then** $C_2 \mathrel{+}= 1$
      **else if** $p_{i,r} > p_{i,j} > -\log(0.05)$ **then** $C_3 \mathrel{+}= 1$
      **else if** $-\log(0.05) > p_{i,j} > p_{i,r}$ **then** $C_4 \mathrel{+}= 1$
      **else if** $-\log(0.05) > p_{i,r} > p_{i,j}$ **then** $C_5 \mathrel{+}= 1$
      **else if** $p_{i,r} > -\log(0.05) > p_{i,j}$ **then** $C_6 \mathrel{+}= 1$
      $C_7 \mathrel{+}= p_{i,j}$
      $C_8 \mathrel{+}= p_{i,r}$
      **if** $p_{i,j} > -\log(0.05)$ **then** $C_9 \mathrel{+}= 1$
      **else** $C_{10} \mathrel{+}= 1$
      **if** $p_{i,r} > -\log(0.05)$ **then** $C_{11} \mathrel{+}= 1$
      **else** $C_{12} \mathrel{+}= 1$
    **for** k=1 to 12 **do**
      $C_k^i = C_k^i \cup C_k$
  **for** k=1 to 12 **do**
    $Stat_k \mathrel{+}= median(C_k^i)$
Output $Stat_k$ for $k = 1 \ldots 12$

exclusive gene pairs where each gene pair is comprised of a reference *CGC* gene and a random gene from the control group. The number of such pairs is the same as those considered for $Stat_9$. Finally $Stat_{12}$ counts the number of pairs found not to be significantly mutually exclusive among the group of pairs considered for $Stat_{11}$. Note that based on the premise that cancer driver genes interacting in the PPI network are likely to exhibit epistasis, $Stat_9$ and $Stat_{11}$, respectively correspond to True Positives and False Positives, whereas $Stat_{10}$ and $Stat_{12}$, respectively correspond to the False Negatives and True Negatives. Thus precision, sensitivity, specificity, and the F1 scores are computed based on these 4 statistics.

For the network-centric epistasis evaluations we employ two different definitions for the control groups. For the first one, the control group $\mathscr{X}_1(g_i)$ consists of *CGC* genes that do not interact with $g_i$ in the PPI network. For the second one, $\mathscr{X}_2(g_i)$ consists of neighbors of $g_i$ in the PPI network that are not in *CGC*.

#### 4.1.4. Network-centric Epistasis Corrections in Relation to MLA

Some statistical mutual exclusivity tests are based on the assumpton that genes' alterations across tumors are identically distributed. Among the approaches considered in this study Fisher's Exact Test and MEGSA [25] belong to this category. However it has been observed that the number of alterations per tumor can vary quite considerably, even in tumors of the same type; colorectal tumors with microsatellite stability have a median of 66 non-synonymous mutations, but colorectal tumors with microsatellite instability have a median of 777 mutations [43, 24]. It has been shown that under such settings the mutual exclusivity tests relying on identical alteration probabilities across tumors may lead to reduced sensitivity for mutual exclusivity analysis [23]. The effects of varying alteration probabilities on pairwise mutual exclusivity calculations have been formalized within the context of the so-called *mutation load confounding (MLC)* in a recent study by [7]. MLC is a correlation between the number of statistically significant mutual exclusivity findings and the *mutation load association (MLA)* of a gene, where logistic regression is used to compute MLA as a standardized score of association between the mutation likelihood of each gene and the *mutation load*, that is the genome-wide number of somatic mutations observed in a tumor. Note that negative MLA values correspond to higher mutation frequencies in tumors with low mutation loads, whereas positive values correspond to

higher mutation frequencies in tumors with high mutation loads. Strong negative correlations between the MLA of a gene and the number of statistically significant pairwise mutual exclusivities have been observed, implicating the finding that the more negative a gene's MLA, the higher the number of other genes that show mutual exclusivity with that particular gene [7].

However, such a negative correlation does not always imply true epistasis since a gene highly mutated in tumors wih low mutation loads, naturally has a better chance of forming mutually exclusive pairs with other genes. Thus extra sources of information are necessary to filter out the pairs with true epistasis relations among a set of statistically significant pairwise mutual exclusivities postulated by some exclusivity test. [7] make use of the subtype information for such a purpose and show that the mutation load confounding can be reduced by correcting via tumor subtype stratification. Such a correcton greatly reduces the number of gene pairs reported to show mutual exclusivity, especially for pairs that include genes with low MLA. A major drawback is the absence of subtype information for many tumors. As part of our network-centric epistasis framework, we suggest that such a correction can be efficiently done with the interaction network data, rather than or better yet on top of the subtype information. For this purpose we calculate the correlation between the number of statistically significant pairwise mutual exclusivity findings and the MLA for two settings; one where pairwise mutual exclusivities are sought between a CGC gene and all other CGC genes, and the other where a CGC gene is checked against only its PPI neighbors that are in CGC. The computations of the two settings are repeated with the subtype-stratified data as well, to see the added value of the network-centric epistasis corrections on top of the subtype-based corrections on statistically significant pairwise mutual exclusivities.

### 4.1.5. Relating MLA to Network-centric Epistasis

van de Haar et. al. define Tumor Mutation Load (TML) for a single tumor sample as the sum of all mutations within that sample. They particularly focus on established tumor subtypes and claim that the mutations with low MLA that drive tumor subtypes may high mutual exclusivity and thus show epistatic interactions with a vast majority of other genes within the cohort. In order to display genetic dependencies they calculate a score called "Mutation Load Association (MLA)" that indicates the strength of association between

the mutation profile of a gene and TML for all tumor samples. MLA is calculated using the mutation profiles of the tumors by applying logistic regression on TML. The MLA score is 0 for genes that show no association, whereas positive or negative values indicate that the gene is more frequently mutated in tumors with high or low mutation loads, respectively. They also show that there is a strong negative correlation between the gene's MLA and the number of significant findings in mutual exclusivity tests. This strong negative correlation is defined as Mutation Load Confounding (MLC), which suggests that, the more negative the MLA of a gene, the higher the number of other genes that show mutual exclusivity with that particular gene.

However, the mutation threshold, *t=20* is too strict and allows rarely mutated drivers to be ignored. The strong negative correlation is not observed when the mutation threshold is changed to *t=5*.

## 4.2. Results

Similar to MEXCOwalk, we implement the algorithms in Python. We evaluate 5 different algorithms, DISCOVER, Fisher's Exact Test, MEGSA, MEMO and WExT. Necessary adjustments were made to all algorithms that do not naturally provide pairwise mutual exclusivity results.

### 4.2.1. Input Data

The somatic mutation data from TCGA was preprocessed and provided by [7]. The 8 different cancer types and their corresponding tumor samples within the dataset is as follows: BLCA (411), BRCA (1026), COADREAD (498), LUAD (568), LUSC (485), SKCM (468), STAD (438) and UCEC (531). The preprocessing step involves the removal of all mutations with 'variant_classification' of 'Silent', '3'UTR', 'Intron', '5'UTR', 'RNA', '3'Flank' and '5'Flank' from the TCGA data. The input data is then further filtered by mutation frequency threshold, $t$, to include genes with ¿ $t$ mutations across the cohort. Namely, $t = 20$ means that we include the genes that are mutated in more than 20 samples within that cancer type. We try the values 5, 10 and 20 for $t$.

Regarding subtypes, we download subtype information for BRCA from cBioPortal and for COADREAD from [44].

For the PPI, we use a filtered version of the IntAct network [45]. We process the original PPI by removing duplicate edges and removing edges below the confidence threshold of 0.35. After these steps, the final network contains 15,079 nodes and 103,520 edges.

### 4.2.2. Network-centric Epistasis Evaluations of Alternative ME Tests

Table 4.1, 4.1, 4.3, 4.3 shows the results of evaluating the five ME detection methods on BRCA and COADREAD data respectively. For BRCA we use molecular stratification of 1026 patients. For COADREAD we use the 498 patients for which CMS classification was available. The control group is defined as $X_1$ and $X_2$ in Table 4.1 and 4.3 and $X_2$ in Table 4.2 and 4.4. Among the provided statistics, precision, sensitivity, specificity and F1 score metrics depend on the significance of the p-values reported for the CGC neighbor and for the random gene sampled from the control group. On the other hand, *Stat* values are more detailed and also assess the relation between these pair of p-values. As such, we first discuss the former set of metrics.

For COADREAD we observe that DISCOVER-Strat gives the highest precision. The ranking of the other methods from best to worst in terms of precision is as follows: MEGSA, WexT, DISCOVER and Fisher's Exact Test. Compared to precision, we observe much higher differences among the sensitivity values output by the employed methods. We can group the methods into two where the first group contains WExT and DISCOVER, and the second group contains the remaining three methods. The former group of methods give much larger sensitivity values than the latter. For instance, the best sensitivity value provided by WExT is an order of magnitude higher than the lowest sensitivity value obtained with Fisher's Exact Test. This difference in sensitivity values is also carried over to F1 scores where WExT results in the best F1 score which is followed by DISCOVER. The remaining three methods give much smaller F1 scores and they rank as follows from highest to lowest: MEGSA, DISCOVER Strat and Fisher's Exact Test.

Table 4.1: BRCA control group $X_1$ (34 COSMIC-COSMIC pairs)

| Method | Stat1 | Stat2 | Stat3 | Stat4 | Stat5 | Stat6 | Stat7 | Stat8 | Precision | Sensitivity | Specificity | F1 Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DISCOVER | 11.0 | 1.0 | 3.0 | 8.0 | 5.0 | 4.0 | 2.937 | 2.076 | 0.652 | 0.469 | 0.750 | 0.545 |
| DISCOVER Strat | 13.0 | 2.0 | 1.0 | 9.0 | 3.0 | 4.0 | 3.126 | 1.945 | 0.696 | 0.500 | 0.781 | 0.582 |
| Fisher's Exact Test | 2.0 | 0.0 | 0.0 | 21.0 | 9.5 | 0.0 | 0.531 | 0.287 | 1.000 | 0.062 | 1.000 | 0.116 |
| MEGSA | 2.0 | 0.0 | 0.0 | 7.0 | 25.0 | 0.0 | 1.169 | 0.970 | 1.000 | 0.059 | 1.000 | 0.111 |
| MEMO | 10.5 | 1.0 | 2.0 | 8.0 | 3.5 | 4.0 | 3.113 | 2.175 | 0.659 | 0.466 | 0.759 | 0.545 |
| WExT | 9.0 | 2.0 | 3.0 | 6.0 | 3.0 | 4.0 | 3.790 | 2.679 | 0.609 | 0.519 | 0.667 | 0.560 |

Table 4.2: BRCA control group $X_2$ (9 COSMIC-COSMIC pairs)

| Method | Stat1 | Stat2 | Stat3 | Stat4 | Stat5 | Stat6 | Stat7 | Stat8 | Precision | Sensitivity | Specificity | F1 Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DISCOVER | 4.0 | 1.0 | 0.0 | 1.0 | 1.0 | 2.0 | 3.043 | 1.753 | 0.625 | 0.556 | 0.667 | 0.588 |
| DISCOVER Strat | 5.0 | 0.0 | 0.0 | 2.0 | 0.0 | 2.0 | 3.351 | 1.633 | 0.714 | 0.556 | 0.778 | 0.625 |
| Fisher's Exact Test | 0.0 | 0.0 | 0.0 | 6.0 | 3.0 | 0.0 | 0.469 | 0.245 | NaN | 0.000 | 1.000 | NaN |
| MEGSA | 0.0 | 0.0 | 0.0 | 4.0 | 5.0 | 0.0 | 1.138 | 0.777 | NaN | 0.000 | 1.000 | NaN |
| MEMO | 4.0 | 1.0 | 0.0 | 2.0 | 0.0 | 2.0 | 3.238 | 1.923 | 0.625 | 0.556 | 0.667 | 0.588 |
| WExT | 4.0 | 1.0 | 0.0 | 1.0 | 1.0 | 2.0 | 3.753 | 2.450 | 0.625 | 0.556 | 0.667 | 0.588 |

Table 4.3: COADREAD control group $X_1$ (196 COSMIC-COSMIC pairs)

| Method | Stat1 | Stat2 | Stat3 | Stat4 | Stat5 | Stat6 | Stat7 | Stat8 | Precision | Sensitivity | Specificity | F1 Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DISCOVER | 28.0 | 7.0 | 7.0 | 84.0 | 54.5 | 8.0 | 2.06 | 1.46 | 0.656 | 0.223 | 0.883 | 0.333 |
| DISCOVER Strat | 8.0 | 0.0 | 0.0 | 108.0 | 74.0 | 3.0 | 1.12 | 0.9 | 0.727 | 0.041 | 0.984 | 0.078 |
| Fisher's Exact Test | 6.0 | 0.0 | 0.0 | 114.5 | 67.0 | 5.0 | 0.43 | 0.26 | 0.545 | 0.031 | 0.974 | 0.059 |
| MEGSA | 11.0 | 0.0 | 1.0 | 16.0 | 162.0 | 4.0 | 1.04 | 0.90 | 0.706 | 0.062 | 0.974 | 0.114 |
| MEMO | 37.0 | 9.0 | 17.0 | 71.0 | 47.0 | 8.5 | 3.026 | 1.947 | 0.646 | 0.332 | 0.818 | 0.439 |
| WExT | 48.0 | 12.0 | 14.0 | 57.0 | 43.0 | 9.0 | 3.99 | 2.63 | 0.679 | 0.404 | 0.809 | 0.507 |

Table 4.4: COADREAD control group $X_2$ (107 COSMIC-COSMIC pairs)

| Method | Stat1 | Stat2 | Stat3 | Stat4 | Stat5 | Stat6 | Stat7 | Stat8 | Precision | Sensitivity | Specificity | F1 Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DISCOVER | 12.0 | 9.0 | 8.0 | 40.0 | 27.0 | 10.0 | 2.27 | 1.89 | 0.518 | 0.274 | 0.745 | 0.358 |
| DISCOVER Strat | 6.0 | 0.0 | 0.0 | 52.0 | 43.0 | 6.0 | 1.13 | 1.05 | 0.500 | 0.056 | 0.944 | 0.101 |
| Fisher's Exact Test | 3.5 | 0.0 | 0.0 | 56.5 | 41.5 | 5.0 | 0.57 | 0.47 | 0.412 | 0.033 | 0.953 | 0.061 |
| MEGSA | 6.0 | 0.0 | 1.0 | 11.0 | 82.0 | 5.0 | 1.16 | 1.05 | 0.538 | 0.067 | 0.943 | 0.119 |
| MEMO | 16.0 | 8.5 | 18.0 | 34.0 | 25.0 | 6.0 | 3.435 | 2.699 | 0.567 | 0.395 | 0.698 | 0.466 |
| WExT | 18.0 | 14.0 | 15.0 | 29.0 | 24.5 | 6.0 | 4.61 | 3.71 | 0.573 | 0.441 | 0.671 | 0.499 |

For BRCA, we obtain fewer number of CGC pairs. We observe Fisher's Exact Test and MEGSA having the highest precision, but this is largely due to the fact that only a few significant mutually exclusive pairs of CGC genes were estimated by these algorithms. Among the remaining methods, Discover Strat has the highest precision. Looking at 4.2, both methods fail to get any significant mutually exlcusive pairs. Ranking the other methods through sensitivity, we observe wext to attain the highest rank. This doesn't reflect on the F1 scores as for both table 4.1 and 4.2, Discover Strat obtains the best F1 score, followed by WExT, Memo and Discover.

When we look at *Stat* values, we first observe a large variance across the number of significant p-values output by the methods. This is evident from the large differences in $Stat_1$, $Stat_2$, $Stat_3$ and $Stat_6$ statistics that count the events where either the CGC neighbor or the random gene from the control group results in a significant p-value. Overall, we observe that DISCOVER-Strat, Fisher's exact test and MEGSA are more conservative compared to DISCOVER and WExT, where from the latter group, WEXT outputs notably larger number of significant p-values. Among the different *Stat* values $Stat_1$ and $Stat_6$

are particularly important since the former counts the number of best outcomes and the latter counts the number of worst outcomes. Namely, $Stat_1$ counts the cases where CGC neighbor gives a significant p-value whereas the random gene from the control group does not. $Stat_6$ counts exactly the opposite type of cases where the random gene gives a significant p-value whereas the CGC neighbor does not. We observe that WExT gives the largest $Stat_1$ value. On the other hand, WExT also gives the largest $Stat_3$ and $Stat_6$ value where the random neighbor's p-value is more significant than that of CGC neighbor. However, the difference between WEXT and the second ranking method in terms of $Stat_1$ value (48 vs 28) is much larger than the difference in terms of $Stat_3$ value (14 vs 7) or $Stat_6$ value (9 vs 8). Namely, Wext's superiortiy in recall is worth the small increase in false positives. Another interesting result is MEGSA's large $Stat_5$ value (162) , which is more than twice the value obtained with the second ranking method, DISCOVER Strat. This pattern differs from that of the other two conservative methods Fisher's exact test and Discover-Strat where $Stat_4$ values are much larger than $Stat_5$. $Stat_4$ and $Stat_5$ together correspond to the cases where both p-values are non-significant. MEGSA's large $Stat_5$ indicates its poor performance in ranking the p-values of the CGC neighbor and that of the random gene from the control group. Lastly, $Stat_7$ and $Stat_8$ values correspond to the average magnitude of the p-values obtained for the CGC neighbors and for the random genes, respectively. A good performing approach should result in a much larger value for $Stat_7$ compared to $Stat_8$. First of all, for both statistics, WExT and Fisher's Exact Test gives the largest and smallest values, respectively. In terms of the difference between $Stat_7$ and $Stat_8$, WExT ranks the first which is followed by DISCOVER, DISCOVER Strat, Fisher's Exact Test and MEGSA.

The ranking of the methods in Table 4.4 with respect to F1 score and sensitivity remain the same as Table 4.3. However, there are differences in the ranking with respect to other metrics. For instance, WExT ranks best in terms of precision whereas the best ranking method in Table 1, DISCOVER Strat, ranks the fourth. Compared to Table 4.3, the precision and specificity values of all the methods are smaller in Table 4.4. We see the opposite trend for sensitivity values. These changes are in parallel with the increase in percent significant p-values output by the methods. For instance, the percentage of significant p-values output by DISCOVER is 12% in Table 4.3 and 18% in Table 4.4.

We observe that the top F1 score is obtained by DISCOVER Strat on BRCA dataset. This shows the benefit of considering subtype information for BRCA. Additionally, Fisher's Exact Test and MEGSA remain to be conservative, and they output no significant p-values when the control set is $X_2$.

Table S1-to-S12 show the corresponding results in the other cancer types. We observe that the sensitivity values decrease dramatically for BLCA data compared to the results on COADREAD data. In parallel, specificity values increase. In fact, Fisher's Exact Test and MEGSA have specificity values of 1 for both types of control sets since they are notably conservative. When we look at the other cancer types, we observe that WExT gives the best F1 score for all cancer types except for LUSC where MEGSA ranks the top. The second ranking method is DISCOVER for all cancer types except for LUSC. MEGSA and MEMO results are not available for some cancer types since we are not able to run them due to memory issues.

#### 4.2.3. Network-centric MLA Evaluations of Alternative ME Tests

Having compared the ME Tests with respect to our novel network-centric evaluation framework, we now assess the effect of incorporating network knowledge on mutation load confounding (MLC) introduced by [7]. van de Haar et al looked at the relation between the MLA of a gene and its percent significant findings in mutual exclusivity tests . They computed these statistics for 341 genes from an established cancer gene panel [46] where for each gene mutual exclusivity tests are performed to all the other genes in this panel. Here, we perform a similar analysis where we use the COSMIC CGC database [33] to define reference cancer genes as it is more comprehensive and up to date. Figure 4.1-A shows the MLA of the reference cancer genes vs the percent significant findings in mutual exclusivity tests performed with DISCOVER for the TCGA COADREAD cohort (498 tumors). We observe a strong negative correlation between MLA values and percent significant findings in mutual exclusivity tests (Pearson correlation -0.82, p-value $1.2e-42$) similar to van de Haar et al. In Figure 4.1-B, we take into account the PPI information to calculate percent significant findings. Namely, for each CGC gene, we perform mutual exclusivity tests only with its PPI neighbors that are in CGC. Note that genes with no CGC neighbors are removed from this analysis. We now observe a much smaller correlation between MLA and ME detection rate (Pearson correlation -0.4, p-value $4.91e-4$). As

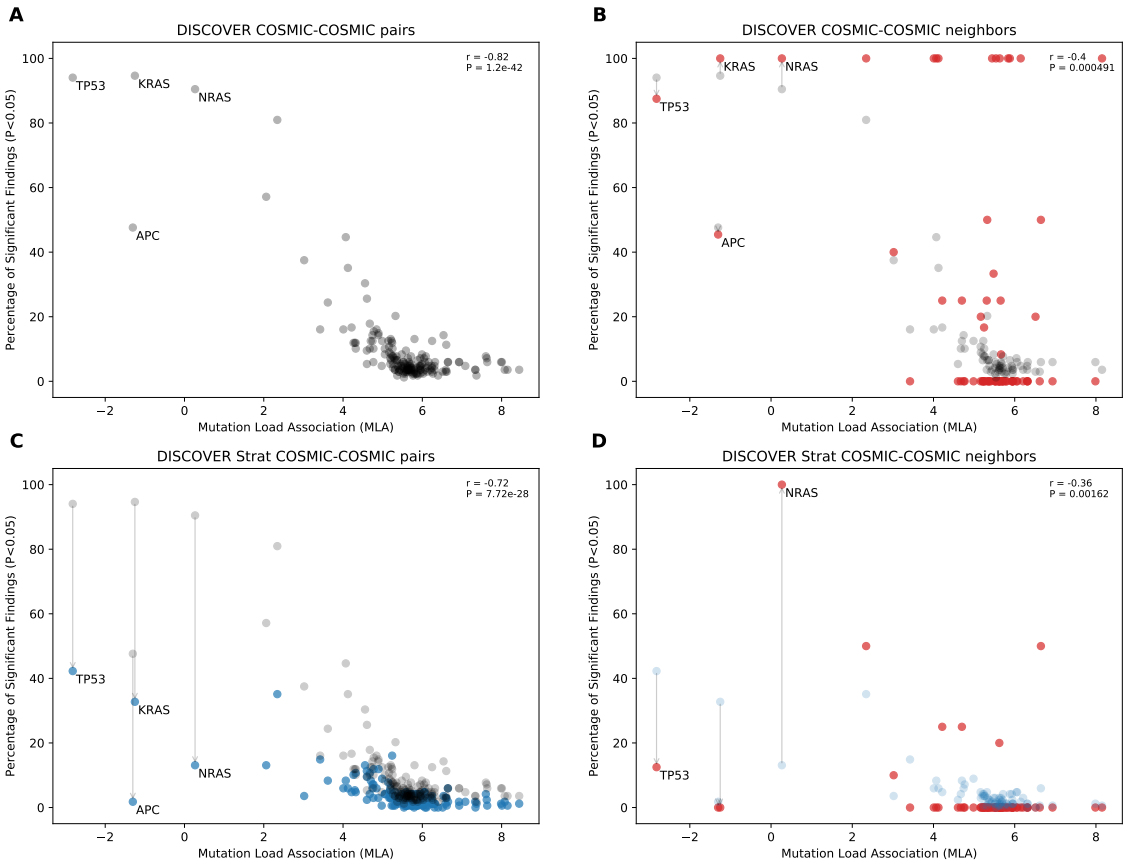


Figure 4.1: Comparison of mutual exclusivity results of DISCOVER and DISCOVER Strat on COADREAD cohort (498 samples) (A) The scatterplot of percentage significance of mutual exclusivity runs (p-value¡0.05) of DISCOVER on COADREAD data where tests are performed between a CGC gene and all other CGC genes . (B) The scatter plot of percentage significance of mutual exclusivity runs of DISCOVER where tests are performed between a CGC gene and its PPI neighbors that are in CGC (red) compared with (A) in gray. (C) The scatterplot of percentage significance of mutual exclusivity runs of DISCOVER Strat where where tests are performed between a CGC gene and all other CGC genes (blue) compared with the results from (A) in gray. (D) The scatterplot of percentage significance of mutual exclusivity runs of DISCOVER Strat where tests are performed between a CGC gene and its PPI neighbors that are in CGC (red) compared with (C) in blue.

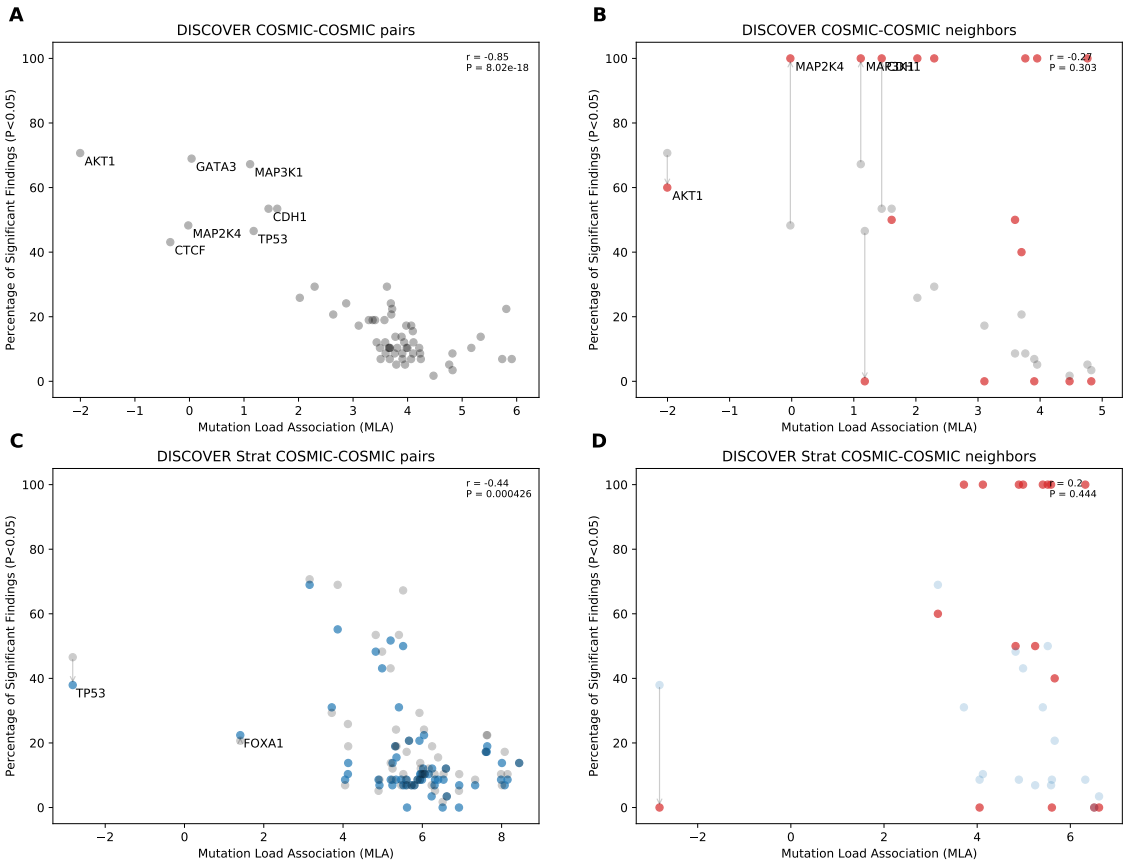


Figure 4.2: Comparison of mutual exclusivity results of DISCOVER and DISCOVER Strat on BRCA cohort (1026 samples) (A) The scatterplot of percentage significance of mutual exclusivity runs (p-value¡0.05) of DISCOVER on BRCA data where tests are performed between a CGC gene and all other CGC genes . (B) The scatter plot of percentage significance of mutual exclusivity runs of DISCOVER where tests are performed between a CGC gene and its PPI neighbors that are in CGC (red) compared with (A) in gray. (C) The scatterplot of percentage significance of mutual exclusivity runs of DISCOVER Strat where where tests are performed between a CGC gene and all other CGC genes (blue) compared with the results from (A) in gray. (D) The scatterplot of percentage significance of mutual exclusivity runs of DISCOVER Strat where tests are performed between a CGC gene and its PPI neighbors that are in CGC (red) compared with (C) in blue.

such, when ME is checked within neighbors of CGC genes, we observe a less strong MLC. We also run DISCOVER Strat where stratification is based on CMS subtypes [44]. We plot these results in Fig 4.1-C. Interestingly the negative correlation between MLA and ME detection rate is still strong when subtype information is incorporated (Pearson correlation -0.72, p-value $7.72e-28$). Lastly, in Fig 4.1-D we incorporate both subtype and PPI information where we use DISCOVER Strat and compute ME tests between CGC genes that are neighbors. Compared to Fig 4.1-B, correlation decreases from -0.4 to -0.36 indicating that including subtype information is still useful when used in addition to PPI. For BRCA, including subtype information does not decrease the correlation as opposed to what we observe for COADREAD. Namely, the correlation value that we calculate in 4.2-C is larger than that of 4.2-A. Similarly, adding subtype information on top of PPI information does not help (4.2-D vs 4.2-B).

We repeat the same analysis for other cancer types and also using the other ME detection methods as well (Supplementary S4 - S10). For all the other cancer types, we see a decrease in Pearson correlation coefficient values with the addition of PPI information for all ME detection methods except Fisher's Exact Test. Fisher's Exact Test results show an increased correlation with the addition of PPI information for LUSC and SKCM. Also, the orrelation value can not be computed for LUAD and STAD since Fisher's Exact Test gives a value of 0 for the percent significant findings of all considered genes. Another interesting observation is the variance in magnitude of decrease in correlation values across different tumor types. In particular, we observe a rather smaller decrease in correlation values for LUAD.

When we look at the change in percent significant findings for individual genes, we observe interesting patterns. For BLCA, we observe an increase in percent significant findings with the addition of PPI for low MLA such as CDKN1A and SMARCA4 in DISCOVER, WExT and MEGSA results. The inactivation of CDKN1A is considered a bladder specific change [47] whereas SMARCA4 is associated with poor prognosis in multiple cancer types including BLCA [48]. Similarly, MAP2K4, MET, NFE2L2 and TP53 show the same pattern for BRCA, LUAD, LUSC and UCEC tumors, respectively.

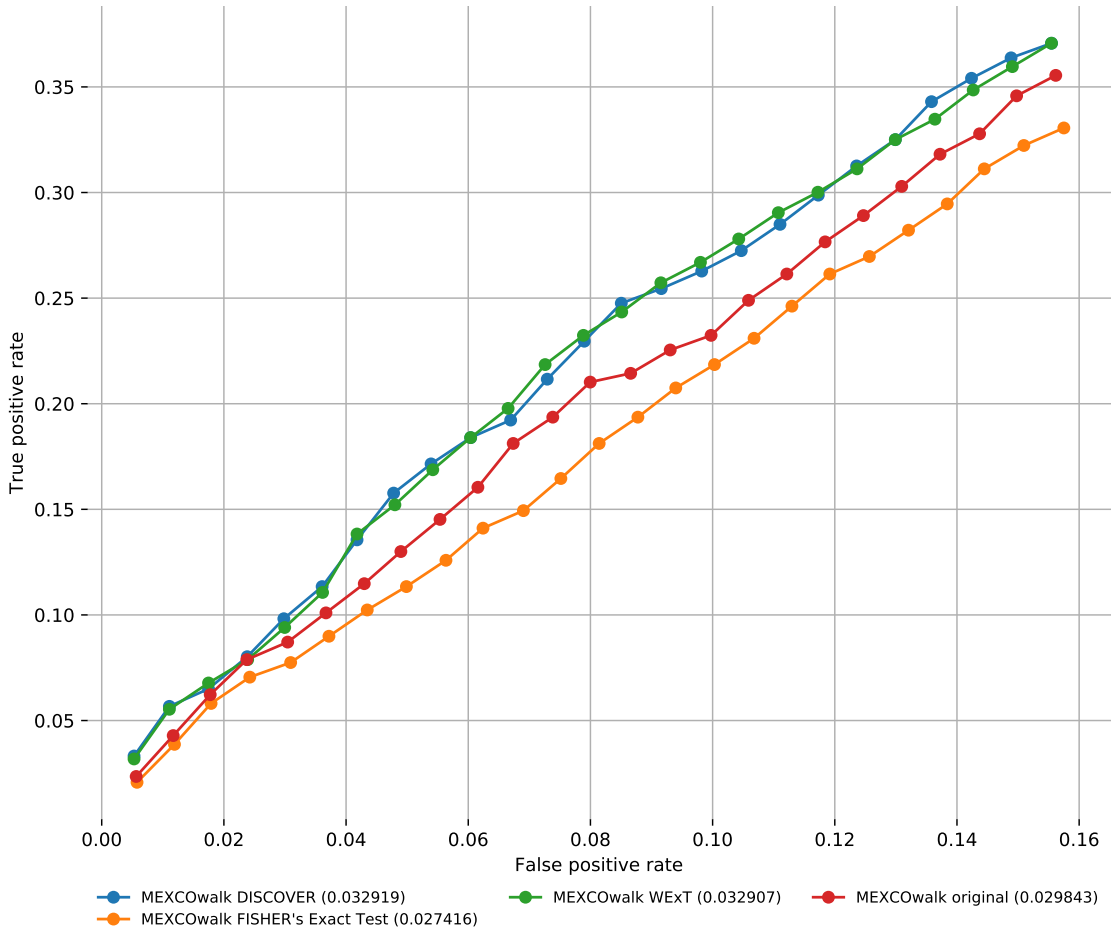


Figure 4.3: The figure shows the area under the ROC curve for MEXCOwalk runs on COADREAD t5 using the mutual exclusivity p-values as the MEX edge weight for each model and the filtered IntAct PPI network. In order to apply similar parameters to MEXCOwalk, number of edges assigned 0 weight is based on the density of the original MEXCOwalk run on HINT network. The original MEXCOwalk algorithm was run on COADREAD t5 with different threshold which are reflected through the model names. Note that, t5 was used because t20 didn't provide 2500 genes

### 4.2.4. Network-centric Epistasis in Identifying Driver Modules

We assess whether mutual exclusivities estimated by alternative ME methods improve the performance of driver identification methods that utilize mutual exclusivity information. To this end, we compare the original version of MEXCOWalk with its alternatives where mutual exclusivity estimates are provided by the employed ME methods. Assuming that $g_i$ and $g_j$ genes are mutated in patient sets $S_i$ and $S_j$, respectively; MEXCOWalk simply computes the mutual exclusivity between these two genes with the following formula: $|S_i \cup S_j/(|S_i| + |S_j|)$. MEXCOwalk uses the estimated mutual exclusivity values as part of edge weights. As such, to utilize p-values output by ME detection methods in MEXCOwalk, we first compute -log(p-value)'s and then convert them to a range between 0 and 1. To this end, we replace all -log(p-value)'s larger than 10 with 1. We then find the maximum -log(p-value) less than 10 and divide all other -log(p-value)'s with this value. The reason why we set a threshold for finding the maximum is the large differences across the smallest p-values output by different ME methods. For instance, WEXT outputs very small p-values and if we scale by dividing by the -log of this p-value, all other -log(p-value)s will be scaled to values that are very close to 0.

In MEXCOWalk, a threshold of 0.7 is applied to mutual exclusivity values such that all values $\leq 0.7$ are clamped to 0. This correspond to removing those edges from the graph. To find the corresponding threshold with the current ME values we computed the percent reduction in density when 0.7 threshold is applied in original MEXCOwalk. We then identified a threshold value for each ME detection method such that applying that threshold value results in the same percent density reduction.

Figure 4.3 shows the number of recovered CGC genes for fixed output gene sizes from 100 to 2500 as a ROC curve for original MEXCOwalk with versions of MEXCOWalk where mutual exclusivity values are estimated with DISCOVER, Fisher's Exact Test and WExT. We observe that MEXCOWalk with WExT's ME values results in the best AUROC value for COADREAD. For LUSC, STAD and UCEC, MEXCOwalk with DISCOVER's ME values gives the best AUROC. Similarly, MEXCOWalk with Fisher's Exact Test's ME values gives the top AUROC value for BLCA, LUAD and SKCM. Though MEXCOwalk with Fisher's Exact Test's ME values performs well for these tumor types,

it performs worse than original MEXCOwalk for three out of the four remaining tumor types. As such, using Fisher's Exact Test in place of MEXCOwalk's original ME values could decrease the performance.

# CHAPTER 5

# 5. Conclusion

## 5.1. Conclusion

Mutual exclusivity can be utilized as additional information in order to discover cancer driver genes. In this study, we introduce a novel method, MEXCOWalk, that incorporates network connectivity, mutual exclusion, and coverage information to identify cancer driver modules. We compare MEXCOwalk with existing cancer driver gene module finding algorithms. We follow up further by performing a network centric epistasis evaluation on existimg mutual exclusivity algorithms and applying them to MEXCOwalk for further evaluation.

The optimization function employed by MEXCOWalk combines the mutual exclusion and coverage scores of modules after normalizing with suitable functions of module size. MEXCOWalk employs a vertex-weighted, edge-weighted random walk strategy where the edge weights reflect a novel combination of mutual exclusion and coverage. MEXCOwalk is able to output a set of modules with a predefined size, that is $total\_genes$. This flexibility avoids *ad hoc* selection of an edge weight threshold and when applied to the other existing methods, it enables a robust comparison across different number of output genes. Another main contribution is to be able to split large modules in a systematic way, which becomes critical for large $total\_genes$ values. Finally, apart from methodological contributions, the proposed modularity-based metrics fill an important gap in the literature.

Our results indicate that MEXCOWalk outperforms several state-of-the-art computational methods on TCGA pan-cancer data in terms of recovering known cancer genes, providing modules that are capable of classifying normal and tumor samples, and that are enriched for mutations in specific cancer types. Additionally, the risk scores determined with output modules can stratify patients into low-risk and high-risk groups in multiple cancer types. We also show that MEXCOwalk is robust against different settings of its parameters.

We follow the results from MEXCOwalk by developing a network centric CGC oriented epistasis evaluation metic that we apply to existing mutual exclusivity finding algorithms. We then apply pairwise mutual exclusivity values estimated by the existing algorithms as the input to MEXCOwalk. Our results show a significant improvement in the recovery of known cancer driver genes.

In summary, in this research we were able to successfully implement a novel cancer gene module finding algorithm by utilizing mutual exlusivity values and further improve the results based on a network-centric approach.

## 5.2. Future Work

The research work has several future directions. Further analysis can be done on tissue specific networks in order to relate the recovered cancer genes from MEXCOwalk to specific tissues. We observed an increase in percentage significance of certain genes in specific cancer types. These were recently discovered as significantly related to these cancer types. Further investigations can be done in order to relate and discover similar driver genes from specific cohorts.

# Bibliography

[1] J. Weinstein, E. Collisson, G. Mills, K. Shaw, B. Ozenberger, K. Ellrott, C. Sander, and et al., "The cancer genome atlas pan-cancer analysis project," *Nature Genetics*, vol. 45, pp. 1113–1120, 10 2013.

[2] J. Dopazo and C. Erten, "Graph-theoretical comparison of normal and tumor networks in identifying brca genes," *BMC Systems Biology*, vol. 11, p. 110, Nov 2017.

[3] S. Erten, G. Bebek, and M. Koyuturk, "Vavien: An algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks," *J Comput Biol*, vol. 18, pp. 1561–74, 11 2011.

[4] M. Lawrence, P. Stojanov, P. Polak, G. V Kryukov, K. Cibulskis, A. Sivachenko, S. L Carter, C. Stewart, C. H Mermel, S. Roberts, A. Kiezun, P. S Hammerman, A. McKenna, Y. Drier, L. Zou, A. H Ramos, T. J Pugh, N. Stransky, E. Helman, and G. Getz, "Mutational heterogeneity in cancer and the search for new cancer-associated genes," *Nature*, vol. 499, 06 2013.

[5] H. Yang, Q. Wei, X. Zhong, H. Yang, and B. Li, "Cancer driver gene discovery through an integrative genomics approach in a non-parametric bayesian framework," *Bioinformatics*, vol. 33, no. 4, pp. 483–490, 2017.

[6] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, "Associating genes and protein complexes with disease via network propagation," *PLOS Comput Biol*, vol. 6, pp. 1–9, 01 2010.

[7] J. van de Haar, S. Canisius, M. K. Yu, E. E. Voest, L. F. Wessels, and T. Ideker, "Identifying epistasis in cancer genomes: A delicate affair," *Cell*, vol. 177, pp. 1375–1383, May 2019.

[8] C. H. Yeang, F. McCormick, and A. Levine, "Combinatorial patterns of somatic gene mutations in cancer," *FASEB J.*, vol. 22, pp. 2605–2622, Aug 2008.

[9] Y. Deng, S. Luo, C. Deng, T. Luo, W. Yin, H. Zhang, Y. Zhang, X. Zhang, Y. Lan, Y. Ping, Y. Xiao, and X. Li, "Identifying mutual exclusivity across cancer genomes: computational approaches to discover genetic interaction and reveal tumor vulnerability.," *Briefings in Bioinformatics*, 2017. Exported from https://app.dimensions.ai on 2019/01/21.

[10] C. M. Dimitrakopoulos and N. Beerenwinkel, "Computational approaches for the identification of cancer genes and pathways," 2017. Published online 11 November 2016.

[11] J. Zhang and S. Zhang, "The discovery of mutated driver pathways in cancer: Models and algorithms," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, pp. 988–998, May 2018.

[12] M. D. M. Leiserson, D. Blokh, R. Sharan, and B. J. Raphael, "Simultaneous identification of multiple driver pathways in cancer," *PLOS Comput Biol*, vol. 9, pp. 1–15, 05 2013.

[13] B. Liu, C. Wu, X. Shen, and W. Pan, "A novel and efficient algorithm for de novo discovery of mutated driver pathways in cancer," *Ann. Appl. Stat.*, vol. 11, pp. 1481–1512, 09 2017.

[14] C. A. Miller, S. H. Settle, E. P. Sulman, K. D. Aldape, and A. Milosavljevic, "Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors," *BMC Medical Genomics*, vol. 4, p. 34, Apr 2011.

[15] F. Vandin, E. Upfal, and B. J. Raphael, "*De Novo* discovery of mutated driver pathways in cancer," in *Research in Computational Molecular Biology - 15th Annual International Conference, RECOMB 2011, Vancouver, BC, Canada, March 28-31, 2011. Proceedings*, pp. 499–500, 2011.

[16] F. Vandin, E. Upfal, and B. Raphael, "Algorithms for detecting significantly mutated pathways in cancer," *J Comput Biol*, vol. 18, pp. 507–22, 03 2011.

[17] M. D. M. Leiserson, F. Vandin, H.-T. Wu, J. R. Dobson, J. V. Eldridge, J. L. Thomas, A. Papoutsaki, Y. Kim, B. Niu, M. McLellan, M. S. Lawrence, A. Gonzalez-Perez, D. Tamborero, Y. Cheng, G. A. Ryslik, N. Lopez-Bigas, G. Getz, L. Ding, and B. J. Raphael, "Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes," *Nature Genetics*, vol. 47, pp. 106–114, Dec. 2014.

[18] M. A. Reyna, M. D. M. Leiserson, and B. J. Raphael, "Hierarchical HotNet: identifying hierarchies of altered subnetworks," *Bioinformatics*, vol. 34, pp. i972–i980, Sept. 2018.

[19] G. Ciriello, E. Cerami, C. Sander, and N. Schultz, "Mutual exclusivity analysis identifies oncogenic network modules," *Genome Research*, vol. 22, pp. 398–406, Sept. 2012.

[20] Ö. Babur, M. Gönen, B. A. Aksoy, N. Schultz, G. Ciriello, C. Sander, and E. Demir, "Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations," *Genome Biology*, vol. 16, p. 45, Feb 2015.

[21] P. Dao, Y.-A. Kim, D. Wojtowicz, S. Madan, R. Sharan, and T. M. Przytycka, "Be-with: A between-within method to discover relationships between cancer modules via integrated analysis of mutual exclusivity, co-occurrence and functional interactions.," *PLoS Computational Biology*, vol. 13, no. 10, 2017.

[22] Y.-A. Kim, D.-Y. Cho, and T. M. Przytycka, "MEMCover: integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types," *Bioinformatics*, vol. 31, no. 12, pp. i284–i292, 2015.

[23] S. Canisius, J. W. M. Martens, and L. F. A. Wessels, "A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence," *Genome Biology*, vol. 17, Dec. 2016.

[24] M. D. Leiserson, M. A. Reyna, and B. J. Raphael, "A weighted exact test for mutually exclusive mutations in cancer," *Bioinformatics*, vol. 32, pp. i736–i745, Sept. 2016.

[25] X. Hua, P. L. Hyland, J. Huang, L. Song, B. Zhu, N. E. Caporaso, M. T. Landi, N. Chatterjee, and J. Shi, "MEGSA: A powerful and flexible framework for analyzing mutual exclusivity of tumor mutations," *The American Journal of Human Genetics*, vol. 98, pp. 442–455, Mar. 2016.

[26] M. Hofree, J. P. Shen, H. Carter, A. Gross, and T. Ideker, "Network-based stratification of tumor mutations," *Nature Methods*, vol. 10(11), pp. 1108–1115, 2013.

[27] M. Bersanelli, E. Mosca, D. Remondini, G. Castellani, and L. Milanesi, "Network diffusion-based analysis of high-throughput data for the detection of differentially enriched modules," *Scientific Reports*, vol. 6, no. 1, p. 34841, 2016.

[28] C. Yang, S.-G. Ge, and C.-H. Zheng, "ndmaSNF: cancer subtype discovery based on integrative framework assisted by network diffusion model," *Oncotarget*, vol. 8(51), pp. 89021–89032, 10 2017.

[29] C. Lei and J. Ruan, "A novel link prediction algorithm for reconstructing protein–protein interaction networks by topological similarity," *Bioinformatics*, vol. 29, pp. 355–364, Dec. 2013.

[30] F. Alkan and C. Erten, "RedNemo: topology-based PPI network reconstruction via repeated diffusion with neighborhood modifications," *Bioinformatics*, p. btw655, Oct. 2017.

[31] H. Yu, L. Tardivo, S. Tam, E. Weiner, and F. e. a. Gebreab, "Next-generation sequencing to generate interactome datasets," *Nature Methods*, vol. 8, pp. 478–480, 2011.

[32] J. Das and H. Yu, "Hint: High-quality protein interactomes and their applications in understanding human disease," *BMC Systems Biology*, vol. 6, p. 92, 2012.

[33] S. Forbes, D. Beare, H. Boutselakis, S. Bamford, N. Bindal, J. Tate, C. Cole, S. Ward, E. Dawson, and L. e. a. Ponting, "Cosmic: somatic cancer genetics at high-resolution," *Nucleic Acids Res*, vol. 45, pp. D777–D783, 2017.

[34] T. E. Taylor, F. Furnari, and W. Cavenee, "Targeting egfr for treatment of glioblastoma: Molecular basis to overcome resistance," *Curr Cancer Drug Targets*, vol. 12(3), pp. 97–209, 2012.

[35] S. Singel, C. Cornelius, K. Batten, G. Fasciani, W. Wright, L. Lum, and J. Shay, "A targeted rnai screen of the breast cancer genome identifies kif14 and tln1 as genes that modulate docetaxel chemosensitivity in triple-negative breast cancer.," *Clin Cancer Res*, vol. 19(8), pp. 2061–2070, 2013.

[36] K. Fang, W. Dai, Y. Ren, Y. Xu, S. Zhang, and Y. Qian, "Both talin-1 and talin-2 correlate with malignancy potential of the human hepatocellular carcinoma mhcc-97 l cell," *BMC Cancer*, vol. 16(45), pp. 2076–2079, 2016.

[37] D. Beer, S. Kardia, C. Huang, T. Giordano, A. Levin, D. Misek, L. Lin, C. G., T. Gharib, and D. e. a. Thomas, "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nature Medicine*, vol. 8(8), pp. 816–824, 2002.

[38] R. Shrestha, E. Hodzic, T. Sauwewald, P. Dao, K. Wang, J. Yeung, S. Anderson, F. Vandin, G. Haffari, C. Collins, and C. Sahinalp, "Hit'ndrive: patient-specific multidriver gene prioritization for precision oncology.," *Genome Res*, vol. 27(9), pp. 1573–1588, 2017.

[39] B. Karakas, K. Bachman, and B. Park, "Mutation of the pik3ca oncogene in human cancers," *British Journal of Cancer*, vol. 94, pp. 455–459, 2006.

[40] J. K. Kim and J. A. Diehl, "Nuclear cyclin d1: An oncogenic driver in human cancer," *J Cell Physiol*, vol. 220(2), pp. 292–296, 2010.

[41] M. Malumbres and M. Barbacid, "Cell cycle, CDKs and cancer: a changing paradigm," *Nat Rev Cancer*, vol. 9, pp. 153–166, 2009.

[42] M. G. Lee, R. Villa, P. Trojer, J. Norman, K.-P. Yan, D. Reinberg, L. D. Croce, and R. Shiekhattar, "Demethylation of h3k27 regulates polycomb recruitment and h2a ubiquitination," *Science*, vol. 318, pp. 447–450, Oct. 2007.

[43] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler, "Cancer Genome Landscapes," *Science*, vol. 339, pp. 1546–1558, Mar. 2013.

[44] J. Guinney, R. Dienstmann, X. Wang, A. de Reyniès, A. Schlicker, C. Soneson, L. Marisa, P. Roepman, G. Nyamundanda, P. Angelino, B. M. Bot, J. S. Morris, I. M. Simon, S. Gerster, E. Fessler, F. D. S. E. Melo, E. Missiaglia, H. Ramay, D. Barras, K. Homicsko, D. Maru, G. C. Manyam, B. Broom, V. Boige, B. Perez-Villamil, T. Laderas, R. Salazar, J. W. Gray, D. Hanahan, J. Tabernero, R. Bernards, S. H. Friend, P. Laurent-Puig, J. P. Medema, A. Sadanandam, L. Wessels, M. Delorenzi, S. Kopetz, L. Vermeulen, and S. Tejpar, "The consensus molecular subtypes of colorectal cancer," *Nature Medicine*, vol. 21, pp. 1350–1356, Oct. 2015.

[45] S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N. H. Campbell, G. Chavali, C. Chen, N. del Toro, M. Duesbury, M. Dumousseau, E. Galeota, U. Hinz, M. Iannuccelli, S. Jagannathan, R. Jimenez, J. Khadake, A. Lagreid, L. Licata, R. C. Lovering, B. Meldal, A. N. Melidoni, M. Milagros, D. Peluso, L. Perfetto, P. Porras, A. Raghunath, S. Ricard-Blum, B. Roechert, A. Stutz, M. Tognolli, K. van Roey, G. Cesareni, and H. Hermjakob, "The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases," *Nucleic Acids Research*, vol. 42, pp. D358–D363, Nov. 2014.

[46] D. T. Cheng, T. N. Mitchell, A. Zehir, R. H. Shah, R. Benayed, A. Syed, R. Chandramohan, Z. Y. Liu, H. H. Won, S. N. Scott, A. R. Brannon, C. O. Reilly, J. Sadowska, J. Casanova, A. Yannes, J. F. Hechtman, J. Yao, W. Song, D. S. Ross, A. Oultache, S. Dogan, L. Borsu, M. Hameed, K. Nafa, M. E. Arcila, M. Ladanyi, and M. F. Berger, "Memorial sloan kettering-integrated mutation profiling of actionable cancer targets (MSK-IMPACT)," *The Journal of Molecular Diagnostics*, vol. 17, pp. 251–264, May 2015.

[47] J. B. Cazier, , S. R. Rao, C. M. McLean, A. K. Walker, B. J. Wright, E. E. M. Jaeger, C. Kartsonaki, L. Marsden, C. Yau, C. Camps, P. Kaisaki, J. Taylor, J. W. Catto, I. P. M. Tomlinson, A. E. Kiltie, and F. C. Hamdy, "Whole-genome sequencing of bladder cancers reveals somatic CDKN1a mutations and clinicopathological associations with mutation burden," *Nature Communications*, vol. 5, Apr. 2014.

[48] J. A. Guerrero-Martínez and J. C. Reyes, "High expression of SMARCA4 or SMARCA2 is frequently associated with an opposite prognosis in cancer," *Scientific Reports*, vol. 8, Feb. 2018.

[49] T. C. G. A. R. Network, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, pp. 1061–1068, 2008.

[50] F. Vandin, E. Upfal, and B. J. Raphael, "De novo discovery of mutated driver pathways in cancer," *Genome Res.*, vol. 22, pp. 375–385, Feb 2012.

[51] M. D. Leiserson, D. Blokh, R. Sharan, and B. J. Raphael, "Simultaneous identification of multiple driver pathways in cancer," *PLoS Comput. Biol.*, vol. 9, no. 5, p. e1003054, 2013.

[52] C.-H. Yeang, F. McCormick, and A. Levine, "Combinatorial patterns of somatic gene mutations in cancer," *The FASEB Journal*, vol. 22, no. 8, pp. 2605–2622, 2008.

[53] X. Liu, J. Xi, C. Zhang, H. Feng, A. Li, and M. Wang, "Identification of driver network modules in protein-protein interaction network using patient mutation profiles," in *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, (Shanghai), pp. 1–6, IEEE, Oct. 2017.

[54] B. J. Raphael, J. R. Dobson, L. Oesper, and F. Vandin, "Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine," *Genome Medicine*, vol. 6, no. 1, p. 5, 2014.

[55] H. Wu, L. Gao, and N. K. Kasabov, "Network-based method for inferring cancer progression at the pathway level from cross-sectional mutation data," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 13, pp. 1036–1044, Nov. 2016.

[56] R. Ahmed, I. Baali, C. Erten, E. Hoxha, and H. Kazan, "MEXCOwalk: mutual exclusion and coverage based random walk to identify cancer modules," *Bioinformatics*, vol. 36, pp. 872–879, 08 2019.

[57] C.-H. Yeang, F. McCormick, and A. Levine, "Combinatorial patterns of somatic gene mutations in cancer," *The FASEB Journal*, vol. 22, no. 8, pp. 2605–2622, 2008.

[58] M. Reyna, M. Leiserson, and B. Raphael, "Hierarchical HotNet: identifying hierarchies of altered subnetworks. ," *Bioinformatics*, vol. 34, no. 17, pp. i972–980, 2018.

# Appendix A

# Supplementary

## A.1. MEXCOwalk with different parameter settings

### A.1.1. Effects of Mutual Exclusivity Threshold $\theta$

Almost all edges from the distribution of $MEX_n$ scores are larger than 0.5. Therefore, we experiment with $\theta$ values greater than or equal to 0.5: $\{0.5, 0.6, 0.7, 0.8, 0.9\}$. Figure S1 show the results of running MEXCOwalk with different $\theta$ values. In parts A and B, we observe that changing $\theta$ has a minimal effect in recovering known cancer genes with $\theta = 0.7$ giving the largest area under the ROC curve. Figure S1 -C reveals a strong correlation between $\theta$ scores and Mutual Exclusivity (MS) scores. Since large values of $\theta$ clamp a larger set of edge weights to 0, the resulting modules are only those with really large mutual exclusivity scores. We observe the largest coverage scores when $\theta$ is set to 0.7 (Figure S1-D), whereas 0.9 results in significantly lower coverage scores across all *total_genes* values. This is likely due to mutual exclusivity dominating over coverage in edge weights. Finally, we observe that Driver Module Set Scores are mosty in parallel with Coverage Scores; see Figure S1-B.

### A.1.2. Effects of *min_module_size*

We experiment with a range of values for *min_module_size*; see Figure S2. We see minimal differences in overlaps with CGC database where $min\_module\_size = 3$ results

in the largest area under ROC curve. As expected, there is an inverse correlation between *min_module_size* and mutual exclusion scores (Figure S2-C). On the other hand, we observe a positive correlation between *min_module_size* and coverage scores. Again, we observe that Driver Module Set Scores are in parallel with Coverage Scores (Figure S2-B).

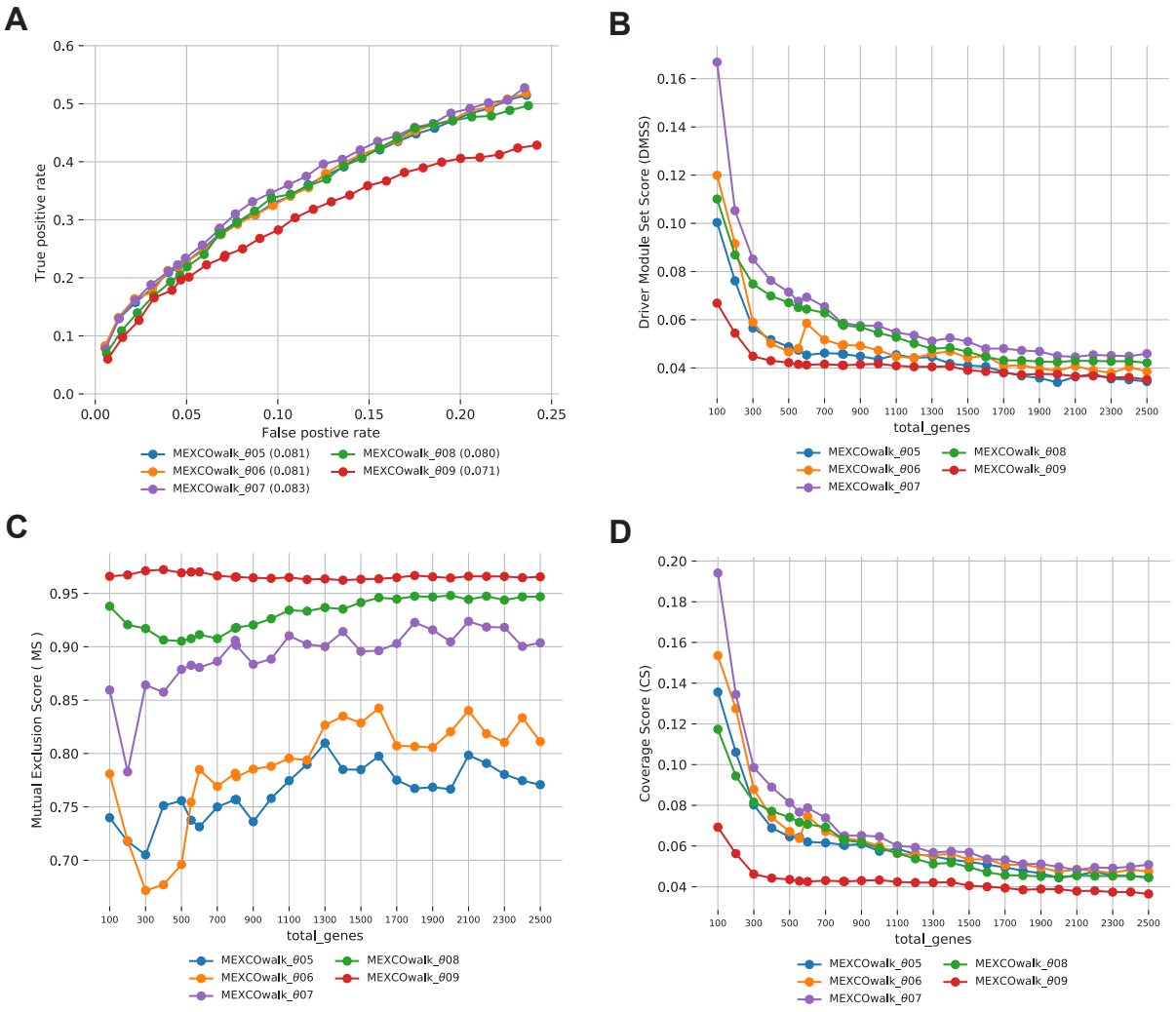


Figure S1: Comparison of MEXCOwalk models with different mutual exclusion score thresholds ($\theta$): 0.5, 0.6, 0.7, 0.8 and 0.9 A) ROC plots and AUROC values written in parentheses. B) Driver Modules Set Score (DMSS). C) Mutual Exclusion Score (MS). D) Coverage Score (CS).

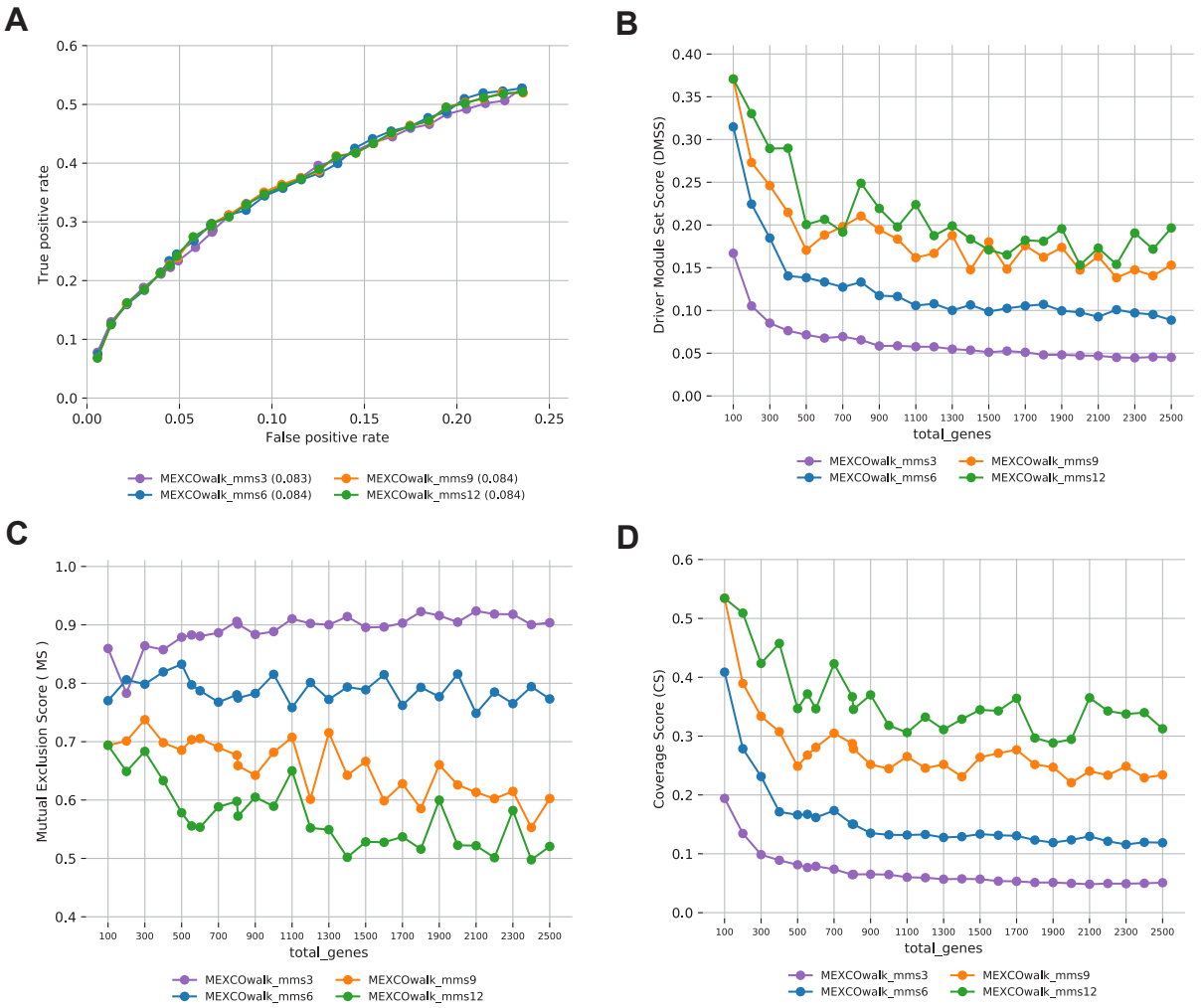


Figure S2: Comparison of MEXCOwalk models with different *min_module_size*: 3, 6, 9, 12 A) ROC plots and AUROC values written in parentheses. B) Driver Modules Set Score (DMSS). C) Mutual Exclusion Score (MS). D) Coverage Score (CS).

## A.2. Network-centric epistatic evaluation framework with control group $X_1$ and $X_2$ for mutation threshold 20

Following are tables containing the network centric epistasis evaluation results for different cancer types.

Table S1: BLCA control group $X_1$ (56 COSMIC-COSMIC pairs)

| Method | Stat1 | Stat2 | Stat3 | Stat4 | Stat5 | Stat6 | Stat7 | Stat8 | Precision | Sensitivity | Specificity | F1 Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DISCOVER | 4.0 | 0.0 | 0.0 | 21.0 | 26.0 | 1.0 | 1.073 | 0.994 | 0.800 | 0.077 | 0.981 | 0.140 |
| Fisher's Exact Test | 2.0 | 0.0 | 0.0 | 23.0 | 30.0 | 0.0 | 0.371 | 0.283 | 1.000 | 0.036 | 1.000 | 0.070 |
| MEGSA | 4.0 | 0.0 | 0.0 | 10.0 | 42.0 | 0.0 | 0.947 | 0.803 | 1.000 | 0.071 | 1.000 | 0.133 |
| MEMO | 4.0 | 0.0 | 0.0 | 20.0 | 25.0 | 2.0 | 1.182 | 1.059 | 0.667 | 0.078 | 0.961 | 0.140 |
| WExT | 4.0 | 0.0 | 0.0 | 22.0 | 25.0 | 3.0 | 1.378 | 1.240 | 0.571 | 0.074 | 0.944 | 0.131 |

Table S2: BLCA control group $X_2$ (24 COSMIC-COSMIC pairs)

| Method | Stat1 | Stat2 | Stat3 | Stat4 | Stat5 | Stat6 | Stat7 | Stat8 | Precision | Sensitivity | Specificity | F1 Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DISCOVER | 2.0 | 0.0 | 0.0 | 8.0 | 14.0 | 0.0 | 0.992 | 0.914 | 1.000 | 0.083 | 1.000 | 0.154 |
| Fisher's Exact Test | 1.0 | 0.0 | 0.0 | 10.0 | 13.0 | 0.0 | 0.354 | 0.297 | 1.000 | 0.042 | 1.000 | 0.080 |
| MEGSA | 2.0 | 0.0 | 0.0 | 3.0 | 19.0 | 0.0 | 0.967 | 0.824 | 1.000 | 0.083 | 1.000 | 0.154 |
| MEMO | 2.0 | 0.0 | 0.0 | 8.0 | 13.0 | 1.0 | 1.112 | 1.006 | 0.667 | 0.083 | 0.958 | 0.148 |
| WExT | 2.0 | 0.0 | 0.0 | 10.0 | 10.0 | 2.0 | 1.300 | 1.204 | 0.500 | 0.083 | 0.917 | 0.143 |

Table S3: LUAD control group $X_1$ (92 COSMIC-COSMIC pairs)

| Method | Stat1 | Stat2 | Stat3 | Stat4 | Stat5 | Stat6 | Stat7 | Stat8 | Precision | Sensitivity | Specificity | F1 Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DISCOVER | 7.0 | 0.0 | 2.0 | 46.0 | 32.0 | 1.0 | 1.354 | 1.078 | 0.750 | 0.102 | 0.966 | 0.180 |
| Fisher's Exact Test | 0.0 | 0.0 | 0.0 | 54.5 | 35.5 | 1.0 | 0.465 | 0.324 | 0.000 | 0.000 | 0.989 | NaN |
| MEGSA | 2.0 | 0.0 | 0.0 | 28.5 | 59.5 | 1.0 | 1.045 | 0.912 | 0.667 | 0.022 | 0.989 | 0.043 |
| MEMO | 10.0 | 1.5 | 2.0 | 42.5 | 29.0 | 3.5 | 1.681 | 1.294 | 0.659 | 0.153 | 0.921 | 0.248 |
| WExT | 12.0 | 2.0 | 2.0 | 39.5 | 30.0 | 5.5 | 1.902 | 1.520 | 0.627 | 0.176 | 0.896 | 0.275 |

Table S4: LUAD control group $X_2$ (59 COSMIC-COSMIC pairs)

| Method | Stat1 | Stat2 | Stat3 | Stat4 | Stat5 | Stat6 | Stat7 | Stat8 | Precision | Sensitivity | Specificity | F1 Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DISCOVER | 5.0 | 1.0 | 2.0 | 29.0 | 18.0 | 3.0 | 1.495 | 1.082 | 0.571 | 0.138 | 0.897 | 0.222 |
| Fisher's Exact Test | 0.0 | 0.0 | 0.0 | 36.0 | 22.0 | 0.0 | 0.538 | 0.347 | NaN | 0.000 | 1.000 | NaN |
| MEGSA | 2.0 | 0.0 | 0.0 | 19.0 | 35.0 | 3.0 | 1.139 | 1.039 | 0.400 | 0.034 | 0.949 | 0.062 |
| MEMO | 9.0 | 1.0 | 1.0 | 25.0 | 18.0 | 3.0 | 1.845 | 1.328 | 0.688 | 0.193 | 0.912 | 0.301 |
| WExT | 10.0 | 1.0 | 2.0 | 25.0 | 17.0 | 3.0 | 2.083 | 1.463 | 0.684 | 0.224 | 0.897 | 0.338 |

Table S5: LUSC control group $X_1$ (38 COSMIC-COSMIC pairs)

| Method | Stat1 | Stat2 | Stat3 | Stat4 | Stat5 | Stat6 | Stat7 | Stat8 | Precision | Sensitivity | Specificity | F1 Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DISCOVER | 2.0 | 0.0 | 0.0 | 21.0 | 14.5 | 0.0 | 1.154 | 0.720 | 1.0 | 0.053 | 1.000 | 0.101 |
| Fisher's Exact Test | 2.0 | 0.0 | 0.0 | 19.0 | 17.0 | 0.0 | 0.690 | 0.392 | 1.0 | 0.053 | 1.000 | 0.100 |
| MEGSA | 6.0 | 0.0 | 0.0 | 10.0 | 22.0 | 0.0 | 1.276 | 0.834 | 1.0 | 0.158 | 1.000 | 0.273 |
| MEMO | 4.0 | 0.0 | 0.0 | 18.0 | 14.0 | 0.0 | 1.292 | 0.875 | 1.0 | 0.111 | 1.000 | 0.200 |
| WExT | 4.0 | 0.0 | 0.0 | 18.0 | 14.0 | 1.0 | 1.470 | 1.014 | 0.8 | 0.108 | 0.973 | 0.190 |

Table S6: LUSC control group $X_2$ (22 COSMIC-COSMIC pairs)

| Method | Stat1 | Stat2 | Stat3 | Stat4 | Stat5 | Stat6 | Stat7 | Stat8 | Precision | Sensitivity | Specificity | F1 Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DISCOVER | 0.0 | 0.0 | 0.0 | 14.0 | 7.0 | 0.0 | 0.982 | 0.747 | NaN | 0.000 | 1.000 | NaN |
| Fisher's Exact Test | 0.0 | 0.0 | 0.0 | 13.0 | 9.0 | 0.0 | 0.571 | 0.406 | NaN | 0.000 | 1.000 | NaN |
| MEGSA | 3.0 | 0.0 | 0.0 | 5.0 | 14.0 | 0.0 | 1.184 | 1.009 | 1.0 | 0.136 | 1.000 | 0.240 |
| MEMO | 2.0 | 0.0 | 0.0 | 12.0 | 7.5 | 0.0 | 1.034 | 0.808 | 1.0 | 0.093 | 1.000 | 0.170 |
| WExT | 1.0 | 1.0 | 0.0 | 11.5 | 7.5 | 1.0 | 1.243 | 1.044 | 0.5 | 0.091 | 0.909 | 0.154 |

Table S7: SKCM control group $X_1$ (458 COSMIC-COSMIC pairs)

| Method | Stat1 | Stat2 | Stat3 | Stat4 | Stat5 | Stat6 | Stat7 | Stat8 | Precision | Sensitivity | Specificity | F1 Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DISCOVER | 19.0 | 1.0 | 0.0 | 214.5 | 210.5 | 4.0 | 1.223 | 0.818 | 0.800 | 0.045 | 0.989 | 0.084 |
| Fisher's Exact Test | 2.0 | 0.0 | 0.0 | 236.5 | 219.5 | 0.0 | 0.491 | 0.112 | 1.000 | 0.004 | 1.000 | 0.009 |
| MEGSA | 8.0 | 0.0 | 0.0 | 65.0 | 383.0 | 1.0 | 1.126 | 0.731 | 0.889 | 0.018 | 0.998 | 0.034 |
| WExT | 48.0 | 2.0 | 2.0 | 181.0 | 181.5 | 16.5 | 1.923 | 1.184 | 0.717 | 0.121 | 0.952 | 0.207 |

Table S8: SKCM control group $X_2$ (313 COSMIC-COSMIC pairs)

| Method | Stat1 | Stat2 | Stat3 | Stat4 | Stat5 | Stat6 | Stat7 | Stat8 | Precision | Sensitivity | Specificity | F1 Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DISCOVER | 14.0 | 0.0 | 0.0 | 149.5 | 144.0 | 4.5 | 1.369 | 0.783 | 0.757 | 0.045 | 0.986 | 0.085 |
| Fisher's Exact Test | 2.0 | 0.0 | 0.0 | 160.5 | 150.5 | 0.0 | 0.641 | 0.115 | 1.000 | 0.006 | 1.000 | 0.013 |
| MEGSA | 6.0 | 0.0 | 0.0 | 43.0 | 263.0 | 0.5 | 1.278 | 0.731 | 0.923 | 0.019 | 0.998 | 0.038 |
| WExT | 32.0 | 3.0 | 1.0 | 126.5 | 129.0 | 13.0 | 2.213 | 1.135 | 0.679 | 0.118 | 0.944 | 0.201 |

Table S9: STAD control group $X_1$ (140 COSMIC-COSMIC pairs)

| Method | Stat1 | Stat2 | Stat3 | Stat4 | Stat5 | Stat6 | Stat7 | Stat8 | Precision | Sensitivity | Specificity | F1 Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DISCOVER | 11.0 | 2.0 | 3.0 | 65.0 | 47.5 | 3.0 | 1.357 | 1.047 | 0.667 | 0.122 | 0.939 | 0.206 |
| Fisher's Exact Test | 0.0 | 0.0 | 0.0 | 83.0 | 56.0 | 1.0 | 0.167 | 0.105 | 0.000 | 0.000 | 0.993 | NaN |
| MEGSA | 2.0 | 0.0 | 0.0 | 14.0 | 123.0 | 1.0 | 0.783 | 0.752 | 0.667 | 0.014 | 0.993 | 0.028 |
| WExT | 18.0 | 3.0 | 5.0 | 60.0 | 44.0 | 7.0 | 2.015 | 1.469 | 0.634 | 0.190 | 0.891 | 0.292 |

Table S10: STAD control group $X_2$ (70 COSMIC-COSMIC pairs)

| Method | Stat1 | Stat2 | Stat3 | Stat4 | Stat5 | Stat6 | Stat7 | Stat8 | Precision | Sensitivity | Specificity | F1 Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DISCOVER | 5.0 | 3.0 | 1.0 | 38.0 | 21.0 | 2.0 | 1.525 | 0.973 | 0.600 | 0.129 | 0.914 | 0.212 |
| Fisher's Exact Test | 0.0 | 0.0 | 0.0 | 46.0 | 24.0 | 0.0 | 0.206 | 0.116 | NaN | 0.000 | 1.000 | NaN |
| MEGSA | 1.0 | 0.0 | 0.0 | 9.0 | 60.0 | 0.0 | 0.806 | 0.758 | 1.000 | 0.014 | 1.000 | 0.028 |
| WExT | 10.0 | 4.0 | 2.0 | 32.5 | 20.5 | 1.0 | 2.329 | 1.427 | 0.696 | 0.229 | 0.900 | 0.344 |

Table S11: UCEC control group $X_1$ (1356 COSMIC-COSMIC pairs)

| Method | Stat1 | Stat2 | Stat3 | Stat4 | Stat5 | Stat6 | Stat7 | Stat8 | Precision | Sensitivity | Specificity | F1 Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DISCOVER | 141.5 | 49.0 | 45.0 | 567.0 | 494.0 | 32.5 | 2.322 | 1.521 | 0.651 | 0.177 | 0.905 | 0.279 |
| Fisher's Exact Test | 10.0 | 0.0 | 0.0 | 705.0 | 638.5 | 2.0 | 0.102 | 0.034 | 0.833 | 0.007 | 0.999 | 0.015 |
| MEGSA | 11.0 | 0.0 | 0.0 | 42.0 | 1299.0 | 3.0 | 0.778 | 0.720 | 0.786 | 0.008 | 0.998 | 0.016 |
| WExT | 217.0 | 76.5 | 66.0 | 456.0 | 397.0 | 82.0 | 4.601 | 2.901 | 0.616 | 0.278 | 0.827 | 0.383 |

Table S12: UCEC control group $X_2$ (1179 COSMIC-COSMIC pairs)

| Method | Stat1 | Stat2 | Stat3 | Stat4 | Stat5 | Stat6 | Stat7 | Stat8 | Precision | Sensitivity | Specificity | F1 Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DISCOVER | 125.0 | 55.0 | 36.0 | 498.5 | 431.0 | 26.0 | 2.456 | 1.421 | 0.649 | 0.184 | 0.900 | 0.287 |
| Fisher's Exact Test | 10.0 | 0.0 | 0.0 | 573.5 | 593.5 | 2.0 | 0.114 | 0.027 | 0.833 | 0.008 | 0.998 | 0.017 |
| MEGSA | 11.5 | 0.0 | 0.0 | 41.0 | 1123.0 | 3.0 | 0.788 | 0.710 | 0.793 | 0.010 | 0.997 | 0.019 |
| WExT | 191.0 | 86.5 | 52.5 | 398.0 | 351.0 | 74.5 | 4.881 | 2.747 | 0.607 | 0.286 | 0.815 | 0.389 |

## A.3. Percentage significance finding of different cancer types

Following figures contain the percentage significance findings from mutual exclusivity results for different cancer types.

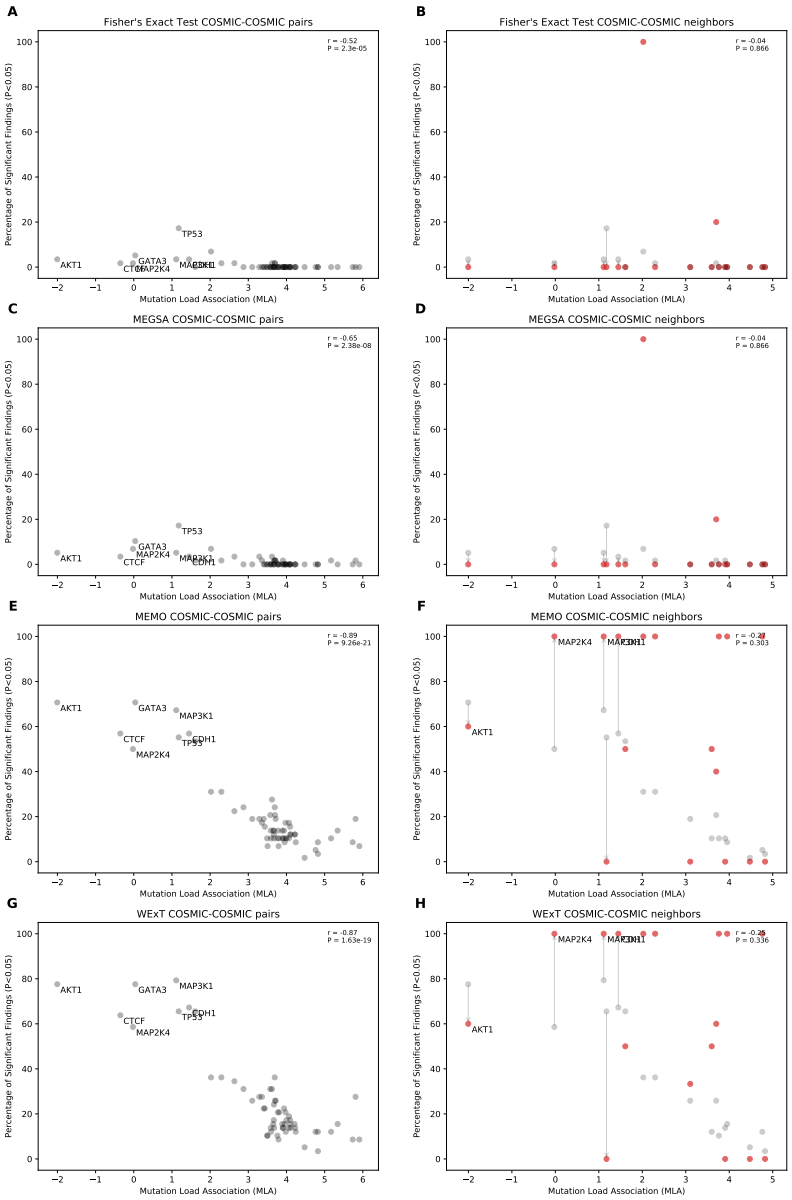


Figure S3: BRCA

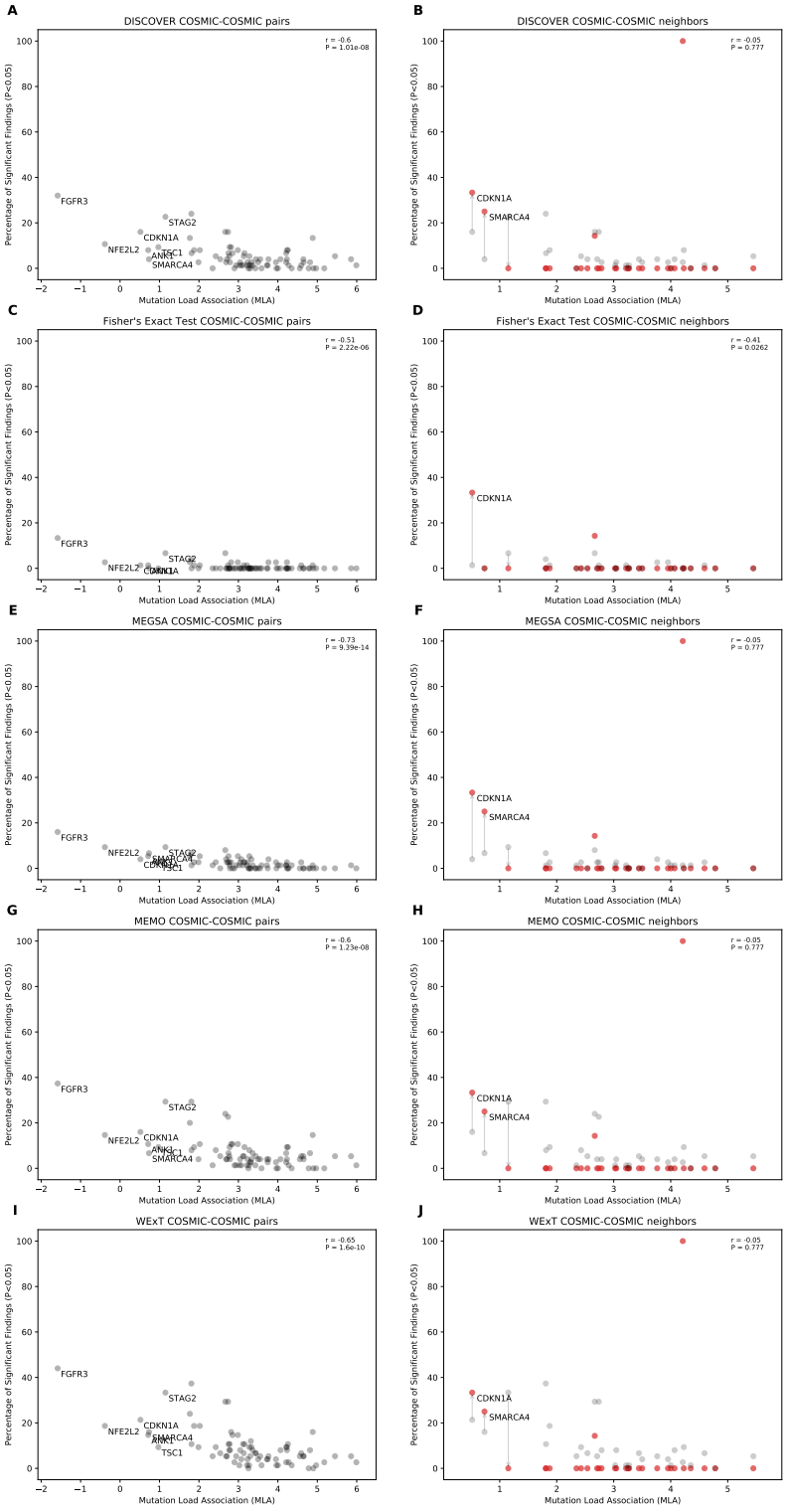


Figure S4: BLCA

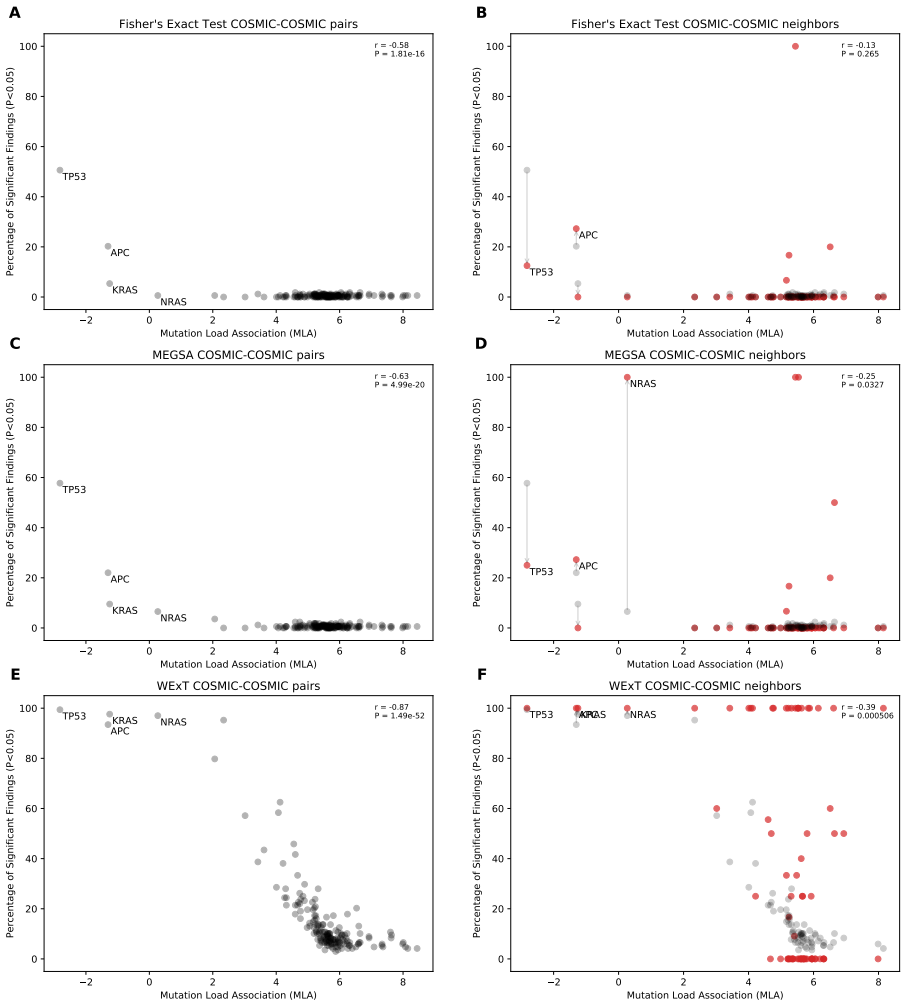


Figure S5: COADREAD

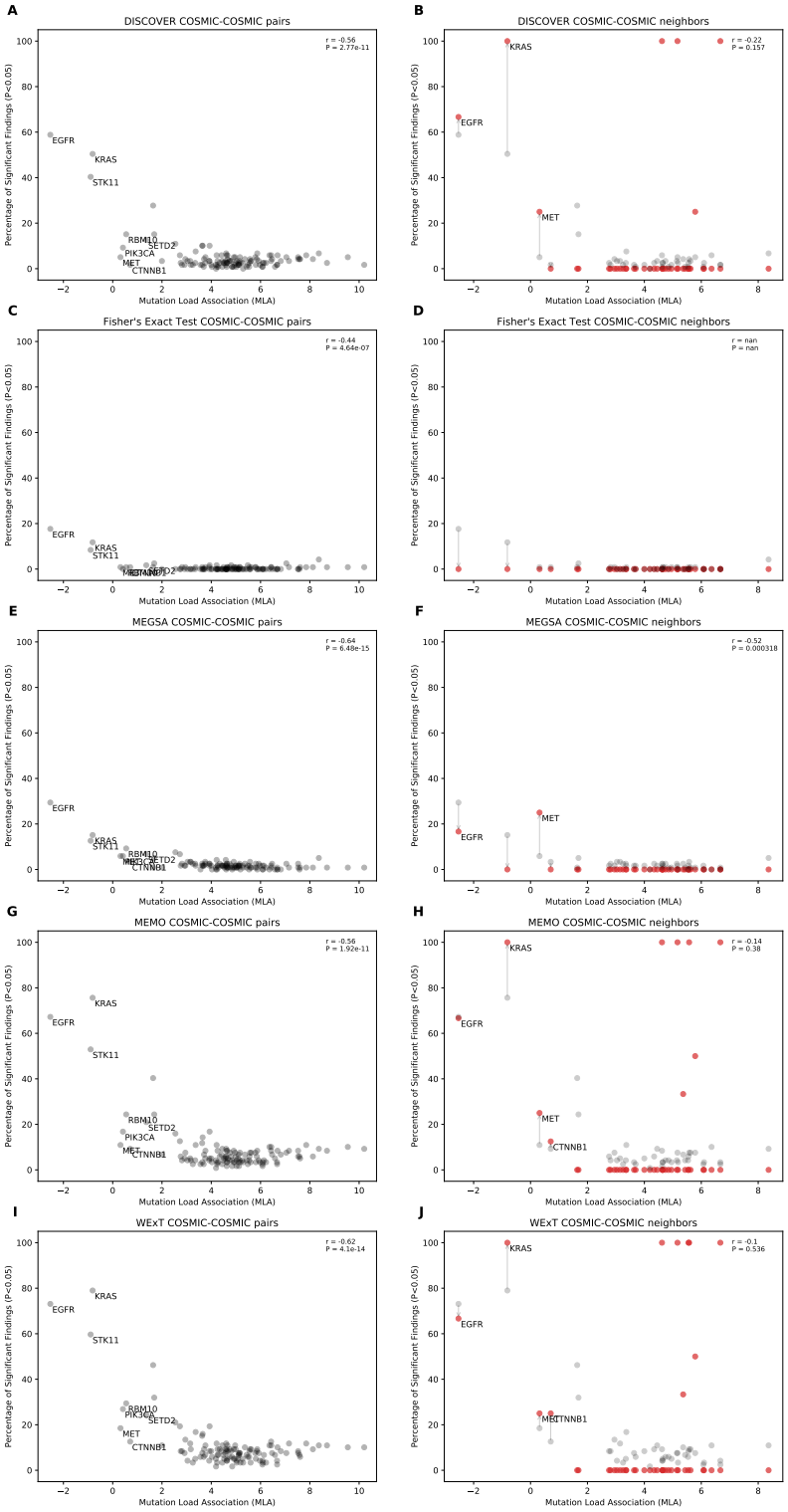


Figure S6: LUAD

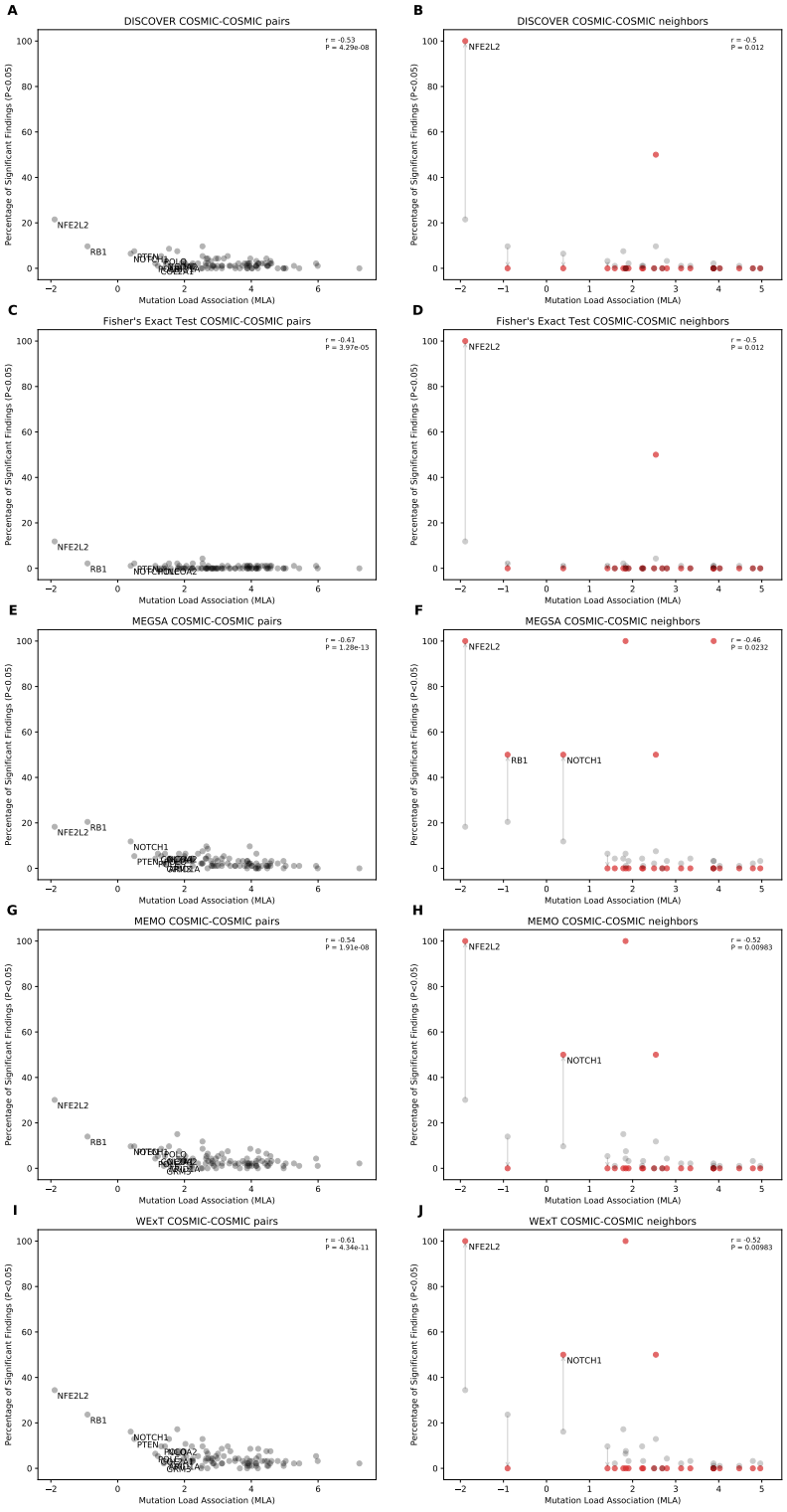


Figure S7: LUSC

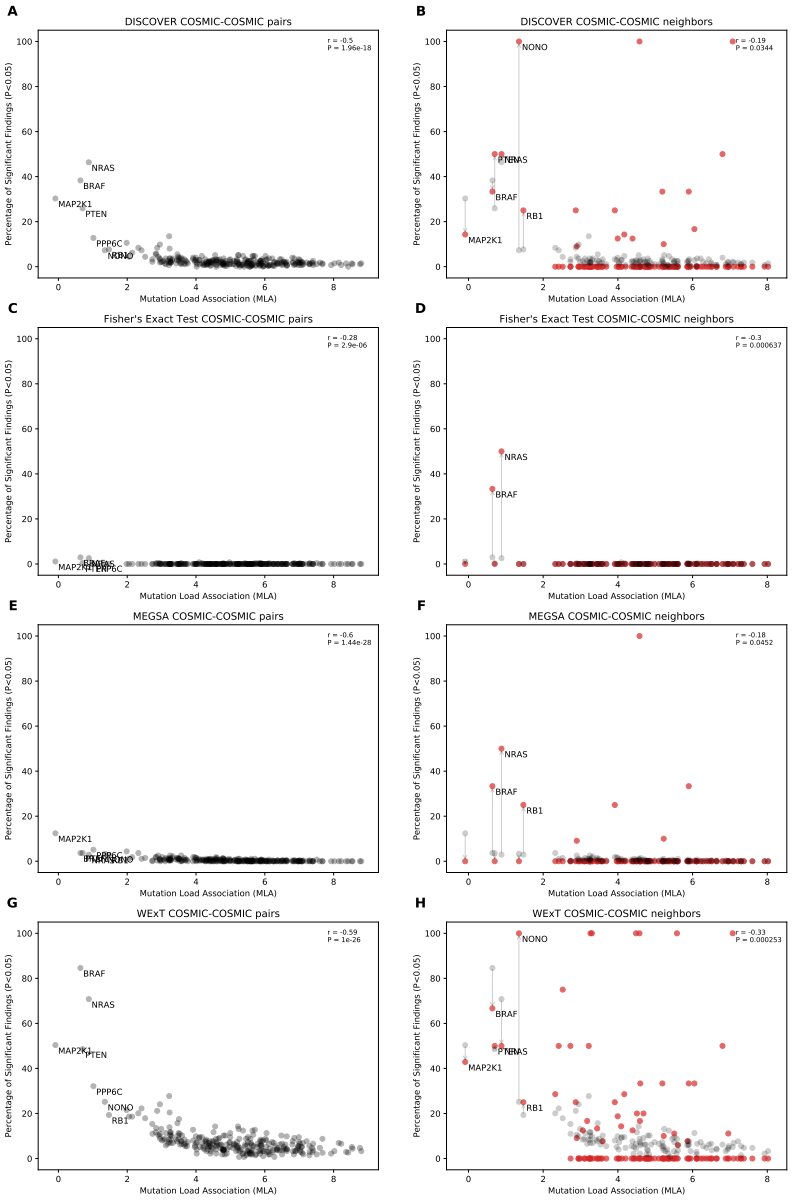


Figure S8: SKCM

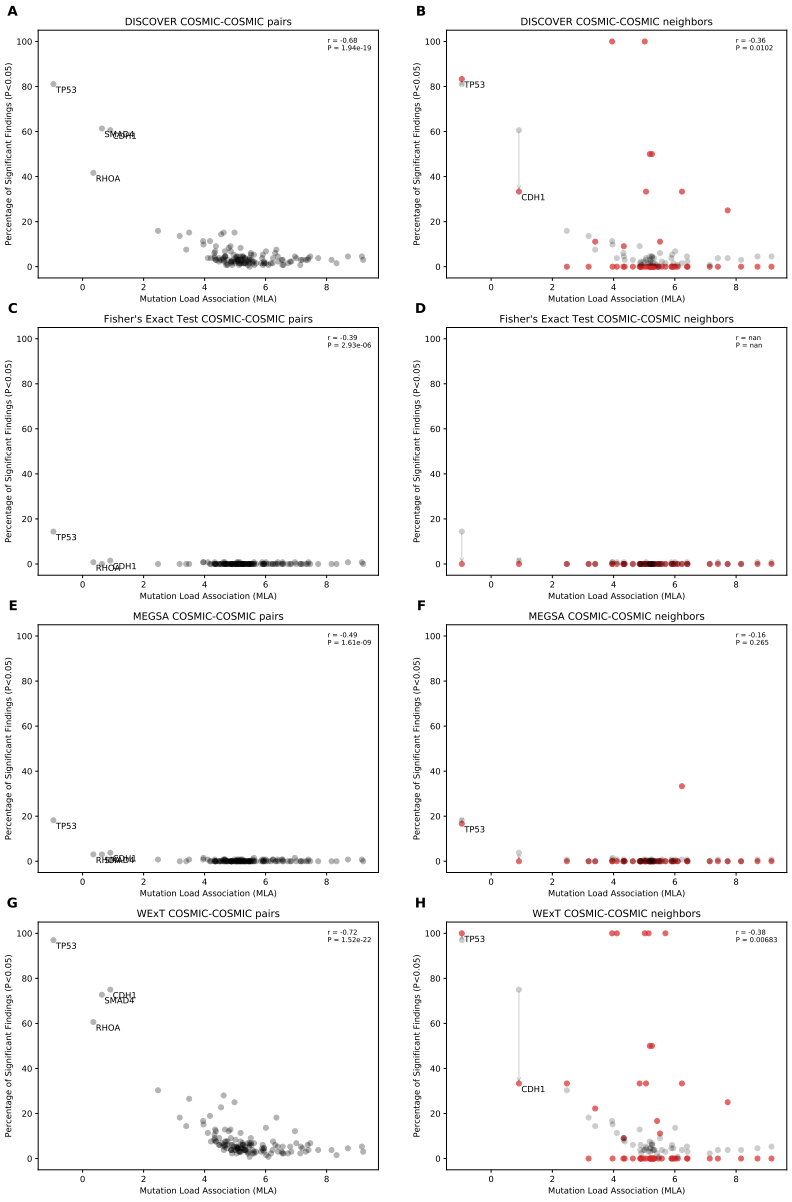


Figure S9: STAD

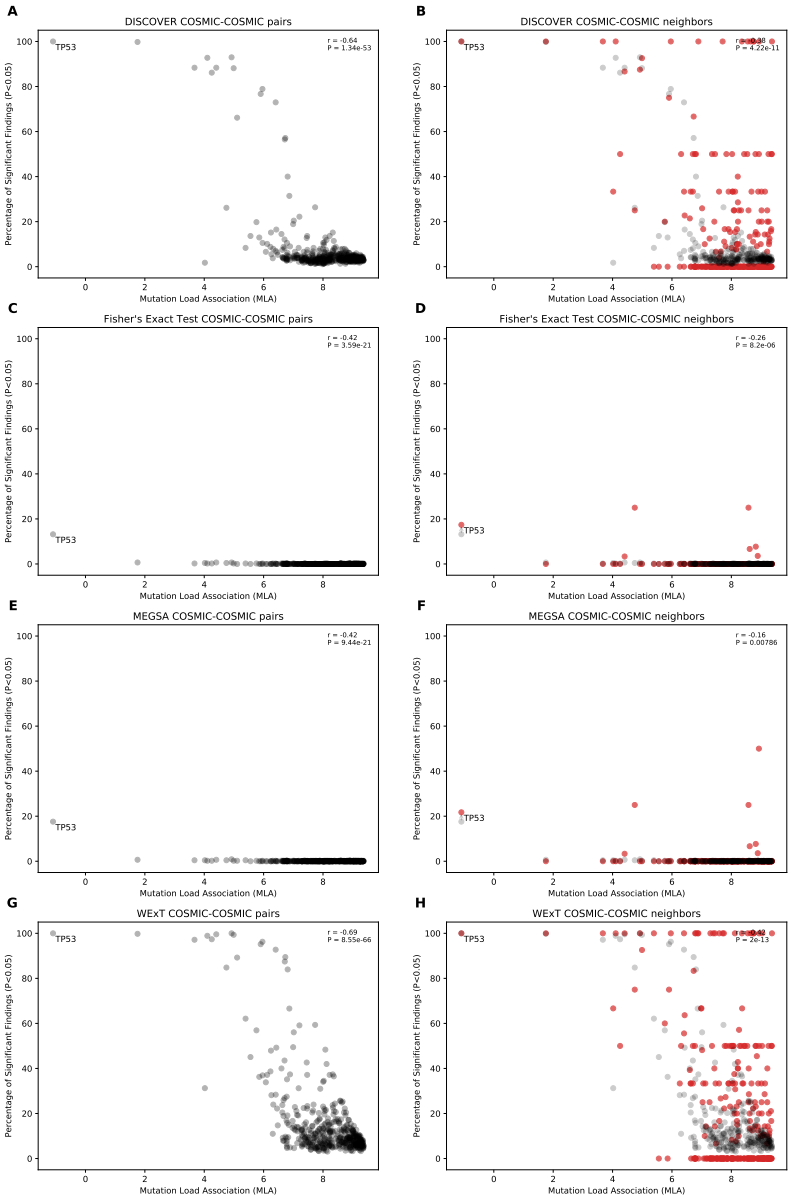


Figure S10: UCEC

## A.4. AUROC for different cancer types

The following figure contains the AUROC curves from MEXCOwalk results on different cancer types.

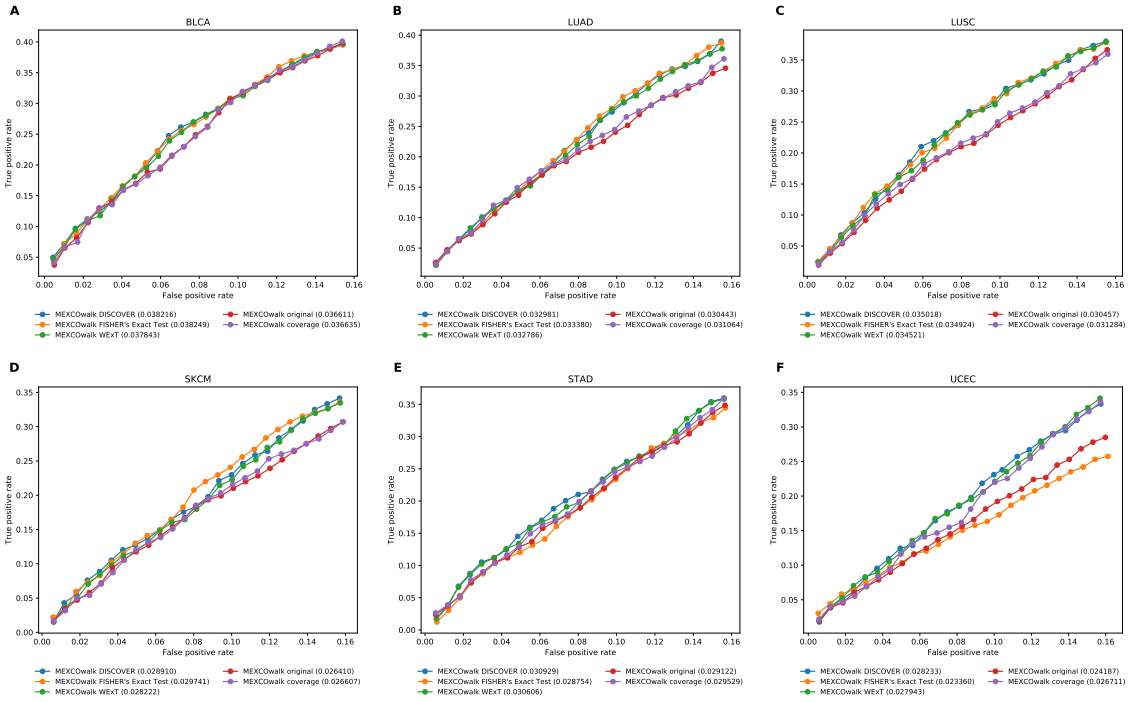


Figure S11: AUROC for different cancer types

# Appendix B

# Code