# Methods for Identifying Combinations of Driver Mutations in Cancer

by
Mark D.M. Leiserson

B.Sc., Tufts University; Medford, MA, 2011
M.S., Brown University; Providence, RI, 2013

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Computer Science and Center for Computational Molecular Biology at Brown University

PROVIDENCE, RHODE ISLAND

May 2016

This dissertation by Mark D.M. Leiserson is accepted in its present form
by the Department of Computer Science and Center for Computational Molecular
Biology as satisfying the
dissertation requirement for the degree of Doctor of Philosophy.

Date_____ _____

Benjamin J. Raphael, Ph.D., Advisor

Recommended to the Graduate Council

Date_____ _____

David H. Laidlaw, Ph.D., Reader

Date_____ _____

Sohini Ramachandran, Ph.D., Reader

Date_____ _____

Peter J. Park, Ph.D., Reader

Approved by the Graduate Council

Date_____ _____

Peter M. Weber, Dean of the Graduate School

# Acknowledgements

This dissertation partially fulfills the requirements for my Ph.D., but also represents the result of years of mentoring and collaboration. I could not have completed my dissertation without the support of many faculty, peers, family, and friends.

I have had the privilege of working with a first-class advisor and mentor, Benjamin J. Raphael, who seems to always have the answers and, better yet, knows the right questions. I have learned more from working with Ben than I would have thought possible, from research and collaboration to mentorship and communication.

In addition to Ben, I have had many excellent collaborators and co-authors without whom this work would have been impossible. I want to particularly acknowledge four of my closest collaborators with whom I have worked throughout my Ph.D. – Connor Gramazio, Matthew Reyna, Fabio Vandin, and Hsin-Ta Wu, and my co-authors in the Raphael research group, Jason Dobson, Jonathan V. Eldridge, Jason Hu, Alexandra Papoutsaki, Jacob Thomas, and Younhun Kim. I am also indebted to my collaborators outside of Brown, including the labs of Li Ding, Gaddy Getz, David H. Laidlaw, Nuria Lopez-Bigas, and Roded Sharan, as well as the multiple The Cancer Genome Atlas analysis working groups with whom I worked.

The biggest advantage I have had in my Ph.D. is the support of past and current members of the Raphael research group, including (in addition to the co-authors mentioned above) Ashley Conard, Mohammed El-Kebir, Rebecca Elyanow, Dora Erdos, Iman Hajirasouliha, Gunnar Klau, Ahmad Mahmoody, Layla Oesper, Anna Ritz, John Shen, Jeremy Watson, and Simone Zaccaria. I am forever grateful to Layla for her advice and support, which began before I started at Brown and continued well after she joined the faculty at Carleton, and from which I derived no small part of the confidence I needed to get through the trying times (to say nothing of our adventures in China, Germany, Ireland, and Poland, or with the Watson Cup!). I also want to thank Hsin-Ta and Ahmad for being great friends and helping me see the big picture.

I also want to acknowledge the contributions of my thesis committee: David Laidlaw, Sohini Ramachandran, and Peter Park. Their support and insight have helped me push through the final hurdles in completing my dissertation.

I will always be proud to be an alumnus of Brown University, and I owe a deep debt of gratitude to the staff, my peers, and my friends. I want to acknowledge the technical staff – especially John Bazik, Donald Johwa, and Frank Pari – for supporting my (many and varied) needs throughout the years. I also want to thank the administrative staff – especially Lauren Clarke, Nathaniel Gill, and Jane Martin – for gracefully fielding many trivial requests and questions, and just generally making my life easier. I truly believe the community of graduate students in the computer science department is remarkable and has vastly improved my quality of life, and I will always remember the trips to the GCB with INEB, the frisbee games, the intramural basketball, and the foosball. I have been particularly fortunate to have friends like Connor Gramazio, John Oberlin, Nickolai Riabov (honorary CS!), and Bryce Richards.

Finally, I want to acknowledge my fiancee and my family. I would not have even made it *to* Brown without Robin, Mom, Pops, or Nick – let alone finish my Ph.D. – and they have helped me in far too many ways to enumerate here. Thank you for your support and advice, and believing in me even when I did not. I hope this dissertation makes you proud!

Abstract of "Methods for Identifying Combinations of Driver Mutations in Cancer" by Mark D.M. Leiserson, Ph.D., Brown University, May 2016

Recent improvements in DNA sequencing technology have opened new paths towards understanding cancer biology and designing personalized cancer treatments. Cancer is caused in part by somatic mutations to the DNA of healthy cells that can be identified with DNA sequencing technology. A major challenge in analyzing the mutations in cancer is distinguishing the handful of driver mutations that cause cancer from the multitude of passenger mutations that play no role in cancer. In part to address this challenge, consortia such as The Cancer Genome Atlas (TCGA) have generated massive catalogues of somatic mutations in thousands of tumors.

The new wealth of mutation data has shown that identifying driver mutations is a difficult computational problem. Many driver mutations are rare, even in these large tumor cohorts, because different combinations of mutations cause cancer in different patients, even those of the same cancer (sub)type. Part of the reason for this phenomenon is that driver mutations target key genetic pathways that perform vital cell functions. Each pathway consists of a set of interacting genes, and can be perturbed in numerous ways. Therefore, we have developed computational methods to search for combinations of driver mutations targeting pathways.

We describe several methods for identifying combinations of putative driver mutations. We first present two methods for identifying combinations of mutations that are mutually exclusive in a cohort of tumors, a pattern expected for genetic pathways. We then present an algorithm to identify significant clusters of mutations in an interaction network. We show that these algorithms outperform previous methods on simulated and real mutation data from TCGA, and identify potentially novel combinations of mutations. We also describe the Mutation Annotation and Genome Interpretation (MAGI) web application, which displays interactive visualizations as well as crowd-sourced and text-mined mutation annotations in order to help prioritize likely driver mutations. These methods contribute towards overcoming the computational challenge of identifying driver mutations in cancer sequencing data.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

The recent advances in DNA sequencing technology promise to improve cancer treatment radically. The cost of DNA sequencing has declined rapidly such that sequencing a cancer patient's genome is sometimes cheaper than drug regimens. Researchers can use this DNA sequencing data to identify the mutations in a patient's tumor. This has the potential to be crucial for designing better cancer therapies, as cancer is caused in part by combinations of mutations to healthy cells that result in the cells' uncontrolled growth into a tumor. Biologists can use the list of mutations to design personalized therapies that target the patient's cancer cells.

A key challenge to realizing this goal of personalized cancer therapy is distinguishing the mutations in a given tumor that drive cancer from the random mutations that have no consequence for cancer [2]. The vast majority of mutations in most cancers are somatic, meaning they occur during the lifetime of an individual and cannot be passed from parent to child. The somatic mutations in a given tumor can be broadly categorized as *drivers* – mutations responsible for cancer – and *passengers* – mutations with no consequence for cancer [2]. Identifying the driver mutations is a key first step in understanding the genetic mechanisms that when perturbed cause uncontrolled cell growth, and can open avenues to improved cancer treatment [3].

The advent of cheap DNA sequencing technology has brought the research community closer to the goal of identifying driver mutations in cancer. Large consortia such as The Cancer Genome Atlas [4, 5, 6, 7, 8, 9, 10, 11, 12, 13] and International Cancer Genome Consortium [14] have sequenced the genomes or exomes of matched tumor-normal pairs in thousands of tumors. Researchers then ran algorithms on these massive datasets to identify the somatic mutations in each tumor, creating a compendium of somatic mutations.

However, despite the unprecedented size and scope of these mutation datasets, identifying the driver mutations remains a significant computational challenge [15]. The challenge derives from two observations about driver mutations. First, driver mutations are relatively rare compared to passenger mutations. In their review of cancer genomics research, Vogelstein, *et al.* [16] estimate that a typical tumor has 3-8 driver mutations, but can have orders of magnitude more passenger mutations. Second, there is vast inter-tumor heterogeneity, meaning that no two tumors have the same collection of mutations [17]. This means that there are driver mutations that may occur only a handful of times, even in a large cohort of tumors. While the ultimate way to determine if a mutation is a driver is to perform experiments to characterize its function, these experiments are often expensive and time-consuming. Furthermore, there are too many combinations of mutations to test. Consequently, algorithms to accurately identify and prioritize driver mutations are an urgent priority.

In this thesis, I present several methods for identifying combinations of driver mutations from a cohort of tumors. In this chapter, I present background information on driver mutations in cancer. In Chapters 2-4, I present work on identifying groups of mutations occurring in the same pathway, or group of genes responsible for some cellular function. In Chapter 5, I present a web application for the visualization and annotation of mutations in cancer genomes that seeks to reduce the computational burden of identifying the driver mutations in a given (cohort of) tumor(s). I conclude in Chapter 6 with a discussion and my conclusions.

## 1.1 Driver mutations in cancer

Mutations in cancer cause uncontrolled growth in a population of cells by targeting genes responsible for key biological functions [18, 2]. In the simplest case, driver mutations either activate or inactivate a cancer gene. Activating mutations give cells an ability required for cancer (e.g. proliferation), and inactivating mutations turn off a function of healthy cells that keeps cell growth in check (e.g. DNA damage repair). The genes targeted by driver mutations are often called *driver genes*, though it is important to note that not all mutations in driver genes are driver mutations, since most mutations in the genome (even in cancer genes) have no functional effect.

Thus, a key challenge in identifying driver mutations lies in predicting their effect. It is particularly difficult to predict the effect of putatively activating mutations. In this vein, researchers have developed sophisticated methods to predict a mutation's functional impact. Given a mutation at a particular locus, algorithms such as SIFT [19] and PolyPhen [20] predict its impact and the structural effect of the protein sequence change. MutationAssessor [21] and others predict functional impact

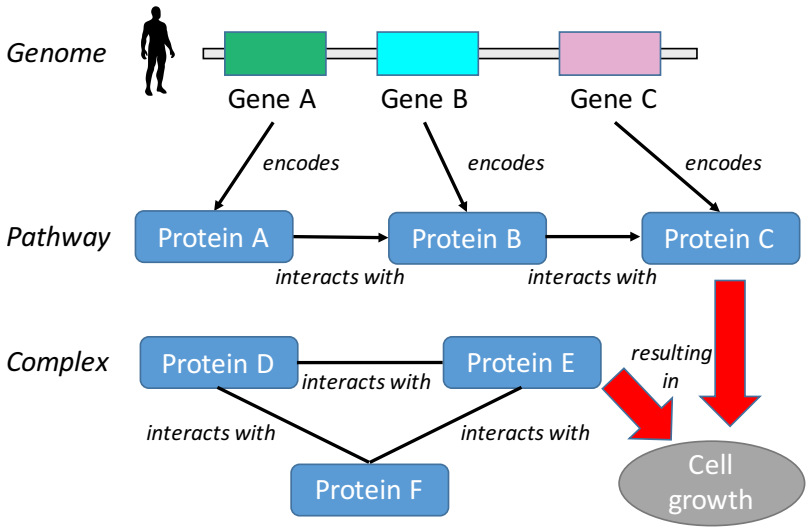



**Figure 1.1:** Cartoon representation of genetic pathways and protein complexes. (top) Genes encode proteins, which interact as part of pathways (middle) or protein complexes to perform cellular functions (e.g. cell growth).

by determining how conserved the locus is across related species.

A second key challenge in identifying driver mutations is that they are relatively rare; a typical tumor generally has orders of magnitude more passenger than driver mutations [22]. To overcome this challenge, researchers have introduced methods to search for significantly recurrent mutations in a cohort of tumors. Most work has focused on analyzing three types of mutations affecting genes (see details in Section 1.2.1): (1) single nucleotide variants, where a single base is changed in the genome; (2) small insertions/deletions, where few bases are added/deleted from a gene, and, (3) copy number aberrations, where (part of) a gene is duplicated or deleted. For example, GISTIC2 [23] and RAIG [24] look for significantly recurrent regions of copy number change in a cohort of tumors. Other methods look for *significantly mutated genes* by evaluating all mutations that target a particular gene. Oncodrive-FM [25] uses mutation prediction algorithms to evaluate each of the mutations in a given gene. MuSiC [26] and MutSigCV [17] use the concept of a background mutation rate (BMR) – the per base rate at which new mutations occur – to compute the significance of the number of mutations in each gene. MuSiC optionally uses a per sample BMR, as different tumors can have vastly different mutation rates depending on its mutations and cancer type. MutSigCV [17] takes this a step further, computing the BMR for a given gene and sample using covariates (CV) that capture the replication timing or expression level of the gene.

These approaches have revealed that most cohorts of tumors tend to have relatively few significantly mutated genes, with a "long tail" of rarely mutated genes [22]. The "long tail" phenomenon results from cancer's significant inter-tumor hetero-

geneity, where different combinations of driver mutations cause cancer in different individuals. Thus some driver mutations may occur in few tumors, even in large tumor cohorts. One explanation for the observed inter-tumor heterogeneity is that mutations in cancer target pathways or complexes, since genes do not act in isolation and it is widely believed cancer requires mutations to multiple pathways or hallmarks [18]. For the purposes of this work, a genetic pathway or protein complex is a group of proteins (each encoded by a gene) that interact to perform some function. Each of these pathways and complexes can be perturbed in numerous ways, which is a common explanation for why each cancer has a different combination of mutations and why there are driver mutations in the rarely mutated genes in the "long tail."

To identify the drivers in the "long tail", researchers have developed methods to identify these mutations by searching for combinations of mutations targeting pathways and protein complexes (Figure 1.1). Because there are too many combinations to test, these approaches utilize different amounts of *prior knowledge* to constrain the search space, and can be broken down into three general categories based on what information they use as prior knowledge.

**Known pathway or gene set databases.** The approach that uses the most prior knowledge is to test known pathway or gene set databases, such as KEGG [27, 28], PINdb [29], or MSigDB [30], for enrichment for mutations. Each of these databases contain on the order of ten-thousand pathways or gene sets. Gene Set Enrichment Analysis (GSEA) [31, 30] and DAVID [32, 33] are two examples of these type of enrichment tests. In general, these approaches are constrained by their input; they cannot identify new pathways or *crosstalk* between multiple pathways, and the datasets often include large and overlapping gene sets.

**Protein-protein interaction networks.** Algorithms such as HotNet [34], EnrichNet [35], and others, use prior knowledge in the form of protein-protein interaction networks, instead of pathway databases. This is much less prior knowledge than the known pathway database approaches. These algorithms search for highly mutated groups of interacting genes or proteins.

**None.** More recently, researchers have introduced methods to identify sets of mutations that are mutually exclusive in a cohort of tumors, a pattern often observed in pathways [36, 37]. These algorithms use no prior knowledge, but instead use the mutual exclusivity pattern to restrict the search space. Algorithms such as Dendrix [38], MEMo [39], and RME [40] search for groups of genes with exclusive mutations. In general cancer perturbs *multiple* pathways [18], and our Multi-Dendrix [1] algorithm

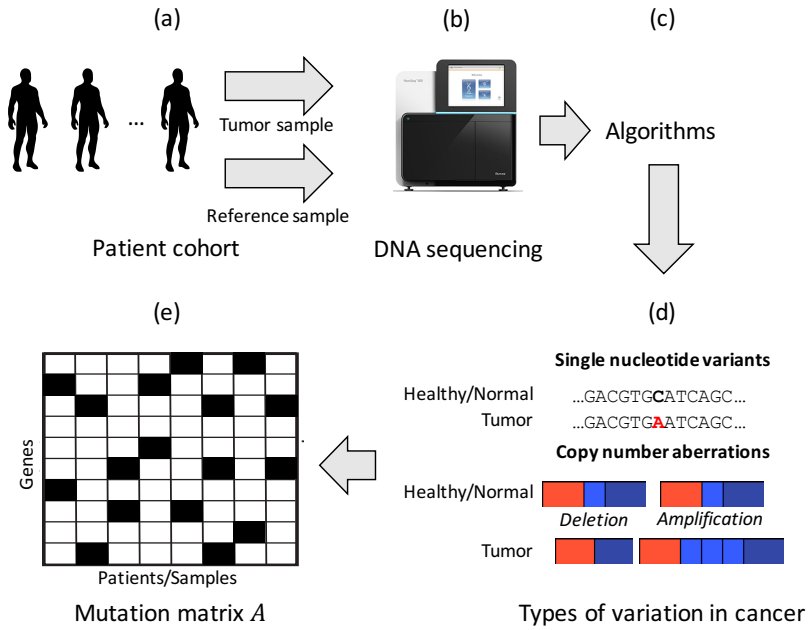



**Figure 1.2:** Schematic for constructing the binary mutation matrix $A$. (a) Tumor and reference samples are taken from a cohort of patients, and (b) are passed through a DNA sequencer. (c) Algorithms predict the somatic mutations in each tumor. (d) These mutations are either single nucleotide variants or copy number aberrations, and (e) are then combined to form the binary mutation matrix $A$.

was the first algorithm to search for multiple sets of mutually exclusive mutations *simultaneously*, which provides advantages over searching for multiple sets iteratively.

## 1.2 Identifying combinations of driver mutations from a binary mutation matrix

The problem we are concerned with in this thesis is identifying combinations of driver mutations from the mutations in a cohort of cancer patients. For much of this work, we will be using as input an $m \times n$ mutation matrix $A$, where

$$A_{ij} = \begin{cases} 1 & \text{if gene } i \text{ has} \geq 1 \text{ mutation in patient } j, \\ 0 & \text{otherwise,} \end{cases}$$

for sets of $m$ genes and $n$ patients (Figure 1.2). We construct this matrix using the output of other algorithms for identifying two types of somatic mutations, described in detail below.

### 1.2.1 Types of somatic mutations in cancer

The mutations in cancer occur at vastly different scales, from a single base in a genome to entire chromosome arms, and can be broadly classified as single nucleotide variants and small insertions and deletions (SNVs), copy number aberrations (CNAs), and structural variants (SVs) (see Figure 1.2b).

**Single nucleotide variants.** At the smallest scale of mutations in cancer are single nucleotide variants (SNVs) – where a single base is changed DNA sequence – and small insertions or deletions (indels) – where a small sequence of DNA is inserted or deleted in the genome. If these mutations occur in the coding region of the genome, they can be further classified as synonymous or nonsynonymous. A synonymous mutation is one that does not change the protein sequence encoded by the mutated region of DNA, while a nonsynonymous mutation does cause a protein sequence change. Nonsynonymous mutations are further classified by the type of protein sequence change they cause. Common classifications include missense mutations, which cause a protein sequence change and nonsense mutations, which insert a premature stop codon into the protein sequence and are therefore putatively inactivating. We refer to nonsynonymous SNVs and indels as SNVs throughout the text, except where explicitly noted.

The most common way to call the SNVs in a tumor is to use next-generation DNA sequencing technology. Most often, a matched tumor-normal pair is sequenced, and variants that are in the tumor genome and not in the matched normal genome are considered to be somatic. Calling somatic variants is an active area of research, and many sophisticated methods have been introduced [41, 42, 43]. In general, these methods output variants with a minimum number of reads supporting the variant.

**Structural variants and copy number aberrations.** Structural variants move, reverse, swap, copy, or delete entire regions (from hundreds of bases to whole chromosome arms) of DNA. Copy number aberrations are a special type of structural variant, and will be the only structural variant discussed in this thesis. The copy number of a region of DNA in a healthy human cell is two, because the human genome is diploid (has two copies of each chromosome). Copy number aberrations (CNAs) occur when a region of DNA is either amplified (copied) or deleted, thus changing the copy number. Copy number aberrations are also often called using next-generation sequencing technology [44, 45], although SNP arrays were popular until relatively recently as well [46, 47]. Because CNAs can span multiple genes, some of these methods perform *target selection* to identify the gene that is the most likely target of CNAs in a given region [23, 24].

For the purposes of this thesis, except where explicitly noted, we will use the output of the SNV or CNA mutation calling algorithms as the input data to our methods.

## 1.3 Contributions

We provide an overview of the research contributions in this thesis towards the problem of identifying combinations of driver mutations. Three of these works improve upon the state-of-the-art in identifying mutually exclusive combinations of mutations (Chapters 2 and 3) and identifying significantly mutated subnetworks (Chapter 4). The fourth work is a web application that reduces the computational burden for biologists to explore mutation data to identify combinations of driver mutations (Chapter 5).

### 1.3.1 Identifying combinations of mutually exclusive alterations

In Chapter 2, we present the Combinations of Mutually Exclusive Alterations (CoMEt) [48, 49] algorithm for identifying collections of sets of mutually exclusive aberrations *with no prior knowledge* (*de novo*). CoMEt uses a novel statistical score for mutually exclusivity that operates on $2 \times 2 \times \cdots \times 2 = 2^k$ contingency tables that is less biased towards the most frequently altered genes than previous methods. CoMEt identifies a collection of multiple sets of mutually exclusive mutations using a Markov chain Monte Carlo (MCMC) algorithm and a novel method to summarize the posterior distribution. We demonstrate that CoMEt outperform other methods using simulated data and mutation data from TCGA. CoMEt identifies groups of exclusive mutations targeting well-known cancer pathways, as well as novel combinations of mutations, in the real mutation data.

In Chapter 3, we present a weighted exact statistical test that uses per gene, per sample alteration probabilities to score mutually exclusive mutations [50]. We use per gene, per sample alteration probabilities because the number of mutations per sample can vary significantly, which can confound the exclusivity signal. We provide an exact formula and a fast and efficient approximation for the weighted test. We analyze hundreds of colorectal and endometrial samples from The Cancer Genome Atlas which have large variation in alteration rates. In both cancer types, the weighted test identifies mutually exclusive alterations in cancer genes with fewer false positives than earlier approaches.

### 1.3.2 Searching for significantly mutated subnetworks

In Chapter 4, we present the HotNet2 (HotNet diffusion oriented subnetworks) algorithm [51] for identifying significantly mutated subnetworks in a protein-protein interaction (PPI) network. HotNet2 uses an insulated heat diffusion process to identify subnetworks of high-scoring (i.e. recurrently mutated) nodes in a protein-protein interaction network. The algorithm utilizes the *direction* of the heat, making it less biased towards high-degree, high-scoring nodes. We demonstrate HotNet2 is superior to other network and pathway analysis methods using simulated and real mutation data. We apply HotNet2 to the TCGA Pan-Cancer [52] dataset of 3,271 tumor samples, where HotNet2 identifies subnetworks overlapping cancer pathways, pathways and complexes with recently characterized roles in cancer, and potentially novel groups of genes.

### 1.3.3 Visualization and annotation of mutations in cancer

In Chapter 5, we present the Mutation Annotation and Genome Interpretation (MAGI; `http://magi.brown.edu`) web application [53]. MAGI lets users visualize, annotate with references from the literature, and combine public and private cancer genomics datasets. MAGI's visualizations and annotation platform integrates public and private mutation data. MAGI is one of the first cancer genomics web application to allow users to upload and combine their own cancer genomics data with *no local installation required*. MAGI includes $\sim 40,000$ literature annotations mined from PubMed Central (PMC), and from the Database of Curated Mutations (DoCM; `http://docm.genome.wustl.edu`). MAGI's tumor sample view shows the mutations in individual tumor samples, combined with the mutation annotations from MAGI's database. The tumor sample view reveals mutations with many literature annotations (presumably drivers), as well as mutations with few annotations that may require additional study.

# Chapter 2

# Identifying Combinations of Mutually Exclusive Alterations

In this chapter, we present the Combinations of Mutually Exclusive Alterations (CoMEt) algorithm for identifying groups of mutually exclusive mutations, a pattern often observed for mutations in the same pathway, from a cohort of tumor samples. This approach is an initial step toward addressing the problem presented by the widespread inter-tumor heterogeneity in cancer for identifying the driver mutations responsible for cancer. One explanation for inter-tumor heterogeneity is that driver mutations target genetic pathways, each of which can be perturbed in numerous ways. CoMEt searches for combinations of mutations that are likely to be targeting the same pathway.

We published most of the material in this chapter in [49] where we introduced CoMEt, and originally presented CoMEt at the $19^{th}$ Annual International Conference on Research in Computational Molecular Biology (RECOMB) [48]. I am a co-first author on both of these publications, and my major contributions were in developing the exact test and MCMC summarization procedure, and performing simulation experiments. I also led the CoMEt analysis of somatic and germline mutations in a pan-cancer dataset [54], and contributed to analysis with a preliminary version of CoMEt as part of the The Cancer Genome Atlas's acute myeloid leukemia study [5].

We apply CoMEt to simulated data, and real mutation data from The Cancer Genome Atlas. On glioblastoma, breast cancer, stomach cancer, and leukemia datasets, we demonstrate that CoMEt finds collections of gene sets with mutually exclusive mutations, including sets that sets that overlap well-known pathways with roles in cancer, and novel sets of exclusive mutations. We emphasize that CoMEt achieves these results solely by looking at a pattern of mutations in a cohort of tumors, with no prior information of interactions or pathways. We also show that

CoMEt outperforms similar methods on both simulated and real data.

## 2.1 Background and related work

A major goal of large-scale cancer genomics projects such as The Cancer Genome Atlas (TCGA) [6, 7, 5, 8, 52, 55], the International Cancer Genome Consortium (ICGC) [14, 56], and others is to identify the genetic and epigenetic alterations that drive cancer development. These projects have generated whole-genome/exome sequencing data measuring the somatic mutations in thousands of tumors in dozens of cancer types. Interpreting this data requires one to distinguish the *driver* mutations that play a role in cancer development and progression from *passenger* mutations that have no consequence for cancer. Identifying driver mutations directly from sequencing data is a significant challenge since individuals with the same cancer type typically harbor different combinations of driver mutations [17, 57].

The observed mutational heterogeneity in cancer has motivated the development of methods to examine *combinations* of mutations. Since driver mutations typically target genes in a small number of key pathways [16], several methods have been introduced to examine mutations in known pathways or networks (reviewed in [15, 58]). However, most pathway databases and interaction networks are incomplete, lack tissue specificity, and do not accurately represent the biology of a particular cancer cell. Thus, *de novo* methods for examining combinations of mutations are of particular interest as they require no prior biological knowledge and enable the discovery of novel combinations. Unfortunately, the number of possible combinations is too large to test exhaustively and achieve statistically significant results. Current *de novo* approaches to identify putative combinations of mutations use the observation that mutations in the same pathway are often mutually exclusive [37]. This observation follows from the observation that there are relatively few driver mutations in a tumor sample, and these are distributed over multiple pathways/hallmarks of cancer [18].

In 2011, three algorithms for identifying sets of genes with mutually exclusive mutations were introduced simultaneously: the De Novo Driver Exclusivity (Dendrix) [38], Recurrent Mutually Exclusive aberrations (RME) [40], and Mutual Exclusivity Modules (MEMo) [39] algorithms. Dendrix and RME are both *de novo* algorithms for identifying gene sets with mutually exclusive mutations, while MEMo examines mutual exclusivity on a protein-protein interaction network. The Dendrix algorithm identifies sets $M$ of $k$ genes with high coverage (many samples have a mutation in the set) and approximate exclusivity (few samples have a mutation in more than one gene in the set). Dendrix combines these two criteria into a weight $W(M)$, which is equal to the coverage of $M$ minus the coverage overlap (co-occurring mutations) of $M$. Finding the set of maximum weight is an NP-hard problem [38].

Dendrix uses a Markov chain Monte Carlo (MCMC) algorithm to sample high weight gene sets; more recently other optimization methods have been used to find high weight sets [59, 60]. Leiserson *et al.* [1] introduced the Multi-Dendrix algorithm to identify multiple mutually exclusive gene sets simultaneously using an integer linear program. In contrast, RME defines the exclusivity weight as the percentage of covered samples that contain exactly one mutation within a gene set, and uses an online-learning linear threshold algorithm to identify groups of genes with high pairwise exclusivity. However, both the RME and MEMo algorithms were shown not to scale to reasonably sized datasets [1], requiring extensive filtering of input data [40, 61].

One limitation of the combinatorial weight function used in Dendrix and subsequent algorithms is that genes with high mutation frequencies (high coverage) can dominate the mutual exclusivity signal, thus biasing the algorithms towards identifying gene sets where the majority of the coverage comes from one gene (Figure 2.1a). These observations motivated the development of probabilistic models of mutual exclusivity. These include the Dendrix++ algorithm (an early version of the approach that we present in this paper) and the muex algorithm [62]. Dendrix++ uses a statistical score and was used in TCGA acute myeloid leukemia study [5]. The muex algorithm [62] uses a generative model of mutual exclusivity and a likelihood ratio test to score the mutual exclusivity of combinations of mutations. However, we find that the muex score is sensitive to high frequency mutations (see section 2.3.3). Moreover, both of these approaches exhaustively enumerate gene sets to find those with high score, limiting their applicability to larger datasets. In addition, they do not identify multiple gene sets simultaneously, a feature that has proved useful with the Dendrix weight [1]. The mutex algorithm [63] also uses a probabilistic model of mutual exclusivity, and was published after this manuscript was submitted. We provide further details of mutex below. Finally, no current method identifies overlapping gene sets[1] — although cancer genes have been shown to participate in multiple pathways [6] — or considers additional sources of mutual exclusivity such as cancer subtype-specific mutations.

[1]We note that while Multi-Dendrix [1] and mutex [63] can identify overlapping gene sets, this feature was not explored in the corresponding publications.

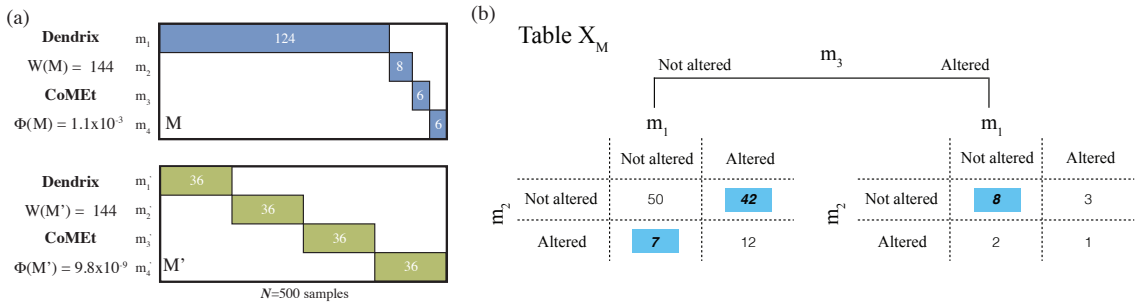



**Figure 2.1:** **(a)** Alteration matrices illustrating differences between the combinatorial weight function $W(M)$ introduced in Dendrix and the probabilistic score $\Phi(M)$ used in CoMEt. Both matrices contain 4 mutually exclusive alterations whose alteration frequencies are indicated inside each bar. The samples without alterations are not shown in either matrix. Since both sets are exclusive and have the same total alteration frequency, the Dendrix weight function does not distinguish between these sets. Sets like $M$ (blue) are common in cancer genome studies which often have a small number of recurrently mutated genes and a long tail of rarely mutated genes. The score used in CoMEt conditions on the observed frequencies of each alteration, giving more significance to the set $M'$ (green). **(b)** An example of $2 \times 2 \times 2$ contingency table $\mathbf{X}_M$ for the set $M = \{m_1, m_2, m_3\}$, illustrating how samples are cross-classified into exclusive, co-occurring, or absent for each alteration. The test statistic $\phi(M)$ used by CoMEt is the sum of the highlighted exclusive cells

We introduce the Combinations of Mutually Exclusive Alterations (CoMEt) algorithm to overcome the challenges outlined above. CoMEt includes the following contributions.

1. We develop an exact statistical test for mutual exclusivity *conditional* on the observed frequency of each alteration. This approach is less biased towards high frequency alterations, and enables the discovery of combinations of lower frequency alterations. We derive a novel tail enumeration procedure to compute the exact test, as well as a binomial approximation.
2. CoMEt simultaneously identifies collections consisting of *multiple* combinations of mutually exclusive alterations, and samples from such collections using an MCMC algorithm. We summarize the resulting distribution by computing the marginal probability of pairs of alterations in the same sets. This enables CoMEt to identify sets of any size, including overlapping sets of alterations, without testing many parameter settings.
3. Given prior knowledge of cancer-types/subtypes, CoMEt analyzes alterations and subtypes simultaneously, allowing the discovery of mutually exclusive alterations across cancer types, while avoiding the identification of spurious mutually exclusive sets of (sub)type-specfic mutations.

We demonstrate that CoMEt outperforms earlier approaches on simulated and real cancer data. We apply CoMEt to acute myeloid leukemia (AML), glioblastoma (GBM), gastric (STAD), and breast cancer (BRCA) data from TCGA, and to a

smaller study of intracranial germ tumors. In each cancer type, we identify combinations of mutated genes that overlap known cancer pathways and also contain potentially novel cancer genes including *IL7R* and the EphB receptor *EPHB3* in STAD, and the scavenger receptor *SRCRB4D* in GBM. On the gastric and breast cancer data, we demonstrate how CoMEt simultaneously identifies mutual exclusivity resulting from pathways and from subtype-specific mutations. CoMEt is available at `https://github.com/raphael-group/comet` and as the `cometExactTest` R package available in CRAN at `https://cran.r-project.org`.

## 2.2 CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer

### 2.2.1 CoMEt algorithm

We consider that a set $\mathcal{E}$ of $m$ *alterations* have been measured in $n$ samples. An alteration may be the somatic mutation of a particular gene, a specific single nucleotide mutation (for example, V600E mutations in the *BRAF* gene), an epigenetic change such as hypermethylation of a promoter, or a variety of other changes. We assume that alterations are binary, such that alterations are either present or absent in each sample. We represent the set of measured alterations with an $m \times n$ binary alteration matrix $A = [a_{ij}]$, where $a_{ij} = 1$ if alteration $i$ occurs in sample $j$, and $a_{ij} = 0$ otherwise. Our goal is to identify *one or more* sets $M_1, M_2, \ldots, M_t$ where the alterations in each $M_i$ are surprisingly mutually exclusive across the $n$ samples. We introduce the CoMEt algorithm for this purpose (see Figure 2.2), a preliminary version of which was presented at the RECOMB conference [64].

We derive a score $\Phi(M)$ for a set $M$ of $k$ alterations using an exact test of mutual exclusivity. Specifically, we examine a $2 \times 2 \times \cdots \times 2 = 2^k$ contingency table $\mathbf{X}_M$ (Figure 2.1(b)) whose entries indicate the number of samples where each combination of alterations occurs. For example, the entry $x_{(24)}$ of $\mathbf{X}_M$ equals the number of samples where the second and fourth alterations in $M$ occur, but the first and third alterations do not occur. The score $\Phi(M)$ is the $P$-value of the observed mutual exclusivity in the table $\mathbf{X}_M$, where the margins of the table (determined by the number of samples where each alteration occurs) is fixed. That is, the score $\Phi(M)$ is *conditional* on the observed frequencies of alterations in $M$. This statistical score reduces the effect of the most frequent alterations having an unduly large contribution to the score. See section 2.2 for further details.

CoMEt scores a collection $\mathbf{M} = (M_1, \ldots, M_t)$ of $t$ alteration sets by taking the product of the scores of each set $M_i$:

$$\Phi(\mathbf{M}) = \prod_{i=1}^{t} \Phi(M_i). \tag{2.1}$$

This score follows from the null hypothesis that exclusivity is independent across sets.

Since the number of possible collections of alteration sets grows exponentially with the number of alterations, it is typically impossible to enumerate and compute the weight of all alteration sets. We derive a Markov chain Monte Carlo (MCMC) algorithm to sample collections $\mathbf{M}$, each consisting of $t$ sets of alterations, in proportion to their significance. We summarize this distribution by computing the marginal probability $p(e, e')$ for each pair of alterations in $A$. We summarize these probabilities using the *marginal probability graph*, a complete, undirected weighted graph $G = (V, E)$ where $V = \mathcal{E}$ (the set of observed alterations) and where each edge $e \in E$ connects a pair of vertices $u, v$ with weight $p(u, v)$. We identify the most exclusive alteration sets by first removing all edges from the graph weight below a threshold $\delta$. CoMEt outputs $C(\delta)$, the connected components in the resulting graph, which we call *modules*. The summarization via the marginal probability graph allows CoMEt to output collections of alteration sets different in number and size than specified by the input parameters. Further details are given in the section 2.2.

**Scoring mutual exclusivity** CoMEt uses a novel statistical score based on an exact test for mutual exclusivity. Figure 2.1 motivates the development of the new score, showing two sets $M$ and $M'$, each with four alterations. The alterations in both sets are perfectly exclusive (no sample has more than one alteration), and the total number of altered samples is the same. The Dendrix weight function $W(M)$ introduced in [38] (and used in later publications [59, 60, 1]) is defined as the *coverage*, the number of samples with at least one mutation in $M$, minus the *coverage overlap*, the number of samples with more than one mutation in $M$. In this case, $W(M) = W(M')$. However, given the frequencies of each alteration, we are more surprised to observe mutual exclusivity among alterations in the set $M'$, which are each altered in 7 % of samples, than we are to observe mutual exclusivity among the alterations in set $M$, where a single alteration has very high frequency (25 %) and three alterations have relatively low frequency ($< 2\,\%$). Sets like $M$ are common in many cancer datasets where highly recurrent alterations (such as mutations in *TP53* or amplification of *EGFR*) occur and can be combined with low frequency, spurious alterations.

We first describe a statistical score $\Phi(M)$ for a tuple $M = (m_1, \ldots, m_k)$ of alterations. The score measures the surprise of the observed exclusivity of these alterations *conditional* on the rate of occurrence of each alteration. Since these rates are

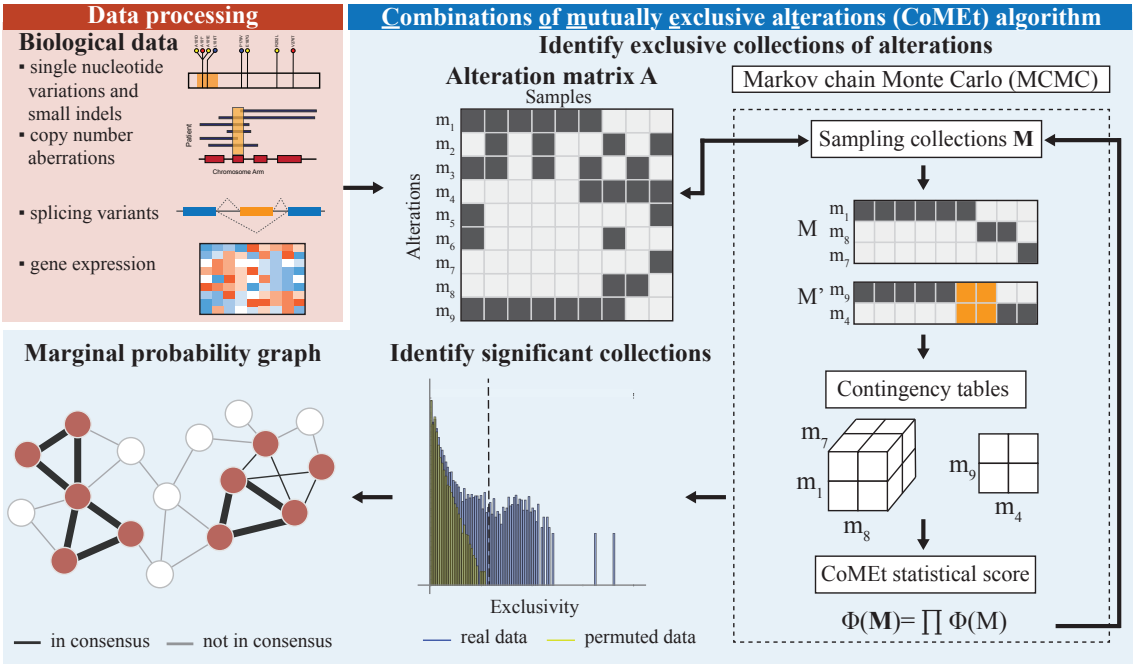



**Figure 2.2:** Overview of the CoMEt algorithm. First, we transform alteration data from different measurements into a binary alteration matrix $A$. Second, we use a Markov chain Monte Carlo (MCMC) algorithm to sample collections $\mathbf{M}$, containing $t$ sets of $k$ alterations, in proportion to the weight $\Phi(\mathbf{M})^{-\alpha}$. Here we show a collection containing sets $M$ and $M'$ with three and two alterations, respectively. We identify all collections whose weight exceeds the maximum observed in randomly permuted datasets. We summarize the alterations in these significant collections with a *marginal probability graph*, whose edge weights indicate the fraction of significant collections with the corresponding pair of alterations. Finally, we remove low-weight edges in the graph, obtaining the output modules

generally unknown (for example, the background mutation rate for single nucleotide mutations varies greatly across genes and samples [65]), we use the *exact distribution* obtained from the observed data as the null distribution. Under this distribution, the status of the $k$ alterations in $n$ samples is described by selecting uniformly a $k \times m$ binary alteration matrix $B$ with the constraint that the number of 1's in row $i$ of $B$ equals the number of 1's in row $m_i$ of the alteration matrix $A$. This distribution is equivalent to the sampling distribution on $2 \times 2 \times \cdots \times 2 = 2^k$ contingency tables under the hypergeometric distribution, where dimension $i$ of the table gives the cross-classification of the number of samples where alteration $i$ occurs or not. For example, three alterations are described by a $2 \times 2 \times 2$ table with margins equal to the frequency of each alteration (Figure 2.1(b)).

We introduce notation to describe the statistical test. Given a set $M$ of alterations, let $x^+_{(j)}$ be the number of samples where alteration $m_j$ occurs. It follows that $n - x^+_{(j)}$ is the number of samples where $m_j$ does not occur. Similarly, for $\mathbf{v} \subseteq [k] = \{1, \ldots, k\}$, let $x_{\mathbf{v}}$ denote the number of samples where alterations only occur in $m_{\mathbf{v}}$. The values $x_{\mathbf{v}}$ for all $v \subseteq [k]$ give the entries of a $2^k$ contingency table

$\mathbf{X}_M$ with fixed margins $\mathbf{x}^+ = (x^+_{(1)}, \ldots, x^+_{(k)})$. Thus, the probability of observing a $2^k$ contingency table $\mathbf{X}_M$ with fixed margins $\mathbf{x}^+$ and whose sum of entries equals $n$ follows the multivariate hypergeometric distribution

$$p_{\mathbf{X}_M} = \Pr(\mathbf{X}_M|\mathbf{x}^+, k, n) = \frac{\prod_{j=1}^{k} x^+_{(j)}! \left(n - x^+_{(j)}\right)!}{(n!)^{k-1} \prod_{\mathbf{v}\subseteq[k]} x_{\mathbf{v}}!}. \tag{2.2}$$

To characterize the mutual exclusivity of alterations in a contingency table, we define the test statistic as the sum of the entries in the contingency table where *exactly* one alteration occurs, that is, $T(\mathbf{X}_M) = \sum_{j=1}^{k} x_{\{j\}}$, where $x_{\{j\}}$ is the number of samples where alterations occur only in $m_j$. We compute a $P$-value for the observed value $T(\mathbf{X}_M)$ of the test statistic as the tail probability of observing tables with the same margins whose exclusivity is at least as large as observed:

$$\Pr(T \geq T(\mathbf{X}_M)|\mathbf{x}^+, k, n) = \sum_{\substack{\mathbf{Y}\in\mathcal{T}(\mathbf{x}^+):\\ T(\mathbf{Y})\geq T(\mathbf{X}_M)}} \Pr(\mathbf{Y}|\mathbf{x}^+, k, n), \tag{2.3}$$

where $\mathcal{T}(\mathbf{x}^+)$ is the set of $2^k$ contingency tables with margins $\mathbf{x}^+$. Note that for $k = 2$, the test statistic $T(\mathbf{X}_M)$ is equivalent to a one-sided Fisher's exact test. $2 \times 2$ contingency tables have only one degree of freedom, and thus there are essentially only two ways in which the corresponding pair of random variables can be non-independent: having too many co-occurrences or too much exclusivity (Figure 2.1(b)). However, $2^k$ tables have $2^k - k - 1$ degrees of freedom and there are many ways in which the corresponding random variables can be non-independent. The $T(\mathbf{X}_M)$ test statistic measures whether the alterations are surprisingly *mutually exclusive*, rather than non-independent in some other way.

We define the score $\Phi(M)$ using the mid $P$-value [66], which is the the average of the probability of observing a value at least as extreme as the observed value and observing a value more extreme than observed:

$$\begin{aligned} \Phi(M) = \quad & \frac{1}{2}(Pr(T \geq T(\mathbf{X}_M)|\mathbf{x}^+, k, n) \\ & + Pr(T > T(\mathbf{X}_M)|\mathbf{x}^+, k, n)). \end{aligned} \tag{2.4}$$

We use the mid $P$-value because the tail probability from exact tests is typically overly conservative, due to the discreteness of the exact distribution [66]. Finally, since cancer is driven by mutations in multiple pathways [18], we define a score $\Phi(\mathbf{M})$ for a collection $\mathbf{M} = (M_1, M_2, \ldots, M_t)$ of $t$ gene sets as $\Phi(\mathbf{M}) = \prod_{i=1}^{t} \Phi(M_i)$. The product results from our assumption that under the null hypothesis mutations in different sets $M_i$ are independent.

### 2.2.2 Computing the mutual exclusivity score $\Phi(M)$

To compute the mutual exclusivity score $\Phi(M)$, one must compute (2.3). This requires computing the probability of all tables $\mathbf{Y}$ with the same margins as $\mathbf{X}_M$ and with exclusivity statistic $T(\mathbf{Y})$ at least as large as the observed value $T(\mathbf{X}_M)$. Unfortunately, no algorithm is known to enumerate such tables. In general the problem of counting contingency tables with fixed margins is #P-complete [67], and thus it is unlikely they can be enumerated efficiently. Several methods have been proposed to solve the problem of counting contingency tables, including using the network algorithm [68, 69] for Fisher's exact test in $r \times c$ contingency tables, or extensions to consider the joint effect of two contingency tables (that is, $2 \times r \times c$) [70]. Branch and bound heuristics have also been used in some specialized cases [71]. However, these approaches still consider at most three-dimensional contingency tables, and the problem of enumerating $2^k$ tables does not seem to have been considered. Even for small $k$ the enumeration problem is intractable: the number of $2^k$ tables with fixed margins grows exponentially in $k$. The work [72] presented an exhaustive algorithm to enumerate all $2^3$ and $2^4$ contingency tables with fixed margins, demonstrating for example that for $n = 36$, there are $> 100$ million $2^4$ tables. Randomized and approximate counting methods for contingency tables have been developed (see, for example, [73, 74] and references therein), although these generally do not provide a rigorous guarantee on the error in the approximation.

We derive a novel *tail enumeration* algorithm to efficiently compute the tail probability in Eq. (2.3) for tables with high values of the exclusivity statistic $T$. The motivation for our approach is that the sets $M$ of interest will have extremely high values of $T(\mathbf{X}_M)$, near the maximum possible value. For example, in the degenerate case of perfect exclusivity (no sample with more than one alteration in $M$) there are no more extreme tables to enumerate, and the algorithm needs only to evaluate the hypergeometric probability of Eq. (2.2) for this single table. Thus, if we enumerate tables starting from the highest possible values for $T$, we can obtain highly accurate $P$-values for the most interesting cases. Furthermore, we can stop the enumeration procedure when the $P$-value becomes sufficiently large and use approximations for these larger $P$-values (see below).

Algorithm 1 is the tail enumeration strategy to enumerate contingency tables in approximate order from most to least exclusive. Briefly, let $\mathbf{C} = (\mathbf{v} \subseteq [k] : |\mathbf{v}| \geq 2)$ be the vector of co-occurring (not exclusive) cells. The basic strategy employed by Algorithm 1 is to generate a table $\mathbf{Y}$ that is more exclusive than $\mathbf{X}_M$ (that is, $T(\mathbf{Y}) > T(\mathbf{X}_M)$ by iterating through the possible values of each cell in $\mathbf{C}$, using the following facts:

- When all values in $\mathbf{C}$ are fixed, the other values in the contingency table are uniquely determined (see Procedure CompleteContTbl in Algorithm 1).

- We can set and update exact upper and lower bounds for each cell in $\mathbf{C}$. The values of each cell are bounded by two values (lines 10–11 in TailEnumeration): the first is how many more co-occurrences are allowed in the current table ($T_{REM}$) before $\mathbf{Y}$ is less exclusive than $\mathbf{X}_M$; the second is given by the constrained marginal ($MarRem$) for that variable in $\mathbf{X}_M$.

We find that Algorithm 1 performs well on real data, evaluating the test statistic $T(\mathbf{X}_M)$ in a few seconds for sets with $k \leq 7$ that have a small number of co-occurrences.

**Algorithm 1** Tail enumeration for any $k > 1$

**Input:** $2^k$ contingency table $\mathbf{X}$.
**Output:** Set $\mathcal{S}$ of contingency tables at least as exclusive as $\mathbf{X}$: $\mathcal{S} = \mathbf{Y} \in \mathcal{T}(x^+) : T(\mathbf{Y}) \geq T(\mathbf{X})$.

1: $\mathcal{S} \leftarrow \{\}$
2: $N \leftarrow 2^k$
3: $\mathbf{C} \leftarrow$ Sorted($\{\mathbf{v} \subseteq [k] : |\mathbf{v}| \geq 2\}$) ▷ Sorted descending vector of co-occurring cells
4: $y_{\mathbf{v}} \leftarrow 0, \forall \mathbf{v} \subseteq [k]$
5: $T_{\max} \leftarrow \sum_{i=1}^{k} x^+_{(i)}$ ▷ Sum of alteration frequencies
6: TailEnumeration($\mathbf{Y}, \mathbf{C}, T_{\max} - T(\mathbf{X})$)
7: **procedure** TailEnumeration($\mathbf{Y}, \mathbf{C}, T_{REM}$) ▷ $T_{REM}$: count of allowed co-occurrences remaining
8: $\quad \mathbf{v} \leftarrow$ Head($\mathbf{C}$)
9: $\quad$ **if** $\mathbf{v} \neq$ Null **then**
10: $\quad\quad MarRem \leftarrow \min_{i \in \mathbf{v}} \{y_i^+\}$ ▷ Minimum margin remaining
11: $\quad\quad$ **for** $(i \leftarrow L, \ldots, \min\{MarRem, \lfloor \frac{T_{REM}}{|\mathbf{v}|} \rfloor\})$ **do**
12: $\quad\quad\quad \mathbf{Y}' \leftarrow$ Copy($\mathbf{Y}$)
13: $\quad\quad\quad y'_{\mathbf{v}} \leftarrow i$ ▷ Set value of cell $\mathbf{v}$ of $\mathbf{Y}'$ to $i$
14: $\quad\quad\quad$ TailEnumeration($\mathbf{Y}'$,Tail($\mathbf{C}$), $T_{REM} - |\mathbf{v}| \times i$)
15: $\quad$ **else** ▷ If all "co-occurring" cells have been set
16: $\quad\quad \mathcal{S} = \mathcal{S} \cup \{$CompleteContTbl($\mathbf{Y}$)$\}$
17: **procedure** CompleteContTbl($\mathbf{Y}$) ▷ Fill in remainder of contingency table $x'$
18: $\quad$ **for** $\mathbf{v} \subseteq [k] : |\mathbf{v}| = 1$ **do** ▷ Iterate over exclusive cells
19: $\quad\quad y_{\mathbf{v}} \leftarrow x^+_{\mathbf{v}} - y^+_{\mathbf{v}}$
20: $\quad \mathbf{Y}_{(0,0,\ldots,0)} \leftarrow n - \sum_{y \in \mathbf{Y}} y$ ▷ Fill in cell with no alterations
21: $\quad$ **return** $\mathbf{Y}$

**Binomial approximation.** We can approximate the distribution of the exclusivity statistic using the binomial distribution, which is a well-known approximation of

the hypergeometric distribution. Under the null hypothesis that alterations occur independently in the samples, let $p_e = \sum_{j=1}^{k} \frac{x_{(j)}}{n}$ be the probability of an exclusive alteration; that is, a sample contains exactly one alteration in $M$. Given a set $M$ of alterations $M$, then the probability of observing $T(\mathbf{X}_M)$ or more exclusive alterations in $n$ samples is given by the binomial tail probability $1 - \sum_{i=0}^{T(\mathbf{X})-1} \binom{n}{i} p_e^i (1-p_e)^{n-i}$.

We find that the binomial provides a good approximation of the exact test $P$-value for sets $M$ with a large number of co-occurring mutations, and consequently a higher $P$-value (see Figure 2.3). Conveniently, these are precisely the cases where the tail enumeration algorithm is slow.

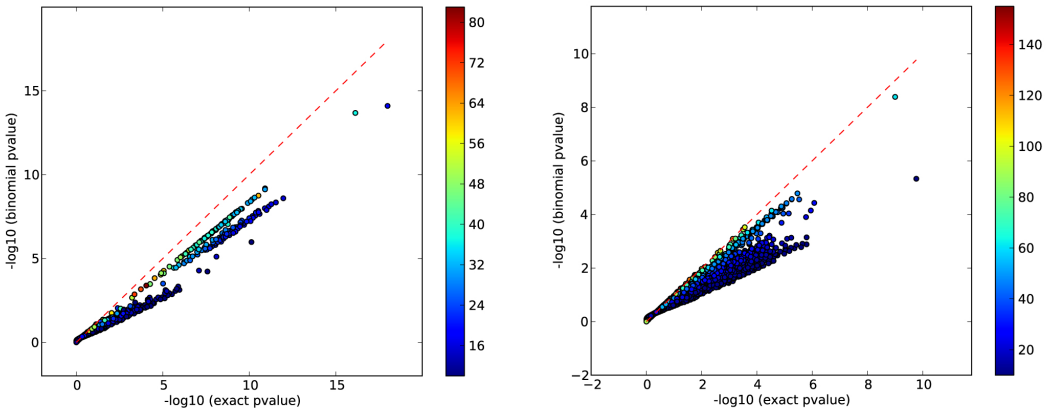



**Figure 2.3:** Scatter plot between negative log of exact and binomial P-values for all sets of $k = 3$ alterations on the GBM dataset (left) and BRCA dataset (right). The color of each dot represents the number of co-occurring alterations according to the scale at the right. Note that the $P$-values for the exact test much smaller than the binomial only in cases with relatively low number of co-occurrences. These cases are the fastest to compute with the tail enumeration algorithm.

**Permutation approximation.** Another approximation to the exact test is obtained using a permutation test. We sample $L$ tables with fixed margins uniformly from the space of all tables and compute the proportion of such tables whose exclusivity value $T$ exceeds the observed value $T(\mathbf{X}_M)$. Of course, sampling uniformly from the set of tables with fixed margins is not straightforward. We use an MCMC approach as described in [39], although we do not fix the number of alterations per sample. Interestingly, while the MEMo algorithm [39] uses a permutation test, the test statistic is the coverage $\Gamma(M)$, rather than the exclusivity $T(M)$ used in CoMEt. While these are equivalent when $k = 2$ (since there is only one degree of freedom), they produce different results for $k > 2$. See further discussion in the section 2.2.9.

In our implementation, we use the exact test, binomial approximation, or permutation approximation to compute $\Phi(M)$ according to the following procedure. First, we calculate the $P$-value from the binomial approximation and compute the

number of co-occurring alterations in $M$. If the number of co-occurring alterations is higher than a fixed threshold $\kappa$ or the binomial $P$-value is larger than a fixed value $\psi$, we set $\Phi(M)$ to be the binomial $P$-value. Otherwise, we perform the tail enumeration procedure to compute the exact test $P$-value, stopping the enumeration if the accumulated tail probability becomes larger than a threshold $\epsilon$. If we stop, then we compute the permutation approximation with $\lceil \frac{1}{\epsilon} \rceil$ samples, such that we expect to sample at least one table with $T > T(\mathbf{X}_M)$. This procedure focuses the time to perform tail enumeration in those cases where high accuracy is needed for small $P$-values.

### 2.2.3 Sampling collections of mutually exclusive alterations with MCMC

Our goal is to identify a collection $\mathbf{M}$ of $t$ alteration sets with low (highly significant) values of $\Phi(\mathbf{M})$. Since it is typically not possible to enumerate all such collections (except for test datasets with small $m$, $n$, $t$, and $k$), we derive a Markov Chain Monte Carlo (MCMC) approach to sample from the space of possible collections. We use the Metropolis-Hastings algorithm [75, 76] to derive an MCMC algorithm to sample collections $\mathbf{M}$ in proportion to the weight $\Phi(\mathbf{M})^{-\alpha}$, where higher values of $\alpha$ increase the sampling frequency of the most mutually exclusive sets (see Appendix A.2 for additional details). We use $\alpha = 2$ except where noted.

**Choosing values for $t$ and $k$** Ideally, CoMEt should be run with the largest values of $k$ and $t$ that are biologically meaningful for a particular dataset. If smaller values of $k$ and $t$ are best supported by the data, the summarization procedure will demonstrate this. We see examples of this in glioblastoma, where the ten most significant collections identified by CoMEt include a set with *TP53*, *MDM2*, *MDM4*, and one of five other alterations (Additional file 2: Table S5).

In practice, using large values of $k$ and $t$ might lead to long run times and slow convergence of the MCMC algorithm, since the space of possible collections will be very large. Thus, an alternative approach that we use to generate results is to run with small values of $t$ and $k$ (for example, $t = 3, 4$ and $k = 3, 4$) and examine the resulting marginal probability graph. If there are $t$ or more cliques or approximate cliques in the graph, this suggests the use of larger values of $t$ and $k$. We used this approach to find larger collections in the AML dataset (see details in section 2.3.2).

### 2.2.4 Marginal probability graph

We now present a method to extract a collection of highly exclusive alteration sets (*with no prescribed size*) from the posterior distribution obtained from the MCMC algorithm. Typically, there are multiple collections with significant scores. This might occur for interesting reasons such as different sets of alterations with similar scores or alterations that appear in multiple mutually exclusive sets. However, the reason might also be suboptimal parameter selection; for example, there may be a significant set of $k = 3$ alterations in the data, but running the algorithm with $k = 4$ will return many sets with the same three genes and a fourth "spurious" gene. To distinguish such cases, we summarize the posterior distribution on collections using a *marginal probability graph* $G$. For a pair $(i, j)$ of alterations, let $p(i, j)$ denote the posterior probability that $i$ and $j$ are found in the same set. We compute $p(i, j)$ using the samples from the MCMC algorithm (see Appendix A.2).

Let $G = (V, E)$ be a complete, undirected weighted graph whose vertices are the alterations and where each edge $e \in E$ connects a pair of vertices $u, v$ with weight $p(u, v)$. Connected subgraphs of $G$ with many high-weight edges are the most exclusive alteration sets in $A$. We identify these most exclusive alteration sets by first removing all edges with weight below a threshold $\delta$ (see Appendix A.2). Let $C(\delta)$ be the connected components of size $\geq 2$ in the resulting graph. The output of CoMEt is the $C(\delta)$ alteration sets. We choose connected components as the output — as opposed to some other partition of the graph such as cliques — in order to be able to identify other topologies such as overlapping pathways (alteration sets), where two sets of alterations are connected by a cut node.

### 2.2.5 Statistical significance

While the score $\Phi(\mathbf{M})$ measures our surprise of observing exclusivity within each of the sets in $\mathbf{M}$ conditional on the observed frequencies of each alteration, there is a large number of possible collections, and thus we might observe a high score by chance. We evaluate the statistical significance of the collection $\mathbf{M}$ by comparing to a null distribution of scores obtained on permuted alteration matrices $A$ with the sample and alteration frequencies (sums of rows and columns of $A$) fixed [39, 77]. Let $\Phi^*$ be the minimum score obtained over $N$ permutations. We use the collections $\mathbf{M}$ satisfying $\Phi(\mathbf{M}) \leq \Phi^*$ (thus each such collection has $P$-value $< \frac{1}{N}$) to compute the marginal probability graph except where noted.

### 2.2.6 Simultaneous analysis of alterations and cancer subtypes

An important confounding factor in identifying cancer pathways *de novo* by analyzing exclusive alterations is that certain alterations primarily occur in particular cancer subtypes [78]. If we analyze a mixed set of samples with multiple subtypes, these subtype-specific alterations will be mutually exclusive in the data, even if they are not in the same biological pathway. When the subtypes are known in advance, one solution is to analyze subtypes separately; unfortunately, this reduces sample numbers, thus reducing power to identify combinations of alterations that are shared across subtypes. CoMEt addresses this problem by adding one new "subtype row" to the alteration matrix $A$ for each subtype. This subtype row contains an alteration in all samples *excluding* those of the given subtype. Thus, the sets of alterations that are surprisingly exclusive with these subtype rows are the ones primarily *altered* in that subtype. Note that when running CoMEt with subtype rows, we do not allow multiple subtypes to be placed in the same set. Because CoMEt simultaneously analyzes multiple alteration sets, it can identify exclusive sets containing subtype-specific alterations, general alterations, or any combination of these.

When analyzing the cancer dataset that included sample subtype classifications, we perform two runs of CoMEt. First we run CoMEt on the alteration matrix $A$. Then we run CoMEt on the alteration matrix with "subtype rows" as we described. We summarize the ensemble of statistically significant collections sampled by the MCMC algorithm in the two CoMEt runs by normalizing and combining the sampling frequencies of each collection across the two runs, and then computing the marginal probability graph on the merged collection.

### 2.2.7 Visualization of results

We created a web application for interactive visualization of the CoMEt results (`http://compbio-research.cs.brown.edu/comet/`; see Figure 2.4). For each dataset, the website shows the modules in the CoMEt marginal probability graph. Users can change the minimum edge weight parameter $\delta$, which dynamically updates the modules. Edges in each module are labeled with the marginal probability. Users can view the rows of the alteration matrix that correspond to a given module, and also view, sort, and search through the collections sampled by CoMEt that include alterations in a given module.

### 2.2.8 Somatic mutation datasets

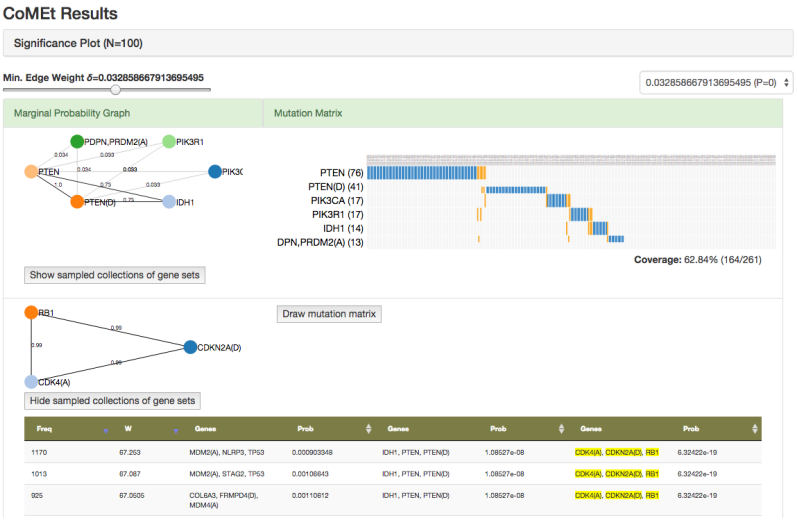



**Figure 2.4:** Screenshot of the web application for interactive visualization of CoMEt results.

**Acute myeloid leukemia (AML)** The AML dataset contains whole-exome and copy number array data in 200 AML patients from The Cancer Genome Atlas (TCGA) [5]. Using the annotations in [5], we have categorized multiple genes together based on expert knowledge, which results in 9 categories including spliceosome, cohesin complex, MLL-X fusions, other myeloid transcription factors, other epigenetic modifiers, other tyrosine kinase, serine/threonine kinase, protein tyrosine phosphatase, and RAS protein. More details are given in [5]. This results in 51 genes and 200 patients.

**Glioblastoma multiforme (GBM)** We analyzed three GBM datasets:

1. TCGA GBM dataset from Leiserson *et al.* [1]. This dataset contains whole-exome and copy number array data in 261 GBM patients and 398 genes from TCGA [6]. Data preparation for GBM can be found in [1]. Note that in section 2.3.2 we included amplifications in *EGFR* which were not considered in [1]. Also, we mapped deletions in *FAF1* to *CDKN2C*, since these genes are adjacent on chromosome 1, and *CDKN2C* is the likely target of the aberration.

2. TCGA GBM dataset from Szczurek *et al.* [62]. This dataset contains 83 alterations in 236 samples from [6], including single nucleotide variants in genes identified as significantly mutated by MutSigCV [17] and copy number aberrations caelled by GISTIC2 [23] then restricted to those with significantly concordant gene expression (higher for amplifications, lower for deletions).

3. TCGA GBM dataset from the TCGA Pan-Cancer project [52]. We analyzed the non-silent mutations (single nucleotide variants and small indels) from the

mutation annotation format (MAF) file and focal copy number aberrations from GISTIC2 output. This dataset contains 509 genes in 291 samples. Moreover, we removed genes with non-silent mutations in $< 1\,\%$ of samples and with mutations in $> 2.5\,\%$ of samples with MutSigCV [17] $q$-value $> 0.1$. This dataset contains 406 genes in 291 samples.

**Gastric cancer (STAD)** We analyzed the non-silent mutations (single nucleotide variants and small indels) from the MAF file in 289 gastric cancer samples. We also included focal driver copy number aberrations from GISTIC2 output via Firehose, fusion genes, rearrangements and splicing events [55]. We removed 74 hypermutators and genes with non-silent mutations in $< 2.5\,\%$ of samples and with mutations in $> 3\,\%$ of samples with MutSigCV [17] $q$-value $> 0.25$. This process results in 217 STAD patients and 397 genes with mutations. We considered four subtypes identified by TCGA [55], including tumors positive for the Epstein-Barr virus (EBV), tumors with high microsatellite instability (MSI), genomically stable (GS) tumors with a low level of somatic copy number aberrations, and chromosomally unstable (CIN) tumors with a high level of somatic copy number aberrations that were called. We do not analyze the MSI subtype since samples in MSI are hypermutated.

**Breast cancer (BRCA)** The BRCA dataset contains whole-exome and copy number array data in 507 BRCA patients and 375 genes from TCGA [8]. Data preparation for BRCA can be found in [1]. We downloaded subtype information of BRCA from TCGA [8]. We considered four subtypes — basal-like, HER2-enriched, luminal A, and luminal B — that each contain at least $10\,\%$ of the total samples.

We list the barcodes of the TCGA samples in each of the datasets in Additional file 2: Table S12.

### 2.2.9 Comparison to MEMo

The MEMo algorithm [39] uses a permutation test to approximate the probability of observing exclusive mutations in a gene set $M$ with contingency table $X$. The permutation test works by permuting the rows in $A$ corresponding to the genes in $M$, and then determining if the permutation has a higher test statistic than $M$. This is then repeated $N$ times to obtain an empirical $P$-value.

The crucial difference between MEMo and CoMEt is that MEMo uses the coverage $\Gamma(M)$ as the test statistic, while CoMEt uses the test statistic $T(X)$. (For ease of exposition, let $\Gamma(X)$ also be defined as the coverage for a contingency table $X$.)

The reasoning behind using the coverage as the test statistic is the idea that a gene set with mutually exclusive alterations will also have the highest coverage possible, for fixed frequencies of individual alterations. While this is true for pairs of genes (which follows from the fact that $2 \times 2$ contingency tables have only one degree of freedom), when one examines three or more genes, maximizing coverage is not the same as maximizing exclusivity. In fact, we can see that for a given contingency table $X$ it is possible to find another contingency table $X'$ with the same margins (gene frequencies) as $X$, but that has:

1. Higher exclusivity ($T(X') > T(X)$) and lower coverage ($\Gamma(X') < \Gamma(X)$), which could result in a deflated $P$-value for MEMo.
2. Lower exclusivity ($T(X') < T(X)$) but the same coverage ($\Gamma(X') = \Gamma(X)$), which would result in an inflated $P$-value for MEMo.[2]

See examples of both cases in Figure 2.5.

### 2.2.10 Comparison of CoMEt and mutex methods

The recently introduced mutex [63] algorithm uses an iterative version of the one-sided Fisher's exact test to evaluate combinations of mutually exclusive alterations. Thus, the tests used in mutex and CoMEt are identical when evaluating the exclusivity of a pair of alterations. However, for $k > 2$ alterations, mutex and CoMEt are quite different. CoMEt directly assesses the exclusivity of a $2^k$ contingency table. In contrast, mutex computes a series of $2 \times 2$ tests examining the exclusivity of alterations in one gene compared to the alterations in all $k - 1$ other genes in the set. For a set with $k > 2$ genes, mutex returns the least significant (highest) $P$-value of these $2 \times 2$ tests. While mutex's method is faster to compute than CoMEt, it is not as powerful at detecting mutual exclusivity in sets of $k > 2$ alterations, as shown in in section 2.3.1. Furthermore, while mutex searches for sets with $k > 2$ genes using a greedy approach to gradually expand mutually exclusive pairs, CoMEt uses an MCMC algorithm to simultaneously sample a collection of mutually exclusive sets. Searching for multiple sets simultaneously was shown to have advantages over the greedy approach in [1]. Finally, CoMEt summarizes the posterior distribution of the significant collections. Typically, CoMEt output contains multiple distinct modules. In contrast, mutex tends to produce results with many more genes, requiring prior

[2]We have not found a case where $T(X') < T(X)$ and $\Gamma(X') > \Gamma(X)$, and conjecture that such a case does not exist.

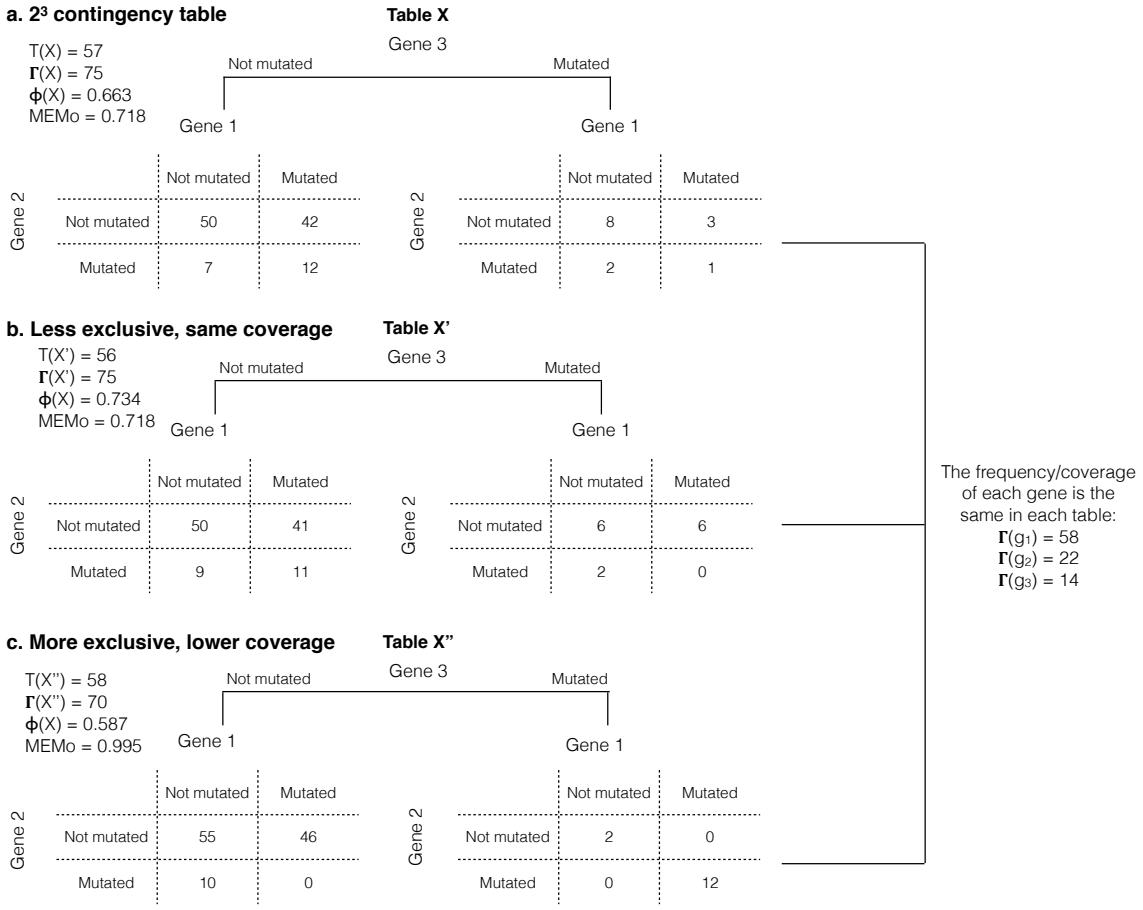



**Figure 2.5:** Two cases where the MEMo permutation test statistic $\Gamma$ (the coverage, or number of altered samples) deflates or inflates the $P$-value compared to the CoMEt test statistic $T$ (the number of samples with exclusive mutations).

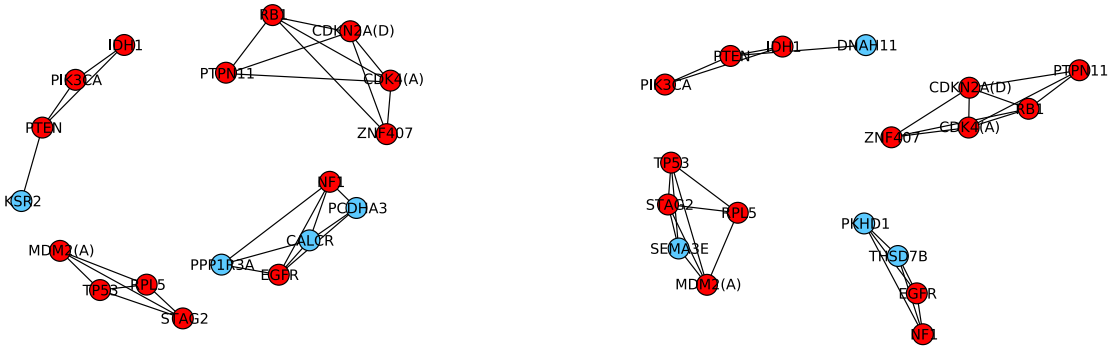
knowledge in the form of an interaction network to reduce the search space.

### 2.2.11 Comparison of methods with and without mutation filtering

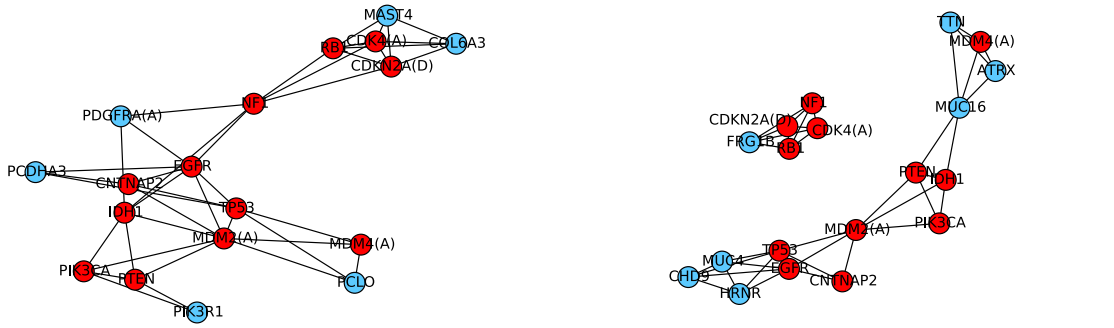
Because CoMEt conditions on the observed alteration frequencies, we argue that it is less biased towards genes that have high frequencies of passenger mutations, such as long genes. To illustrate this point, we compared CoMEt, Multi-Dendrix, and mutex on glioblastoma (GBM) data with and without the MutSigCV [17] filter that requires that frequently mutated genes have low MutSigCV $q$-values (see section 2.2.8 for details). We ran CoMEt with $k = 4$ and $t = 4$, ran Multi-Dendrix with its default parameters of $t$ ranging from 2 to 4 and $k$ ranging from 3 to 5, and ran mutex with default parameters except that we set the maximum size of a result group to 4 and did not include a signaling network. We used mutation data from the TCGA Pan-Cancer dataset [52] which contains whole-exome and copy number array data, and downloaded MutSigCV output from the corresponding Synapse repository (syn2812925). We used different TCGA GBM datasets here than in section 2.3.2 because of the availability of MutSigCV results on the Pan-Cancer dataset. For each cancer, we generated two datasets. In one dataset, we applied a MutSigCV filter to remove highly altered genes (altered in $> 2.5\,\%$ of samples) but insignificant by MutSigCV ($q$-value $< 0.1$). The second dataset did not include any MutSigCV filter.

We found that CoMEt identifies the key combinations of mutated genes with or without the mutation filtering on the GBM dataset (Table 2.1 and Figure 2.6a). These key combinations include genes from the Rb signaling (*CDK4*, *RB1*, *CDKN2A*), p53 signaling (*TP53*, *MDM2*), and PI(3)K signaling (*PIK3CA*, *PTEN*, *IDH1*) pathways, as well as *EGFR* and *NF1*. The CoMEt results were also largely stable: the core members of each module were unchanged, while four genes with less clear roles in GBM were lost and four genes were gained when we removed the mutation filtering.

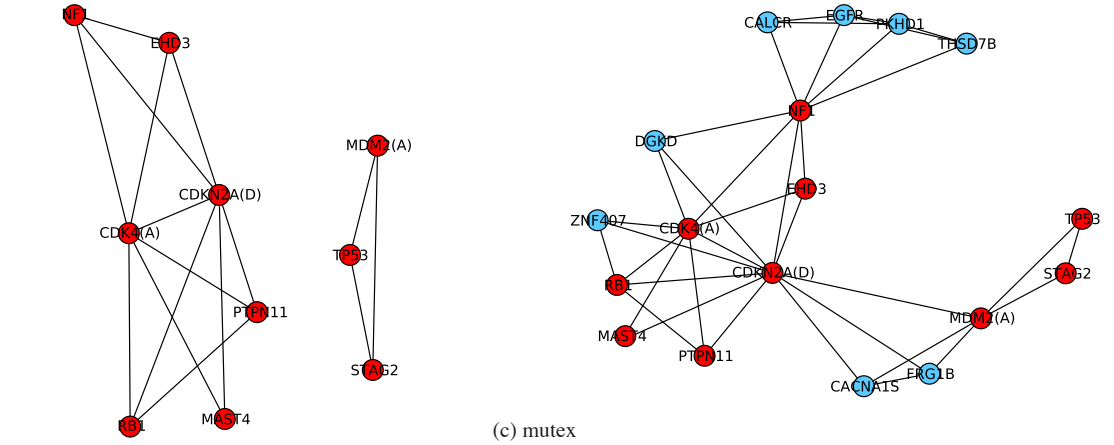
In contrast, Multi-Dendrix and mutex results change more substantially, with and without mutation filtering. The Multi-Dendrix modules are shuffled considerably, including the group of key GBM cancer genes (Table **??** and Figure 2.6(b)). In addition, six genes are lost and seven genes are gained after we remove mutation filtering. Furthermore, many of the genes that are added without mutation filtering are known to have elevated mutation rates, including *TTN*, *MUC16*, and *MUC4* [17]. This demonstrates a deficiency of the Dendrix weight function, also used by Multi-Dendrix, in that high coverage (frequently altered genes) may dominate over mutual exclusivity. The modules output by mutex also change considerably with and without mutation filtering (Table **??** and Figure 2.6c). Without mutation filtering, the number and composition of each module change, and eight genes are added. Moreover, mutex did not report the strong exclusive set from the PI(3)K signaling pathway (*PTEN*, *PIK3CA*, *IDH1*) found by CoMEt.




(a) CoMEt




(b) Multi-Dendrix




(c) mutex

**Figure 2.6:** **(a)** CoMEt modules, **(b)** Multi-Dendrix consensus, and **(c)** Mutex groups from the TCGA Pan- cancer GBM datasets [5] with (left) and without (right) mutation filtering with the MutSigCV algorithm. Red and blue circles represent genes that are common and different between the two results, respectively.

| Algorithm | Without filtering | With filtering |
|---|---|---|
| CoMEt | | |
| | 1. *IDH1, PIK3CA, PTEN,* ***KSR2*** | 1. *IDH1, PIK3CA, PTEN,* ***DNAH11*** |
| | 2. *MDM2*(A), *RPL5, STAG2, TP53* | 2. *MDM2*(A), *RPL5, STAG2, TP53,* ***SEMA3E*** |
| | 3. *EGFR, NF1,* ***CALCR, PCDHA3, PPP1R3A*** | 3. *EGFR, NF1,* ***PKHD1, THSD7B*** |
| | 4. *CDK4*(A), *CDKN2A*(D), *PTPN11, RB1, ZNF407* | 4. *CDK4*(A), *CDKN2A*(D), *PTPN11, RB1, ZNF407* |
| Multi-Dendrix | | |
| | 1. *CNTNAP2, CDKN2A*(D), *CDK4*(A), *EGFR, IDH1, MDM2*(A), *MDM4*(A), *NF1, PIK3CA, PTEN, RB1,* ***COL6A3, MAST4, PCDHA3, PCLO, PDGFRA*(A), *PIK3R1*** | 1. *CNTNAP2, EGFR, IDH1, MDM2*(A), *MDM4*(A), *PTEN, TP53,* ***ATRX, CHD9, HRNR, MUC4, MUC16, TTN*** |
| | | 2. *CDKN2A*(D), *CDK4*(A), *NF1, RB1,* ***FRG1B*** |
| mutex | | |
| | 1. *CDK4*(A), *CDKN2A*(D), *EHD3, MAST4, NF1, PTPN11, RB1* | 1. *CDK4*(A), *CDKN2A*(D), *EHD3, MAST4, MDM2*(A), *NF1, PTPN11, RB1, STAG2, TP53,* ***CACNA1S, CALCR, DGKD, EGFR, FRG1B, PKHD1, THSD7B, ZNF407*** |
| | 2. *MDM2*(A), *STAG2, TP53* | |

**Table 2.1:** Comparison of CoMEt, Multi-Dendrix, and mutex on the TCGA GBM dataset from the TCGA Pan-Cancer project [52] with and without mutation filtering. The consensus modules output by each algorithm are shown for the dataset with and without mutation filtering. Bolded genes indicate differences in output with and without mutation filtering. The (A) and (D) following the gene names indicate amplifications and deletions, respectively

## 2.3 Results

### 2.3.1 Comparison to other methods on simulated data

We compared CoMEt on two simulated mutation datasets to four other published methods for finding mutually exclusive gene sets: Dendrix [38], Multi-Dendrix [1], muex [62], and mutex [63]. In addition, we performed a separate comparison to MEMo [39] (see details in section 2.2.9).

**Benchmarking of methods for individual gene sets** We first compared the mutual exclusivity scores used by each of the methods on single gene sets using simulated datasets that represent key features of cancer sequencing data. In particular, each simulated dataset contains: (1) one implanted pathway $P$ with $k = 3$ genes that is altered in a fraction $\gamma_P$ samples with highly exclusive mutations; (2) a set $C$ of 5 highly altered genes whose alterations are not necessarily exclusive; (3) other genes containing only passenger mutations that were altered at rate $q$. The set $C$ models the highly recurrently altered genes that often appear in real cancer datasets, and can confound methods for identifying exclusive mutations. Further details of the simulation are given in Appendix A.3.

We compared CoMEt to the other methods on datasets with $n = 500$ samples and with implanted pathways with coverages $\gamma$ ranging from 0.1 to 1.0. We ran CoMEt with $k = 3$ for ten million iterations with 100 permutations, identifying modules in sets with $P < 0.05$. We ran mutex with default parameters except with maximum group size set to 3 and with 1,000 permutations, and reported the gene sets above mutex's suggested cutoff. We ran Dendrix and muex with $k = 3$ and reported the highest scoring significant ($P < 0.05$) set as neither algorithm outputs a consensus. We used coverage (parameter $\gamma$ in [62]) and the weight $W$ as the score for muex and Dendrix, respectively.

We computed the precision and recall for each algorithm across 25 simulated datasets for each coverage $\gamma$ (Additional file 2: Table S1). We summarized the results across the datasets using the $F$-measure, which is the harmonic mean of precision and recall. All the methods performed poorly ($F < 0.1$) with coverage $\gamma = 0.1$, and all the methods except mutex performed very well ($F > 0.9$) for coverage $\gamma \geq 0.8$ (Figure 2.7a). However, CoMEt outperformed the other methods for $\gamma = 0.2$ to 0.6. Both muex and Dendrix struggled to identify the implanted pathway ($F < 0.4$) with coverage $\gamma < 0.5$. In comparison, CoMEt had $F \geq 0.6$ for $\gamma > 0.2$. While mutex's performance was only slightly below that of CoMEt with $\gamma < 0.5$, mutex performed poorly compared to CoMEt and the other methods with $\gamma \geq 0.6$. Interestingly, the reason mutex performed poorly is because it identified many false positives resulting

in a low precision ($\leq 0.6$) even though its recall was 1.0. These false positive gene sets often include at least one gene from $C$ (the set of highly altered genes), indicating a problem with mutex's mutual exclusivity score. These simulations demonstrate the advantages of CoMEt's mutual exclusivity score in identifying mutually exclusive sets of genes (even when rarely mutated) in the presence of highly altered genes.

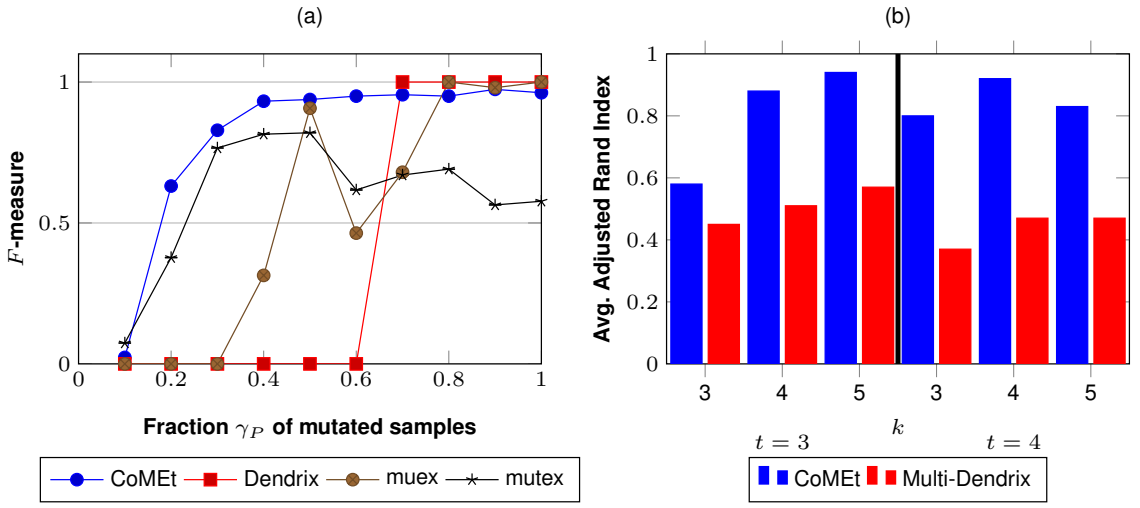



**Figure 2.7:** Comparison of CoMEt with other methods on simulated data with $n = 500$ samples. **(a)** The average $F$-measure of each method over 25 simulated datasets with varying coverage of the implanted pathway: CoMEt (blue), mutex (black), muex (brown), and Dendrix (red). **(b)** Comparison of CoMEt and Multi-Dendrix in identifying an implanted *collection* containing multiple sets of alterations. Bars indicate average of adjusted Rand index between reported and implanted collection across 25 simulated datasets

**Benchmarking identification of collections of gene sets** We compared CoMEt to Multi-Dendrix [1], an earlier method that also simultaneously finds collections containing more than one mutually exclusive set. We compared these two algorithms on two types of simulated datasets: one containing collections of gene sets with no overlapping genes, and the other containing overlapping gene sets. We generated simulated data using a procedure similar to that above with three important differences. First, we implanted a collection $\mathbf{P} = (P_1, P_2, \ldots, P_t)$ of $t$ pathways, each with exclusive mutations with total coverage $\gamma_{\mathbf{P}}$. Second, all genes in each implanted pathway are mutated in the same number of samples. Third, we include $m = 20,000$ genes and remove those mutated in fewer than 1 % of total samples (Figure 2.8). We generated datasets varying $t$ from 2 to 4 and $k$ from 3 to 5 with coverages $\gamma_P$ between 0.40 and 0.70 (Additional file 2: Table S2). We also generated datasets with *overlapping* implanted pathways with $t = 3$, $k$ from 3 to 5, with $\gamma_{\mathbf{P}} = (0.75, 0.75, 0.60)$.

On each dataset, we ran CoMEt using $k = 4, t = 3$, and Multi-Dendrix using its default parameters of $t$ ranging from 2 to 4, and $k$ ranging from 3 to 5. We compared the consensus sets output by Multi-Dendrix with the modules output by CoMEt, using the adjusted Rand index (ARI) [79], to score how well each algorithm identified

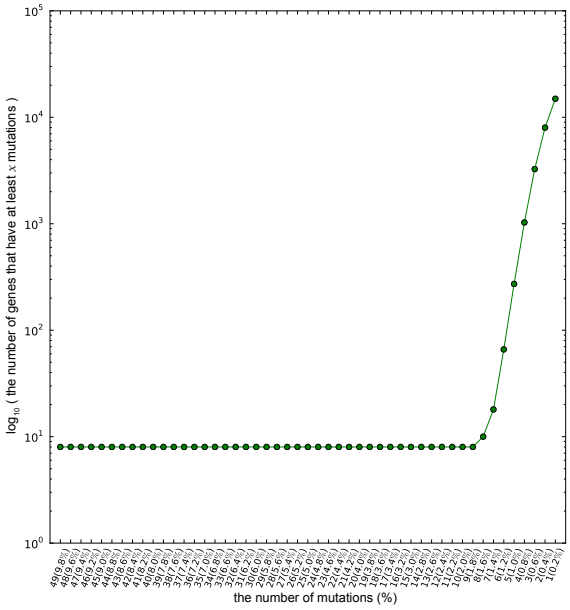



**Figure 2.8:** The distribution of the number of genes with $\geq x$ mutations in simulated data. We removed those genes mutated in fewer than 1% of mutations, i.e. genes mutated in fewer than 5 samples.

the implanted pathways. The ARI measures the agreement between two partitions, with ARI = 1 indicating that two partitions are identical and ARI = -1 indicating that two partitions are maximally dissimilar. CoMEt outperformed Multi-Dendrix in 11/12 simulated datasets (each containing 25 replicates) (Figure 2.7b and Additional file 2: Table S3). CoMEt found a much larger fraction of the implanted pathways (difference in ARI was $> 0.2$ for 8/12 datasets). Furthermore, CoMEt had an ARI $> 0.5$ for all 12 datasets, and ARI$> 0.8$ for 7/12 datasets. We emphasize that we ran CoMEt with a *single* value of $t$ and a *single* value of $k$ over all datasets even though the size and number of implanted pathways varied across datasets. In contrast, Multi-Dendrix was run with a range of parameter values. This demonstrates that CoMEt is much less sensitive to parameter choices than Multi-Dendrix.

We also compared the output of CoMEt and Multi-Dendrix using the true values of $t$ and $k$. We found that CoMEt outperformed Multi-Dendrix on 11/12 datasets (Additional file 2: Table S3). This shows that the statistical score used by CoMEt and the MCMC sampling are important features, even on simulated datasets where the implanted collections are fairly strong signals in the data.

### 2.3.2 CoMEt results on real cancer datasets

We ran CoMEt on four mutation datasets from TCGA: glioblastoma (GBM) [6], breast cancer (BRCA) [8], gastric cancer (STAD) [55], and acute myeloid leukemia (AML) [5]. We also analyzed the dataset of intracranial germ tumors from Wang *et al.* [80]. Because CoMEt can analyze any type of binary alterations, we include many types of alterations in these datasets: small indels and single nucleotide variations, copy number aberrations, aberrant splicing events, gene fusions, and (for BRCA and STAD) cancer subtype. See section 2.2.8 for details on these datasets and Appendix A.2 for details on parameters.

**Acute myeloid leukemia (AML)** We first ran CoMEt with $t = 4$ alteration sets, each of size $k = 4$. The CoMEt output contains four mutually exclusive modules that include 18 alterations (Figure 2.9). These four modules are: (1) *TP53*, *RUNX1*, *NPM1*, *PML-RARα* (52.5 % of samples); (2) *KDM6A*, *FLT3*, tyrosine kinases, *RAS* proteins, serine/threonine kinases, *DNMT3A*, *MLL*-X fusions, *MYH11-CBFβ*, and *RUNX1-RUNX1T1* fusion (70 % of samples); (3) cohesin complex, other myeloid transcription factors, and other epigenetic modifiers (33 % of samples); (4) *TET2* and *IDH2* (18.5 % of samples).

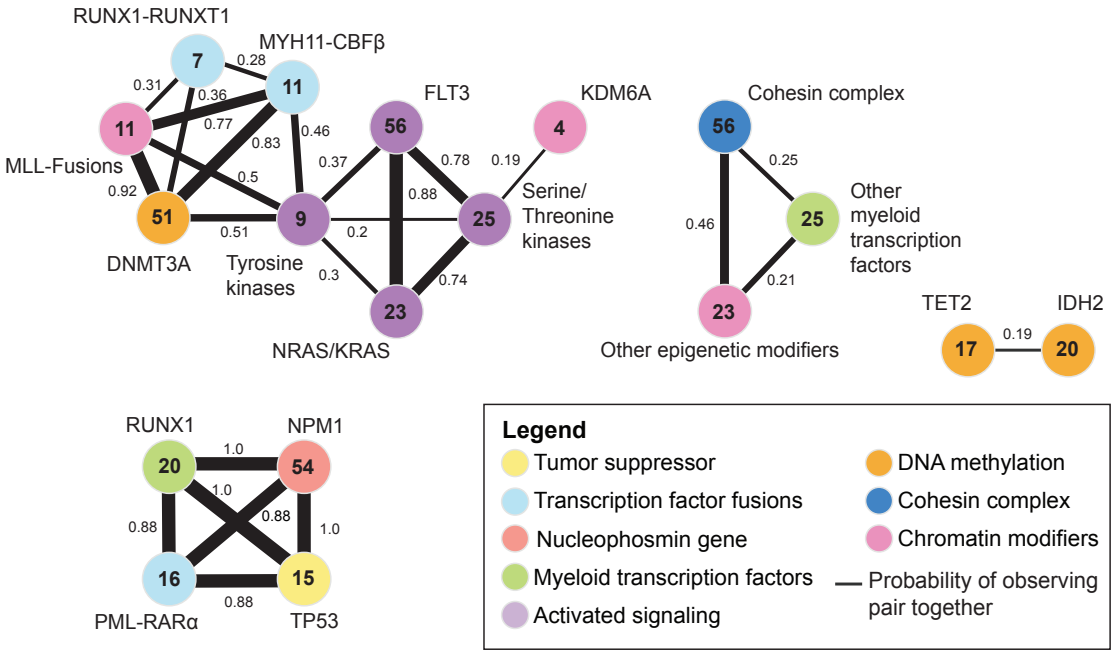



**Figure 2.9:** CoMEt results on AML dataset with $t = 4$ and $k = 4$ in the same style as the Figure 2.10.

The recent TCGA AML publication [5] reported strong mutual exclusivity (using

an earlier version of the CoMEt algorithm, called Dendrix++) across several expert-defined classes. Thus, we increased the value of $k$ to identify $t = 4$ gene sets with sizes $k = 6, 4, 4, 3$. Because of the larger values of $k$, we increased the number of MCMC iterations to 200 million (Additional file 2: Table S4). The resulting marginal probability graph ($\delta = 0.179$) contained four mutually exclusive modules with a total of 19 genes (Figure 2.10).

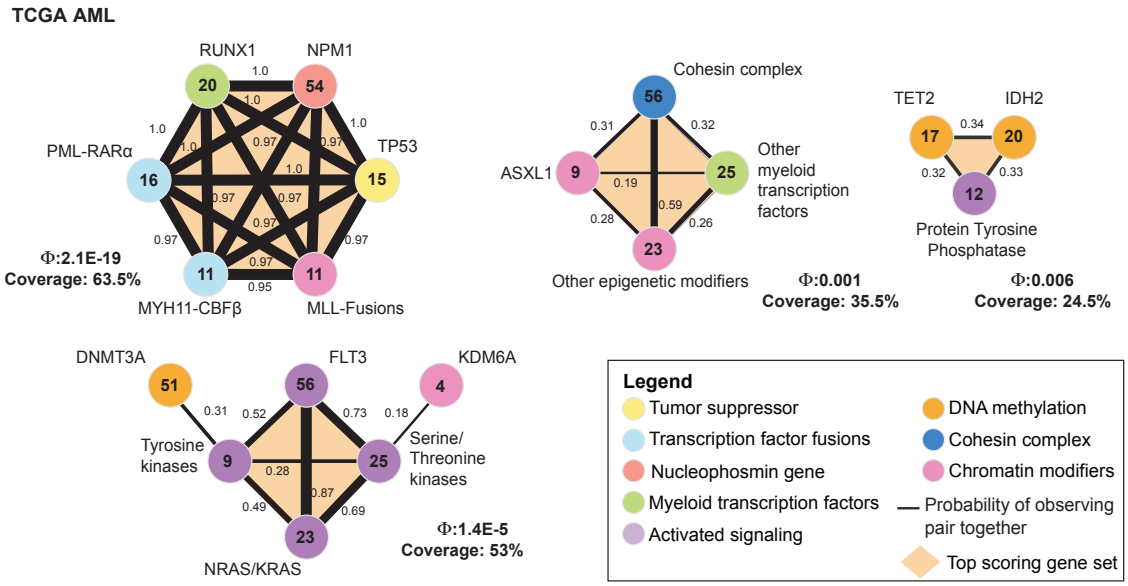



**Figure 2.10:** CoMEt results on TCGA AML consisting of four modules. Each circle represents the alterations in a gene or genomic region. The number in the circle indicates the number of samples in which the alteration occurs. Black lines are edges in the marginal probability graph with indicated probabilities. Orange polygons indicate the sets in the collection $\mathbf{M}$ with the most significant value $\Phi(\mathbf{M})$. Below each most significant set (orange) are the corresponding score $\Phi$ and coverage

---

The first module contains six perfectly mutually exclusive alterations. These six alterations include: mutations in *TP53*, *RUNX1*, *NPM1*; *PML-RARα*, *MYH11-CBFβ* fusion genes, and other *MLL* fusions, which we denote as *MLL*-X fusions, following [5]. These six alterations are known to be drivers in AML, and together are found in 63.5 % of the samples. These fusion genes are defining aberrations for certain subtypes of AML, as *PML-RARα*, *MYH11-CBFβ*, and *MLL* fusions are associated with acute promyelocytic leukemia, acute monoblastic or monocytic leukemia, and acute megakaryoblastic leukemia, respectively. The second module (altered in 63 % of samples) contains receptor tyrosine kinases (RTKs) and their downstream RAS target proteins. These include mutations in the *FLT3* tyrosine kinase, other tyrosine kinases, serine/threonine kinases, and *RAS* proteins. Two additional genes, *DNMT3A* and *KDM6A*, are also included in this set. These genes are involved in DNA/histone methylation, and their interactions with the other RTK/RAS genes in the set are less clear. Notably, the marginal probability graph (Figure 2.10) shows that the connection between *DNMT3A* and other genes in the set is largely due to its mutual exclusivity with other tyrosine kinases, and in fact a number of samples

have mutations in both *FLT3* and *DMNT3A* (Figure 2.11(b)). Thus, the patterns of exclusivity/co-occurrence between alterations may be subtle, demonstrating the advantages of CoMEt's approach to simultaneously examine multiple collections of sets of alterations.

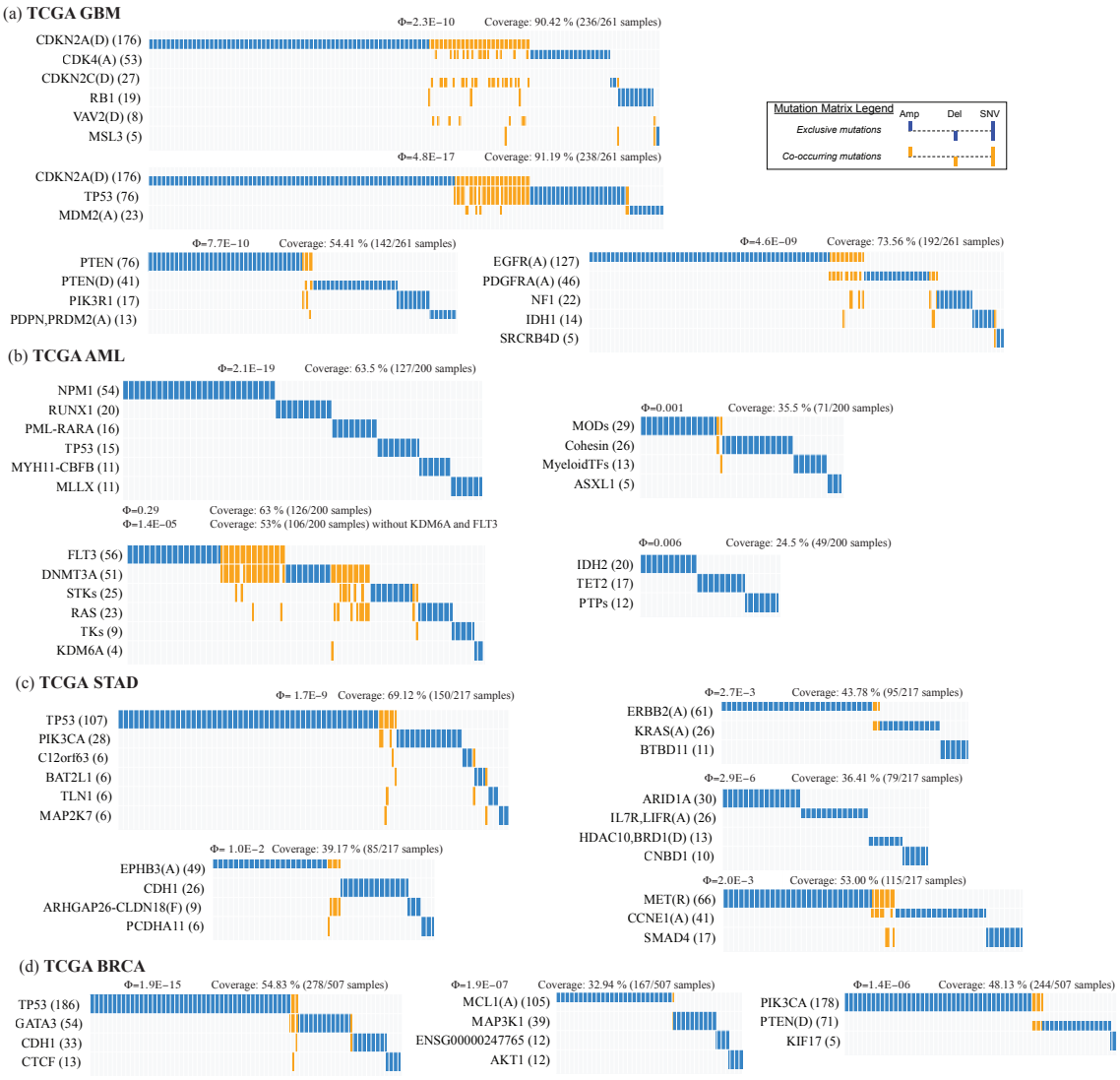



**Figure 2.11:** Mutation matrices for the CoMEt results on **(a)** TCGA GBM, **(b)** TCGA AML, **(c)** TCGA STAD, and **(d)** TCGA BRCA datasets. The matrices have alterations as rows, and samples as columns. Each cell indicates whether or not an alteration occurred in a particular sample, where grey indicates the sample was not altered. Samples with co-occurring alterations in the same set are colored orange, while exclusive alterations are colored blue.

The third module (altered in 35.5 % of samples) contains genes related to chromatin modification and gene regulation including *ASXL1*, the cohesin complex, other myeloid transcription factors, and other epigenetic modifiers. Finally, the fourth module (altered in 24.5 % of samples) contains genes related to DNA methylation

including *TET2*, *IDH2*, and protein tyrosine phosphatases. Mutual exclusivity between *TET2* and *IDH2* in AML has been previously reported [81, 82, 83]. Moreover, recent work provides a mechanistic explanation for this observed exclusivity: Figeroa et al. [81] show that mutant *IDH1 2* inhibits *TET2*'s function in demethylation of 5-methylcytosine. These results demonstrate that CoMEt is able to extract multiple functional modules directly from alteration data.

**Glioblastoma multiforme (GBM)** We ran CoMEt on the TCGA GBM dataset from Leiserson *et al.* [1] with $t = 4$ and $k = 4$ (Additional file 2: Table S5). While Leiserson *et al.* [1] removed amplifications in *EGFR* because they were so frequent it confounded their analysis, we added these amplifications back when running CoMEt, treating *EGFR* amplifications and *TP53* as subtypes so they could not be sampled in the same set (see section 2.2.6 for details). The resulting marginal probability graph ($\delta = 0.263$) includes two mutually exclusive modules (Figure 2.12a).

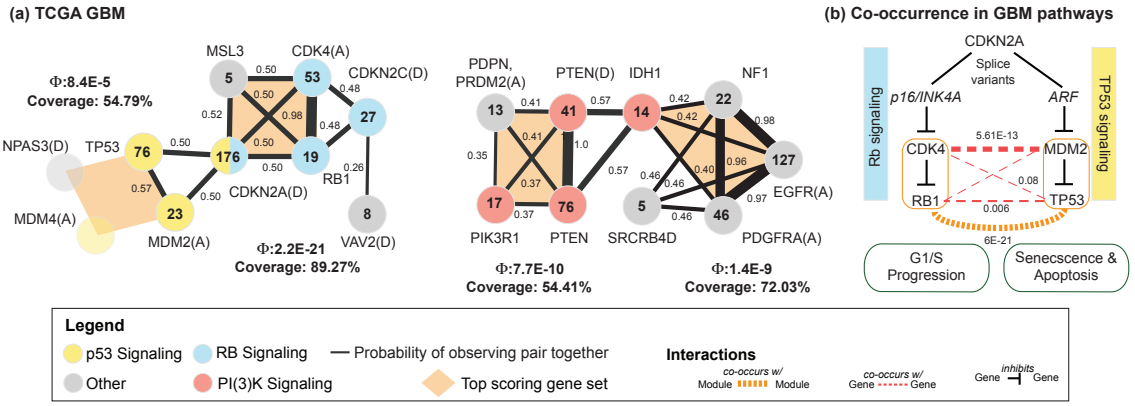



**Figure 2.12:** CoMEt results on TCGA GBM. **(a)** Two output modules from CoMEt are shown in the same style as in Figure 2.10. Characters in parentheses following gene name indicate copy number aberrations: **(D)** is a deletion, and **(A)** is an amplification. **(b)** Different splice variants of *CDKN2A* are part of both the Rb signaling (left) and p53 signaling (right) pathways. CoMEt recovers this relationship as two separate mutually exclusive gene sets. The gene sets {*RB1*, *CDK4*} and {*MDM2*, *TP53*} have a statistically significant number of co-occurring mutations ($P = 6 \times 10^{-21}$, dotted orange line), which is much more significant than the co-occurrence between pairs of genes in these sets (dotted red lines with corresponding $P$-values)

The first module includes alterations in three genes in the Rb signaling pathway (*CDK4*, *RB1*, *CDKN2C*) and in three genes in the p53 signaling pathway (*TP53*, *MDM2*, and *MDM4*), as annotated by the original TCGA GBM publication [6]. This module also contains deletions in *CDKN2A*, which is a member of both the Rb signaling and p53 signaling pathways. Indeed, it is well known that different isoforms of the *CDKN2A* gene are involved in the Rb and p53 signaling pathways (see Figure 2.12b and also [6]) and that the genomic deletion of *CDKN2A* affects both isoforms. Moreover, we find that the pairs *CDK4*-*RB1* and *MDM2*-*TP53* have surprisingly co-occurring alterations ($P = 6 \times 10^{-21}$; see Figure 2.12(b)). This

co-occurrence is stronger than the co-occurrence of alterations in individual genes. This pattern indicates that glioblastomas can alter the function of the Rb and p53 signaling pathways either by deleting *CDKN2A*, *or* by altering one gene in each of the pairs (*CDK4*, *RB1*) and (*TP53*, *MDM2*). We emphasize that CoMEt identified this overlapping module by sampling *nonoverlapping* exclusive sets. Finally, this module contains alterations in three additional genes: *NPAS3*, *VAV2*, and *MSL3*. *NPAS3* has been studied as a novel late-stage acting progression factor in gliomas with tumor suppressive functions [84, 85]. *VAV2* has been reported to regulate *EGFR*, and knockdown of *VAV2* enhanced EGFR degradation and further reduced cell proliferation [86]. MSL3 is a member of the male-specific lethal (MSL) complex and is thought to play a role in transcriptional regulation. As reported in [1], the MSL complex also includes MOF, which regulates p53 in the cell cycle and may be involved in cancer [87].

The second module includes alterations in the PI(3)K signaling pathway — including *PIK3R1*, *PTEN*, deletion of *PTEN* and *IDH1* — as well as amplifications in the genes (*EGFR*, *PDGFRA*) and in a region containing *PRDM2* and *PDPN*. Additional genes in this module are *NF1* and *SRCRB4D*. The PI(3)K signaling pathway genes overlap the results reported by Multi-Dendrix on this dataset in [1]; the important differences are that CoMEt includes *NF1* and amplifications in *EGFR* (which were not analyzed by [1]). In this module, we also found one mutually exclusive gene set (from the highest weight collection) that includes *EGFR*, *IDH1*, *NF1*, and *PDGFRA*. Alterations in these genes have strong association with individual subtypes in GBM [78]: *EGFR* amplification is associated with the Classical GBM subtype, *IDH1* and *PDGFRA* amplification are associated with the Proneural GBM subtype, and *NF1* is associated with the Mesenchymal GBM subtype. This shows that mutually exclusive gene sets can result from subtype-specific mutations.

Finally, *SRCRB4D* is a scavenger receptor with no known associations with cancer. However, two other scavenger receptor genes have previously reported roles in glioblastoma. Homozygous deletions of *DMBT1* were reported in glioblastomas and astrocytomas [88, 89]. *CD36* was recently reported to be involved in cancer stem cell maintenance in glioblastoma [90].

These results show that CoMEt can automatically find large portions of the pathways that were manually curated in TCGA GBM publication [6], including overlapping pathways. Moreover, CoMEt identifies additional genes with putative roles in glioblastoma and significant exclusivity with other known glioblastoma genes.

**Gastric cancer (STAD)** We performed two runs of CoMEt on the TCGA gastric cancer (STAD) dataset, and then merged the runs (described in section 2.2.6). We first ran CoMEt with $t = 4$ and $k = 4$. We then ran CoMEt on a STAD dataset that

included sample subtype classifications. TCGA recently classified gastric cancers into four subtypes based on integration of different molecular data [55]. To examine the relationships between subtypes and other alterations, we introduce "subtype alterations" for the three subtypes from [55] (we did not include the hypermutated samples from the MSI subtype in our analysis). As described in section 2.2.6, these "subtype alterations" are marked as altered in samples that are *not* members of the subtype, so that mutual exclusivity between a "subtype alteration" and another alteration indicates that the alteration is enriched in the subtype. We ran CoMEt on the STAD dataset with subtype alterations using $k = 4$ and $t = 3$ (the number of subtypes).

CoMEt identified five mutually exclusive modules from the marginal probability graph ($\delta = 0.132$) in the STAD dataset (Additional file 2: Table S6 and Figure 2.13a). Each of these modules includes known cancer genes and novel candidate genes. Two modules indicate subtype-specific altered genes and pathways. The first module (altered in 69 % (150/217) of the STAD samples) includes two genes, *TP53* and *PIK3CA*, that are enriched for alterations in the CIN and EBV subtypes, respectively. TCGA gastric study reported that 80 % of EBV tumors contain an alteration in *PIK3CA*, and suggested that EBV tumors might respond to PI3-kinase inhibitors [55]. Given this strong signal, it is not surprising that these two genes appear in CoMEt results. However, these signals do not dominate the CoMEtWe introduce the CoMEt algorithm for identifying collections of mutually exclusive alterations in cancer *de novo*, that is, with no prior biological knowledge. CoMEt uses a novel statistical score for exclusive alterations that conditions on the frequency of each alteration and thus can detect exclusivity of rare mutations. CoMEt overcomes large computational challenges in computing the score using a new algorithm for contingency table analysis, and in optimizing the score in genome-scale data using the first Markov chain Monte Carlo (MCMC) algorithm for identifying collections of multiple sets of exclusive alterations.

We demonstrate that CoMEt is superior to earlier *de novo* methods — Dendrix [38], muex [62], Multi-Dendrix [1], and mutex [63] — on simulated and real data. We then apply CoMEt to large mutation datasets from multiple TCGA cancer types [6, 5, 8, 55]. On each dataset, CoMEt identifies significantly exclusive collections of alterations that overlap well-known cancer pathways and also implicates novel cancer genes. In addition, CoMEt illustrates subtle relationships between mutual exclusivity resulting from cancer subtypes and exclusivity resulting from pathways or protein interactions. These findings provide testable hypotheses for further downstream analysis or experimental validation.

The input to CoMEt is a matrix of binary alterations, and thus can be used to analyze a variety of alterations including point mutations and indels, copy number aberrations (amplifications and deletions) and complex rearrangements, splice-site mutations, gene fusions, and subtype annotations. CoMEt may be useful in the

analysis of other types of alterations, such as germline variants.

Another application for CoMEt is pan-cancer analysis, such as the recently published TCGA study [52] and the upcoming ICGC Pan-Cancer Project. Since pan-cancer datasets have many cancer-type-specific alterations, CoMEt's ability to simultaneously analyze type-specific and other types of exclusive alterations should prove useful for this analysis. Finally, we anticipate that the novel tail enumeration strategy used in CoMEt may be of broader interest, both for examining mutual exclusivity in other datasets, including non-biological data, as well as for adapting for other types of exact statistics. CoMEt results, and four other interesting modules are also output. There are six other mutated genes in this first module including *MAP2K7*, *TLN1*, *BAT2L1*, *C12orf63* (recently renamed *CFAP54*), *MYOM3*, and *PTPRJ*. Given the rarity of these mutations, their significance is unclear.

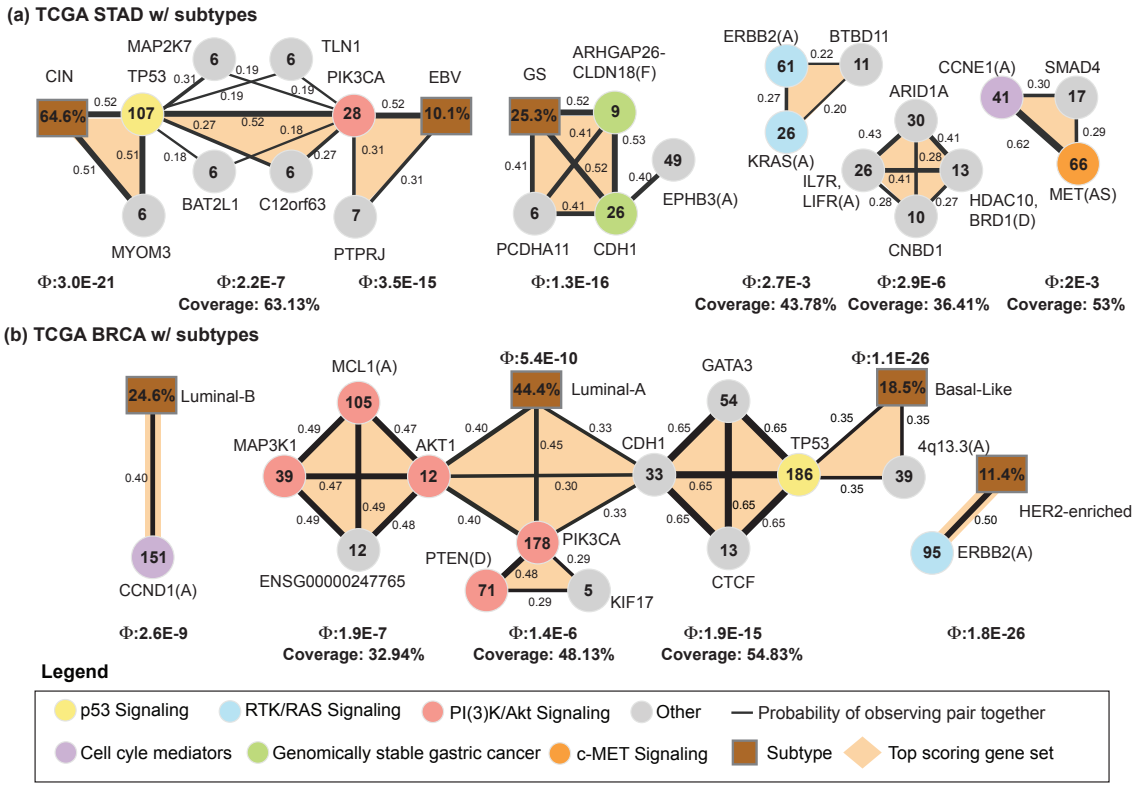



**Figure 2.13:** CoMEt results on (**a**) TCGA STAD subtypes, (**b**) TCGA BRCA subtypes. Style is the same as in Figure 2.12, except for the addition of subtype alterations (brown) and additional characters in parentheses following gene name: **(AS)** is an alternative splicing event, and **(F)** is a fusion gene. Note that an edge between a subtype (brown vertex) and an alteration indicates that the alteration occurs frequently in the subtype

---

The second STAD model includes the genomically stable (GS) subtype, mutations in *CDH1*, mutations in *PCDHA11*, *ARHGAP6-CLDN18* fusions, and amplification of a region containing *EPHB3*. *CDH1* somatic mutations and *ARHGAP6-CLDN18* fusions were reported to be mutually exclusive and enriched in the genomically stable

subtype in gastric cancer [55], and CoMEt recapitulates this result. *EPHB3* is the member of Eph/ephrin signaling which controls the compartmentalization of cells in epithelial tissues. A recent study [91] demonstrated EphB receptors (for example, *EPHB1* and *EPHB3*) interacting with *CDH1* in epithelial intestinal cells which regulates the formation of E-cadherin-based adhesions. This interaction explains the perfect mutual exclusivity between *CDH1* and *EPHB3*, which to our knowledge is the first report of this relationship. This demonstrates that mutual exclusivity between pairs of alterations/subtypes may have subtle explanations, further underscoring the need for analysis of collections of multiple alterations.

The third module (altered in 95/217 of samples) includes amplifications of *KRAS* and *ERBB2*, and mutations in *BTBD11*. *KRAS* and *ERBB2* are members of the RTK/RAS signaling pathway, and their role in cancer is well-documented. Little is known about the function of *BTBD11*, and thus the significance of the mutations is unclear.

The fourth STAD module (115/217 of samples) contains three altered genes, including amplifications of *CCNE1*, mutations in *SMAD4* and splice-site mutations in *MET*. *CCNE1* is a well-known cell cycle mediator, *SMAD4* is a member of the TGF-$\beta$ pathway, and *MET* participates in the RTK/RAS signaling pathway [55].

The fifth STAD module (79/217 of samples) contains four altered genes, including amplifications in a region with *IL7R* and *LIFR*, deletions in a region with *HDAC10* and *BRD1*, mutations in *ARID1A*, and mutations in *CNBD1*. *ARID1A* is a well-known cancer gene shown to be significantly mutated in gastric cancer [55]. Moreover, inhibition of *HDAC10* has been reported to be associated with human gastric cancer cells [92]. Gain-of-function mutations in *IL7R* have been reported to be associated with childhood acute lymphoblastic leukemia [93]. Our CoMEt results suggest that *IL7R* mutations may have a role in gastric cancer as well.

**Breast cancer (BRCA)** We performed two runs of CoMEt on the TCGA breast cancer (BRCA) dataset, and then merged the runs. We first ran CoMEt with $k = 4$ and $t = 3$. We then introduced subtype alterations for four subtypes from [8] (as described in section 2.2.6). Breast cancers are traditionally classified into multiple subtypes based on mRNA expression. Here we analyze four subtypes: luminal A, luminal B, basal-like, and HER2-enriched. We ran CoMEt on a BRCA dataset that included sample subtype classifications with $k = 4$ and $t = 4$ (the number of subtypes).

CoMEt identified three subtype-specific modules and three modules with mutated genes (Additional file 2: Table S7 and Figure 2.13b) in the marginal probability graph ($\delta = 0.287$). The first module shows the strong association between amplification of

*CCND1* and the luminal B subtype as previously reported [94]. Similarly, the third module shows the strong association between *ERBB2* amplification and the HER2 (*ERRB2*)-enriched subtype.

The second module shows a complicated relationship between: (1) subtype-associated alterations in the luminal A and basal-like subtypes, and (2) mutual exclusivity resulting from alterations in the same pathway(s). This module contains five sets of genes (highlighted in orange in Figure 2.13b) in the highest scoring collection **M** output by CoMEt. Consistent with TCGA study [8], we find that *CDH1*, *AKT1*, and *PIK3CA* are associated with the luminal A subtype, and they form a set in the CoMEt output. Similarly, *TP53* and amplification of chromosome region 4q13.3 are associated with the basal-like subtype, and they also form a set in the CoMEt output. Two of the other sets contains genes in the same pathway. *PTEN* is a known inhibitor of *PIK3CA*, explaining the observed exclusivity between *PTEN* deletion and *PIK3CA* mutation. Moreover, *MCL1*, *MAP3K1*, *AKT1* are all part of the PI(3)K/Akt signaling pathway. Together, these sets contain five genes that are annotated as part of the PI(3)K/Akt signaling pathway in TCGA study [8] (red circles in Figure 2.13b).

The final set in this module includes mutations in the genes *TP53*, *CDH1*, *GATA3*, and *CTCF*. These four genes are altered in 54.83 % (278/507) of the BRCA samples. *TP53* is a member of the p53 signaling pathway, while *CDH1*, *GATA3*, and *CTCF* all have been reported as potential driver genes in breast cancer. *CDH1* is a tumor supressor that is well-known to play multiple roles in cancer [95], including invasion and proliferation in breast cancer [96]. *GATA3* is a transcription factor that has long been known to be involved in breast cancer tumorigenesis [97]. Recently, *GATA3* has been reported to promote differentiation, suppress metastasis, and alter the tumor microenvironment in breast cancer [98]. As noted by Leiserson *et al.* [1], *GATA3* has also been reported to suppress tumor metastases through inhibition of *CDH1* promoters [99], which suggests that the mutations in *GATA3* are an alternate way to downregulate *CDH1* and may explain the exclusivity of the mutations in *GATA3* and *CDH1*. Moreover, *GATA3* is enriched for mutations in both luminal A and luminal B; that is, 32 of the 54 mutations in *GATA3* occur in luminal A ($P = 0.0207$) and 19 of the 54 mutations in *GATA3* occur in luminal B ($P = 0.065$). This might suggest that *GATA3* mutations mainly occur in patients with luminal breast cancer. *CTCF* neighbors *CDH1* on chromosome 16q22.1 and has been reported with *CDH1* to be a tumor suppressor in breast cancer [100, 101]. Interestingly, both *CDH1* and *CTCF* have most of their mutations in samples of the luminal A subtype. *CDH1* is enriched for mutations in luminal A (as reported in [8]) and 9 of the 13 mutations in *CTCF* occur in luminal A ($P = 0.0891$), suggesting that these two genes are in a pathway specifically targeted in luminal A. Furthermore, 4 of the 9 mutations in *CTCF* in luminal A are missense mutations in zinc finger domains, suggesting a possible functional role for these mutations [102].

Together, these results demonstrate CoMEt's ability to simultaneously identify alterations that are mutually exclusive due to interactions between genes in pathways or due to subtype-specific alterations. This allows a more refined interpretation of mutually exclusive alterations than simple pairwise analyses.

**Intracranial germ tumors** To investigate CoMEt's performance on a smaller dataset that is less intensively studied than TCGA datasets, we ran CoMEt on a dataset of somatic and germline mutations in intracranial germ cell tumors (IGCTs) from [80]. This dataset consists of somatic single nucleotide variants and indels in 163 genes from 53 patients (combining both the discovery and validation cohorts). Given the small size of this dataset, we first ran CoMEt to identify $t = 1$ set of $k = 3$ genes (Additional file 2: Table S8). CoMEt found that the alterations in the set of $k = 3$ genes *KIT* (16 mutations), *KRAS* (9), and *NRAS* (3) were the most exclusive ($\Phi = 0.002$). Wang *et al.* [80] identified this triple using Fisher's exact test comparing mutations *KIT* with the union of mutations in *KRAS* and *NRAS* ($P = 0.018$). Notably, the exact test gives the triple a more significant $P$-value. Mutual exclusivity between these three genes is consistent with the RAS genes being downstream of *KIT* in the signaling receptor tyrosine kinase (RTK) signaling pathway.

The top-ranked *KRAS*, *NRAS*, *KIT* triple was closely followed by several other gene sets including *KIT*, *KRAS*, and a third gene (*FLT3*, $\Phi = 0.004$; *KDM2A*, $\Phi = 0.004$; *LAMA4*, $\Phi = 0.004$; *SPRY4*, $\Phi = 0.004$). Notably, *KIT* and *FLT3* (2 mutations) are both receptor tyrosine kinases (RTKs); the mutual exclusivity of their mutations suggests that *FLT3* mutations may substitute for *KIT* mutations in some samples. In addition, *SPRY4* is a negative regulator of RAS signaling and was recently shown to inhibit RAS signaling in AML [103]. *SPRY4* was not discussed in the Wang *et al.* study, and thus is a novel discovery by CoMEt. Intriguingly, the observed mutual exclusivity that we see in the high-scoring gene triples from CoMEt (Additional file 2: Table S8) are similar to relationships seen between RTK and RAS signaling in AML [5, 103].

CoMEt summarized the mutually exclusive sets into two statistically significant modules ($P < 0.01$). The first module includes *KIT*, *KRAS*, *NRAS*, *TP53*, and *LAMA4* (Figure 2.14), which are collectively mutated in 62 % (33) of the 53 patients. All five of these genes were identified as containing significantly recurrent mutations by Wang *et al.* The second module contains perfectly exclusive mutations in *JMJD1C* and *CBL*, which are mutated in 30 % (16) of the 53 patients. *CBL* is the third most somatically mutated gene in the Wang *et al.* study, and Wang *et al.* described a role for *CBL* as a negative regulator of RTKs, including *KIT*. However, mutations in *CBL* and *KIT* are not significantly exclusive ($P = \Phi = 0.253$). This is because *CBL* is mutated in only six samples, one of which also has a mutation in *KIT*. Furthermore, the exclusivity between mutations in the gene triple, *KIT*, *KRAS*, and

*CBL*, is less significant ($\Phi = 0.023$) than the mutations in the gene triple, *KIT*, *KRAS*, and *NRAS* ($\Phi = 0.002$). Interestingly, all the mutations in *JMJD1C* are germline variants (Wang *et al.* noted a significant enrichment of germline variants in *JMJD1C*). Thus, CoMEt identified mutual exclusivity between germline mutations in *JMJD1C* and somatic mutations in *CBL*.

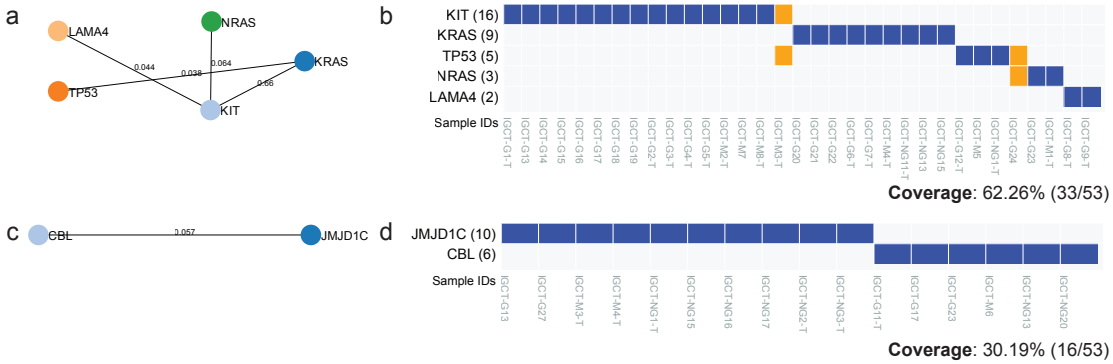



**Figure 2.14:** Statistically significant modules identified by CoMEt (with $k = 3, t = 1$) on ICGTs from Wang et al. [27]. **(a, c)** Marginal probability graphs of the two modules. **(b, d)** Mutation matrices of the two modules. Representation as in Figure 2.10.

We further investigated these modules by running CoMEt with parameters $\alpha = 3, t = 2, k = 3$ in order to to identify multiple gene sets simultaneously (Additional file 2: Table S9). The highest scoring collections included *KIT* and *KRAS* in one gene set, and *JMJD1C* and *CBL* in the other. This suggests that the mutations in these two pairs of genes co-occur, and indeed 6 samples have a mutation in either *JMJD1C* or *CBL* and either *KIT* or *KRAS* (co-occurrence $P = 0.28$ by Fisher's exact test). With only 53 samples in this dataset, it is difficult to identify all of the subtle relationships between mutual exclusivity and co-occurrence in larger sets of genes. Nevertheless, these results show the advantages of CoMEt analysis over pairwise tests of mutual exclusivity.

### 2.3.3 Comparisons to other methods on real data

We compared CoMEt to the Multi-Dendrix [1] and mutex [63] algorithms on the TCGA GBM dataset from Leiserson *et al.* [1] and the TCGA AML [5] dataset. We did not compare on the TCGA BRCA or STAD datasets because CoMEt analyzes subtype-specific mutations, while Multi-Dendrix and mutex do not. We also provide a separate comparison to the muex algorithm [62] on GBM data in Appendix A.2, as muex does not identify multiple sets of alterations simultaneously. We ran Multi-Dendrix and mutex with default parameters, except that we set the maximum size of the mutex result groups to 4. We compared the modules output by CoMEt with

the consensus output by Multi-Dendrix and the default output of mutex. We list the modules identified by Multi-Dendrix and mutex in Tables S10 and S11.

**Glioblastoma multiforme (GBM)** We compared CoMEt's results to the consensus modules reported by Leiserson *et al.* with Multi-Dendrix [1] on the TCGA GBM dataset [6] from Leiserson *et al.* (Additional file 2: Table S10(a)). Both CoMEt and Multi-Dendrix identify modules overlapping the Rb, p53, and PI(3)K signaling pathways. However, there are several key differences. First, CoMEt correctly places *CDKN2A* in a module with both the Rb and p53 signaling pathways, consistent with the figure in the TCGA GBM publication [6], while Multi-Dendrix does not. Second, in the module overlapping the PI(3)K pathway, CoMEt includes *NF1* and amplifications in *EGFR*, the latter alteration not analyzed in the Multi-Dendrix publication [1].

We performed two comparisons with mutex. First, we ran mutex without an input signaling network. Mutex reported a single connected component with 125 genes (Figure 2.15 and Additional file 2: Table S11(a)). Although this component overlaps the four signaling pathways mentioned in the TCGA GBM paper [6], too many genes are included due to pairwise exclusivity with individual genes in the well-known signaling pathways. This makes it difficult to interpret the results. Next, we ran mutex with its default input signaling network to see whether limiting the search space would improve the mutex results (Figure 2.16 and Additional file 2: Table S11(b)). Again, mutex reported a single connected component, this time with 16 genes. The component contains multiple mutually exclusive relationships also reported by CoMEt (for example, exclusivity between mutations in *CDK4*, *RB1*, and *CDKN2A*), but the CoMEt results are much easier to interpret because they include multiple modules. Even without the prior knowledge of protein interactions, the CoMEt results are arguably superior to those of mutex.

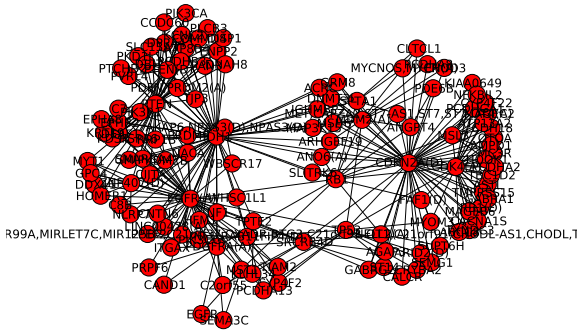


**Figure 2.15:** mutex results on the TCGA GBM dataset from Leiserson *et al.* [1].

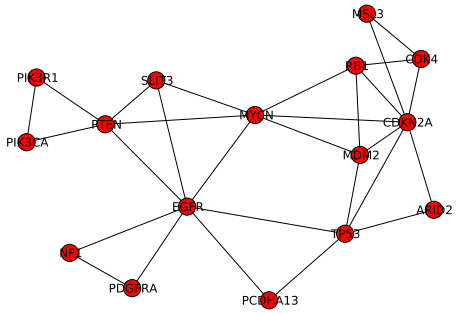



**Figure 2.16:** mutex results on the TCGA GBM dataset from Leiserson *et al.* [1] with mutex's default signaling network.

The comparison between CoMEt and mutex demonstrates several key advantages of our approach. First, although mutex's results were indeed improved when using the signaling network, the massive differences between mutex's results with and without the network indicates that mutex relies heavily on the network for prior knowledge. By not using prior knowledge, CoMEt can identify more novel combinations of mutations. Second, mutex's reliance on the signaling network makes it more difficult for it to handle different types of aberrations compared to CoMEt. This is because when using a signaling network, aberrations must be mapped to single genes. But this is typically difficult for copy number aberrations that span a large region containing many genes. Mapping these aberrations to a signaling network is a difficult computational problem, and may obscure the underlying exclusivity between these copy number aberrations and other alterations. In contrast, CoMEt handles any types of aberrations as separate entries in the alteration matrix.

**Acute myeloid leukemia (AML)** We ran Multi-Dendrix and mutex on the TCGA AML dataset [5]. We did not run mutex with a signaling network because many of the alterations in the AML dataset are for groups of genes (for example, protein tyrosine phosphotases; see [5]). Multi-Dendrix reports a single consensus module that includes 19 genes (Figure 2.17 and Additional file 2: Table S10(b)), and mutex identifies a connected component with 17 genes (Figure 2.18 and Additional file 2: Table S11(c)). The size and complicated topology of these results make them difficult to interpret, especially compared to the CoMEt results, which include four different modules with 3 to 7 alterations each (Figure 2.10). However, it is clear that while both Multi-Dendrix and mutex identify mutually exclusive mutations also identified by CoMEt (for example, mutual exclusivity between mutations in *PML-RARα*, *NPM1*, and *RUNX1*), they also miss key relationships (for example, exclusivity of mutations between *TET2*, *IDH2*, and the protein tyrosine phosphatase

group).

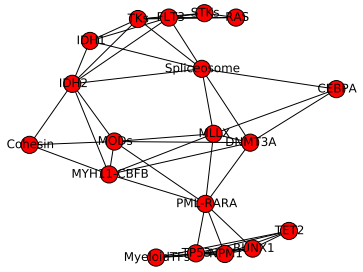



**Figure 2.17:** Multi-Dendrix results on the TCGA AML dataset [5].

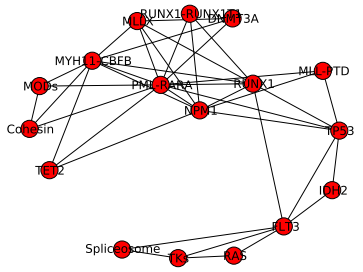



**Figure 2.18:** mutex results on the TCGA AML dataset [5].

### 2.3.4 Robustness of CoMEt results on real data

**Bootstrapping** We used a bootstrapping approach to determine the robustness of the results from CoMEt. We sampled with replacement from the TCGA GBM dataset from Leiserson *et al.* [1] to generate resampled datasets. For each resampled dataset, we ran CoMEt and compared the output modules to the modules obtained on the whole dataset. We recorded the number of genes in common and the number of additional genes found by CoMEt in the resampled datasets (Figure 2.19a). CoMEt recovered an average of 11 from the 17 genes in the modules from the whole dataset, and found an average of 8 additional genes. The genes in the most exclusive triples were recovered the most often (Figure 2.20a): *CDKN2A*(D)-*TP53*-*MDM2*(A) (at least 68 % of datasets), *CDKN2A*(D)-*CDK4*(A)-*RB1* (84 %), and *PTEN*-*PTEN*(D)-*IDH1* (88 %).

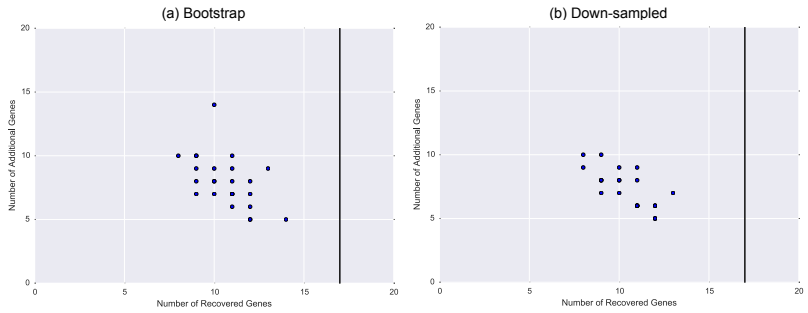



**Figure 2.19:** Robustness of CoMEt on TCGA GBM [6] datasets from Leiserson *et al.* [1]. Datasets were **(a)** bootstrapped and **(b)** down-sampled to include only 50% of the samples. We ran CoMEt on 25 such datasets, and computed the number of recovered genes (x-axis) and number of additional genes (y-axis) compared to the CoMEt results on the full dataset. For comparison, the CoMEt results on the full dataset included 17 genes (black line).

**Downsampling** We also compared the CoMEt results on TCGA GBM dataset from Leiserson *et al.* [1] to the results obtained with only half the samples from this dataset. We created 25 datasets, each containing a random selection of 131 (50%) of the samples. For each 50% dataset, we ran CoMEt on the resampled datasets (Figure 2.19b). Across the 25 datasets, CoMEt recovered an average of 11 from the 17 genes in the modules from the whole dataset, and found an average of 7 additional genes. We also computed how often CoMEt recovered the pairs in the *CDK4-RB1-CDKN2A* triple 84 % of the time, and the pairs in the *TP53-MDM2-CDKN2A* and the *PTEN-PTEN*(D)*-IDH1* triples 48 % of the time (Figure 2.20b). This demonstrates that the results of CoMEt are fairly robust to changes in the number of samples. However, the well-known cancer pathways are found less frequently than in the bootstrapping results above, demonstrating that robust detection of mutual exclusivity does require a sufficient number of samples. Further theoretical analyses of the number of samples required to detect mutually exclusive sets are reported in [104].

## 2.4 Discussion

We introduce the CoMEt algorithm for identifying collections of mutually exclusive alterations in cancer *de novo*, that is, with no prior biological knowledge. CoMEt uses a novel statistical score for exclusive alterations that conditions on the frequency of each alteration and thus can detect exclusivity of rare mutations. CoMEt overcomes large computational challenges in computing the score using a new algorithm for contingency table analysis, and in optimizing the score in genome-scale data using the first Markov chain Monte Carlo (MCMC) algorithm for identifying collections of multiple sets of exclusive alterations.

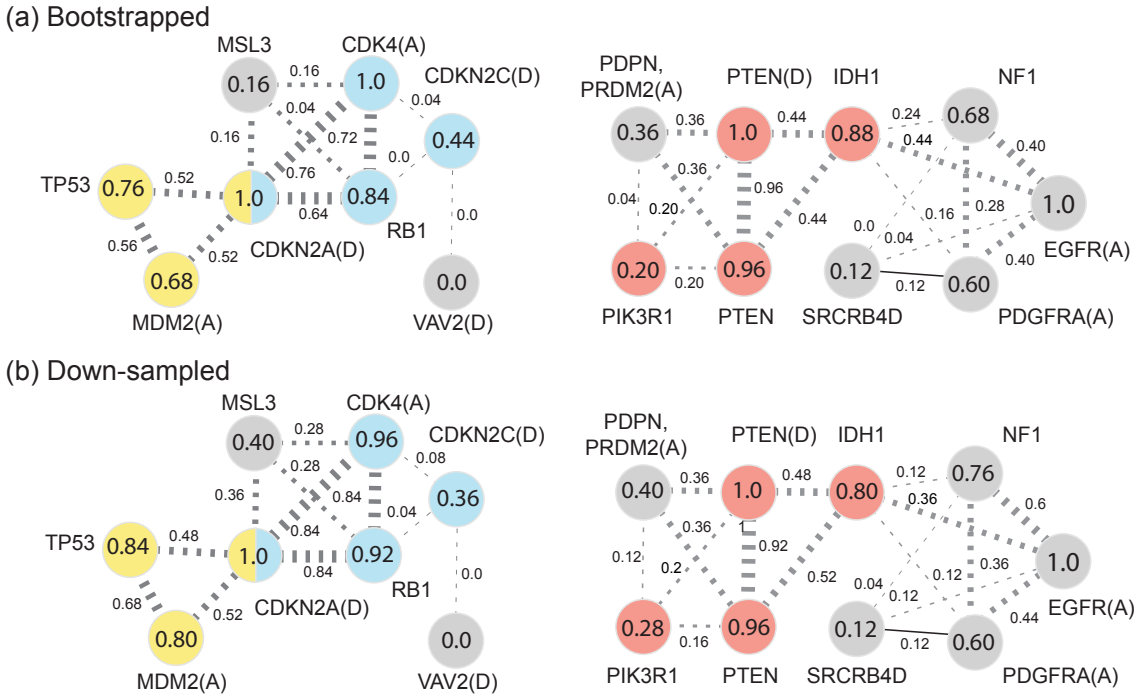



**Figure 2.20:** Robustness of the modules identified by CoMEt on the TCGA GBM [6] dataset from Leiserson *et al.* [1]. We ran CoMEt on TCGA GBM datasets **(a)** bootstrapped (sampled with replacement) and **(b)** down-sampled to include only 50% of the samples. Shown is the marginal probability graph output by CoMEt on the TCGA GBM dataset. Nodes and edges are labeled with the proportion of down-sampled datasets in which they were identified by CoMEt. The most mutated genes in each of the Rb, p53, and PI(3)K signaling pathways were identified on at least 68% (17/25) and 80% (20/25) of the bootstrapped and down-sampled datasets, respectively.

We demonstrate that CoMEt is superior to earlier *de novo* methods — Dendrix [38], muex [62], Multi-Dendrix [1], and mutex [63] — on simulated and real data. We then apply CoMEt to large mutation datasets from multiple TCGA cancer types [6, 5, 8, 55]. On each dataset, CoMEt identifies significantly exclusive collections of alterations that overlap well-known cancer pathways and also implicates novel cancer genes. In addition, CoMEt illustrates subtle relationships between mutual exclusivity resulting from cancer subtypes and exclusivity resulting from pathways or protein interactions. These findings provide testable hypotheses for further downstream analysis or experimental validation.

The input to CoMEt is a matrix of binary alterations, and thus can be used to analyze a variety of alterations including point mutations and indels, copy number aberrations (amplifications and deletions) and complex rearrangements, splice-site mutations, gene fusions, and subtype annotations. CoMEt may be useful in the analysis of other types of alterations, such as germline variants.

Another application for CoMEt is pan-cancer analysis, such as the recently published TCGA study [52] and the upcoming ICGC Pan-Cancer Project. Since pan-cancer datasets have many cancer-type-specific alterations, CoMEt's ability to simultaneously analyze type-specific and other types of exclusive alterations should prove useful for this analysis. Finally, we anticipate that the novel tail enumeration strategy used in CoMEt may be of broader interest, both for examining mutual exclusivity in other datasets, including non-biological data, as well as for adapting for other types of exact statistics.

# Chapter 3

# A Weighted Exact Test for the Significance of Mutually Exclusive Mutations

In this chapter, we present a weighted test for mutual exclusivity that uses per gene, per sample mutation probabilities. We set these mutation probabilities based on the observed number of mutations in each gene and sample, and we apply this weighted test to tumors from two cancer types – colorectal and endometrial – with high and a highly variable number of mutations per sample. At the time of writing, most of the material in this chapter is in submission at the *European Conference on Computational Biology* [50]. I am the first author on the submission, and want to acknowledge Matthew Reyna (a co-author on the paper) for deriving the saddlepoint approximation presented in this chapter.

## 3.1 Background and related work

As we have discussed throughout the first two chapters, a key challenge in cancer genomics is distinguishing the small number of somatic mutations that drive cancer from the vast majority of mutations that accumulate randomly. The ability to distinguish these *driver* mutations from the random *passenger* mutations may lead to better understanding of cancer biology and personalized therapies customized to a tumor's mutational profile. However, large scale cancer sequencing efforts such as The Cancer Genome Atlas (TCGA) [105, 106, 107] and the International Cancer Genome Consortium (ICGC) have shown that many driver mutations are rare across patient cohorts and thus distinguishing the driver mutations from the passengers by

their frequency of occurrence is a difficult problem.

Driver mutations are hypothesized to group into a small number of pathways or hallmarks [18], and this hypothesis is a widely accepted explanation for the observed mutational heterogeneity of cancer [17]. Thus, researchers have developed methods to identify combinations of mutations using varying levels of prior knowledge, from pathway databases [108, 109, 30], protein-protein interaction networks [110, 111, 112, 113], or *de novo* methods that use no prior knowledge [114, 115, 111, 116, 117, 118, 119, 120, 121]. De novo methods are advantageous because prior knowledge of pathways and interactions is often noisy or unavailable.

The vast number of possible combinations of mutated genes makes complete de novo discovery of combinations computationally and statistically intractable. Thus, de novo methods to date rely on the observation that mutations within the same pathway are often mutually exclusive across tumors [36, 37]. In recent years, a number of algorithms have been introduced to search for mutually exclusive sets of mutations. These methods differ in how they score mutual exclusivity and in how they identifying the best scoring set(s) of mutations.

The first type of score for mutual exclusivity is a combinatorial score, such as the scores employed in [114, 115, 116]. For example, in the Dendrix algorithm [115], the score for a set $M$ is the difference between the number of samples with a mutation in $M$ (coverage) and the number of mutations in $M$ occurring in more than one sample (coverage overlap). The advantage of a combinatorial score is that it is easy to compute, but it was observed by [120] and others that the score is often biased towards sets with frequently mutated genes.

The second type of score for mutual exclusivity is a statistical score [111, 117, 118, 120, 121, 119]. A particularly useful statistical score for exclusivity is based on exact distribution that conditions on the observed number of mutated samples in each gene [120, 119]. For a pair of mutations, such a test is a one-sided Fisher's exact test [122, 120, 119]. For more than two genes, [120] generalized the exact test to multi-dimensional contingency tables, as presented in Chapter 2. They introduced the CoMEt algorithm that computes a generalization of Fisher's exact test for gene sets of any size using either an exact test or approximation. They showed that conditioning on the number of mutations in each gene reduces bias towards frequently mutated genes compared to combinatorial scores.

Statistical scores that condition only on mutation frequencies do not account for the variation in mutation rate among tumors. It has been observed that the number of mutations in a tumor can vary over several orders of magnitude (e.g., see [16, 17, 123]). For example, colorectal tumors with microsatellite stability have a median of 66 nonsynonymous mutations, but colorectal tumors with microsatellite

instability have a median of 777 mutations [16]. Another example is from [106], who classified a subset of TCGA endometrial cancers as ultramutated or hypermutated.

Another useful statistical test for mutual exclusivity conditions on *both* the number of mutated samples in each gene *and* the number of mutated genes in each sample [111, 121]. Since computing this exact distribution is not computationally efficient, permutation tests are used. The permutation tests, which compare observed results to a number of samples ($\sim 10^4$) drawn from a null distribution, are more tractable than the exact tests on genome-scale data, but the significance of the score is directly limited by the number of permutations. MEMo [111] computes the significance of the *coverage* (number of mutated samples) of $M$ using this permutation distribution for sets of any size $k$. MEMCover [121] computes the significance of the exclusivity of pairs to search for exclusivity within, between, and across cancer types. Both MEMo and MEMCover test sets of genes that interact in a protein-interaction network. To our knowledge, there is no method for quickly computing the significance of mutual exclusivity conditioned on both the observed number of mutations per gene and number of mutations per sample. Moreover, the permutation tests have not been applied across all sets of genes, without restricting to interacting sets.

### 3.1.1 Contributions

We introduce a weighted exact test for exclusivity that conditions on the number of mutations in each gene in $M$ while incorporating the probability that each gene is mutated in each sample. We use this model to approximate the fixed gene and sample frequency permutation test quickly and accurately by estimating the mutation probabilities from the null distribution of the permutational test. We present a formula for computing this test exactly and derive a saddlepoint approximation for arbitrarily sized groups of genes. We show that the saddlepoint approximation is both fast and accurate, and can approximate the permutational distribution. We also demonstrate that the saddlepoint approximation can be used to rapidly compute the CoMEt statistical test, which is a special case of the weighted test where the mutation probabilities for a given gene are the same in each sample.

We use the weighted exact test to identify sets of exclusive mutations in hundreds of colorectal and endometrial cancers. Cancer of these types often have the extremely high mutation rates (e.g., see [16]), which makes them difficult to analyze when conditioning only on the number of mutations per gene. However, our weighted statistical test allows us to effectively condition on the number of mutations per sample, and we identify exclusive patterns of mutations in these cancers that were missed by earlier approaches. We find that the weighted enrichment test identifies more biologically interesting sets than the exact test used by CoMEt [120]. We expect that the weighted test for mutual exclusivity will prove useful for many cancer types

where defects in DNA damage or environmental exposures, e.g., ultraviolet light, lead to very high mutation rates in some samples.

## 3.2 Methods

We observe the presence of mutations across genes in a collection of samples. For a set $M$ of genes, a sample $s$ has a *mutually exclusive* mutation if there exists one and only one gene $g \in M$ with at least one mutation in $s$. Our goal is to identify sets $M$ of genes with statistically significant number of samples with mutually exclusive mutations.

We derive a mutually exclusive mutation score $\Psi$ that incorporates the above observations with a per-gene, per-sample mutation probability. For a set $M$ of $k$ genes, the score $\Psi(M)$ computes the probability of observing at least $t$ samples with mutually exclusive mutations given the number of mutations observed in each gene and the probability of observing a mutation in each gene and each sample. We derive methods for computing this tail probability exactly, and we construct a fast and accurate saddlepoint approximation for the score.

The score $\Psi$ is a generalization of the score $\Phi$ introduced by [120]. For a set $M$ of $k$ genes, the score $\Phi(M)$ is equivalent to computing how often permuting the mutations observed in each gene across other samples results in more exclusivity. We show how our saddlepoint approximation can be used to compute accurate approximations of $\Phi(M)$ much faster than the method used by [120]. We also show how to use $\Psi$ to approximate a variant of the permutation test described above when the number of mutations per gene *and* mutations per sample is fixed, which was used by [111] and others [121] but is too prohibitive to be computed exactly (e.g., both [111, 121] sample 10,000 permuted matrices).

### 3.2.1 Weighted exact test for mutually exclusivity

Let $\{g_i\}_{i=1}^m$ be a set of genes and $\{s_j\}_{j=1}^n$ be a set of samples. For each sample, we observe the the presence of one or more mutations in each gene, and we construct the per-gene, per-sample binary mutation matrix $A \in \{0,1\}^{m \times n}$, where $A = [a_{ij}]$ with $a_{ij} = 1$ if gene $g_i$ has at least one mutation in sample $s_j$ and $a_{ij} = 0$ otherwise.

A set of genes $\{g_i\}_{i \in M}$ has *co-occurring* mutations in sample $s_j$ if every gene is mutated in that sample, i.e., $a_{ij} = 1$ for every $i \in M$. Alternatively, a set of

genes $M$ has a *mutually exclusive* mutation in sample $s_j$ if one and only one gene is mutated in that sample, i.e., there exists $\ell \in M$ such that $a_{ij} = 1$ for $i = \ell$ and $a_{ij} = 0$ otherwise. Our goal is to identify sets $\{g_i\}_{i \in M}$ of $k$ genes with a statistically significant number of samples with mutually exclusive mutations in $A$.

Let $r_i = \sum_{j=1}^{n} a_{ij}$ be the number of samples with mutations in $g_i$, let $c_j = \sum_{i=1}^{m} a_{ij}$ be the number of mutated genes in $s_j$, let $z_M$ be the number of samples with co-occurring mutations in $\{g_i\}_{i \in M}$, and let $t_M$ be the number of samples with mutually exclusive mutations in $\{g_i\}_{i \in M}$.

We derive our model following [124], who developed a two-variable weighted enrichment test for highly expressed genes. We assume that $\{X_{ij}\}_{j=1}^{n}$ is a set of mutually independent Bernoulli trials with success probabilities $P = [p_{ij}]$, i.e.,

$$\Pr(X_{ij} = \ell) = \begin{cases} p_{ij}, & \ell = 1, \\ 1 - p_{ij}, & \ell = 0, \end{cases} \tag{3.1}$$

where $p_{ij}$ is the probability that gene $g_i$ is mutated in sample $s_j$. Let $T_{M,j}$ be a random variable with $T_{M,j} = 1$ if $s_j$ has a mutually exclusive mutation in $\{g_i\}_{i \in M}$ and $T_{M,j} = 0$ otherwise. If $Y_i = \sum_{j=1}^{n} X_{ij}$ and $T_M = \sum_{j=1}^{n} T_{M,j}$, then we want to find the probability of observing at least $t_M$ samples with mutually exclusive mutations in $\{g_i\}_{i \in M}$ given that $g_i$ is mutated in $r_i$ samples, i.e.,

$$\Psi(M) = \Pr(T_M \geq t_M \mid Y = r) = \frac{\Pr(T_M \geq t_m, Y = r)}{\prod_{i=1}^{k} \Pr(Y_i = r_i)}, \tag{3.2}$$

where $Y = [Y_i]_{i \in M}$ and $r = [r_i]_{i \in M}$.

Note that, for any gene $g_i$, the conditional assumption of $Y_i = r_i$ implies that

$$\sum_{j=1}^{n} p_{ij} = \sum_{j=1}^{n} E[X_{ij}] = E\left[\sum_{j=1}^{n} X_{ij}\right] = E[Y_i] = r_i$$

by the definitions of $\{X_{ij}\}_{j=1}^{n}$ and $Y_i$.

### Computing the tail probability

Our test for exclusivity requires computing the tail probability in Eq. 3.2, which can be computationally expensive. We compute the tail probability using two different strategies: an exact test and a saddlepoint approximation.

**Exact test.** We present an exact test for computing the conditional tail probability in Eq. 3.2 for sets $M$ of any size $k$. We compute the marginal probabilities in the denominator using dynamic programming, which is a standard method for computing the Poisson-Binomial probability mass function [125]. We then use the following recursive formula for computing the joint probability $\Pr(T_M \geq t_M \mid X = r)$, where

$$\Pr(T_M \geq t_M \mid X = r) = F(t_M, r_1, \ldots, r_k, n) \tag{3.3}$$

with

$$\begin{aligned} &F(t, x_1, \ldots, x_k, j) = \\ &\sum_{\pi \in \{0,1\}^k} \prod_{i=1}^{k} q_{ij\pi_i} F(w_\pi(t), y_{\pi_1}(x_1), \ldots, y_{\pi_k}(x_k), j-1), \end{aligned} \tag{3.4}$$

where

$$q_{ij\ell} = \begin{cases} p_{ij} & \text{if } \ell = 1, \\ 0 & \text{otherwise,} \end{cases} \tag{3.5}$$

$$w_\pi(t) = \begin{cases} t-1 & \text{if } \sum_{i=1}^{k} \pi_i = 1, \\ t & \text{otherwise,} \end{cases} \tag{3.6}$$

and

$$y_\ell(x) = \begin{cases} x-1 & \text{if } \ell = 1, \\ x & \text{otherwise.} \end{cases} \tag{3.7}$$

The base cases for Eq. 3.4 are:

$$F(t, x_1, \ldots, x_k, j) = \begin{cases} 1, & t = x_1 = \cdots = x_k = j = 0, \\ 0, & \min\{t, x_1, \ldots, x_k, j\} > 0 \text{ or} \\ & t > \sum_{i=1}^{k} x_i \text{ or} \\ & \max_{i=1}^{k} x_i > n. \end{cases} \tag{3.8}$$

**Saddlepoint approximation.** We derive a saddlepoint approximation [126] for computing the conditional tail probability in Eq. 3.2. This approach is similar to [124], which use a saddlepoint approximation for a weighted enrichment test for differentially expressed genes in Gene Ontology categories. The saddlepoint approximation is specifically designed to provide a quick and accurate approximation of the tail probability. We present the key equation in (3.9).

$$\Pr(T_M \geq t_M \mid X = r) \approx 1 - \Phi(\tilde{w}) - \phi(\tilde{w}) \left( \frac{1}{\tilde{w}} - \frac{1}{\tilde{u}} \right), \tag{3.9}$$

where

- $\Phi$ and $\phi$ are the cumulative distribution and density functions, respectively, of the standard normal distribution;
- $\tilde{w}$ is defined by

$$\tilde{w} \coloneqq \sqrt{2}\,\mathrm{sgn}\,(\tilde{y}_{k+1}) \sqrt{\sum_{i \in M} \mathcal{K}_{X_i}(\hat{x}_i) - \mathcal{K}(\tilde{y}) - \tilde{y}^T\,(\hat{x} - \tilde{x})};$$

- $\tilde{u}$ is defined by

$$\tilde{u} \coloneqq 2\sinh\left(\frac{\tilde{y}_{k+1}}{2}\right) \sqrt{\frac{|\mathcal{K}''(\tilde{y})|}{\prod_{i \in M} \mathcal{K}''_{X_i}(\hat{x}_i)}};$$

- the moment generating function of $\{X_i\}_{i \in M}$ is

$$\mathcal{M}_X(\lambda) \coloneqq E[e^{\sum_{i \in M} \lambda_i X_i}];$$

- the joint cumulant generating function of $\{X_i\}_{i \in M}$ is

$$\mathcal{K}_X(\lambda) \coloneqq \log \mathcal{M}_X(\lambda);$$

- the cumulant generating function of $X_i$ is $\mathcal{K}_{X_i}(\lambda)$;
- the gradient vector and the Hessian matrix of $\mathcal{K}_X(\lambda)$ are $\mathcal{K}'_X(\lambda)$ and $\mathcal{K}''_X(\lambda)$, respectively;
- $\tilde{x} \coloneqq \left(r_1, \ldots, r_k, z - \frac{1}{2}\right)$ and $\tilde{y} = (\tilde{y}_1, \ldots, \tilde{y}_{k+1})$, where $\tilde{y}$ is the solution to $\mathcal{K}'(\tilde{y}) = \tilde{x}$, with $\tilde{y}_{k+1} \neq 0$ for (3.9) to be defined; and
- $\hat{x} \coloneqq (\hat{x}_1, \ldots, \hat{x}_k, 0)$, where $\hat{x}$ is the solution to $\mathcal{K}'_{X_i}(\hat{x}_i) = r_i$.

### 3.2.2 Computing per-gene, per-sample mutation probabilities

Our model requires a matrix $P$ of per-gene, per-sample background mutation probabilities with $\Pr(X_{ij} = 1) = p_{ij}$. We estimate $P$ from the permutational distribution where both the row and column sums of $A$ are fixed. This permutational distribution has been used to examine mutual exclusivity in multiple previous studies [111, 121]. Both [111] and [121] sample $N = 10^4$ matrices to approximate tail probabilities from this distribution. By estimating $P$ from the permutational distribution, we can use the weighted test to approximate the tail probability of the permutational test with greater significance.

To estimate the probability an individual gene is mutated in a given sample, we generate an empirical distribution of $N$ permuted matrices with fixed row and

column sums. We then compute $p_{ij}$ as the proportion of permuted matrices in which gene $g_i$ is mutated in sample $s_j$. In the case when a gene $g_i$ is never observed to be mutated in sample $s_j$, we set $p_{ij} = \frac{1}{2N}$ so that $p_{ij} > 0$ for all $i, j$. In other words,

$$\tilde{P}_N = [\tilde{p}_{ij}] = \frac{1}{N} \sum_{l=1}^{N} \Pi_\ell A,$$

where $\Pi_\ell A$ is a random permutation of the mutation matrix $A$ that preserves row and column sums of $A$, and

$$P_N = [p_{ij}] = \begin{cases} \tilde{p}_{ij}, & \tilde{p}_{ij} > 0, \\ \frac{1}{2N}, & \text{otherwise.} \end{cases}$$

### 3.2.3 Comparison to other scores

In the case when $p_{i1} = \cdots = p_{in}$ for each $i$, the weighted score $\Psi(M)$ is equivalent to $\Phi(M)$, the generalization of Fisher's exact test used in CoMEt [120]. We refer to this special case as the unweighted test.

Moreover, the weighted score $\Psi(M)$ with mutation probabilities $P_N$ approximates the permutational $p$-value $H(M)$ conditioned on fixed row and columns sums, where

$$H(M) = \frac{1}{N} \Big| \{ B \in \{\Pi_\ell A\}_{\ell=1}^{N} \, : \, \tau_M^{(B)} \geq t_M \} \Big|,$$

where $\tau_M^{(B)}$ is the number of samples with mutually exclusive mutations in $M$ according to mutation matrix $B$ and $\Pi_\ell A$ is a permutation of the mutation matrix $A$ that preserves row and column sums of $A$; see [111, 121].

### 3.2.4 Searching for sets of mutually exclusive mutations

Our goal is to identify sets of mutations with significant $\Psi(M)$. There has been much work on methods for optimizing scores for mutually exclusive mutations, including Markov chain Monte Carlo methods [115, 120], integer linear programs [116, 127], greedy algorithms [119], and others, but this is beyond the scope of this work. Instead, we search for mutually exclusive sets $M$ of $k$ genes by enumerating all combinations of genes that satisfy the following basic criteria.

1. The number of samples with exclusive mutations must exceed the number of samples with co-occurring mutations, i.e., $t_M > z_M$.

2. Each gene $g_i \in M$ must be exclusively mutated in at least one sample.

We use the Benjamini-Hochberg procedure [128] to control the False Discovery Rate (FDR). In practice, we limit $k$ and the number of genes in the datasets enough that we can enumerate and test all combinations in a reasonable amount of time.

### 3.2.5 Implementation

We implemented CoMEt in C. We implemented the saddlepoint approximation in Python with the SciPy library, which is written in Python, C, C++, and Fortran. For the permutational test, we used the BiRewire package [77], which is implemented in R. Our implementation of the saddlepoint approximation is available on `http://compbio.cs.brown.edu/projects/weighted-exact-test/`.

## 3.3 Results

We compare the results of the weighted test to both the permutational test and CoMEt on real data, and we apply the weighted test to discover mutually exclusive sets of mutations in thyroid, colorectal, and endometrial cancers. The rest of this section is organized as follows. In Section B.2, we describe the data used in our experiments. In Section 3.3.2, we compare the exact and saddlepoint methods for computing the unweighted test, and we find that the saddlepoint approximation is both accurate and fast. In Section 3.3.3, we compare the results of the weighted exact test and saddlepoint approximation with the permutational test, and we show that the weighted test is an accurate approximation of the permutational test. In Section 3.3.4, we show that the weighted test is still an accurate approximation of the permutational test, even with mutation probabilities $P$ estimated from orders of magnitude fewer permuted matrices. Finally, in Sections 3.3.5 and 3.3.6, we present the results of the weighted test on thyroid, colorectal, and endometrial cancers.

### 3.3.1 Data

We analyzed non-synonymous single nucleotide variants (SNVs) and small indels in the 224 colorectal (COADREAD), 402 papillary thyroid carcinoma (THCA), and 248 uterine corpus endometrial carcinoma (UCEC) samples from The Cancer Genome Atlas (TCGA) [105, 106, 129]. We analyzed the mutations in the COADREAD

and UCEC samples from the TCGA Pan-Cancer project [130] by downloading the mutations in Mutation Annotation Format (MAF) from Synapse.[1] We downloaded the mutations in THCA from Firehose[2]. We also downloaded lists of hypermutator samples for COADREAD and UCEC. We created a list of 35 hypermutator samples in COADREAD listed in [105] in their Supplementary Table 3, and 82 hypermutator samples in UCEC listed by [106] as samples labeled "POLE OR MSI" their Supplementary Datafile S1.1. We restrict our analysis to genes mutated in at least 20, 5, and 30 samples in the COADREAD, THCA, and UCEC datasets, leaving 76, 30, and 62 genes in each dataset, respectively.

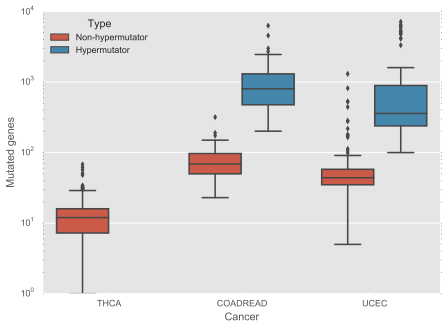



**Figure 3.1:** Boxplots of the number of mutated genes per sample in the THCA, COADREAD, and UCEC datasets. The COADREAD and UCEC samples are further broken down into hypermutators (blue) and non-hypermutators (red).

We show the distribution of the number of mutations per sample in Figure 3.1. In general, COADREAD samples have the most mutations (median: 79), with COADREAD hypermutators mutated in an at least an order of magnitude more samples than non-hypermutators (median for hypermutators: 800; median for non-hypermutators: 69). THCA samples have the fewest mutated genes (median: 12) with no hypermutators, while UCEC samples have more mutations (median: 58) with UCEC hypermutators mutated in approximately an order of magnitude more samples than non-hypermutators (median hypermutators: 358.5; median non-hypermutators: 44).

For each dataset, we estimated the weights $P_N$ using the permutational procedure described Section 3.2.2 using $N = 10^3$ permutations. We show weights for each dataset in Figure 3.2.

[1] `https://doi.org/10.7303/syn1710680.4`

[2] `http://gdac.broadinstitute.org/runs/stddata__2016_01_28/data/THCA/20160128/gdac.broadinstitute.org_THCA.Mutation_Packager_Calls.Level_3.2016012800.0.0.tar.gz`

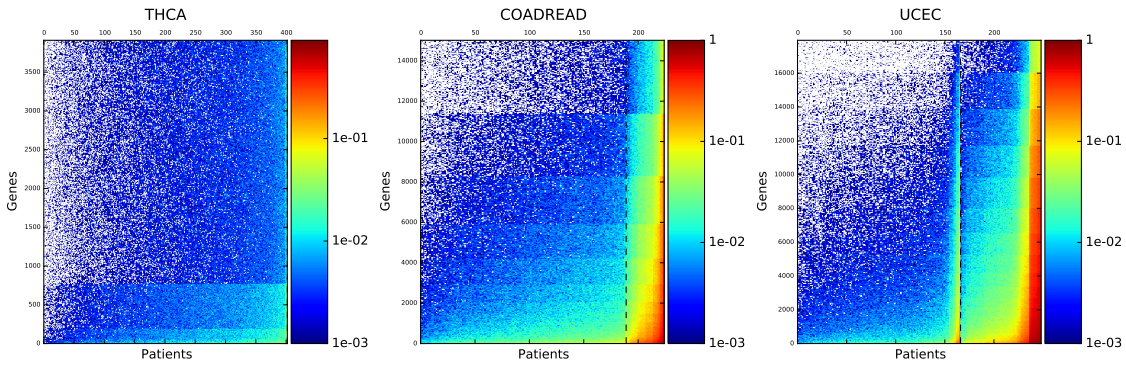



**Figure 3.2:** The weights $P_N$ estimated by sampling $N = 10^3$ permuted matrices on the THCA, COADREAD, and UCEC datasets. Samples ($x$-axis) are sorted by the number of mutated genes in increasing order from left to right, with hypermutators (right) separated from non-hypermutators (left) with a dashed line in COADREAD and UCEC. Genes ($y$-axis) are sorted by the number of mutated samples in increasing order from top to bottom.

| Method | Minimum | Median | Maximum | Total |
|---|---|---|---|---|
| Exact test (CoMEt) | 6.91E-6 | 7.21E-3 | 2.32E+0 | 2.21E+3 |
| Saddlepoint | 3.26E-3 | 4.64E-2 | 5.21E-2 | 5.25E+2 |

**Table 3.1:** Runtimes (seconds) for 111,495 triples from the THCA, COADREAD, and UCEC datasets for the unweighted test.

### 3.3.2 Comparison of methods for computing the unweighted test on real data

First, we investigated the accuracy and speed of the saddlepoint approximation of the unweighted test (i.e., the generalization of Fisher's exact test). We enumerated triples according to the procedure described in Section 3.2.4 in the THCA, COADREAD, and UCEC datasets, and computed the exact test – using the CoMEt software from [120] – and the saddlepoint approximation, assigning $p_{i1} = p_{i2} = \cdots = p_{in}$ for all $i \in M$. We show a comparison of the $p$-values and runtimes given by the two methods in Figure 3.3. On these datasets, the saddlepoint approximation is an extremely accurate approximation of the exact test ($\rho^2 = 0.9999$). Additionally, while the median runtimes of the two tests are similar, the exact test is much slower for for sets with co-occurring mutations while the saddlepoint is largely unaffected; see Table 3.1. We expect the discrepancy between runtimes to grow for sets of larger sizes.

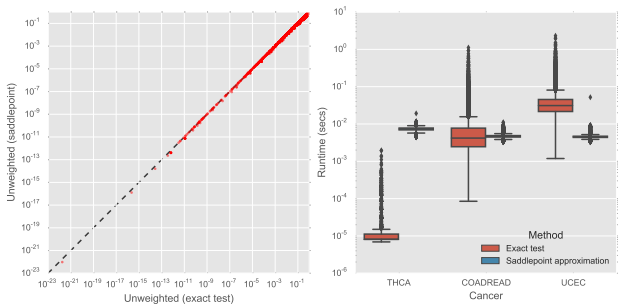



**Figure 3.3:** Comparison of the exact test and saddlepoint approximation for computing the unweighted $p$-value of the enumerated triples in THCA, COADREAD, and UCEC. (left) Scatter plot comparing the $p$-values given by the exact test ($x$-axis) versus the saddlepoint approximation ($y$-axis). (right) Distribution of the runtime (in seconds) required to compute each method on a single triple.

### 3.3.3 Comparison of methods for computing the weighted test on real data

Next, we compared the results of methods for computing the weighted test with the permutational test on pairs of genes from the THCA, COADREAD, and UCEC datasets. We chose pairs instead of triples because of the prohibitive cost of computing the exact and permutational tests. We considered the exact and saddlepoint approximations of the weighted test, the permutational test with $N = 10^4$ permutations, and the CoMEt test as a control.

In Table 3.2, we see that the results of the weighted exact test and the saddlepoint approximation are strongly correlated with the permutational test. The results of the CoMEt test are more weakly correlated with the permutational test, showing that conditioning on the number of mutations in each sample changes the distribution of mutually exclusive mutations. This discrepancy continues when we restrict ourselves to the mode of the distributions ($p \geq 10^{-4}$), where the permutational test has more power.

The saddlepoint approximation of the weighted test is highly correlated with the exact test, with a Pearson's correlation coefficient for the exact and saddlepoint $p$-values of 0.996163 for all $p$. On the tail of the distribution for $p < 10^{-4}$, where the saddlepoint approximation performs best, the correlation increases to 0.999865.

Also, in Table 3.3, we see that the runtime of the weighted exact test varies widely because pairs with co-occurring mutations require more computation, but the runtime of the weighted saddlepoint approximation is more consistent. As a result, testing all pairs with the exact test requires nearly 2 hours, but testing the same pairs with the saddlepoint approximation requires slightly over a minute.

| Pairs | CoMEt | Weighted Exact | Weighted Saddlepoint |
|---|---|---|---|
| All | 0.722329 | 0.999588 | 0.995926 |
| $H(M) \geq 10^{-4}$ | 0.683213 | 0.999628 | 0.995412 |

**Table 3.2:** Pearson's correlation coefficient $\rho^2$ of the $p$-values of pairs of genes from the THCA, COADREAD, and UCEC datasets of different tests with the permutational test $H(M)$. The correlations were computed for two sets of pairs of genes: $p$-values for 5,163 pairs (all) and 4,926 pairs with permutational $H(M) \geq 10^{-4}$.

| Method | Minimum | Median | Maximum | Total |
|---|---|---|---|---|
| Exact | 7.31E-2 | 8.89E-1 | 3.86E+2 | 6.72E+3 |
| Saddlepoint | 2.53E-3 | 4.63E-2 | 1.05E+0 | 6.63E+1 |

**Table 3.3:** Runtimes in seconds of methods for computing the weighted test for 5,163 pairs from the THCA, COADREAD, and UCEC datasets.

### 3.3.4 Approximating the permutational test with the weighted test

We compared the saddlepoint approximation of the weighted test to the permutational test with $N = 10^6$ permutations using gene triples from the COADREAD dataset, again using the unweighted exact test (CoMEt) as a control. We computed the saddlepoint approximation using $P_N$ for $N = 10^3$, three orders of magnitude fewer permutations than we used for the permutational test. The $p$-values predicted by CoMEt and the permutational test are weakly correlated in the tail ($\rho^2 = 0.67$ for permutational $p < 0.001$; see Figure 3.5a). In contrast, the saddlepoint approximation of the weighted test does provide an accurate approximation of the permutational test. The saddlepoint approximation and permutational $p$-values are highly correlated in the tail ($\rho^2 = 0.986$ for permutational $p$-values $p < 0.001$; see Figure 3.5b). Moreover, the saddlepoint approximation of the weighted test is an accurate estimate of the permutation test to within one or more digits for most gene triples and within a factor of two for all triples. Furthermore, despite the much lower number of permutations used to generate $P_N$, the saddlepoint approximation can make more significant predictions than the permutational test, and it is orders of magnitudes faster.

### 3.3.5 Mutually exclusive mutations in thyroid carcinomas

We computed the $p$-values for all triples of genes that were each mutated in at least 5 of the 402 thyroid carcinomas in the THCA dataset. The weighted test identifies 48 triples with significantly exclusive mutations (FDR $< 0.001$), while the CoMEt

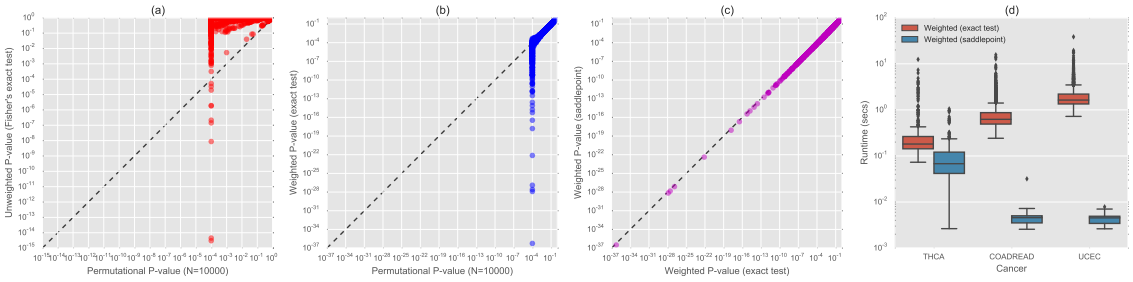



**Figure 3.4:** Comparison of $p$-values and runtimes of different statistical tests on COADREAD, THCA, and UCEC pairs. (a-b) Scatter plots comparing the permutational test with $10^6$ permutations against (a) Fisher's exact test (CoMEt) and (b) the weighted exact test. (c) The weighted exact test plotted against the saddlepoint approximation of the weighted test. (d) Boxplots of the runtimes for computing the weighted exact test (red) and the saddlepoint approximation (blue) for each pair of genes in the datasets.

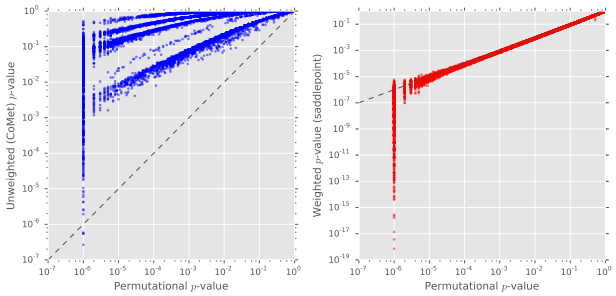



**Figure 3.5:** Comparison of the CoMEt (unweighted) test (left, $y$-axis) and the saddlepoint approximation of the weighted (right, $y$-axis) test to the permutational test ($x$-axis) on triples from the COADREAD dataset. The saddlepoint approximation uses mutation probabilities $P_N$ estimated from $N = 10^3$ permutations, while the permutational $p$-values were computed with $N = 10^6$ permutations.

exact test identifies 38 triples (FDR $< 0.001$).

The top 25 ranked triples by both tests are identical, which is not surprising since THCA samples have low mutation rates compared to most cancer types (see [16] and Figure 3.1). In addition, the $p$-values for the top ranked tripes are all within a few orders of magnitude, demonstrating that the two tests are very similar on this dataset.

The top triples include many known thyroid cancer genes. The top five triples include seven genes, five of which are well-known cancer genes with known roles in thyroid cancer [107]: *BRAF*, *HRAS*, *NRAS*, *EIF1AX*, and *ATM*. The other two genes are *BDP1* and *TG*, both of which may play a role in cancer. [131] describe a role for *BDP1* in AKT signaling, which was also noted in TCGA thyroid publication[107]), although *BDP1* is greater than 11,000 amino acids in length so may accumulate many passenger mutations. *TG* is the thyroglobulin gene, and is used as a tumor

| CoMEt rank | Weighted Rank | Triple | $\Phi$(M) | $\Psi$(M) |
|---|---|---|---|---|
| 1 | 1 | *BRAF*, *HRAS*, *NRAS* | 1.79E-22 | 1.43E-27 |
| 2 | 2 | *BRAF*, *EIF1AX*, *NRAS* | 2.27E-16 | 2.01E-20 |
| 3 | 3 | ***BDP1***, *BRAF*, *NRAS* | 2.70E-14 | 4.72E-18 |
| 4 | 4 | *BRAF*, *NRAS*, *TG* | 3.51E-13 | 4.57E-17 |
| 5 | 5 | ***ATM***, *BRAF*, *NRAS* | 6.58E-13 | 1.63E-16 |

**Table 3.4:** Five most significant triples identified by CoMEt $p$-value $\Phi(M)$ and the weighted test $p$-value $\Psi(M)$ on the THCA dataset. Genes in bold are among the 600 longest genes (at least 9560 amino acids).

marker in papillary thyroid carcinoma[3], which is the same subtype of thyroid cancer analyzed in TCGA.

### 3.3.6 Mutually exclusive mutations in colorectal cancers and endometrial carcinomas

We expect that the difference between the CoMEt exact test and weighted test would be more pronounced on cancer types with higher and highly variable mutation rates. Thus, we computed $p$-values on triples of genes from colorectal cancers (COADREAD) and endometrial carcinomas (UCEC). We find that the weighted test predicts more biologically interesting triples than the exact test used by CoMEt. The weighted test identifies 5,286 and 6,790 triples (many of which overlap) with significantly mutually exclusive mutations (FDR $< 0.001$) in the 224 COADREAD and 248 UCEC samples, respectively. In contrast, CoMEt computes 4 and 130 triples with significantly mutually exclusive mutations (FDR $< 0.001$).

Compared to the CoMEt results, the highest ranked triples by the weighted test include fewer long genes that tend to accumulate random, passenger mutations – especially in samples with high mutation rates (Tables 3.5 and 3.6).

On COADREAD, the weighted test identifies ten different genes in the five most significant triples (Table 3.5). Nine of these genes are well-known cancer genes – *BRAF*, *KRAS*, *NRAS*, *ACV2RA*, *PIK3CA*, *TP53*, *ATM*, *TGFBR2*, and *ARID1A* – while the tenth gene (*ABCA12*) is known to have an association with colorectal cancers [132]. The CoMEt results are similar – two of the top five triples identified by the weighted test are in the top five triples identified by CoMEt – but CoMEt does not identify *ARID1A*, *TGFBR2*, *KRAS*, or *NRAS*. Further, the three additional genes identified by CoMEt – *APC*, *FAT2*, and *WDFY3* – are all in the top 600 longest

[3] `http://www.mayomedicallaboratories.com/test-catalog/Clinical+and+Interpretive/62800`

| CoMEt rank | Weighted Rank | Triple | Φ(M) | Ψ(M) | Hypermutator mutations |
|---|---|---|---|---|---|
| 1 | 2 | *ACVR2A*, *PIK3CA*, *TP53* | 2.65E-07 | 3.59E-18 | 31 |
| 2 | 31 | ***APC***, *BRAF*, *PRDM2* | 5.44E-07 | 2.74E-13 | 33 |
| 2 | 33 | ***APC***, *BRAF*, ***WDFY3*** | 5.44E-07 | 2.84E-13 | 32 |
| 4 | 3 | ***ATM***, *PIK3CA*, *TP53* | 5.87E-07 | 1.37E-17 | 24 |
| 5 | 70 | ***APC***, *BRAF*, ***FAT2*** | 1.93E-06 | 6.28E-12 | 35 |
| 6 | 1 | *BRAF*, *KRAS*, *NRAS* | 2.50E-06 | 7.05E-19 | 26 |
| 1 | 2 | *ACVR2A*, *PIK3CA*, *TP53* | 2.65E-07 | 3.59E-18 | 31 |
| 4 | 3 | ***ATM***, *PIK3CA*, *TP53* | 5.87E-07 | 1.37E-17 | 24 |
| 9 | 4 | *ABCA12*, *TGFBR2*, *TP53* | 4.29E-06 | 1.65E-16 | 28 |
| 10 | 5 | *ARID1A*, *TGFBR2*, *TP53* | 5.89E-06 | 2.26E-16 | 29 |

**Table 3.5:** Five most significant triples identified by CoMEt (upper 5) and the weighted test (lower 5) on the COADREAD dataset. Genes in bold are among 600 longest genes (at least 9560 amino acids).

| CoMEt rank | Weighted Rank | Triple | Φ(M) | Ψ(M) | Hypermutator mutations |
|---|---|---|---|---|---|
| 1 | 20 | ***CACNA1E***, *PTEN*, *TP53* | 3.11E-12 | 4.08E-30 | 77 |
| 2 | 21 | ***LAMA2***, *PTEN*, *TP53* | 4.13E-12 | 5.76E-30 | 77 |
| 3 | 28 | *PTEN*, ***RYR2***, *TP53* | 4.60E-12 | 3.29E-29 | 78 |
| 4 | 24 | ***NBEA***, *PTEN*, *TP53* | 8.40E-12 | 2.21E-29 | 76 |
| 5 | 35 | ***FAT4***, *PTEN*, *TP53* | 1.23E-11 | 2.04E-28 | 75 |
| 22 | 1 | *CTNNB1*, *RPL22*, *TP53* | 2.11E-10 | 1.34E-41 | 47 |
| 44 | 2 | *CTNNB1*, *KRAS*, *TP53* | 3.05E-09 | 2.21E-37 | 48 |
| 55 | 3 | *CTNNB1*, *MLL4*, *TP53* | 4.26E-08 | 9.99E-36 | 42 |
| 57 | 4 | *CTCF*, *CTNNB1*, *TP53* | 4.84E-08 | 5.94E-35 | 43 |
| 60 | 5 | *CTNNB1*, ***RYR1***, *TP53* | 1.14E-07 | 1.66E-34 | 40 |

**Table 3.6:** Five most significant triples identified by CoMEt (top 5) and the weighted test (bottom 5) on the UCEC dataset. Notation as in Table 3.5.

genes in the human genome (at least 9560 amino acids). While mutations in *APC* are well-known to play a role in colorectal cancers, there is currently little evidence for the roles of *FAT2* or *WDFY3* in cancer, and it is likely that these long genes have accumulated many passenger mutations, particularly in hypermutated samples. Also of note is the fact that the number of hypermutator samples that contain mutations in the top triples from the weighted test are not appreciably different from the number of hypermutator samples that contain mutations in the top triples from the CoMEt test. This demonstrates that the weighted test is not systematically excluding hypermutator samples from consideration, but rather weighting the contribution of these samples appropriately in evaluating the significance of mutual exclusivity.

On UCEC, the differences between the weighted test and CoMEt are even more pronounced. The weighted test identifies seven genes in the top five most significant triples (Table 3.6). These include six genes with known roles in cancer – *CTTNB1*, *TP53*, *RPL22*, *KRAS*, *CTCF* and *MLL4* – with only one gene, *RYR1*, with likely spurious mutations. In contrast, the top five triples ranked by CoMEt include *PTEN* and *TP53* – two well-known cancer genes – but also five genes with no known role in cancer that are all longer than 11,000 amino acids: *CACNA1E*, *LAMA2*, *RYR2*, *NBEA*, and *FAT4*. Further, none of the top five triples identified by the weighted test are in the top twenty CoMEt triples. Finally, the CoMEt triples include many more mutations in hypermutator samples (ranging from mutations in 75 to 78 of the 81 hypermutators, versus 40 to 47 for the weighted test triples). This further demonstrates how the results of the CoMEt test are skewed by hypermutator samples,

while the weighted test incorporates the contribution of these samples appropriately in evaluating the significance of mutual exclusivity.

## 3.4 Discussion

We introduce a weighted exact test for the mutual exclusivity of mutations in cancer. We use this statistical test to approximate the permutation test for exclusivity where the number of mutations in each gene and each sample are fixed. To do so, we estimate per-gene, per-sample mutation probabilities directly from the permutational distribution. We derive an exact test and a saddlepoint approximation for computing the tail probability using sets of any size, and we demonstrate the accuracy and efficiency of the saddlepoint approximation on genome-scale mutation datasets. Together, these contributions allow us to overcome the significant computational challenge of finding highly significant sets of mutations according to the permutational distribution.

We then demonstrate the weighted test on three datasets with hundreds of samples from TCGA, including colorectal and endometrial cancers that have high variability in the number of mutations per sample. The weighted test identifies sets of mutually exclusive mutations including known cancer genes in each dataset, and its results include many fewer long genes and mutations in hypermutator samples than the results of the unweighted exact test from CoMEt [120].

There are several avenues for improving analysis with the weighted test. The input to the weighted test is a binary mutation matrix, and while we restricted our study to non-synonymous SNVs and indels, we may want to incorporate other types of aberrations, such as copy number aberrations and gene fusions, into the input to our test. We searched for mutually exclusive mutations by enumerating sets containing the most mutated genes, but the weighted test could easily be used in existing algorithms for optimizing mutual exclusivity scores (e.g., the MCMC from [115] or the greedy approach from [119]) or for searching for multiple sets simultaneously (e.g., from [120]). We estimated the per-gene, per-sample mutation probability weights directly from the permutational distribution, but we also anticipate alternative methods for setting the weights that incorporate different gene or sample attributes, such as gene length, to further reduce the number of false positives.

The weighted exact test may be of broader interest beyond searching for mutually exclusive mutations, both in other areas of computational biology and other disciplines. For example, statistical tests of "presence-absence" matrices with fixed row and column sums are a common tool in ecology for looking at species-associations,

but can be computationally prohibitive (e.g., [133]). The weighted exact test presented here may offer a fast, alternative approach for computing the significance of associations within a reasonable amount of accuracy.

# Chapter 4

# Searching for Significantly Mutated Subnetworks

In this chapter, we present the HotNet2 algorithm for identifying significant clusters of mutations on a PPI network. Protein-protein interaction (PPI) networks are often used to identify groups of functionally related genes, such as protein complexes or pathways. PPI networks are often "small world" graphs, meaning that most nodes remain within a few hops of every other node, even when the graph has thousands of nodes. This complicates the problem of identifying significant clusters of mutations on the network, as most genes are within a few hops of a frequently mutated gene.

We apply HotNet2 to The Cancer Genome Atlas's Pan-Cancer dataset, which includes 3,271 tumor samples from twelve cancer types. We demonstrate that HotNet2 identifies subnetworks overlapping well-known cancer pathways and complexes, as well as subnetworks overlapping complexes with potentially novel roles in cancer. We also demonstrate that HotNet2 identifies potential cancer genes missed by single gene tests, and outperforms related methods on both simulated and real mutation data.

Most of the completed work in this chapter is taken from [51], on which I am a co-first author. My major contributions to the paper were in developing the HotNet2 algorithm (with Benjamin Raphael, Fabio Vandin, and Hsin-Ta Wu) and performing many of the experiments on real and simulated data. I applied HotNet2 as part of the TCGA Pan-Cancer subtypes [134] and papillary kidney cancer [12] projects, in a study of somatic and germline mutations in ovarian cancer [135], and contributed to the HotNet2 analysis in the TCGA stomach cancer project [55]. I also note that HotNet2 was applied as part of the TCGA adrenocortical carcinoma [4] and thyroid cancer [13] projects.

## 4.1 Background and related work

Recent whole-genome and whole-exome sequencing studies have provided an ever-expanding survey of somatic aberrations in cancer, and have identified multiple new cancer genes [8, 6, 10, 5, 9, 136, 2, 11]. At the same time, these studies demonstrated that most cancers exhibit extensive mutational heterogeneity with few significantly mutated genes and many genes mutated in a small number of samples [16, 22]. This "long tail" phenomenon complicates efforts to identify cancer genes by statistical tests of recurrence, as rarely mutated cancer genes may be indistinguishable from genes containing only passenger mutations. Even recent TCGA Pan-Cancer studies [65, 52, 18, 34] have limited power to characterize genes in the long tail leaving an incomplete picture of the functional, somatic mutations in these samples.

A prominent explanation for the mutational heterogeneity observed in cancer is the fact that genes act together in various signaling/regulatory pathways and protein complexes [16, 18]. Clustering of mutations on known pathways is illustrated in many cancer sequencing papers [8, 6, 9, 11], but typically without a measure of statistical significance. While statistical tests of enrichment in known pathways or gene sets exist, such tests do not reveal novel pathways, have limited power to evaluate crosstalk between known pathways, and generally ignore the topology of interactions between genes.

We introduce a novel and complementary approach to identify pathways and protein complexes perturbed by somatic aberrations. This approach combines: (1) a new algorithm, HotNet2, for identification of mutated subnetworks in a genome-scale interaction network; (2) a large TCGA Pan-Cancer dataset of somatic single nucleotide variants, small indels, and copy number aberrations measured in 3,281 samples from 12 cancer types [52]. HotNet2 uses a directed heat diffusion model to simultaneously assess both the significance of mutations in individual proteins and the local topology of interactions among proteins, overcoming limitations of pathway-based enrichment statistics and earlier network approaches.

Our TCGA Pan-Cancer HotNet2 analysis identifies 14 significantly mutated subnetworks that encompass classic cancer signaling pathways, pathways and complexes with more recently characterized roles in cancer, and protein complexes and groups of interacting proteins with less characterized roles in cancer such as the cohesin and condensin complexes. These latter two subnetworks – as well many of the genes in all subnetworks – are rarely mutated in each cancer type, and thus revealed only by the Pan-Cancer network analysis. Many of the rarely mutated genes in the subnetworks have documented physical interactions with well-characterized cancer genes and/or mutational patterns (e.g. clustering in protein sequence/structure or an excess of inactivating mutations) that lend additional support for their role in cancer. Co-occurrence of mutations across these subnetworks supports the hypothesis that

many of the subnetworks correspond to distinct biological functions.

In comparison to single-gene tests of significance, our TCGA Pan-Cancer HotNet2 analysis delves deeper into the long tail of rarely mutated genes and also assembles combinations of individual genes into a relatively small number of interacting networks. The mutational landscape of cancer has been proposed to consist of "mountains" of frequently mutated genes and "hills" of less frequently mutated genes [16]. Our Pan-Cancer network approach provides a richer annotation of this landscape, grouping individual peaks and mountains into mountain ranges and their associated foothills, further enabling diagnostic and therapeutic approaches in cancer care.

## 4.2 Results

### 4.2.1 HotNet2 identifies significantly mutated subnetworks

We assembled a TCGA Pan-Cancer dataset of exome sequencing, array copy number, and RNA-seq data from 3,281 samples from 12 cancer types, analyzing single nucleotide variants (SNVs), small indels, and copy number aberrations (CNAs) in 19,424 transcripts (Figure 4.1a and 4.2). After removing hypermutated samples and genes with low expression in all tumor types (Online Methods), the dataset contained 11,565 mutated genes in 3110 tumors. We observed that the number of samples with a mutation in a gene varied over three orders of magnitude, from 1 to 1291 mutated samples (Figure 4.1b). Moreover, we discovered that this broad spectrum of mutational frequencies – from common to extremely rare mutations – posed a challenge for the identification of significantly mutated subnetworks. Specifically, our goal is to identify subnetworks according to both the frequency of somatic mutations in individual genes/proteins and the topology of the interactions between them. However, the presence of highly mutated and highly connected genes like TP53 presents difficulties for existing algorithms that attempt to achieve this goal; e.g. the HotNet algorithm [34, 115] that was used for cancer network analysis in TCGA and other studies [10, 5, 11, 137], or related network propagation approaches [138]. In the heat diffusion model used in HotNet genes like TP53 are extremely "hot" nodes and propagate this heat to their neighboring nodes. The resulting "star subnetworks" centered on the hot node (Figure 4.4; Section 4.4) contain many neighboring genes that are not mutated at appreciable frequency and are of limited biological interest.

We introduce the HotNet2 (HotNet diffusion oriented subnetworks) algorithm to address the problem of finding significantly mutated subnetworks on large, broad mutation frequency spectrum datasets like Pan-Cancer (Figures 4.1c and 4.3). Hot-

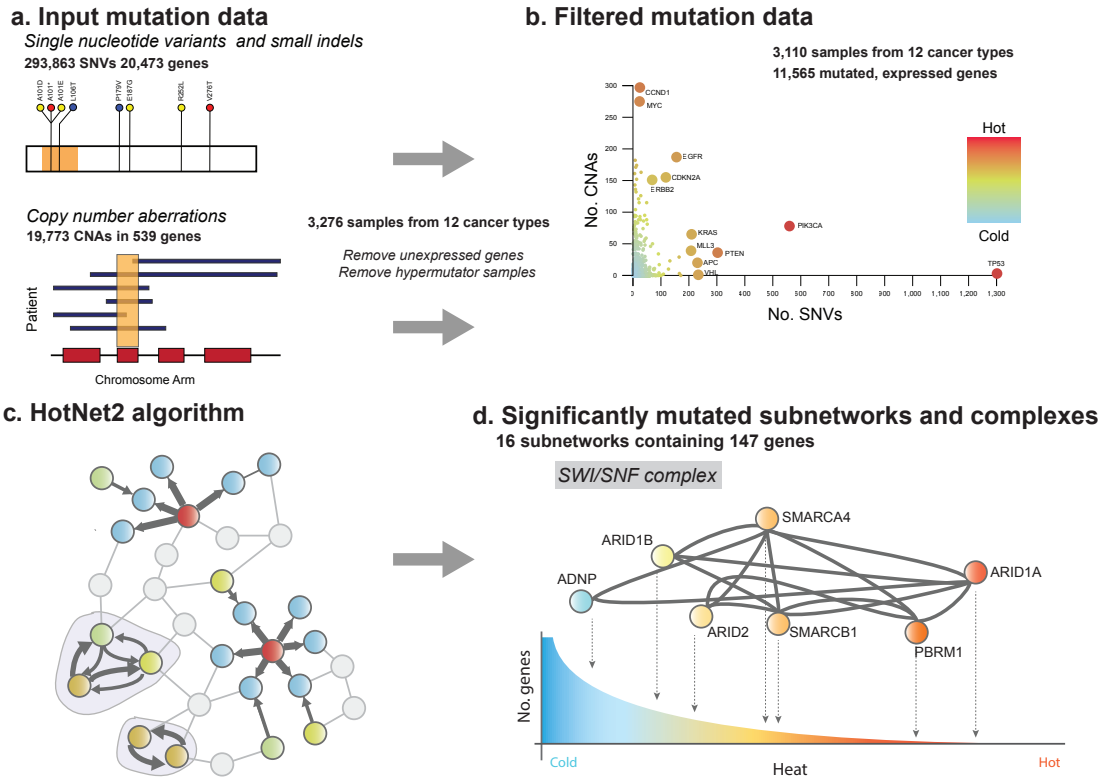



**Figure 4.1:** HotNet2 Pan-Cancer analysis **(a)** The Pan-Cancer mutation data combines SNVs (nsSNVs and small indels) and CNAs (amplifications and deletions) in 19,459 genes in 3,281 samples. The number of samples with SNVs/CNAs is shown for each gene, with points colored by the total. **(b)** Removing hypermutator samples and genes with few RNA-Seq reads in all tumor types leaves 11,565 genes in 3,110 samples for analysis with a wide range in the number of samples having an SNV (x-axis) or CNA (y-axis) in these genes. **(c)** HotNet2 finds significantly mutated subnetworks using a diffusion process on a protein-protein interaction network. Each node (protein) is assigned a score (heat) according to the frequency/significance of SNVs or CNAs in the corresponding gene. Heat diffuses across edges of network. Subnetworks containing nodes that both send and receive a significant amount of heat (outlined) are reported. **(d)** Subnetworks identified by HotNet2 include genes with wide range of heat scores, including both frequently mutated, known cancer genes (hot genes) and rarely mutated genes (cold genes) that are implicated due to their interactions with other cancer types. Thus, HotNet2 delves into long tail of rarely mutated genes by analysis of combinations of interacting genes.

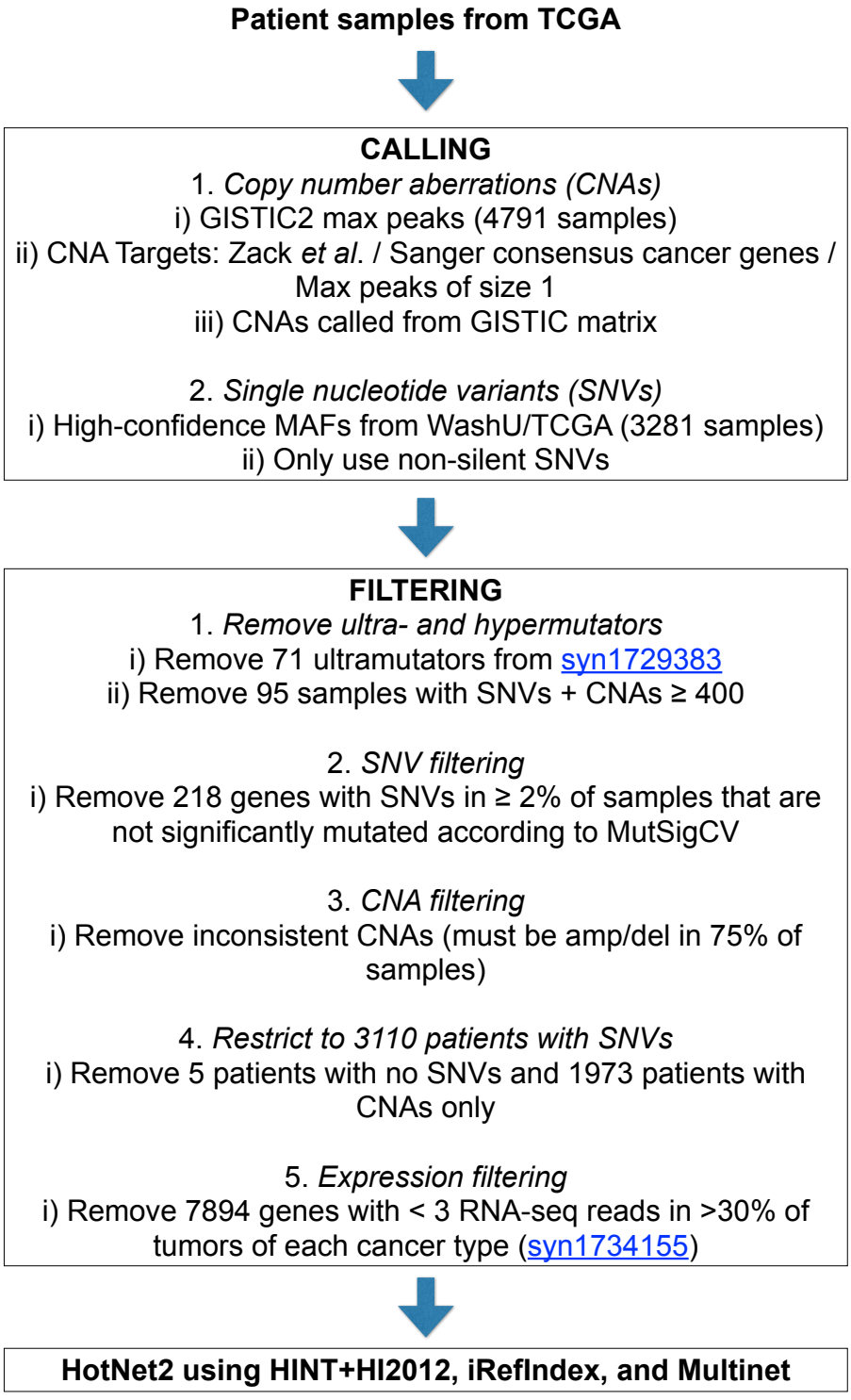



**Figure 4.2:** HotNet2 mutation data processing.

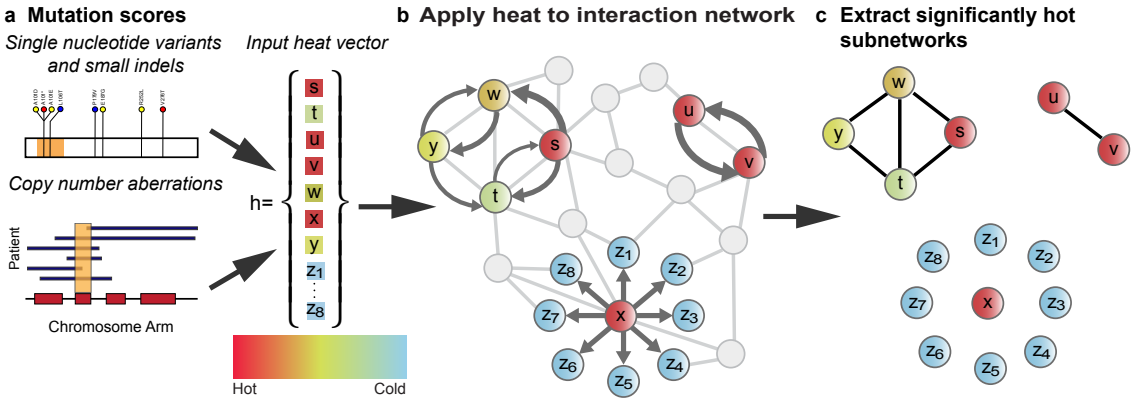



**Figure 4.3:** Overview of HotNet2 algorithm and Pan-Cancer analysis. HotNet2 assigns heat to each gene (node) in an interaction network according to a gene score encoding the frequency and/or predicted functional impact of mutations in the gene. This heat spreads to neighboring nodes using an insulated heat diffusion process. At the equilibrium heat distribution, the network is partitioned into subnetworks according to the amount and direction of heat exchange between pairs of nodes. Thus, the partition depends on both the individual genes scores and the local topology of protein interactions. The statistical significance ($p$-value and FDR) for the resulting subnetworks is computed using the same procedure on random data. In our TCGA Pan-Cancer analysis, gene scores are computed according to single nucleotide variants, small indels, and splice site mutations (from exome sequencing data), copy number aberrations (from SNP array data), and gene expression (from RNA-seq data).

Net2 uses a modified diffusion process and considers the source, or directionality, of heat flow in the identification of subnetworks (Figure 4.5). This approach reduces the artifact of star subnetworks by more than 80%, reducing the false positive rate and enabling the identification of more subtle subnetworks with rare mutations of high biological relevance (see Section 4.4). We compare HotNet2 to other algorithms (Section 4.4.3), and find that HotNet2 has higher sensitivity and specificity on both real and simulated data.

We performed HotNet2 analysis using two approaches to assign heat to individual genes according to recurrence [17], and using three different interaction networks [139, 140, 141, 142] with varying numbers of interactions (Section 4.4). HotNet2 identified a significant number of subnetworks ($P < 0.01$, Supplementary Tables 1-2) for each of the two gene scores and three networks. We combined the resulting subnetworks into 14 consensus subnetworks that were found across different gene scores and networks ($P < 0.004$, Supplementary Table 3), plus the condensin complex and CLASP/CLIP proteins (Figure 4.6) that were significant in individual interaction networks (Supplementary Tables 6,7). Our consensus process also identifies 13 "linker" genes that are members of more than one consensus subnetwork. We developed an online interactive viewer (see URLs and Figure 4.7) for Pan-Cancer HotNet2 subnetworks.

The subnetworks and linker genes (Figure 4.8a) include: portions of well-known

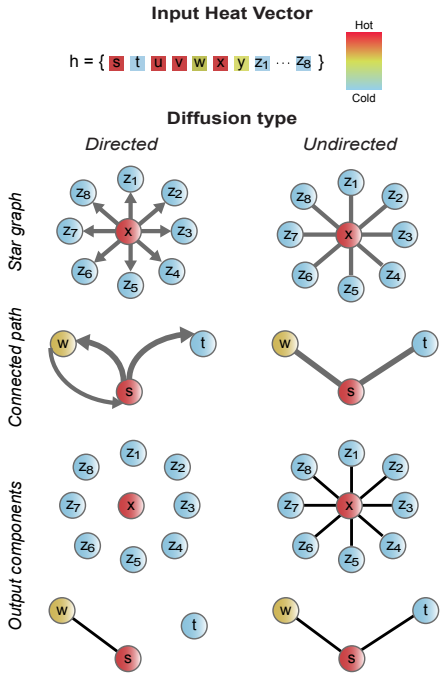



**Figure 4.4:** Comparison of directed and undirected diffusion kernels. (Left) HotNet2 identifies strongly connected components in a directed graph with edges weighted by a directed diffusion kernel. By using directed edges, HotNet2 will not output star graphs around a single "hot" node (*star graph*), nor will it include cold nodes in larger components (*connected path*). (Right) HotNet identifies connected components in an undirected graph with edges weighted by an undirected diffusion kernel, and therefore identifies more components comprised of more cold nodes.

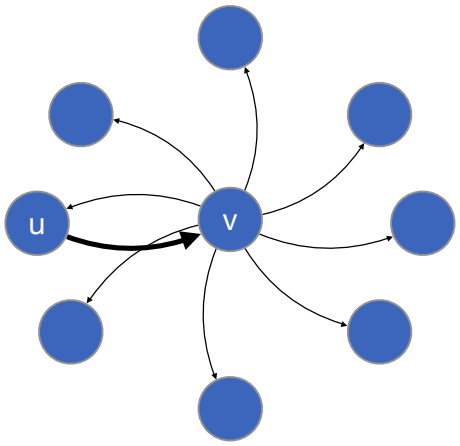



**Figure 4.5:** HotNet2 similarity between neighbors in a small graph. Node $u$ has degree one, so sends most of its heat to its one neighbor $v$. Node $v$ has multiple neighbors, and therefore sends less of its heat to each of its neighbors.

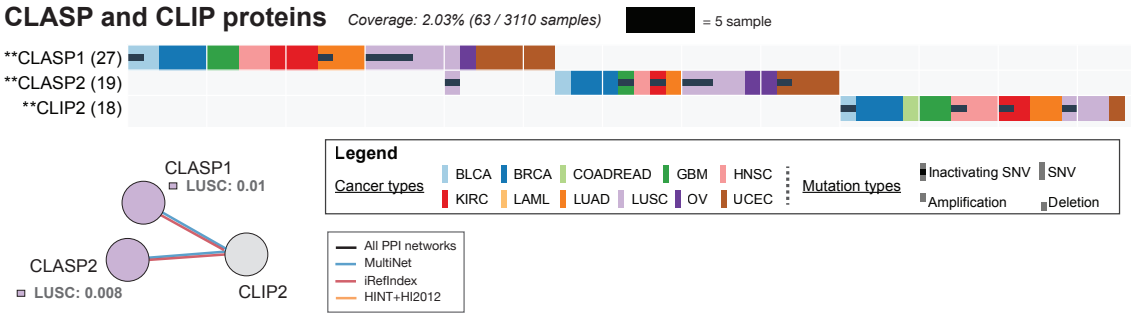



**Figure 4.6:** The CLASP and CLIP proteins subnetwork and its mutation matrix. The mutation matrix shows the mutations in the genes in the subnetworks, represented on the left, across the samples with mutations in the subnetwork. The number of samples with mutations in a gene is in parenthesis. The number of samples with mutations in any of the members of the subnetwork is on the top. Upticks and downticks represent amplifications and deletions, respectively. Full ticks represent SNVs, indels, and splicesite mutations. Colors reflect the cancer type of the samples. Genes with * were significant by exactly one of MuSiC, MutSigCV, or Oncodrive, while genes with ** were not significant by any of these methods. Colors of interactions in the subnetwork represent the interaction network where it was found. *P*-values for cancer type enrichment of mutations in the genes are shown.

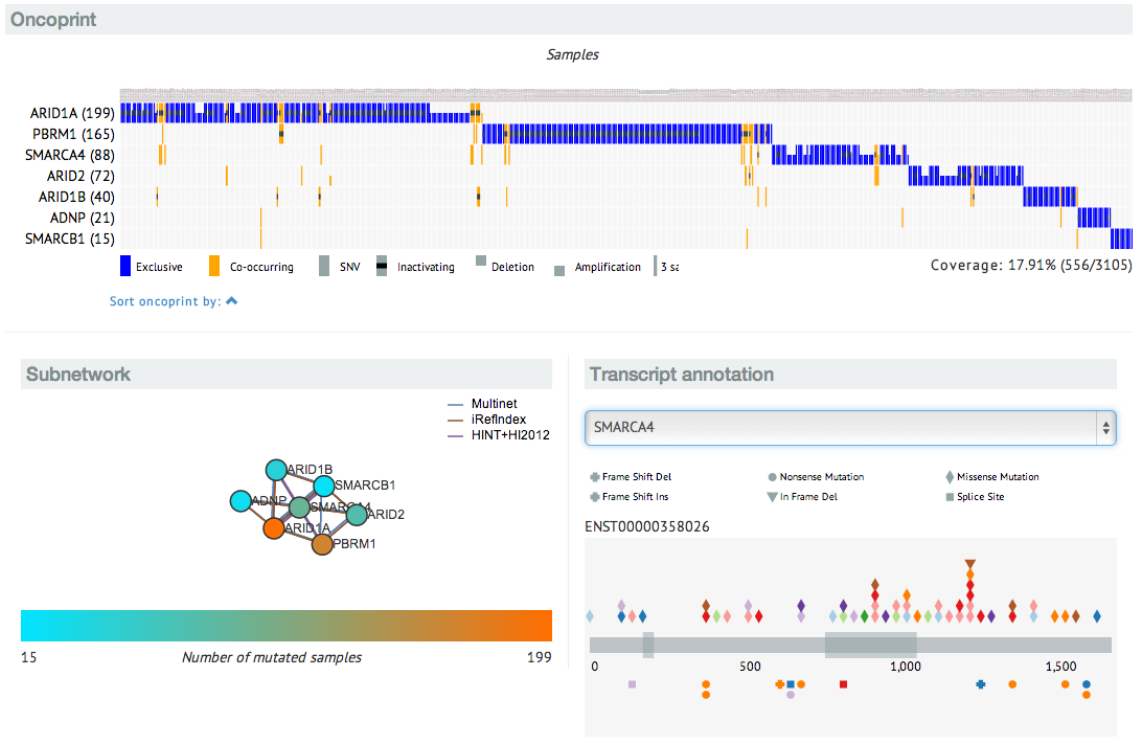



**Figure 4.7:** Example subnetwork visualization from the HotNet2 web output of the SWI/SNF subnetwork.

cancer pathways such as TP53, PI3K, NOTCH, and receptor tyrosine kinases (RTKs; Figure 4.9), as well as pathways and complexes that have more recently been observed to be important in cancer such as SWI/SNF complex, BAP1 complex, NFE2L2-KEAP1 (Figures 4.10 and 4.11), and RUNX1-CBFB core binding complex (Figure 4.12). The fifth most mutated subnetwork (16.9% of samples) consists of MLL2 and MLL3 and the putative interacting protein KDM6A (Figure 4.13), and was highly mutated (28.9% of samples) in TCGA Pan-Cancer squamous integrated subtype25. HotNet2 identified less-characterized and potentially novel subnetworks that may have also important roles in cancer including the cohesin and condensin complexes and MHC Class I proteins. The MHC Class I subnetwork (Figure 4.14) is an example of the ability of HotNet2 ability to identify rarely mutated cancer genes; all of the genes in the subnetwork are mutated in fewer than 35 samples (1.1%), yet four of the five genes have recently been proposed as novel cancer genes13. The sections below further detail a subset of these subnetworks. Additional analyses are in Appendix B.

Many of the subnetworks exhibit a significant enrichment for mutations in a subset of cancer types, including many previously unreported associations (Supplementary Tables 6-18). We also identify genes within these subnetworks enriched for mutations in particular cancer types. In addition, the HotNet2 Pan-Cancer analysis provides a clearer and more robust summary of subnetworks and novel genes than HotNet2 analysis of individual cancer types (Supplementary Table 19).

These subnetworks and linkers include a total of 147 genes, including many well-known cancer genes and pathways, but also including genes with mutations that are too rare to be significant by the single-gene tests (Supplementary Table 20). In total, 92 genes in the HotNet2 subnetworks are not reported by any of five single-gene tests (MutSigCV [17], Oncodrive-FM [25] and –CIS [143], MuSiC [26], or GISTIC2 [23]) or listed as a known driver gene in Vogelstein et al. [16], while an additional 13 genes are reported in only one such list. Many of these genes have literature evidence supporting a potential role in cancer, while others are in biological processes that suggest these genes warrant further study. Table 4.1 lists a subset of promising candidates, with the full list and associated references in Supplementary Table 20.

To obtain additional support for these genes we examined whether they had either an excess of inactivating mutations [16] or a cluster of missense mutations in protein sequence (using NMC30) or in protein structure (using iPAC [144]; Figures 4.15 and 4.16, and Supplementary Tables 21,22). We find that genes in HotNet2 consensus subnetworks are enriched for inactivating mutations ($P < 0.0001$) or mutation clusters ($P < 0.0001$) compared to genes not in subnetworks (Supplementary Table 6-18 and Appendix B.7). Finally, we evaluated a subset of the mutations in these genes using RNA-Seq and whole-genome sequencing (WGS) data from the same samples, and found RNA-Seq and/or WGS reads that validated 39 mutations in these novel genes (Appendix B.6 and Supplementary Table 23). These genes may represent novel

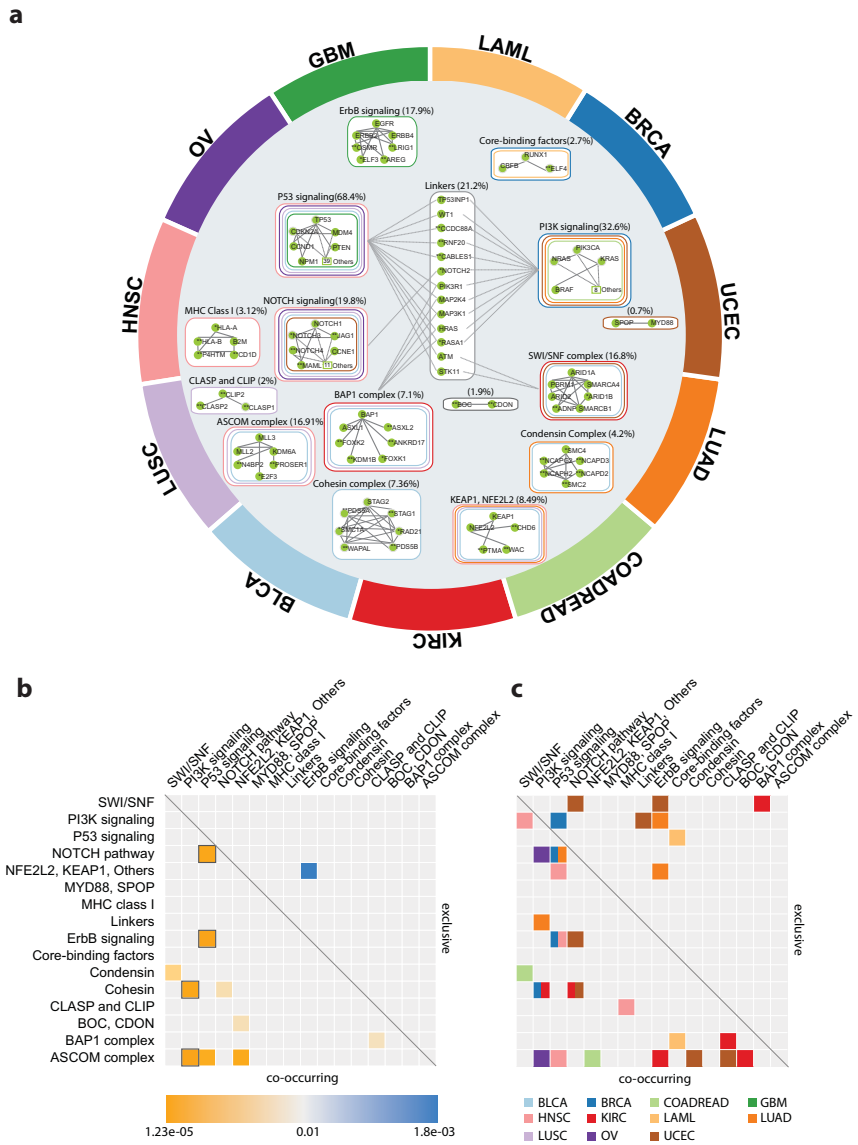



**Figure 4.8:** Overview of HotNet2 Pan-Cancer results. **(a)** Hotnet2 consensus subnetworks are arranged near the cancer types where they are enriched for mutations using a force-directed layout (BLCA=bladder urothelial carcinoma, BRCA=breast invasive carcinoma, COADREAD=colon adenocarcinoma and rectum adenocarcinoma, GBM=glioblastoma multiforme, HNSC=head and neck squamous cell carcinoma, KIRC=kidney renal clear cell carcinoma, LAML=acute myeloid leukemia, LUAD=lung adenocarcinoma, LUSC=lung squamous cell carcinoma, OV=ovarian serous cystadenocarcinoma, UCEC=uterine corpus endometrioid carcinoma). Colored outlines surrounding each network indicate the cancer types that are enriched for mutations (corrected $P < 0.05$). Interactions between proteins in a subnetwork are derived from the three interaction networks used in our Pan-Cancer analysis. In the center, there are 13 "linker" genes that are members of more than one consensus subnetwork; dotted lines between linkers and other consensus subnetworks indicate protein-protein interactions between them. **(b)** Heat map of significant co-occurrence (yellow, lower triangular) and exclusivity (blue, upper triangular) of mutations across all Pan-Cancer samples in the most frequently mutated HotNet2 Pan-Cancer consensus and condensin subnetworks ($P < 0.01$, Cochran-Mantel-Haenszel test). Black outlines indicate pairs of subnetworks that have $P < 0.05$ after multiple hypothesis correction. **(c)** Exclusivity/co-occurrence ($P < 0.01$, Fisher's exact test) within individual cancer types using the same color scheme as part (a).

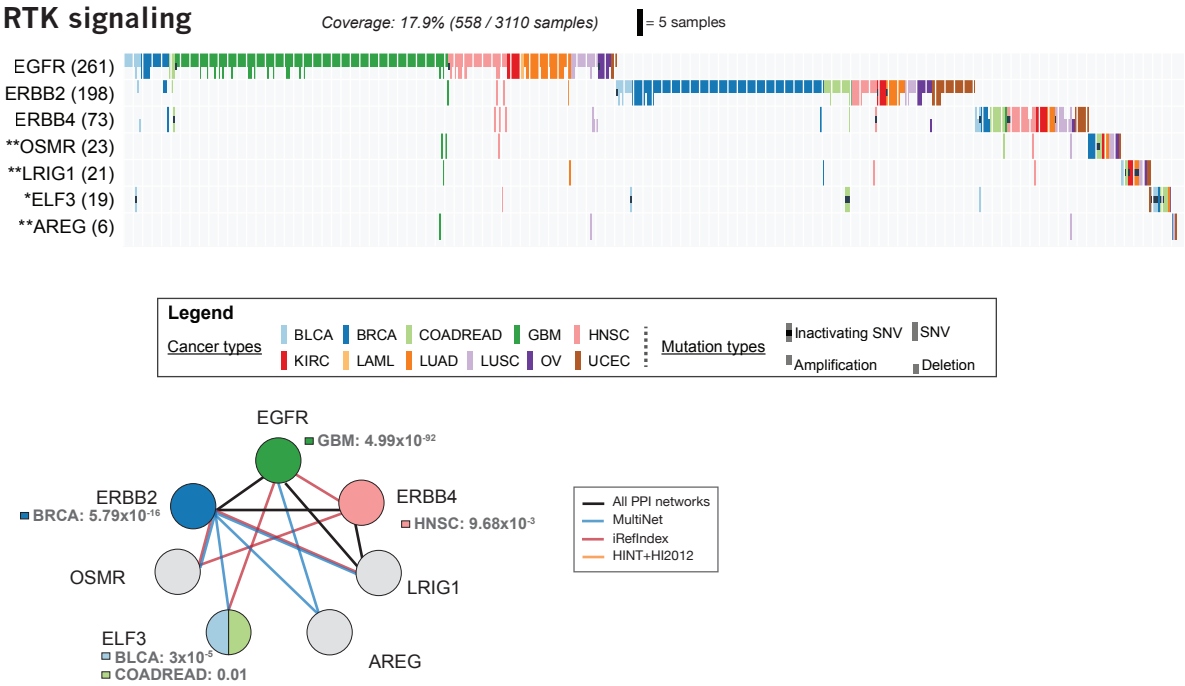



**Figure 4.9:** The RTK subnetwork and its mutation matrix. The mutation matrix shows the mutations in the genes in the subnetworks, represented on the left, across the samples with mutations in the subnetwork. The number of samples with mutations in a gene is in parenthesis. The number of samples with mutations in any of the members of the subnetwork is on the top. Upticks and downticks represent amplifications and deletions, respectively. Full ticks represent SNVs, indels, and splicesite mutations. Colors reflect the cancer type of the samples. Black dots corresponds to inactivating genes, that is genes that contain at least one of the following mutations: frame shift indels, nonsense, nonstop, and splice sites. Genes with * were significant by exactly one of MuSiC, MutSigCV, or Oncodrive, while genes with ** were not significant by any of these methods. olors of interactions in the subnetwork represent the interaction network where it was found. $P$-values for cancer type enrichment of mutations in the genes are shown.

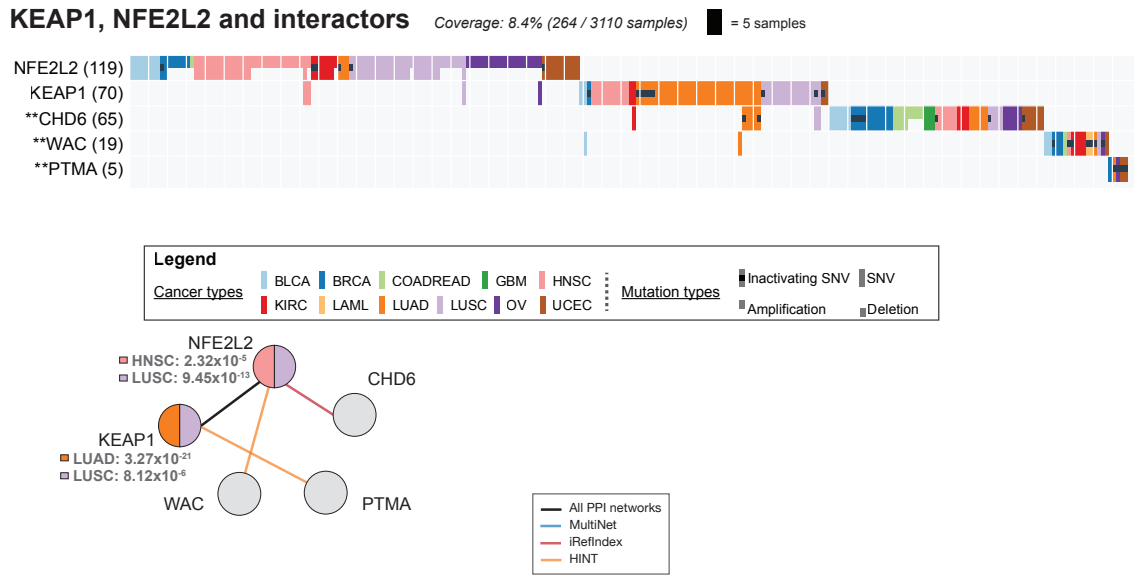



**Figure 4.10:** The KEAP-NFE2L2 subnetwork and its mutation matrix. Representation as in Figure 4.9.

**Figure 4.11:** The KEAP1 structure (PDB ID: 1ZGK) color coded by segment. Residues that are only in the iPAC most significant cluster (amino acids 480 – 524) are colored in light blue. The residue in the NMC most significant cluster (amino acid 470) is shown in red.

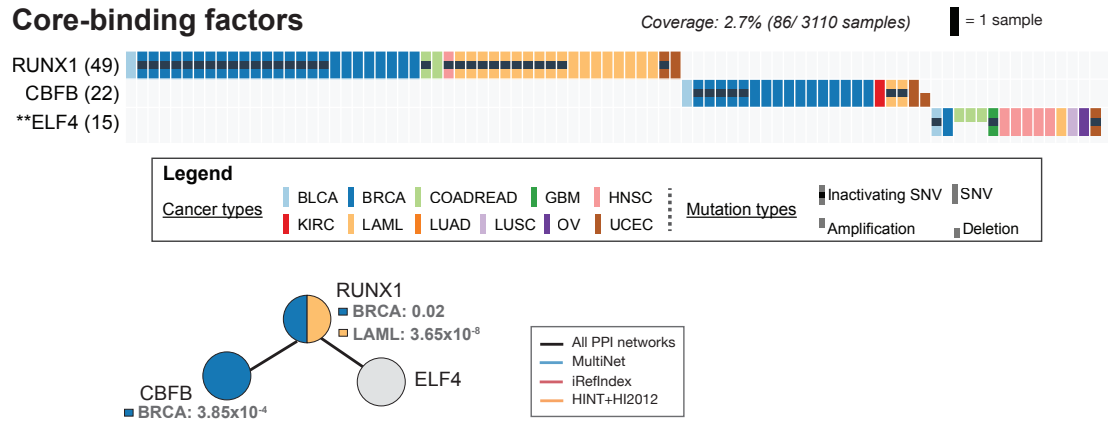



**Figure 4.12:** The RUNX1-CBFB-ELF3 subnetwork and its mutation matrix. Representation as in Figure 4.9.

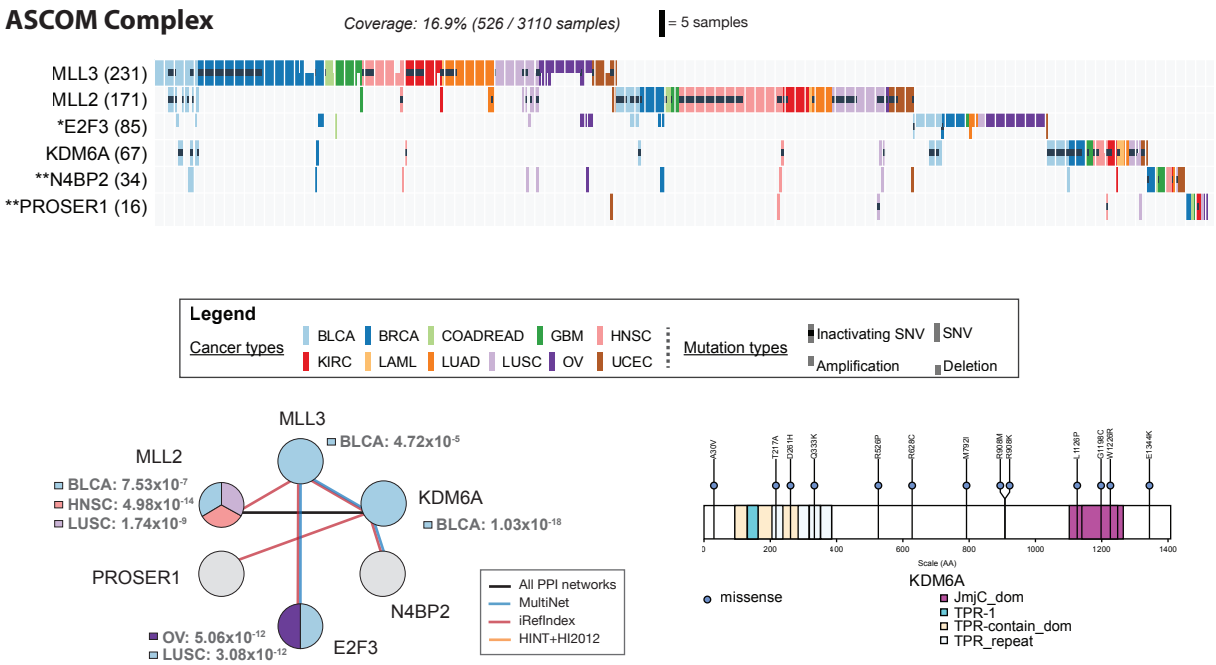



**Figure 4.13:** The ASCOM subnetwork and its mutation matrix. Representation as in Figure 4.9.

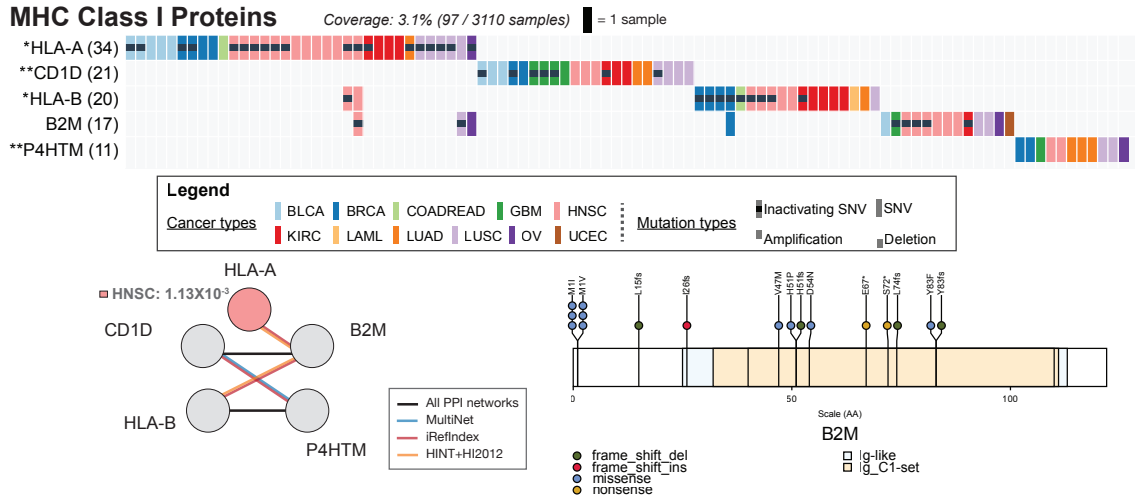



**Figure 4.14:** The MHC Class I subnetwork and its mutation matrix. Representation as in Figure 4.9. (bottom left) The distribution of mutations in *B2M*.

| Gene | SNVs | CNAs | Cancer Enrichment(s) | Function |
|---|---|---|---|---|
| *ADNP* | 21 | 0 | | Homeobox transcription factor with 9 zinc fingers found in the SWI/SNF complex; mediates neuroprotective responses to cellular growth, and regulates cancer cell proliferation. |
| *ASXL2* | 30 | 0 | | *BAP1* complex mediated chromatin modulation and transcriptional regulation; plays an opposing role to *ASXL1*. |
| *CCDC88A* | 38 | 0 | | Girdin family member with a key role in PI(3)K and Akt signaling pathways that may be involved in metastasis when overexpressed. |
| *CHD8* | 49 | 9 | | DNA helicase that acts as a chromatin remodeling factor and suppresses transcription. Suppresses TP53 and negatively regulates $\beta$-catenin in WNT signaling. CHD8 is essential for embryonic development. |
| *CUL9* | 48 | 0 | | Involved with p53 localization; critical regulator of cell cycle and quiescence. |
| *ELF3* | 19 | 0 | BLCA, COADREAD | Transcriptional activator that binds ETS motifs. May be a downstream effector of the *ERBB2* signaling pathway. |
| *EPHA3* | 50 | 3 | | Receptor tyrosine kinase with possible roles in BRCA, COADREAD, GBM, HNSC, lung, and pancreatic cancer. |
| *FOXK2* | 13 | 12 | | Forkhead transcription factor whose functions are cell cycle regulated; recruits AP-1 and functions in DNA mismatch repair. |
| *IWS1* | 16 | 0 | | Involved in transcriptional elongation and transcriptional surveillance. |
| *JAG1* | 24 | 0 | | Ligand for multiple Notch receptors and involved in the mediation of Notch signaling. May play a role in AML, BRCA, COADREAD, GBM, OV, and pancreatic cancer. |
| *KDM1B* | 14 | 0 | | Histone demethylase that acts as a co-repressor; along with *BAP1*, regulates cell growth. |
| *KLF5* | 12 | 36 | BLCA, COADREAD, HNSC | Kruppel-like transcriptional activation factor; regulates pluripotency and cellular growth. |
| *MLL5* | 30 | 0 | | Histone methyltransferase that acts as an important cell cycle regulator. High *MLL5* expression is associated with a favorable outcome in AML. |
| *NCAPH2* | 19 | 0 | | Non-SMC Condensin II subunit; critical for mitotic chromosome assembly. |
| NOTCH3 | 93 | 4 | OV | Receptor for Jagged1/2 and Delta 1 to regulate cell fate through transcriptional activation; mutations in NOTCH3 cause CADASIL. |
| *RNF20* | 27 | 0 | | E3 ubiquitin-protein ligase for H2BK120ub1; putative tumor suppressor. |
| *SHPRH* | 39 | 0 | | E3 ubiquitin-protein ligase for PCNA involved in DNA repair. |
| *SMG1* | 51 | 0 | | mRNA surveillance through nonsense-mediated mRNA decay. |
| *SMG7* | 23 | 0 | LUSC | mRNA surveillance through nonsense-mediated mRNA decay. |
| *STAG1* | 31 | 0 | | Cohesin subunit involved in sister chromatin adhesion following DNA replication. |
| *WAC* | 19 | 0 | | Regulates cell cycle progression by linking transcription to H2BK120ub1. |

**Table 4.1:** A subset of candidate cancer genes identified by HotNet2, but not by single-gene tests of significance (non-italicized genes are listed as a cancer driver by Oncodrive or GISTIC). For each gene, the number of samples with at least one SNV/CNA in the gene and the cancers enriched for mutations ($P < 0.05$, corrected) are listed. More information on these genes – as well as other candidate driver genes – is in Supplementary Table 20.

biomarkers for the classification of patients for treatment regimens.

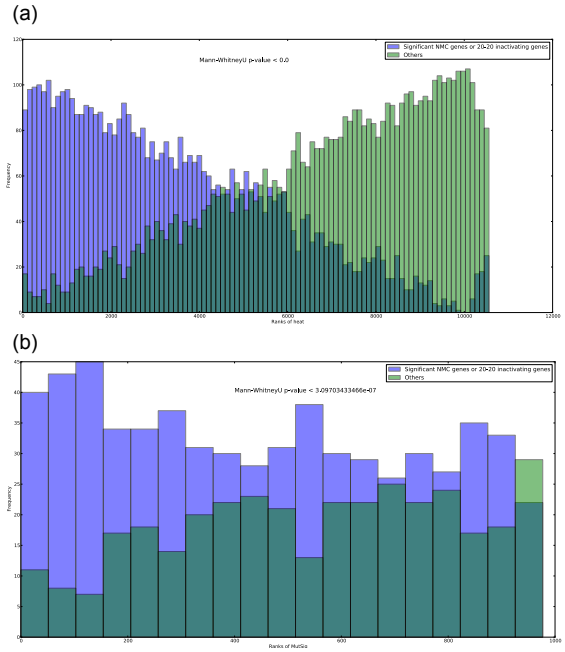



**Figure 4.15:** Histograms of two groups of genes ranked by heat score: genes in blue have either a cluster of missense mutations (NMC cluster $P < 0.05$) or $\geq 20\%$ of mutations inactivating; genes in green have mutations that meet neither criteria. **(a)** *Mutation frequency as heat.* The two distributions are significantly different ($P < 10^{-20}$, Mann Whitney U test), with higher-scoring genes showing enrichment for mutation clustering or high percentages of inactivating mutations. **(b)** *MutSigCV scores as heat.* The two distributions are significantly different ($P = 3.09 \times 10^{-7}$, Mann Whitney U test), with higher-scoring genes showing enrichment for mutation clustering or high percentages of inactivating mutations.

### 4.2.2 Co-occurrence and mutual exclusivity of mutations in subnetworks

Cancer cells are thought to harbor multiple driver mutations that perturb multiple biological functions [18]. Consistent with this model, we find that 4 pairs of subnetworks, including TP53 and NOTCH signaling, TP53 and RTK signaling, PI3K signaling and cohesin complex, and PI3K and ASCOM complex exhibit significant co-occurrence ($P < 0.05$, multiple hypotheses corrected) across the Pan-Cancer cohort (Figure 4.8b) or in individual cancer types (Figure 4.8c). Multiple pairs of genes within these subnetworks show co-occurring mutations (Supplementary Table 24). In contrast, mutual exclusive mutations are typically expected within a pathway, and not across pathways32,33. We observe significant mutual exclusivity within 4 of our subnetworks (Supplementary Table 25). Intriguingly, the RTK signaling and NFE2L2-KEAP1 subnetworks were the only pair with significant mutual exclusivity

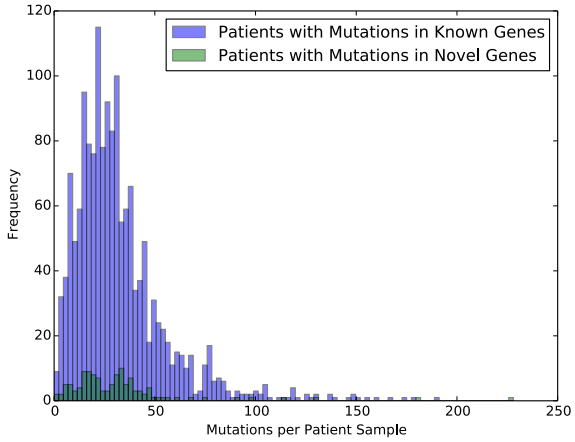



**Figure 4.16:** Histograms of the number of mutations per patient for two gene groups: novel genes identified by HotNet2, and a list of known cancer or significantly mutated cancer genes. The two distributions are not significantly different ($P = 0.4$, Mann Whitney U test).

across the Pan-Cancer cohort. This exclusivity was largely due to LUAD samples with mutually exclusive EGFR and KEAP1 mutations (Figure 4.17). This observation is consistent with reports of exclusivity between EGFR mutations and NFE2L2 expression in LUAD34 and also that NFE2L2 expression is downstream of EGFR signaling35. Examining individual cancers, we find a modest but not statistically significant enrichment for co-occurrence or exclusivity in a few cancer types. Neither within-subnetwork mutual exclusivity nor across-subnetwork co-occurrence is explicitly programmed into the HotNet2 algorithm. These observations support the hypothesis that the HotNet2 subnetworks represent distinct biological functions that are mutated in samples.

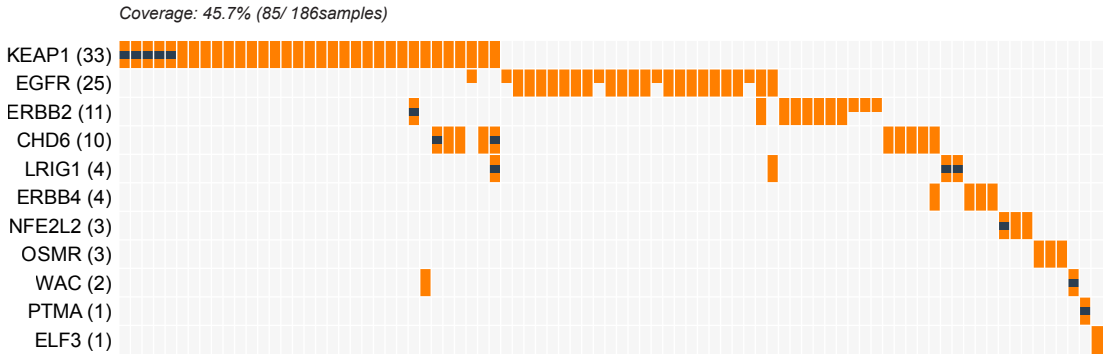



**Figure 4.17:** A mutation matrix in LUAD using genes in subnetworks, ErbB signaling and *KEAP1*, *NFE2L2* and interactors.

### 4.2.3 *TP53*, *PIK3CA*, and NOTCH networks

The three largest subnetworks – including a TP53 subnetwork, a PIK3CA subnetwork, and a NOTCH subnetwork – contain many well-known cancer genes (Supplementary Tables 8-10 and Figures 4.18 and 4.19). Linker genes join these three subnetworks, demonstrating the extensive crosstalk between well-annotated cancer pathways. Most of these linker genes encode signaling proteins that have known cancer-related functions (e.g. WT1, NOTCH2, PIK3R1, MAP2K4, MAP3K1, HRAS, ATM, and STK11). Taken together, 81.9% of the samples contain at least one mutation in these three large subnetworks and linker genes.

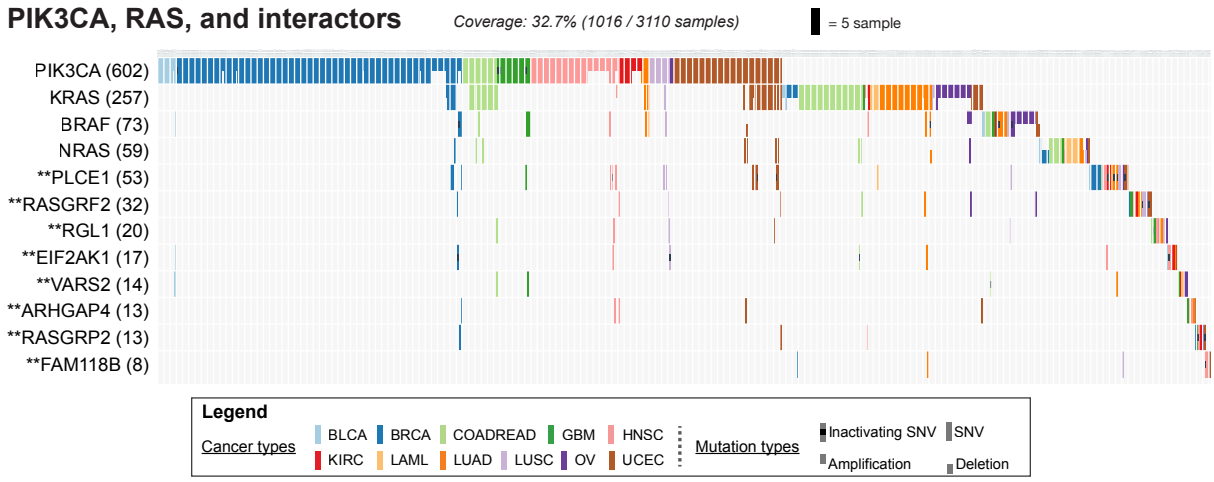



**Figure 4.18:** The PI(3)K/RAS subnetwork and its mutation matrix. Representation as in Figure 4.9.

HotNet2 Pan-Cancer analyses also revealed a number of novel genes (Supplementary Table 20) within these three subnetworks. These genes have documented interactions with well-known cancer genes and similar functions, but with somewhat lower mutational frequency ($\sim$ 1%), and were not marked as significant by single-gene tests [17, 25, 143, 26, 23]. For example, the TP53 subnetwork, includes CUL9. CUL9 sequesters p53 in the cytoplasm, and we find a cluster of 45 missense mutations ($P = 1.32 \times 10^{-8}$) as well as a cluster in protein structure (FDR = 0.025). Another gene of interest is IWS1, which is involved in transcriptional elongation and mRNA surveillance. Half (8/16) of the mutations in this gene are inactivating, and it also has a cluster of mutations ($P = 0.013$). This subnetwork also contains CHD8, an ATP-dependent chromatin-remodeling factor that regulates a wide range of genes36. We find three independent signals of CHD8 inactivation across samples: CHD8 is deleted in 9 samples in a focal peak from GISTIC; 18/58 (31%) of its mutations as inactivating; and has a wide cluster of missense mutations ($P = 6.37 \times 10^{-5}$). In the NOTCH subnetwork, we find rare mutations in JAG1 and DLL1, which interact with the NOTCH receptors and have some reports of a role

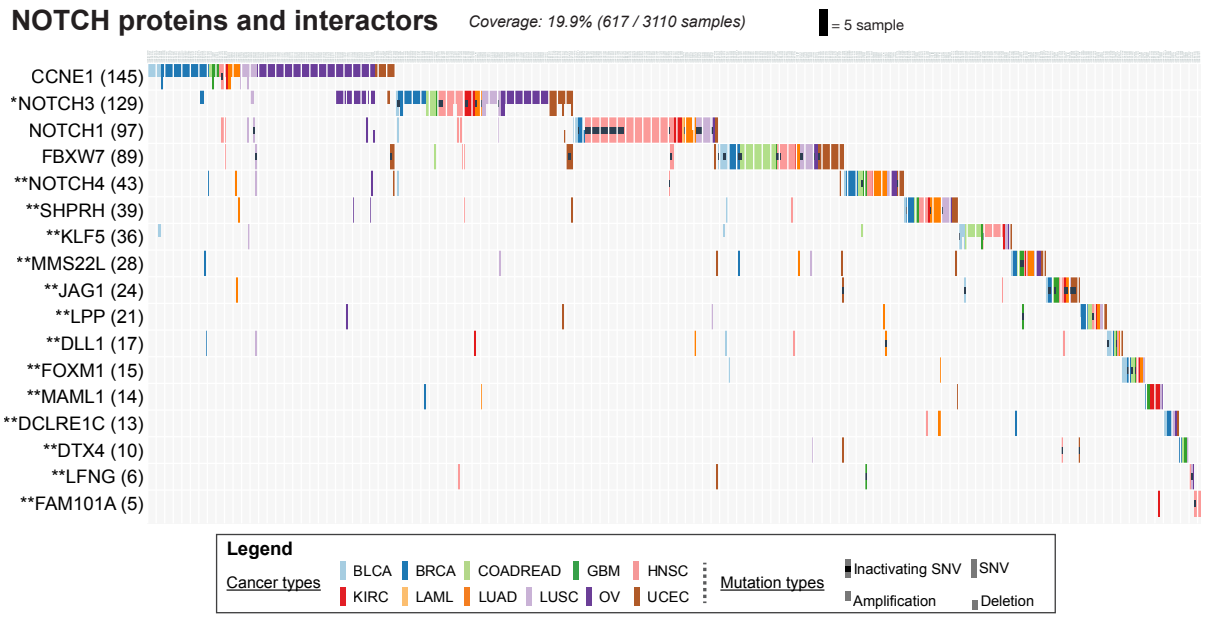



**Figure 4.19:** The NOTCH subnetwork and its mutation matrix. Representation as in Figure 4.9.

in cancer [145]. Moreover, 11/24 mutations in JAG1 are inactivating. The NOTCH subnetwork also includes SHPRH, which has a significant ($P < 8 \times 10^{-5}$) cluster of missense mutations (Figure 4.20).

### 4.2.4 SWI/SNF complex

The sixth most mutated HotNet2 Pan-Cancer subnetwork (16.8% of samples) includes multiple members of the SWI/SNF chromatin-remodeling complex (Figure 4.21a and Supplementary Table 12). Mutations in this complex have previously been reported in several cancers [146, 147], including TCGA samples [148]. Our HotNet2 Pan-Cancer analysis demonstrates the prevalence of mutations in SWI/SNF: at least 1.5% of the samples from each of the 12 cancer types contain a mutation in this subnetwork. KIRC ($P < 10^{-15}$), UCEC ($P = 7 \times 10^{-10}$), and BLCA ($P = 1.8 \times 10^{-8}$) were enriched for mutations in this subnetwork and several genes were enriched for mutations in specific cancer types including PBRM1 in KIRC ($P < 10^{-15}$) and ARID1A in both BLCA ($P = 4.8 \times 10^{-8}$) and UCEC ($P < 10^{-15}$). The subnetwork also contains ARID1B, which is reported to have somatic mutations in juvenile neuroblastoma [149] and germline mutations in Coffin-Siris syndrome [150].

Beyond known members of SWI/SNF, the subnetwork includes ADNP. ADNP mutations have not previously been reported in cancer and were not considered significant by the three individual gene-scoring methods. However, ADNP has a known

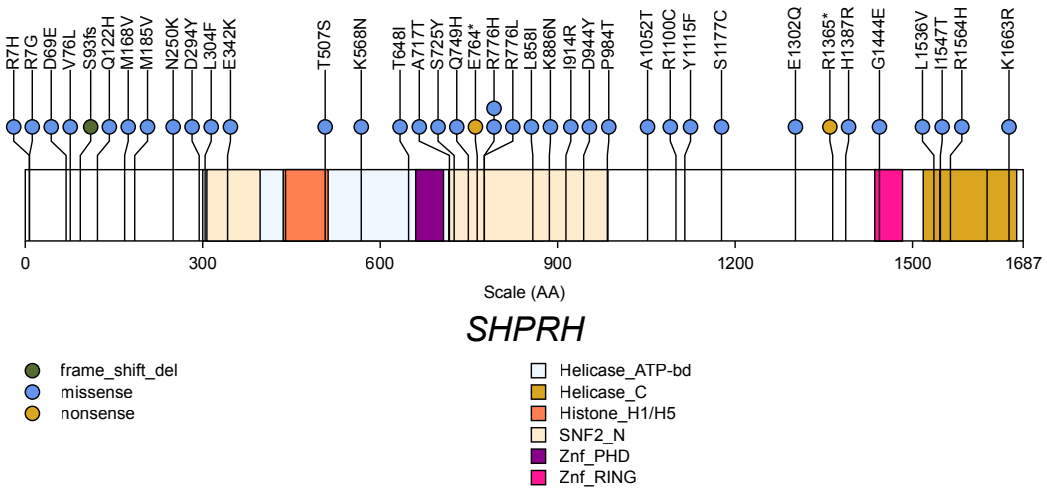



**Figure 4.20:** Pan-Cancer non-silent mutations in the *SHPRH* gene (ENSEMBL transcript ENST00000367503). *SHPRH* had a significant cluster of three mutations in position 776 ($P < 8 \times 10^{-5}$). In LUAD, GBM, and UCEC, these mutations are primarily in the helicase domain, indicating that they are likely inactivating.

interaction with SWI/SNF [151] and protects against oxidative stress in neuronal cells [152], suggesting that in rare cases ADNP mutations contribute to tumorigenesis. Thus, HotNet2 analyses broaden the view of mutations in SWI/SNF to additional cancer types and additional interacting proteins.

### 4.2.5 BAP1 complex and interactors

Another HotNet2 Pan-Cancer subnetwork (mutated in 7.1% of samples) overlaps the BAP1 complex (Figure 4.21b and Supplementary Table 13). This subnetwork includes BAP1, ASXL1, ASXL2, FOXK1, FOXK2, all members of the BAP1 core complex [153], as well as two additional interacting proteins: KDM1B and ANKRD17. Only BAP1 and ASXL1 were significant by individual gene scores – the other genes harbored rare mutations across many cancer types – a subtle signal revealed by HotNet2 Pan-Cancer analysis. This subnetwork is mutated in at least 6 samples from each cancer type, demonstrating the breadth of mutations in the BAP1 complex.

BAP1 inactivation has been reported in several cancers [153]. We find the subnetwork enriched for mutations in KIRC ($P = 2 \times 10^{-4}$), as previously reported46. Consistent with Pea-Llopis *et al.* [154], we find that mutations in the BAP1 gene are mutually exclusive ($P < 7.2 \times 10^{-3}$) of mutations in the PBRM1 gene in KIRC. We find that mutations in the SWI/SNF and BAP1 complexes show even greater mutual exclusivity ($P = 9.4 \times 10^{-5}$) in KIRC because of mutations in additional genes in these complexes besides BAP1 and PBRM1, respectively (Appendix B.5.8).

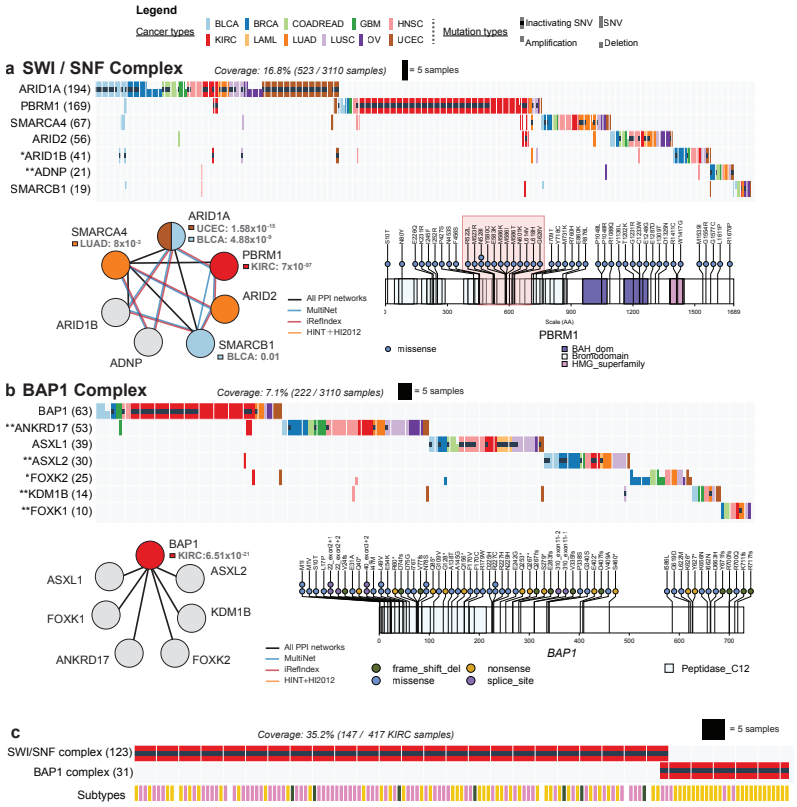



**Figure 4.21:** HotNet2 Pan-Cancer subnetworks overlapping SWI/SNF and BAP1 complexes. **(a)** Subnetwork containing members of the SWI/SNF complex including the BAF proteins ARID1A and ARID1B, PBAF proteins PBRM1 and ARID2, catalytic core member SMARCA4, SMARCB1 and ADNP. (a - Top) Mutation matrix shows the samples (colored by cancer type as shown in legend) with a mutation of the indicated type: full ticks represent SNVs, indels, and splice site mutations; upticks and downticks represent amplifications and deletions, respectively. A black dot corresponds to samples with an inactivating mutation in the gene, that the genes contain at least one of the following mutations: nonsense, frame shift indels, nonstop, or splice site. The number of samples with mutations in a gene is in parenthesis; genes with * were significant by exactly one of GISTIC2, MuSiC, MutSigCV, Oncodrive, or the list of driver genes in [16] while genes with ** were not significant by any of these methods. (a - Bottom left) Interactions between proteins in the subnetwork from each interaction network are colored according to mutually enriched cancer type with corresponding $P$-values. (a - Bottom right) PBRM1 protein sequence exhibited significant clustering of missense mutations ($P = 1.6x10^{-5}$) in a 105 amino acid bromodomain, a region that was reported to be mutated in a different renal clear cell carcinoma cohort39, but not in TCGA KIRC publication3. **(b)** Subnetwork containing members of the BAP1 complex including core PR-DUB complex, comprised of the deubiquinating enzyme BAP1 and the polycomb group proteins ASXL1 and ASXL2, as well as the BAP1-interacting proteins: ANKRD17, FOXK1, FOXK2, and KDM1B. Colors, marks, and panel organization are structured as in panel (a). **(c)** Inactivating mutations across samples (columns) in the SWI/SNF and BAP1 complexes (rows) in KIRC. The bottom row shows the mRNA expression classification of each sample.3 The mutations in these complexes are surprisingly exclusive in KIRC ($P < 3.6 \times 10^{-4}$, Fisher's exact test, corrected), and BAP1 is significantly enriched in mutations in the third expression subtype ($P < 3.4 \times 10^{-8}$, Fisher's exact test).

This mutual exclusivity suggests that mutations in these complexes define different subtypes of kidney cancer. Supporting this hypothesis, we observe that inactivating mutations in the BAP1 complex are enriched ($P < 3.4 \times 10^{-8}$) for samples in the third mRNA expression subtype from [10] (Figure 4.21c).

We find that a large fraction of the mutations in BAP1, ASXL1, and ASXL2 in different cancer types are inactivating mutations, demonstrating alternative strategies for inactivation of the BAP1 complex. In addition, 6/13 missense mutations in FOXK2 are in the forkhead transcription factor domain or forkhead associated domain, which may inactivate the DNA-binding properties of FOXK2. Finally, we examined the mutations in KDM1B, a gene that is involved in H3K4-methylation [155], but not considered a core part of the BAP1 complex. We find that 12/19 mutations in KDM1B (including 10/16 missense mutations) fall in the C-terminal amino-oxidase domain that is important for lysine-specific demethylation of histones [156]. Moreover, 2 of the 3 KDM1B mutations in LUSC and LUAD are inactivating, and these are also exclusive of BAP1 inactivating mutations, suggesting that KDM1B mutations might play a role in cancer.

### 4.2.6 Cohesin and condensin

HotNet2 Pan-Cancer analysis identifies 4/5 members of the cohesin complex as a significantly mutated subnetwork (7.3% of samples, Figure 4.22a and Supplementary Table 15). While named for its role in sister chromatid cohesion, the cohesin complex has recently been implicated more broadly in gene regulation [157, 158, 159], and its role in myeloid leukemia was only recently reported [160]. We found that cohesin was universally mutated across cancer types ($> 4\%$ of samples in each cancer type). Moreover, the mutations in the complex were spread uniformly across the genes with no gene in the complex mutated in more than 1.9% of samples. This pattern of mutations complicates the identification of recurrent mutations in individual genes, and indeed only half of the genes in the complex (STAG2, SMC1A, and RAD21) were significant by at least one of the three gene scores.

Mutations in some of these genes have recently been reported to be significant in several cancers. We find enrichment for mutations in the subnetwork in BLCA ($P = 7 \times 10^{-4}$); this enrichment derives largely from enrichment for mutations in STAG2 in BLCA ($P = 0.005$), which was recently reported53. STAG2 has a significantly higher fraction of inactivating mutations than other genes in the subnetwork (53% for STAG2 compared to 28% for the subnetwork as a whole); these inactivating mutations are not only in BLCA, but also across multiple cancer types with multiple inactivating mutations in LAML and COADREAD. In addition, BLCA samples without STAG2 inactivating mutations harbor rare inactivating mutations in several other cohesin genes. All mutations in RAD21 in LAML samples were inactivating,

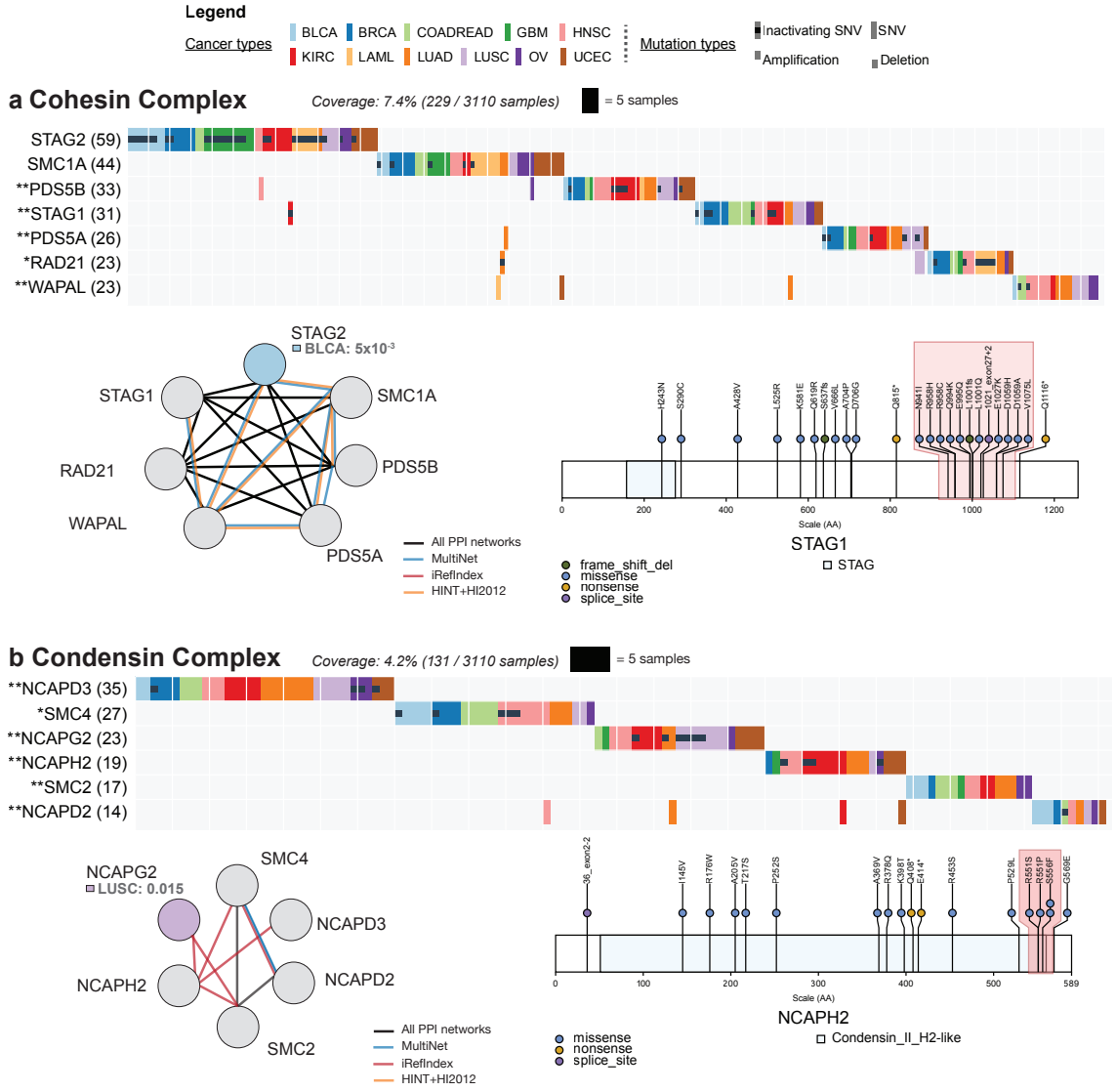



**Figure 4.22:** HotNet2 Pan-Cancer subnetworks overlapping the cohesin and condensin complexes. **(a)** Cohesin consensus subnetwork and its mutations. Colors and marks as in Figure 4.21a. None of the genes is mutated in more than 1.9% of the samples, but the subnetwork is mutated in $> 4\%$ of the samples in each cancer type. STAG1 exhibits significant ($P < 6 \times 10^{-5}$) clustering of missense mutations across 135 residues (highlighted) in the Pfam-B domain (PFAM ID: PB002581), a pattern suggesting inactivation of the corresponding domain. **(b)** Condensin consensus subnetwork, its mutations. (Top) Mutation matrix shows five genes in the condensin I and II complexes. Only one gene, SMC4, was significant by individual gene scores. (Bottom left) A subnetwork consisting of NCAPD2 and SMC4, both members of Condensin I, was significantly mutated in BLCA, while a subnetwork consisting of NCAPD3, NCAPG2 and NCAPH2, all members of Condensin II, was significantly mutated in LUAD and LUSC. At the gene level: NCAPD2 was significantly mutated in BLCA; SMC4 was significantly mutated in BLCA and HNSC; NCAPD3 was significantly mutated in LUAD; and NCAPG2 was significantly mutated in LUSC. (Bottom right) NCAPH2 shows a significant ($P < 2.6 \times 10^{-4}$) cluster of missense mutations between R551 and S556.

and BRCA and KIRC harbor inactivating mutations in STAG1. In addition, we observed a significant clustering of missense mutations in STAG1 ($P = 6 \times 10^{-5}$), and the broad span of the cluster (135 residues) is indicative of inactivation. STAG1 has been shown to function as a transcriptional coactivator [158, 159], and thus mutation of STAG1 may play another role in cancer apart from genome stability. Together, these results show that mutational inactivation of the cohesin complex occurs broadly across cancer types and across genes within the complex.

HotNet2 also identifies two subnetworks containing six proteins in the condensin complex, in HotNet2 runs from individual interaction networks. The combined subnetwork is mutated in 4.2% of samples (Figure 4.22b and Supplementary Table 6). Only SMC4 was reported significant by at least one of the individual gene scores. A subnetwork consisting of NCAPD2, SMC2, and SMC4, both members of Condensin I form of the complex, was significantly mutated in BLCA ($P = 6.2 \times 10^{-6}$). Condensin I is thought to primarily be involved in the sister chromatid condensation during mitosis [161, 162], suggesting that these mutations promote genome instability. In contrast, a subnetwork consisting of NCAPD3, NCAPG2 and NCAPH2, all members of Condensin II form of the complex, was significantly mutated in LUAD ($P = 0.04$) and LUSC ($P = 0.002$) and the majority (4/7) of NCAPG2 mutations in LUSC are inactivating. Condensin II is generally involved in gene regulatory processes [161, 162], suggesting a different phenotype for these mutations. In addition, we found a significant ($P = 0.002$) cluster of missense mutations in NCAPH2 (Figure 4.22b), implying that mutations in this region of unknown function may be important for the deregulation of condensin. We also note that it was recently observed that expression of NCAPD3 was positively associated with recurrence-free survival [163]. Finally, RNA-seq and whole-genome sequencing data from the same samples provide further validation of the somatic mutations in SMC2, SMC4, NCAPD2, NCAPD3, NCAPH2, and NCAPG2 and show that some of these mutations are expressed (Appendix B.6 and Supplementary Table 39). Our HotNet2 Pan-Cancer analysis suggests that multiple cancer types harbor rare mutations in the cohesin and condensin complexes, supporting a proposed tumor suppressor role for these complexes [157, 161, 162].

## 4.3 Discussion

We present a novel approach for identifying combinations of somatic aberrations in different cancer types using our HotNet2 algorithm to analyze a high-quality Pan-Cancer dataset of 3281 samples from 12 cancer types. This analysis represents the largest network analysis of somatic aberrations across multiple cancer types. We recover many classic cancer pathways like TP53, PI3K, NOTCH, and RTK automatically from a large-scale interaction network, demonstrating the power of the

Pan-Cancer network approach. Second, we highlight the extensive crosstalk between these pathways, overlaps that are often overlooked in analyses that treat pathways as distinct gene lists. Third, we find pathways and complexes whose role in cancer was only appreciated recently such as the SWI/SNF chromatin-remodeling complex [146] and BAP1 complex [153]. Fourth, we find that several pairs of HotNet2 subnetworks have co-occurring mutations, while within subnetworks mutations are mostly exclusive. This supports the hypothesis that these subnetworks represent distinct biological functions that are mutated in samples. Finally, we identify a number of novel mutated subnetworks with potential roles in cancer including: the cohesin and condensin complexes [161]; MHC Class I proteins; and the telomerase complex. These subnetworks have rare mutations in nearly all cancer types, making them difficult to detect without a sensitive Pan-Cancer network approach that examines combinations of genes across multiple cancer types.

The HotNet2 subnetworks contain 92 genes that are rarely mutated, both in individual cancer types and across the Pan-Cancer cohort, and are not reported as significant by single-gene tests. Nearly all of the subnetworks contain such genes, which are revealed by the combination of their mutations and interactions across cancer types. Some of these rarely mutated genes are inevitably false positive predictions of the analysis, but many (including SHPRH, CUL9, CHD8, RNF20, JAG1, ELF3, STAG1, NCAPH2, and others) exhibit either mutational clustering or protein interactions that support a role for the observed somatic aberrations (Supplementary Tables 6-18). In addition, we find that well-characterized mutations in a single gene in one cancer type (e.g. inactivating mutations BAP1 in KIRC) are replaced in other cancer types by rare mutations in other members of the same complex (e.g. inactivating mutations in ASXL1, ASXL2, FOXK2, KDM1B). Such observations suggest that Pan-Cancer network analyses may prove useful in translating diagnostic or therapeutic approaches that were developed in one cancer type to other cancer types.

Our analysis complements other recent Pan-Cancer analyses including studies that analyze only one type of aberration [164, 165, 65] or restrict attention to recurrent aberrations [166] (Appendix B.8.3 and Supplementary Table 27). The HotNet2 Pan-Cancer network approach identifies combinations of rare and common mutations in groups of interacting genes; combinations that were not apparent by analysis of single genes, known pathways, or single cancer types. Indeed, we observe that many of the identified subnetworks contain genes altered by both SNVs and CNAs, demonstrating that integrating multiple types of aberrations is beneficial when jointly analyzing multiple cancer types that might have different mutational landscapes. Pan-Cancer network analysis of multiple aberration types thus provides an alternative approach to prioritize rare mutations for further experimental characterization.

As with any computational approach, our findings are limited by the quality and

quantity of input data. Further power is anticipated by including additional samples [65], additional types of genetic and epigenetic aberrations, and better interaction networks. For example, structural variants, non-coding variants and methylation data were not included, the first two being unavailable for most TCGA samples. This lack of data, plus false negatives in the analyzed data (e.g. due to difficulties in identification of indels and subclonal variants) imply that our analysis likely underestimates the number and frequency of mutated subnetworks across cancer types. On the other hand, we note that some genes that are highly significant by individual gene scores are not reported in our network analysis; often this is due to problems with the interaction network. Improved knowledge of the human interactome – including more systematic efforts to record known interactions, measure additional interactions, and determine the tissue specificity of interactions – are needed to increase coverage and reduce possible ascertainment bias.

Finally, the HotNet2 algorithm introduced here is suitable for other applications, both biological and non-biological. In particular, genome-wide association studies (GWAS) and other studies of genetic diseases face an analogous problem of identification of combinations of genetic variants with a statistically significant association to a phenotype. With an appropriate gene score, the HotNet2 algorithm can be applied to such data.

## 4.4 Methods

### 4.4.1 Somatic aberration data

SNVs, indels, and splice-site mutations were extracted from TCGA Pan-Cancer analysis on Synapse (syn1710680), and copy number aberrations (CNAs) from GISTIC2 output via Firehose. We restricted attention to the 3276 samples containing both SNV and CNA data. We removed 71 samples identified as ultramutators in syn1729383 and additional 95 samples with an unusually high number of aberrations ($> 400$ SNVs or CNAs). We selected the threshold of 400 aberrations per sample as the derivative of the number of mutations per sample starts increasing rapidly beyond this value (Figure 4.23). We removed genes without CNAs that contained SNVs in $> 2\%$ of samples but were not identified as significant ($q < 0.05$) by MutSigCV20. Finally, we used only those genes that had at least 3 reads from RNA-seq data in at least 70% of samples of at least one of the cancer types, as described in syn1734155 (See URLs). The resulting dataset contained aberrations in 11,565 genes and 3110 samples (Figure 4.2). We used genes scores from: mutation frequency and MutSigCV -log10 $q$-values. Nonsense, frame shift indels, nonstop, or splice site mutations were classified as inactivating following [164]. We used three

interaction networks: HINT+HI2012, a combination of HINT network [139] and the HI-2012 [140] set of protein-protein interactions; MultiNet [141]; iRefIndex [142]. Additional details of the datasets are in Appendix B.

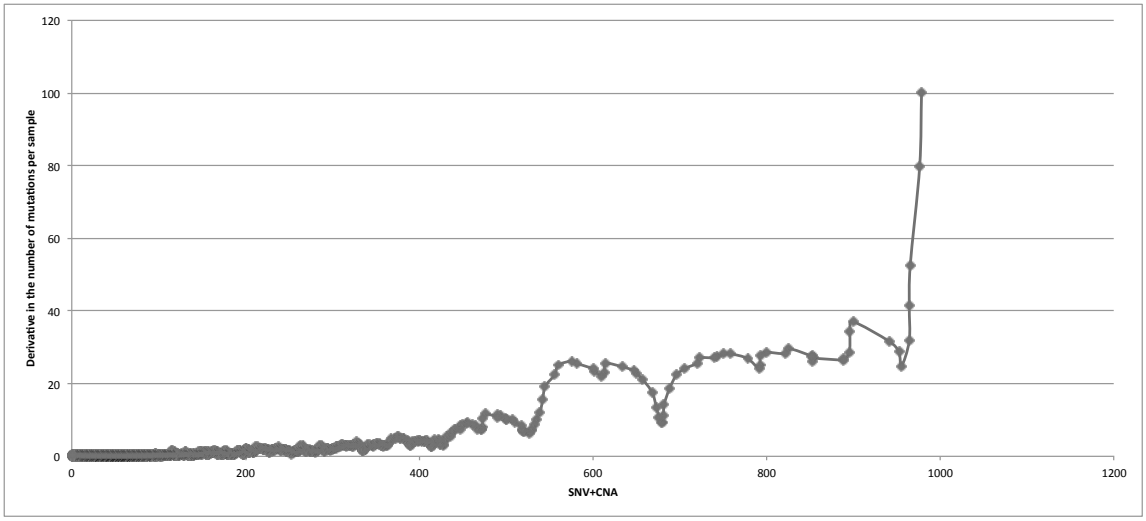



**Figure 4.23:** The number of aberrations (SNV and CNA) in each sample. Samples with $> 400$ aberrations were removed.

### 4.4.2 HotNet2

We developed the HotNet2 (HotNet diffusion oriented subnetworks) algorithm to identify subnetworks of a genome-scale interaction network that are mutated more than expected by chance. While interaction networks have proven useful in analyzing various types of genomic data [167], statistically robust identification of significantly mutated subnetworks is a difficult problem with several major challenges (Appendix B.1.1). HotNet2 addresses these challenges and identifies significantly mutated subnetworks of a genome-scale interaction network, using an insulated heat diffusion process that considers both the scores on individual genes/proteins as well as the topology of interactions between genes/proteins (Figure 4.3).

The input to HotNet2 is: a heat vector $\vec{h}$ that contains the scores (e.g., mutation frequency) for each gene $g$; and a graph $G = (V, E)$, where each node corresponds to a gene/protein and each edge corresponds to an interaction between the corresponding genes/proteins. HotNet2 performs the following steps:

1. *Heat Diffusion.* HotNet2 employs an insulated heat diffusion process [168, 169] that captures the local topology of the interaction network surrounding a protein. At each time step, nodes in the graph pass to and receive heat

from their neighbors, but also retain a fraction of their heat, governed by an insulating parameter . The process is run until equilibrium; the amount of heat on each node at equilibrium thus depends on its initial heat, the local topology of the network around the node, and the value If a unit heat source is placed at node $j$ (e.g. a mutation in in one sample) then the amount of heat on node $i$ is given by the $(i, j)$ entry of the diffusion matrix $F$ defined by:

$$F = \beta(I - (1 - \beta)W)^{-1}, \tag{4.1}$$

where

$$W_{ij} = \begin{cases} \frac{1}{\deg(j)}, & \text{if node } i \text{ interacts with node } j, \\ 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

Thus, $W$ is a normalized adjacency matrix of the graph $G$. We interpret $F(i, j)$ as the influence that a heat source placed on $g_j$ has on $g_i$. The insulated heat model can also described in terms of a random walk with restart (Appendix B.1.2). Note that the insulated diffusion process is generally asymmetric, i.e. $F(i, j) \neq F(j, i)$. The diffusion matrix depends only on the graph $G$, and not the heat vector $\vec{h}$. Therefore the influence (for a given $\beta$) needs to be computed only once for a given interaction network.

2. *Exchanged heat matrix.* The insulated heat diffusion process described above encodes the local topology of the network, assuming unit heat is placed on nodes. To jointly analyze network topology and gene scores given by the initial heat vector $\vec{h}$, we define the exchanged heat matrix $E$:

$$E = FD_{\vec{h}}, \tag{4.3}$$

where $D_{\vec{h}}$ is the diagonal matrix with entries $\vec{h}$. $E(i, j) = F(i, j)\vec{h}(j)$ is the amount of heat that diffuses from node $g_j$ to node $g_i$ on the network when $\vec{h}(j)$ heat is placed on $g_j$, which we interpret as the similarity of $g_j, g_i$. Since the diffusion matrix $F$ is not symmetric and in general $\vec{h}(i) \neq \vec{h}(j)$, the similarity $E(i, j)$ is also not symmetric (Appendix B.1.2).

3. *Identification of hot subnetworks.* We form a weighted directed graph $H$ whose nodes are all measured genes. If $E(i, j) > \delta$, then there is a directed edge from node $j$ to node $i$ of weight $E(i, j)$. HotNet2 identifies *strongly connected components* in $H$. A strongly connected component $C$ in a directed graph is a set of nodes such that for every pair $u, v$ of nodes in $C$ there is a path from $u$ to $v$.

4. *Statistical test for subnetworks.* HotNet2 employs a statistical test to determine the significance of the number and size of the subnetworks determined in the previous step. The statistical test is the same as the two-stage statistical test introduced in the original HotNet algorithm [34, 115] (Appendix B.1.3, Figures 4.24, 4.25, 4.26, and 4.27, and Supplementary Table 28).

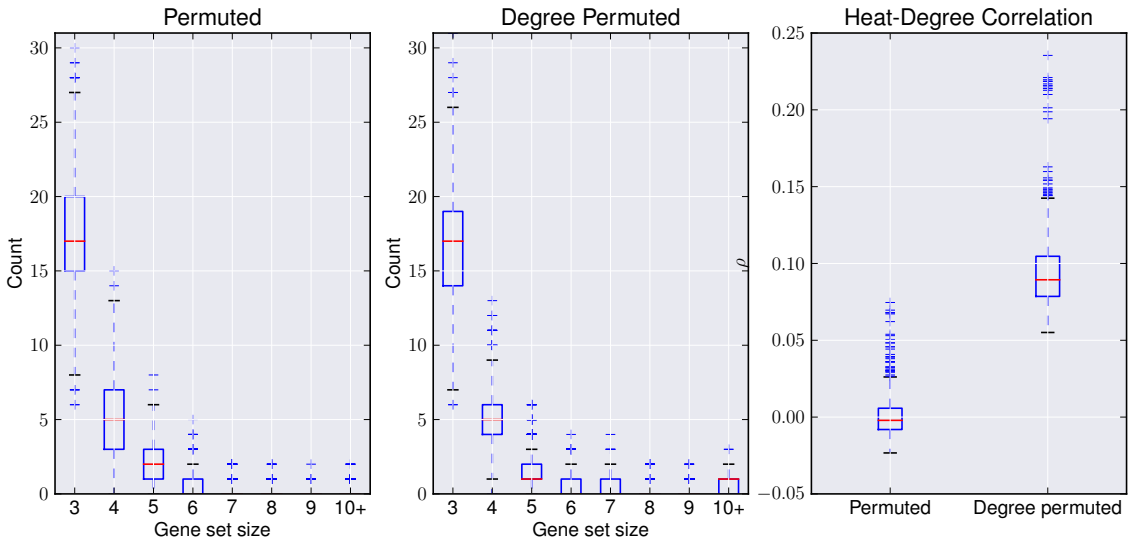



**Figure 4.24:** **(a-b)** The distribution of sizes of subnetworks identified by HotNet2 on 1000 permuted datasets. **(a)** Datasets were generated permuting heat scores uniformly at random. **(b)** Datasets were generated by permuting the top 100 highest heat scores amongst themselves, and permuting the remaining heat scores amongst themselves. **(c)** The distribution of heat-degree correlations from the datasets in (a) and (b).

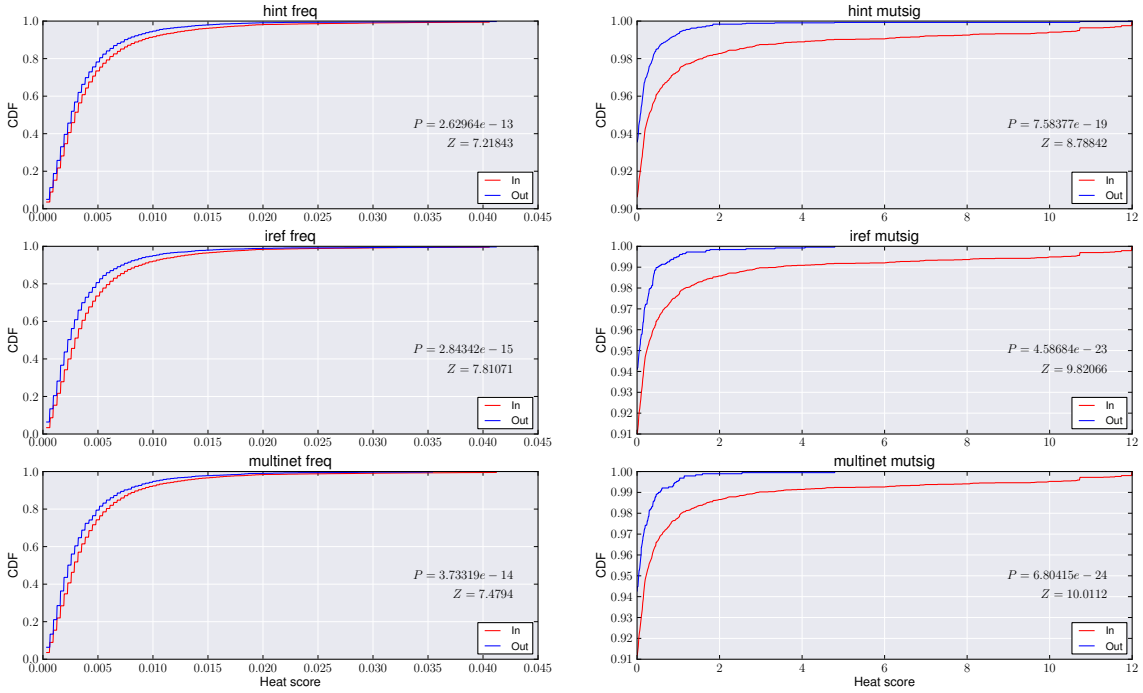



**Figure 4.25:** The CDF of the distribution of heat (mutation frequency or MutSigCV) in each of the three interaction networks. Using a two-sample z-test, the heat scores of genes in the network are significantly higher than the heat scores of genes not in the network ($P$-values and $Z$-scores are shown on each plot). The fifty highest heat scores are removed from the plots to demonstrate the difference visually.

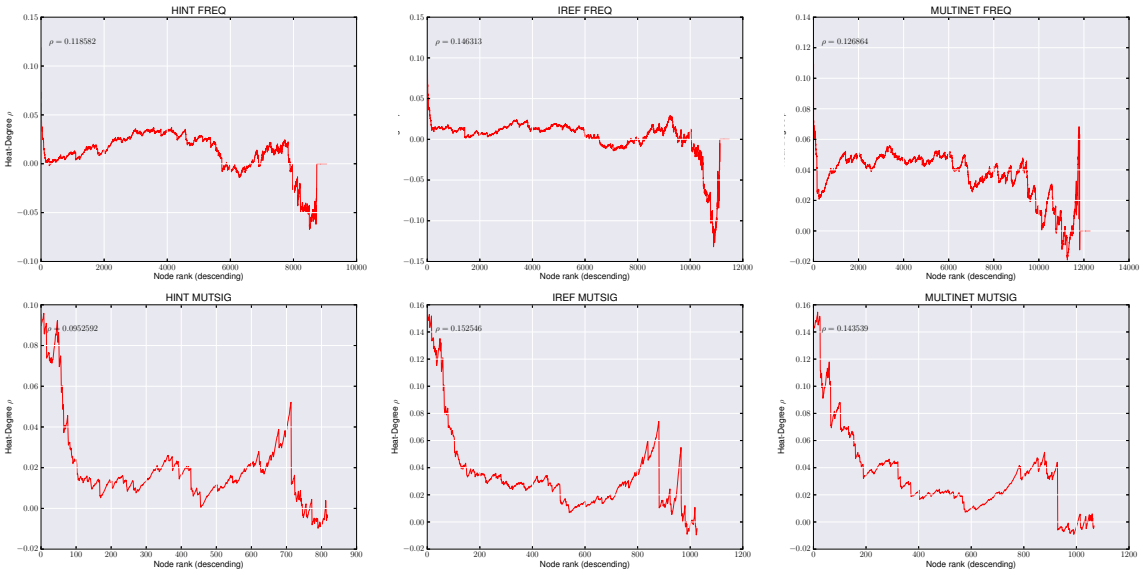



**Figure 4.26:** The correlation between heat and degree (y-axis, Spearman's $\rho$) after removing the top $N$ hottest nodes (x-axis) for each combination of network and heat score. After removing $N = 100$ nodes, the correlation between heat and degree has drooped to below $\rho = 0.05$ for each pair except Multinet-MutSigCV. The correlation statistic before removing any nodes is shown on each plot.

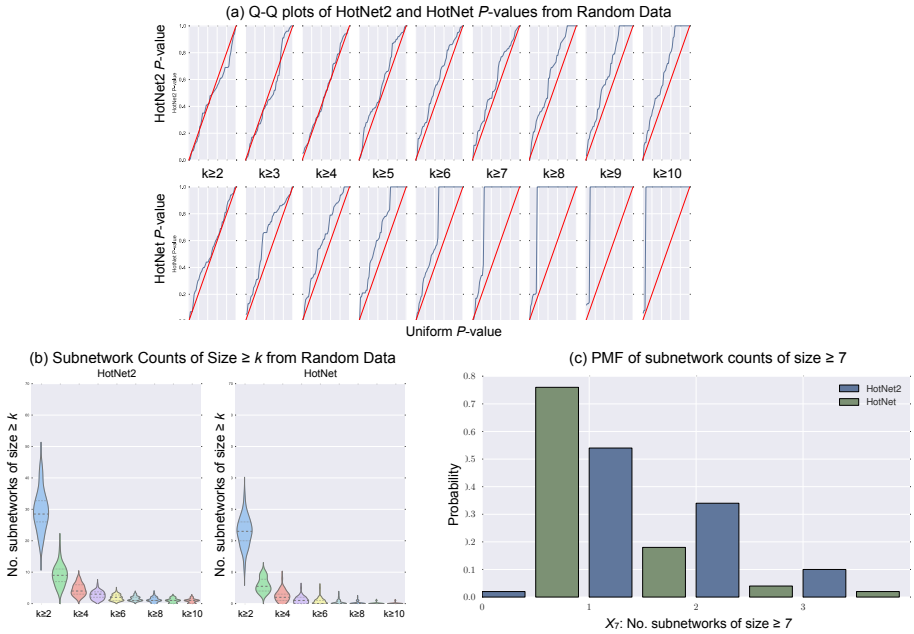



**Figure 4.27:** Results of HotNet2 and HotNet when applied to the $N = 50$ randomized MutSigCV datasets. (a) Q-Q plots of the $P$-values from HotNet2 (top) and HotNet (bottom) for each $k = 2 - 10$. (b) Violin plots of the number of subnetworks of size at least $k$ identified by HotNet2 (left) and Hotnet (right). (c) Histogram of the number of subnetworks of size $\geq 7$ identified by each algorithm.

HotNet2 is available online (See URLs).

HotNet2 has two parameters $\beta$ and $\delta$, and selects values for both of these parameters using automated procedures. $\beta$ is selected from the protein-protein interaction network, independently of any gene scores (Appendix B.1.4, Figure 4.28, and Supplementary Table 29). We evaluated the sensitivity of the HotNet2 results to the value of $\beta$ and found that varying $\beta \pm 10\%$ has only a minor effect on the results, with at most 7 genes (3.8% of total) added/removed from the subnetworks (Supplementary Table 28). The value of $\delta$ is chosen such that large connected components are not found using the observed gene score distribution on random networks with the same degree distribution as the observed network (Appendix B.1.4, Figure 4.29, and Supplementary Table 30). We evaluated the sensitivity of the HotNet2 results to the value of $\delta$, and found that varying $\delta \pm 5\%$ changed at most 35 genes (12.3% of total) in the subnetworks (Supplementary Table 29).

### 4.4.3 Comparison of HotNet2 to other algorithms

HotNet2 extends our previous algorithm HotNet [34, 115] in several directions. First, HotNet2 employs an insulated heat diffusion process that better encodes the local topology of the neighborhood surrounding a protein in the interaction network. Second, HotNet2 uses an asymmetric influence between two proteins to derive a directed measure of similarity between them, while HotNet derives a symmetric influence. Third, HotNet2 identifies strongly connected components in the directed graph $H$, while HotNet computes connected components in an undirected graph. These differences enable HotNet2 to effectively detect significant subnetworks in datasets in which the number of samples is order(s) of magnitude larger than considered by HotNet, and in which the mutational frequencies, or scores, occupy a broad range (from very common to extremely rare). See Figure 4.4.

Expanding on this third point, when undirected diffusion algorithms like HotNet or related network propagation algorithms [138] are run on large datasets containing a wide range of gene scores (e.g. the Pan-Cancer dataset), many of the resulting subnetworks are "hot" star graphs determined by a single high-scoring node and the immediate neighbors of this node (Figure 4.4). Star graphs, or more generally spider graphs, have one central node connected to multiple neighboring nodes that are not interconnected. While the hot, center node in these star graphs is typically a significant gene, the neighboring nodes are often artifacts.

We found that HotNet2 returns $> 80\%$ fewer hot stars/spiders than HotNet on the Pan-Cancer datasets (Supplementary Table 31). This is a major difference between the algorithms and is one of the reasons why HotNet fails to find statistically

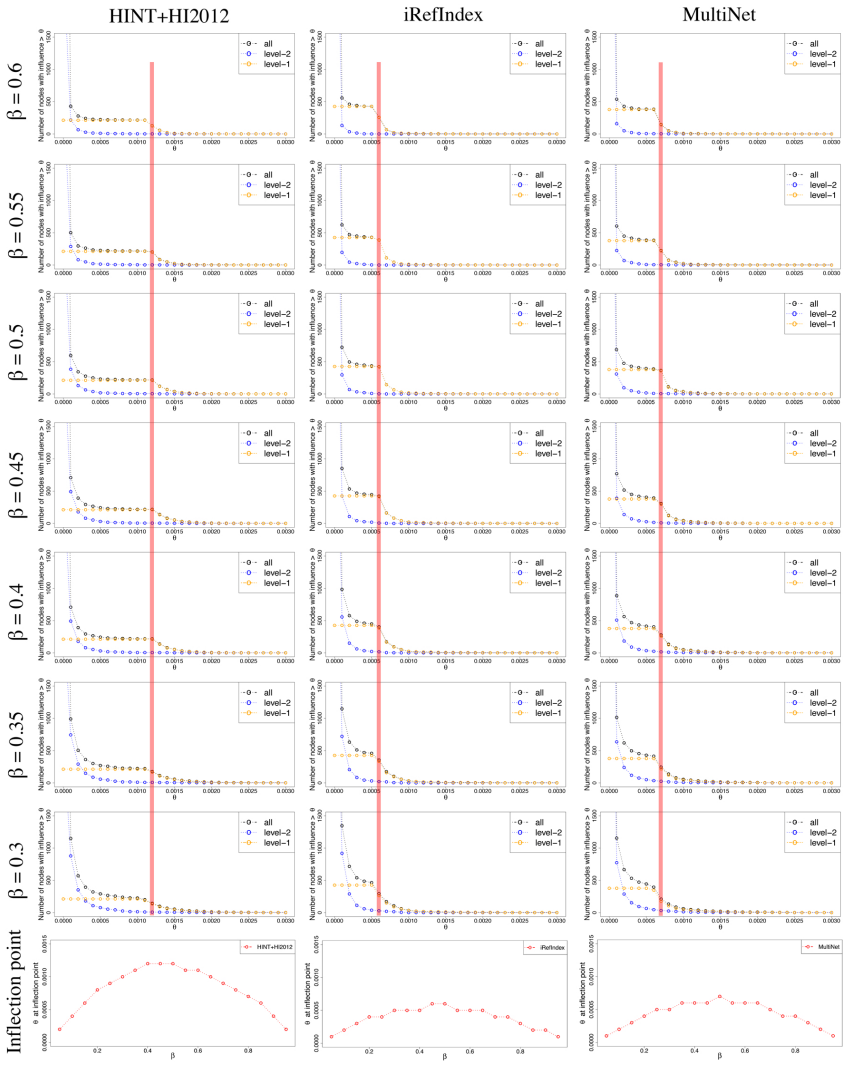



**Figure 4.28:** Distributions used to set the diffusion parameter $\beta$. Figures in different columns and rows represent distributions on HINT+HI2012, iRefIndex, and MultiNet interaction networks across $\beta$ from 0.6 to 0.3 for an example gene *TP53*, which has high betweenness centrality. The $x$-axis of each distribution represents $\theta$, a cut off of influence. The $y$-axis of each distribution represents the number of nodes in the interaction network with influence larger than $\theta$. Respectively, black, orange, and blue dotted circles represent the number of all nodes, level one nodes, and level two nodes with influence larger than $\theta$ across different $\theta$. Three red vertical lines across all distributions in an interaction network represent the location of the inflection point in level one in the $\beta$ we chose for different interaction networks, i.e. HINT+HI2012: $\beta = 0.4$, iRefIndex: $\beta = 0.45$, and MultiNet: $\beta = 0.5$. The bottom plot indicates the distribution of $\theta$ at the inflection point in $\beta$ from 0.05 to 0.95.

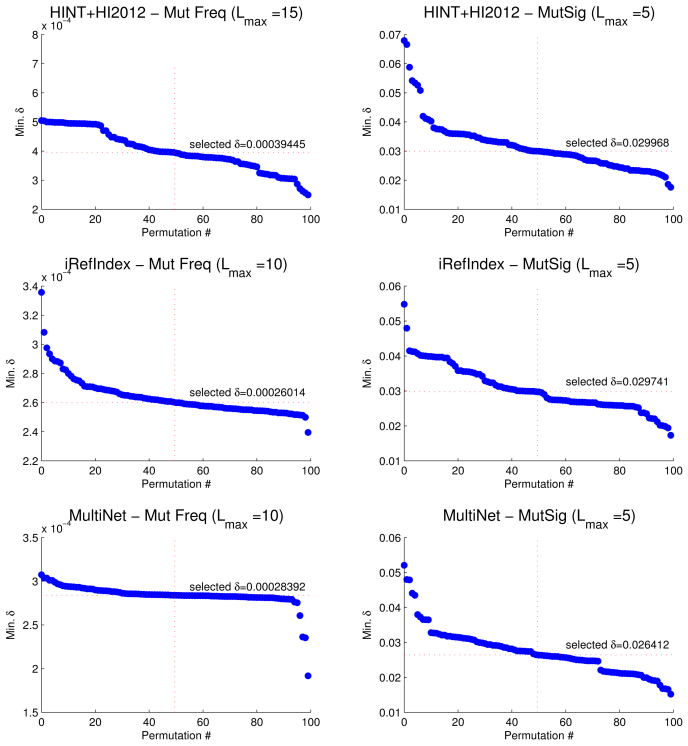



**Figure 4.29:** Distributions used to set the threshold $\delta$. For each combination of interaction network and score, 100 random networks were generated, and the minimum delta resulting in the maximum size of a connected component being $\leq L_{\max}$, for $L_{\max} = 5, 10, 15, 20$, were computed. The median $\delta$ is then used as threshold for HotNet2. For each combination, only the value of $L_{\max}$ used to generate the HotNet2 results is shown. Values of $\delta$ are sorted in decreasing order in each plot. Dashed lines identify the median value of $\delta$, reported in each distribution as well.

significant results ($P \leq 0.01$ for any subnetwork size $k$) on three of six runs (Supplementary Table 32,33), while HotNet2 finds statistically significant results on all six runs. The HotNet2 subnetworks also have a higher fraction of interactions with proteins other than a hot central node (Appendix B.7.1). These differences are explained by the undirected vs. directed heat similarity measures used in HotNet versus HotNet2. We note that the goal of HotNet2 is not to eliminate hot stars/spiders, but rather to reduce the number of such subnetworks that are false positives. We also compared HotNet2 to HotNet on simulated data. In short, the results show that HotNet2 achieves higher sensitivity and specificity than HotNet (Appendix B.7.2 and Figure 4.30).

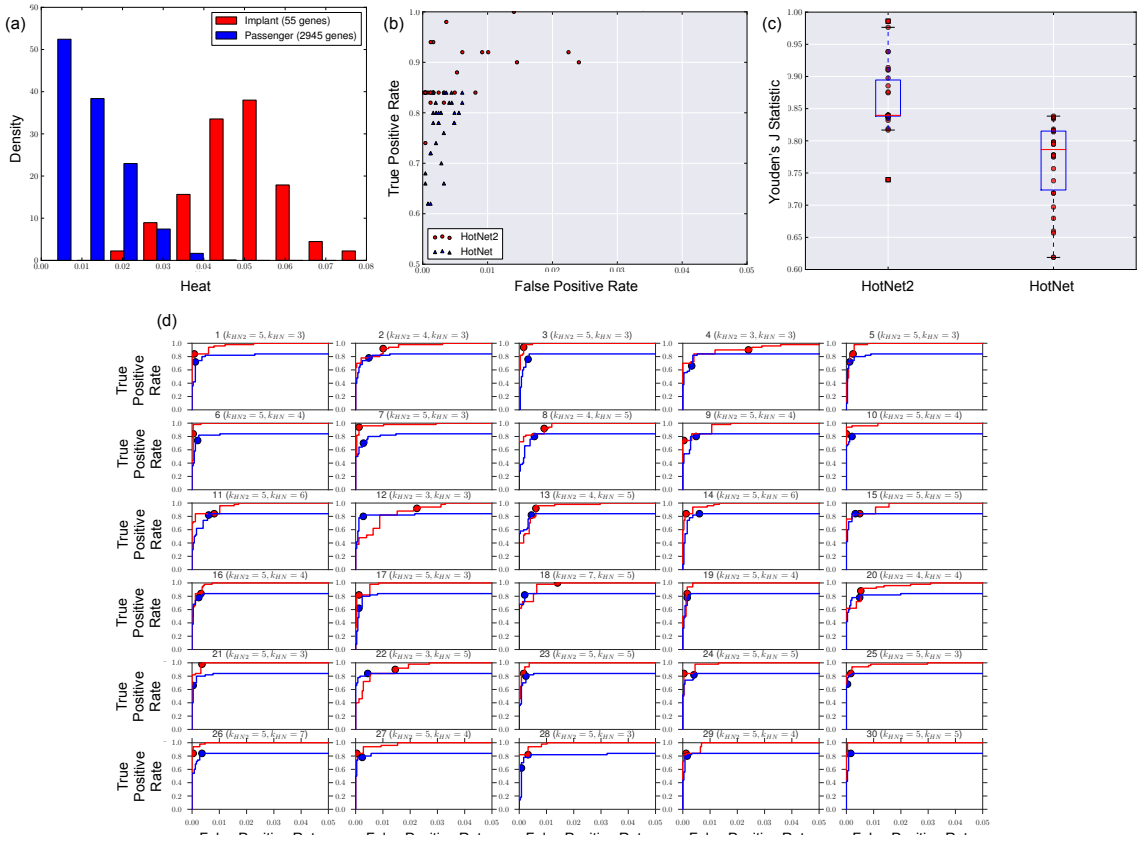



**Figure 4.30:** **(a)** Gene score (heat) distribution for a simulated dataset. The distribution was generated using two overlapping normal distributions, where the implanted genes (red) had a higher mean (0.05) than the mean of the background (passenger; blue) genes (0.01). **(b)** Scatterplot of the false positive rate (x-axis) and true positive rate (y-axis) for HotNet2 (red circles) and HotNet (blue triangles) for identifying the implanted subnetworks in the 30 simulated datasets. **(c)** Boxplots of HotNet2 vs. HotNet's Youden's $J$ = sensitivity + specificity − 1 for identifying the implanted subnetworks in the 30 simulated datasets. **(d)** ROC curves for HotNet2 (red) and HotNet (blue) on the 30 implanted pathway simulation datasets. The true positive and false positive rates were calculated as a function of the minimum edge weight parameter $\delta$. Also shown in the plots are the $\delta$ values automatically selected for each dataset by the HotNet2 and HotNet algorithms.

To further demonstrate the advantages of HotNet2 on the Pan-Cancer mutation frequency dataset, we compared HotNet2 to HotNet and to two standard tests of

pathway enrichment, DAVID [32, 33] and gene set enrichment analysis (GSEA) [31, 30]. We find that HotNet2 provides both new insights and a simpler summary of groups of interacting genes, and is a useful complement (or arguably a replacement for) other pathway tests (Appendix B.8.1). We also show that HotNet2 has much higher specificity than HotNet, DAVID, and GSEA in identifying genes satisfying the 20/20 rule [16] (Appendix B.8.1, Figure 4.31, and Supplementary Tables 34-36). Finally, we find that HotNet2 was more stable than HotNet in identifying 20/20 genes using cross-validation (Appendix B.7.3 and Figure 4.32).

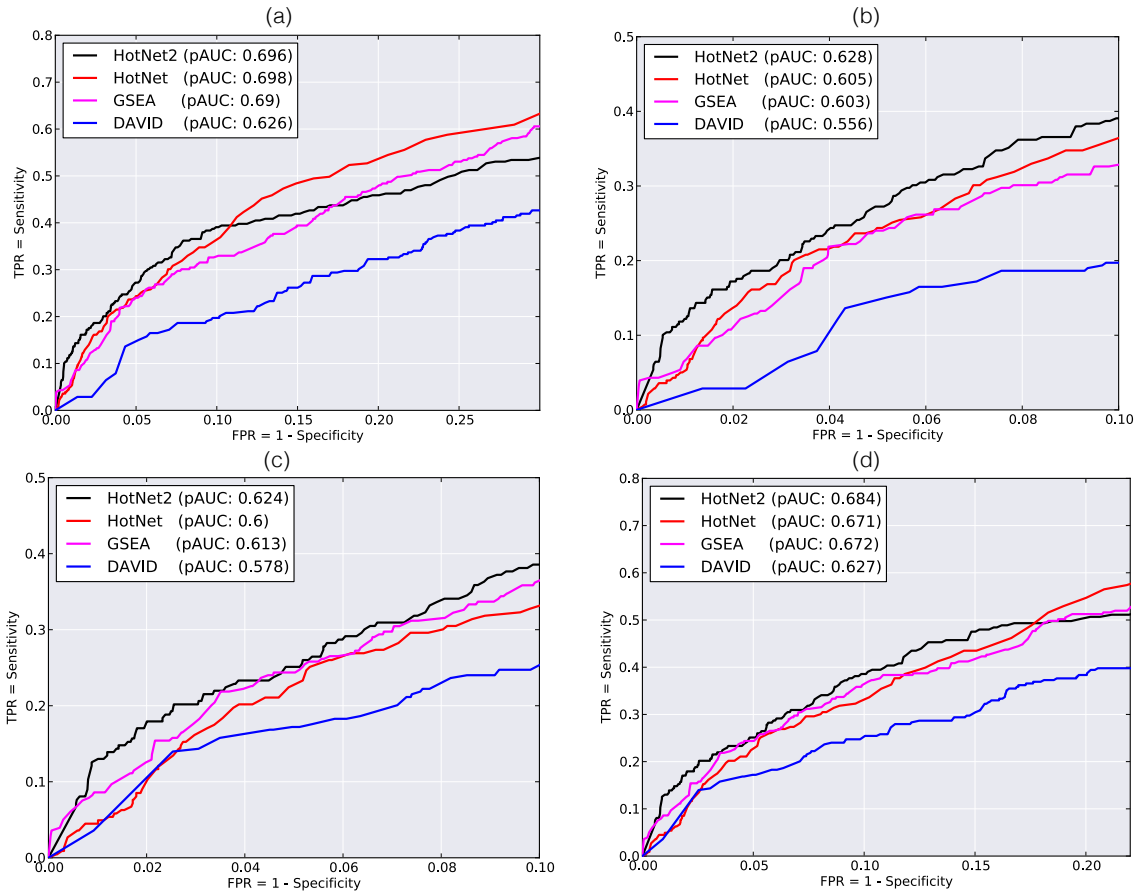



**Figure 4.31:** **(a-b)** The receiver operator characteristic (ROC) curves for HotNet2, HotNet, GSEA, and DAVID in finding 20/20 genes. The ROC is computed using all 11,565 genes in the mutation frequency dataset as input to each algorithm. HotNet2 and HotNet were run on the HINT+HI2012 network. **(a)** ROC restricted to FPR 0.1 (corresponding to $> 1,100$ false positive predictions). **(b)** ROC restricted to FPR 0.3 (corresponding to $> 3,300$ false positive predictions). **(c-d)** The receiver operator characteristic (ROC) curves for HotNet2, HotNet, GSEA, and DAVID in finding 20/20 genes. The ROC is computed using the 6,930 genes in the mutation frequency dataset and the HINT+HI2012 interaction network as input to each algorithm. HotNet2 and HotNet were run on the HINT+HI2012 network. **(c)** ROC restricted to FPR 0.1 (corresponding to around 700 false positive predictions). **(d)** ROC restricted to FPR 0.3 (corresponding to around 2100 false positive predictions).

We attempted to compare HotNet2 to MEMo65, an algorithm to identify groups

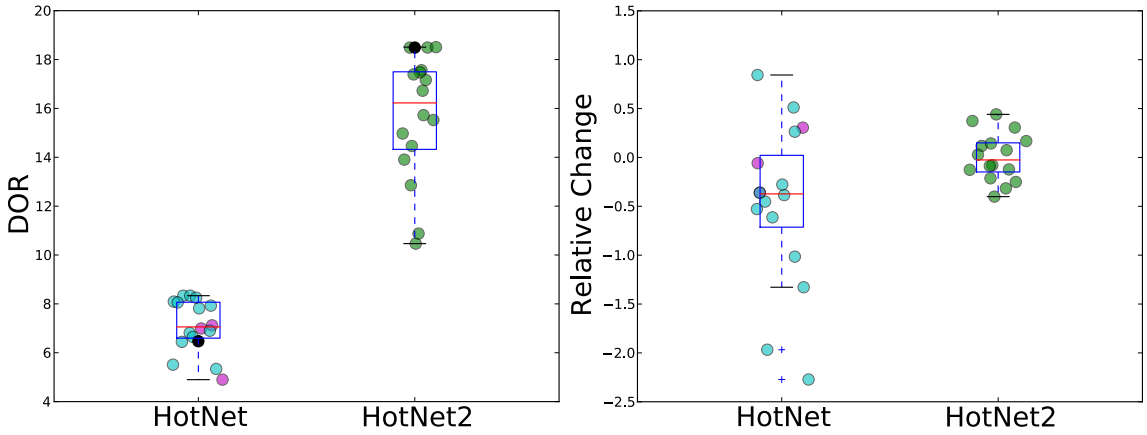



**Figure 4.32:** Two-fold cross-validation comparison of HotNet2 and HotNet. **(a)** The diagnostic odds ratio (DOR) in finding 20/20 genes is higher for HotNet2 than HotNet. Black points indicate DOR on 100% of mutation frequency samples; magenta/cyan points indicate DORs for HotNet results on 50% of mutation frequency samples that were significant/not significant; green dots indicate DORs for HotNet2 on 50% of mutation frequency samples. **(b)** Relative change in DOR across two halves of mutation frequency data, showing higher that HotNet2 has higher stability than HotNet.

of interacting genes with mutually exclusive mutations. First, we note several important difference between HotNet2 and MEMo. Namely, HotNet2 (1) analyzes the mutations and network topology simultaneously; (2) is not restricted to analyzing exclusive mutations and can analyze co-occurring mutations, and (3) can use input heat scores that capture additional information (e.g. functional significance) about the mutations. We found that MEMo was unable to run on the Pan-Cancer mutation frequency dataset, consistent with the authors' recommendation that MEMo should be run only on a small number of significant mutations (details in Appendix B.8.2).

### 4.4.4 Finding consensus subnetworks and linkers

We ran HotNet2 on each combination of gene scores (mutation frequency and MutSigCV20 q-values; see Appendix B.2.2) and interaction networks (HINT+HI201221,22, iRefIndex23, and Multinet24; Appendix B.2.5 and Figure 4.33). We derived "consensus" subnetworks and "linker" genes from the HotNet2 results on the different network and gene scores using an iterative procedure on a weighted graph. This procedure is described in Appendix B.1.5.

We evaluated the statistical significance of the HotNet2 consensus subnetworks using the HotNet2 statistical test on consensus networks found in randomly permuted data. We generate the null distribution of consensus networks by permuting tuples containing the mutation frequency and MutSigCV scores of genes over each

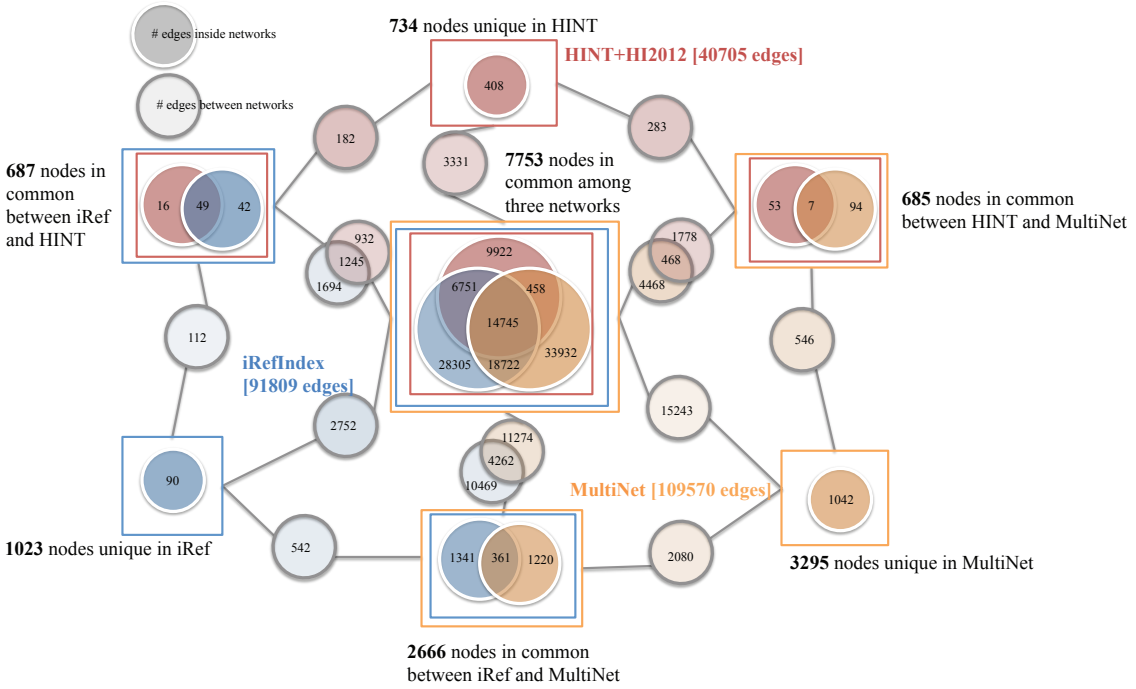



**Figure 4.33:** Overlap between nodes and edges in the HINT+HI2012 [139, 140], iRefIndex [141], and MultiNet [142] interaction networks. Each rectangle encloses the nodes (proteins) that are unique to an interaction networks, shared by two interaction networks, or shared by all three networks (middle). Inside each rectangle is a Venn diagram of the overlap in edges in each interaction network between nodes in the enclosing rectangle. Connections between two rectangles show the number of edges that join nodes in the two rectangles.

of the networks. Thus, the permutation preserves the relationship between the mutation frequency and MutSigCV score. We then ran HotNet2 on the three networks using the permuted mutation frequency and MutSigCV scores forming a "permuted consensus" using the same consensus procedure described above. We used these permuted consensus subnetworks to form an empirical distribution for the statistical test. Additional details of the statistical procedure are in Appendix B.1.3.

### 4.4.5 Expression and germline filtering

Most of the subnetworks (12/14) identified by HotNet2 were also found when we remove the requirement for RNA-Seq expression (Supplementary Table 37). This result demonstrates the robustness and scalability of the HotNet2, as the unfiltered mutation data includes 19,459 genes. Notable among the additional subnetworks identified when we remove the requirement for RNA-Seq expression is a subnetwork (Supplementary Table 25) containing members of the telomerase complex (including TERT and TEP1) that has a well-studied role in cancer66 (Figure 4.34 and Supplementary Table 38). While the lack of RNA-Seq reads from these genes is a concern, we note that the RNA-Seq expression criteria was strict enough to exclude several bona fide cancer genes (See URLs). Thus, the lack of RNA-Seq reads should not automatically exclude these genes from further study. We also ran HotNet2 using a more aggressive criterion to remove potential germline mutations (See URLs). We found only minor differences in the HotNet2 subnetworks (Supplementary Table 39), demonstrating that our reported subnetworks are altered by somatic aberrations in these samples.

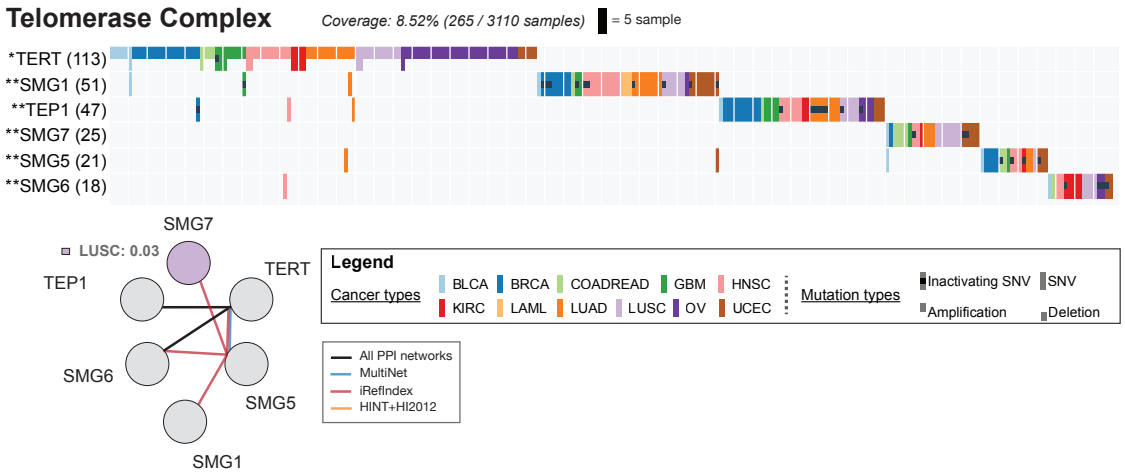



**Figure 4.34:** The telomerase complex subnetwork and its mutation matrix. Representation as in Figure 4.9.

# Chapter 5

# Visualization and Annotation of Aberrations in Cancer

The development of next-generation DNA sequencing technology has led to a super exponential increase in the amount of cancer genomics data. In the public domain, large consortia such as The Cancer Genome Atlas (TCGA) have made available cancer genomics datasets from thousands of tumor samples, while at the same time researchers are now able to sequence their own, smaller tumor cohorts. A major challenge lies in integrating these datasets, such that follow-up studies continually improve our understanding of the large public datasets, and such that researchers can leverage the large public datasets in their analysis of their own cancer genomics data.

In this chapter, we introduce the Mutation Annotation and Genome Interpretation web application for visualization, integration, and annotation of public and private cancer genomics datasets. MAGI creates interactive visualizations of cancer genomics data from different platforms that allow users to view their data at different scales. We show that MAGI is distinct from existing web applications and web portals, first in reducing the computational burden for users to analyze their mutation data with large public datasets, and second in "expert-sourcing" literature annotation of the large public datasets. We also present several case studies for MAGI, demonstrating some key use cases.

The MAGI web application is available online at `http://magi.brown.edu`. We published the MAGI web application in [53], and much of the material in this chapter comes from the publication and its supplementary material.

## 5.1 Overview

We introduce Mutation Annotation and Genome Interpretation (MAGI), an open-source web application (`http://magi.brown.edu/`) for exploration, annotation, and integration of public and private cancer genomics data (Figures 5.1 and 5.2, and Section 5.2). MAGI is the first tool to support interactive, bidirectional user interactions between public and private cancer genomics datasets [170] (Table 5.1). This bidirectional interaction enables researchers to leverage large public datasets in the analysis of their own cancer samples, while also facilitating the expert-sourcing of public datasets through collaborative annotation. The software can also be installed locally (`http://magi.brown.edu/` and Section 5.2.8).

MAGI includes several key features.

1. MAGI generates interactive visualizations of cancer genomics datasets with real-time zooming, panning, and data filtering (Figure 5.1). Visualizations include single nucleotide variants and small indels, copy number aberrations, gene expression, and protein-protein interactions in a query set of genes across samples. The MAGI web application is loaded with genomic data from the TCGA Pan-Cancer study [52] and protein-protein and protein domain annotations from various sources. MAGI also shows categorical and continuous attributes for each sample, and is initialized with survival time, gender, and tumor purity estimates from the Pan-Cancer dataset [52] (Figure 5.1).

2. Users may upload genomics data – including mutations, gene expression, methylation, or sample attributes – directly into MAGI. This private data is available for query, visualization, and annotation in conjunction with public datasets (Figure 5.3). Data upload uses a simple web form with no local software installation required. MAGI automatically generates summaries of user datasets with interactive graphs, sortable tables, and pathway analysis (Figure 5.4).

3. Annotation capabilities facilitate collaborative annotation of mutations and protein-protein interactions in both public and private cancer genomics datasets. Annotations include literature citations, text comments, and votes, and are seamlessly integrated with the interactive visualizations (Figures 5.5 and 5.6). The MAGI website is initialized with 40,000 protein sequence change annotations from the Database of Curated Mutations (`http://docm.genome.wustl.edu/`) and from PubMed Central searches.

4. A sample view shows all sample features and annotations of aberrations in a given (public or private) tumor sample (Figure 5.7). An annotation score prioritizes the display of aberrations with annotations in the MAGI database. This view leverages information across many samples for single sample ("N=1")

analyses. The sample view also reveals unusual samples with few known aberrations, encouraging further investigation and annotation of these samples.

5. Bookmarking features enable sharing of interactive views (optionally with private data) directly with colleagues or collaborators. MAGI exports publication-quality graphics of views.

6. A compute engine provides interactive computation of statistical tests of association between mutations and sample annotations for both public and user-uploaded datasets.

We demonstrated MAGI's features by uploading TCGA stomach adenocarcinoma (STAD) data [55] and analyzing the STAD aberrations together with TCGA Pan-Cancer data [52]. We identified mutations in the TGF-$\beta$ pathway that are strikingly similar to those in TCGA colorectal (COADREAD) data [171] (Figure 5.8) but were not reported in either TCGA publication [171, 55] (Section 5.4).

The Cancer Genome Atlas (TCGA), International Cancer Genome Consortium (ICGC) and other large-scale sequencing efforts have produced valuable datasets of genomic aberrations in many cancers. At the same time, rapidly declining DNA sequencing costs now allow individual investigators to sequence cancer genomes. MAGI reduces the barriers to combining public and private cancer genomics datasets. The advantages of combining data flow in both directions. First, the large number of samples in public datasets can inform the analysis of private datasets; e.g. increasing the power to identify recurrent genomic aberrations [17] or combinations of aberrations in pathways and networks [172]. Second, private datasets can reveal additional insights about the public datasets themselves; e.g. demonstrating a functional role for a rare genomic aberration in TCGA samples. Using MAGI, such insights can be linked directly to the underlying public genomics data, rather than scattered in the literature and disconnected from the data. By connecting the information in large public datasets with insights from smaller private datasets and the literature, MAGI enhances the value of all data sources and improves the characterization of genomic aberrations in individual samples.

In the remainder of this chapter, we provide more detail on MAGI. We first provide more details on the web application and the data loaded into MAGI in Section 5.2. We then describe several case studies in which we apply MAGI to real data in Section 5.4. We finally conclude with a short discussion in Section 5.5.


**Figure 5.1:** Screenshot of the MAGI web application displaying mutations in the Notch signaling pathway from TCGA Pan-Cancer dataset. (i) The aberrations view shows the pattern of mutations in the gene set across tumor samples. (ii) Additional rows in the aberration and heatmap views show attributes for each sample; displayed here are survival time, gender, and estimated tumor purity. (iii) Users can upload and display continuous-valued data (e.g. expression or methylation data) in a heatmap. Shown here are RNA-Seq differential expression values for TCGA Pan-Cancer dataset. (iv) The network view shows the interactions among genes from multiple interaction networks. (v) The transcript view shows the locations and types of the mutations in a transcript of a given gene. (vi) The copy number aberrations view shows the amplified or deleted segments across tumor samples in a given gene. (g) Annotations are added using a simple web form that records mutation/interaction type, a PubMed ID (PMID), and/or free-text comment. Clicking on an element in any of the views populates this form with the corresponding mutation or interaction, simplifying data entry.

| Feature | MAGI | Cancer Regulome | cBioPortal [173, 174] | CGWB | UCSC Cancer Genomics Browser [175] |
|---|---|---|---|---|---|
| Allows private data (upload tool) | Yes (web upload) | No | Yes | Yes (Java Web Tool) | Yes (web upload) |
| Combine private and public data | Yes | No | No | Yes | Yes |
| Collaborative annotation | Yes | No | No | No | No |
| Interactive (zoom, pan, and filter data) | Yes | No | Yes | No | No |
| Aberrations matrix | Yes | No | Yes | Yes | Yes |
| Heatmap | Yes | No | No | No | Yes |
| Transcript plot | Yes | Yes | Yes | Yes | No |
| PPI networks | Yes | Yes | Yes | No | No |
| CNA browser | Yes | No | No | Yes | Yes |
| Sample annotations (e.g. clinical data) | Yes | No | Yes | No | No |
| Statistical tests of association | Yes | Yes | Yes (survival) | No | No |
| Interactive linking of visualizations | Yes | No | No | No | No |

**Table 5.1:** Comparison of features offered by MAGI and similar web tools.


**Figure 5.2:** Screenshot of the MAGI home page and query interface. (i) Users select a combination of (public or private) datasets to query. (ii) Users can enter up to 25 genes to query at once. (iii) Alternatively, users can view the mutations in a single sample (Figure 5.7).

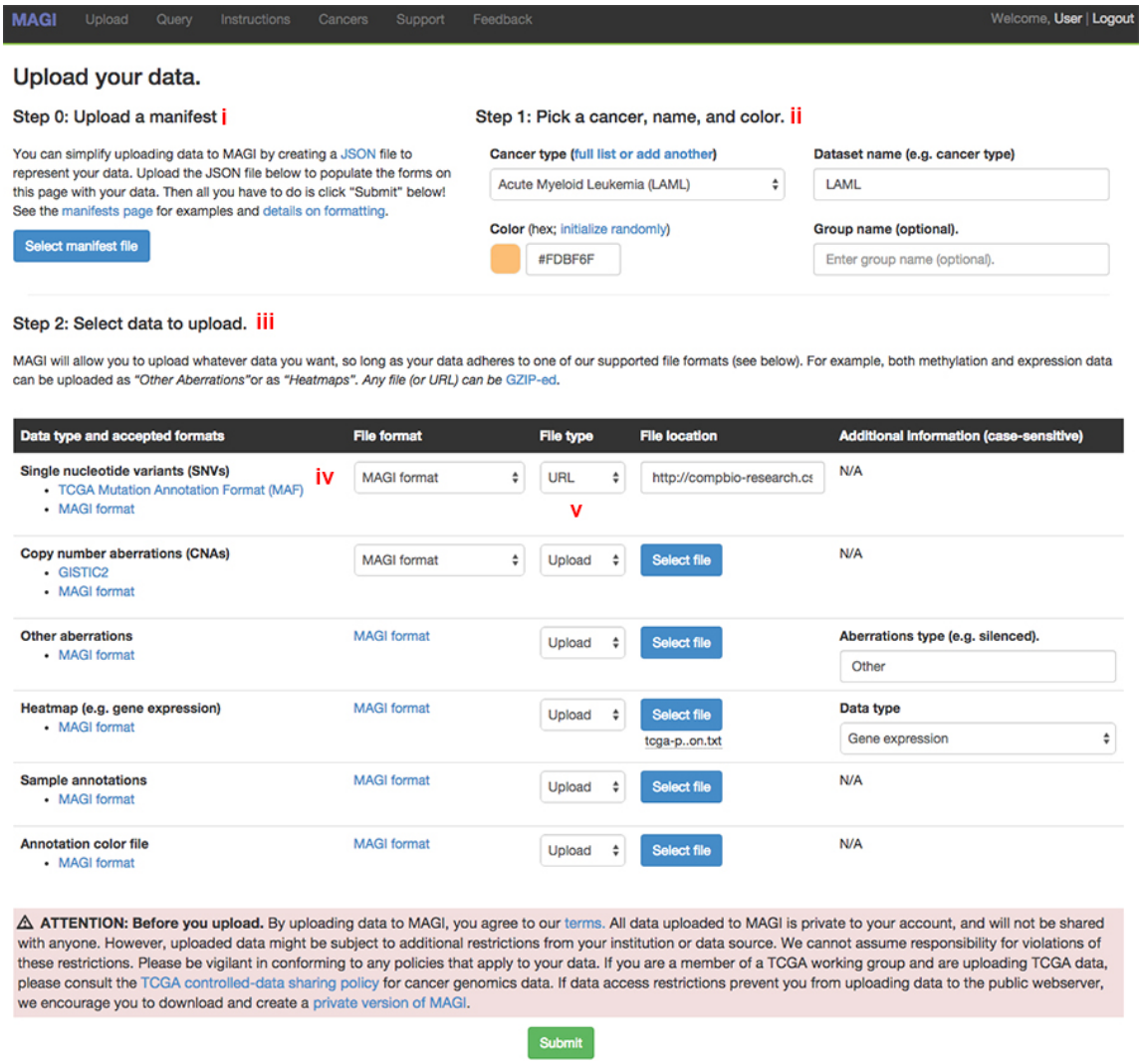



**Figure 5.3:** Screenshot of MAGI data upload interface. (i) Users can choose to upload a single JSON manifest file that includes the URLs for all the data files in a given dataset. (ii) Users select a cancer type for their dataset either from a predefined list of TCGA/ICGC types or by adding their own – ensuring every mutation annotation maps to a particular cancer type. (iii) Users select the data files they want to load into MAGI. MAGI supports six different data types: SNVs, CNAs, other aberrations, heatmap (continuous values), sample annotations, and sample annotation colors. (iv) Users can also upload SNVs in TCGA MAF and CNAs in GISTIC2 format. (v) All data files can be provided as URLs, or uploaded directly to MAGI.

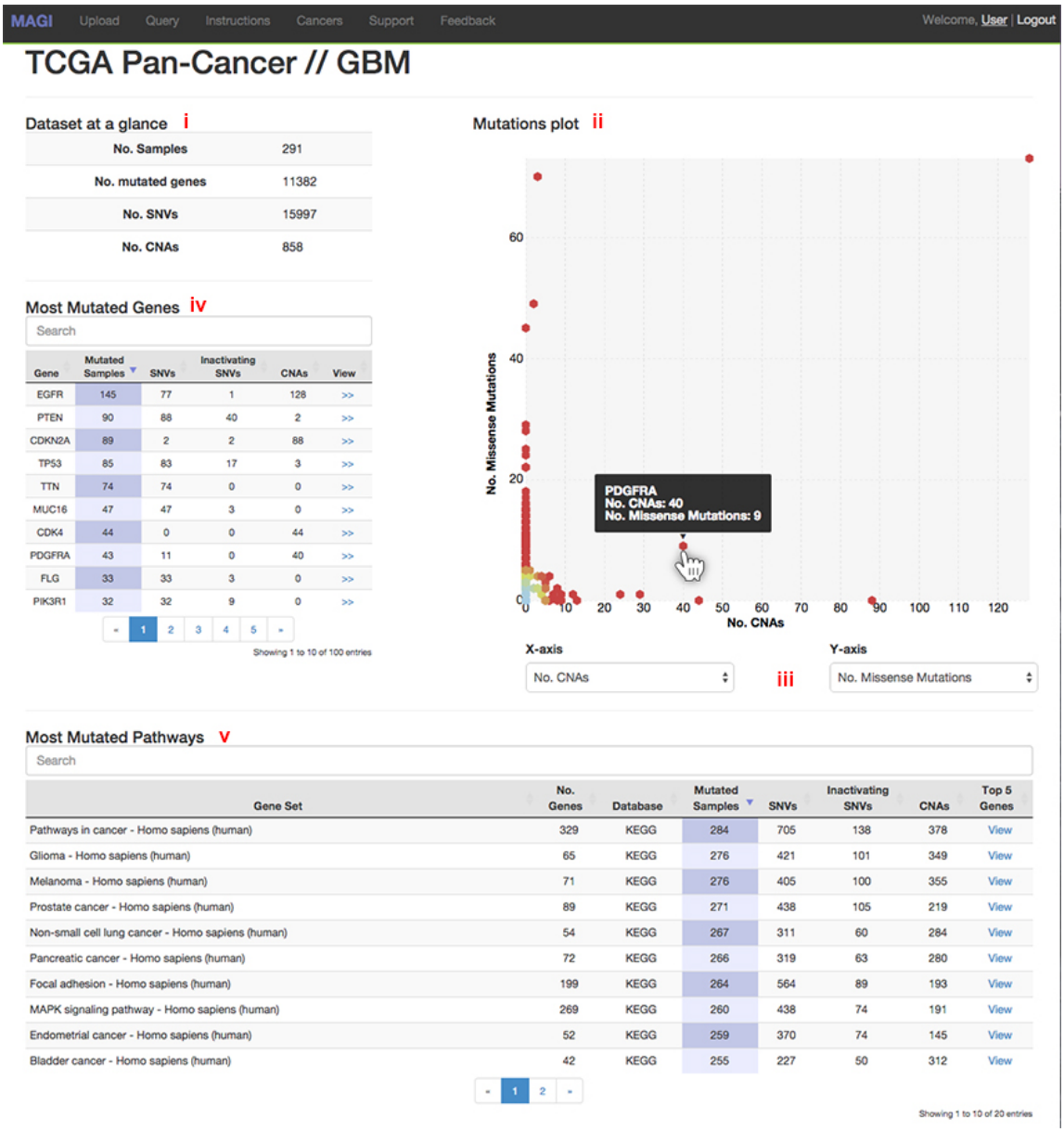



**Figure 5.4:** Summary of the glioblastoma (GBM) samples from the TCGA Pan-Cancer dataset automatically produced by MAGI. Users can view these summary pages for public datasets or their own uploaded private data. (i) Summary statistics for the mutation data. (ii) Plot of the number of the mutations in each gene in the dataset. The plot is interactive, as users can zoom in and out or scroll over points to get additional information. (iii) Users can also choose from 11 mutation categories for the x- and y-axes (number of SNVs and number of CNAs by default) of the mutations plot. (iv) Sortable and searchable table of the genes in the dataset. (v) Sortable and searchable list of pathways with the most mutations in the dataset. Pathways are from the KEGG and PINdb databases.

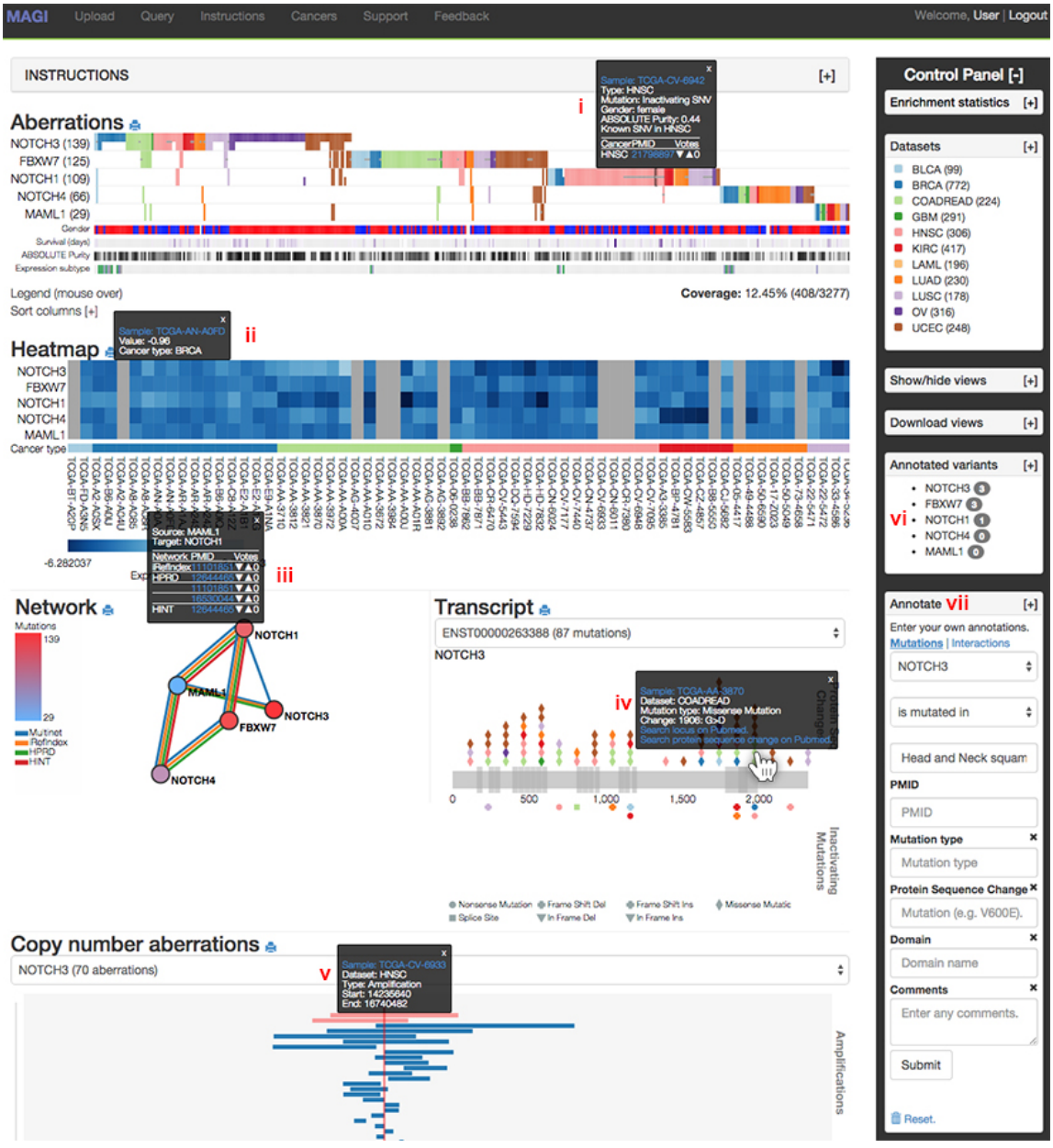



**Figure 5.5:** Screenshot of annotation MAGI annotation interface. Annotations of (i) genes (ii) interactions, (iii) mutated residues, and (iv) copy number aberrations are viewable as tooltips in each view. MAGI displays annotations interactively as users mouse over different views. Users can also upvote or downvote existing annotations in the tool tips. (v) Users can click on individual data points (e.g. interactions in the network view) to pre-load the annotation form on the right.

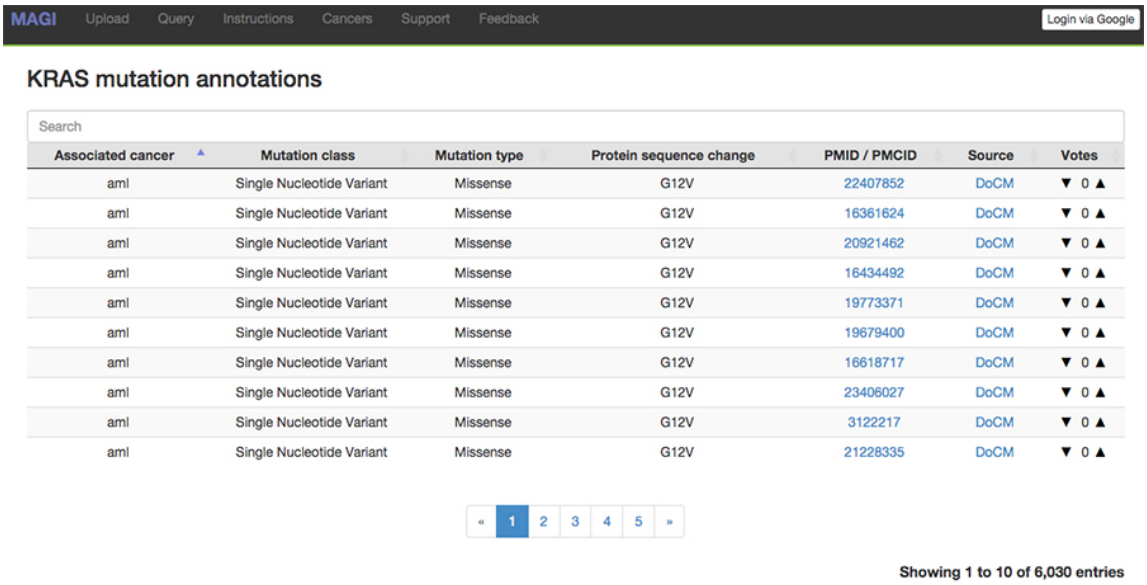



**Figure 5.6:** The MAGI gene annotation page shows a table listing all the annotations in a given gene. The PubMed or PubMed Central ID, the source, and (optionally) the associated cancer, mutation class, mutation type, and protein sequence change are shown for each annotation. Logged-in users can upvote or downvote any of the annotations.

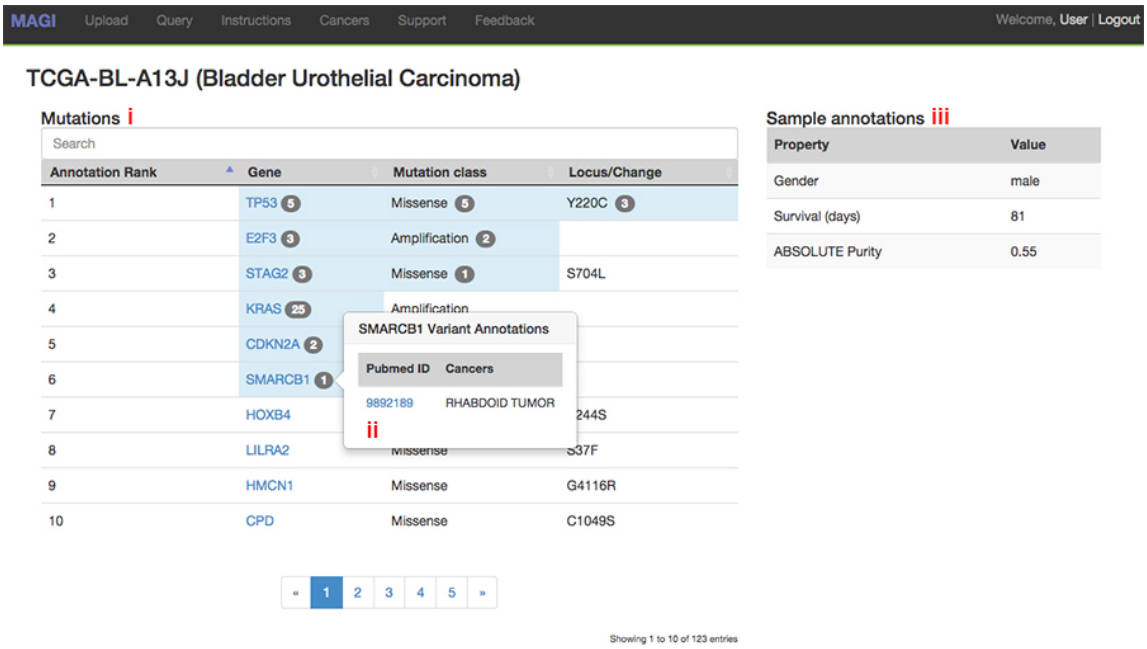



**Figure 5.7:** MAGI's tumor sample view. (i) The tumor sample view lists all the aberrations in the given tumor sample in a table. The aberrations are displayed are represented by the gene, mutation class, and locus or protein sequence change, and are ordered by the score of the annotations for that aberration in the MAGI database. The badges next to highlighted genes/classes/loci list the number of annotations for the gene/class/locus. (ii) Users can click on any of the badges to view the annotations in more detail. (iii) The page also includes a table of the attributes (e.g. gender) for the given sample.


**Figure 5.8:** Transcript plots for mutations in SMAD2 (top) and SMAD4 (bottom) in the TCGA gastric (STAD; dark green) and Pan-Cancer datasets. (top) Three of the four mutations in SMAD2 in the TCGA STAD dataset are inactivating nonsense mutations, while the fourth mutation is a missense mutation that occurs in the same location in the MH2 binding domain as two missense mutations in the TCGA colorectal cancer (COADREAD; light green) dataset. (bottom) All of the mutations in SMAD4 in the TCGA gastric dataset occur in the MH2 binding domain. In addition, 12 of the 49 mutations in SMAD4 are mutations at position 361. (Note that only 7 of these mutations are visible in the screenshot.)

## 5.2 Methods

MAGI (`http://magi.brown.edu` or installed locally) is a web-based platform for creating custom interactive visualizations of cancer mutation data and enabling the collaborative annotation of this data. The visualizations integrate mutation data from one to thousands of tumors with publicly available annotations of genes and proteins including protein domains and protein-protein interactions. MAGI has several key features.

1. MAGI allows the querying of mutation data from TCGA Pan-Cancer project [52] for any combination of genes (Figure 5.2), detailed in Section 5.2.1.

2. MAGI generates multiple visualizations of genomics data across thousands of samples including an aberrations view, heatmap view, transcript view, network view, and copy number aberration view. These views illustrate relationships between mutations across multiple biological scales from genome and protein sequence through protein domains and protein-protein interactions. The visualizations are dynamic and interactive, allowing users to explore their data by filtering, sorting, and rescaling the visualizations. Any displayed view can be exported in vector graphics format for subsequent presentation/publication. Each view is described further in Section 5.2.2.

3. MAGI includes a straightforward system to upload mutation data from additional samples into a private database. This private data can be queried and analyzed in combination with public data. Users can also upload additional aberration data for TCGA samples (e.g. methylation data), and query this data alongside TCGA and/or private mutation data. Users can share links to views generated with public and/or private data to colleagues. This is further detailed in Section 5.2.4.

4. MAGI enables collaborative annotation of aberrations in individual genes or transcripts as well as protein interactions using an interactive, context-aware system that makes it easy to add publication support (PubMed IDs) or free-text annotations. This is further detailed in Section 5.2.5.

5. MAGI provides a sample view that shows all sample features and annotations of aberrations in a given (public or private) tumor sample. An annotation score prioritizes the display of aberrations with annotations in the MAGI database. This leverages information across many samples of "N=1" analyses. Moreover, the view reveals unusual samples with few known aberrations, encouraging users to further investigate and annotate these samples. This is further detailed in Section 5.2.6.

6. MAGI provides interactive computation of statistical tests of association between mutations and sample annotations across a combination of public and

user-uploaded datasets, further detailed in Section 5.2.7.

Table 5.1 provides a comparison of MAGI and other cancer genomics browsers and web applications.

MAGI is publicly available and accessible through any modern web browser with Javascript enabled. No local installation of software or browser plugins is necessary. Thus, MAGI places no computational burden on its users beyond the ability to use a web browser. MAGI generates visualizations directly in the browser using the Data Driven Documents (D3) [176] Javascript framework in scalable vector graphics (SVG) format. The MAGI web server is written in Node.js using a MongoDB database, with web services provided by Nginx (See Table 5.2 for additional details on the implementation). We show a schematic of how MAGI uses these technologies to respond to queries in Figure 5.9. Using these technologies, MAGI can display mutation data for thousands of samples for dozens of genes in under a second. The source code for MAGI is publicly available on GitHub (`http://github.com/raphael-group/magi`; See Section 5.2). We also provide an Amazon machine image containing a preconfigured webserver for easy deployment of private MAGI installations (Details in Section 5.2).

| Component | Description | Technology | Framework | Language |
|---|---|---|---|---|
| Backend | Web server | Node.js | Express | Javascript |
| | Database | MongoDB | Mongoose | Javascript |
| | User authentication | OAuth2 | Passport.js | Javascript |
| | HTML templating | Jade | Express | HTML, CSS |
| Frontend | Visualizations | D3 | N/A | Javascript |
| | Interaction | D3 | N/A | Javascript |
| | | jQuery | N/A | Javascript |

**Table 5.2:** Technologies used by MAGI.

### 5.2.1 Integration of mutation and annotation data

MAGI integrates mutation data, including non-synonymous single nucleotide variants (SNVs), small indels, copy number aberrations (CNAs) and predicted fusion genes, and gene expression data with publicly available annotations of the genome, transcriptome, and proteome. MAGI is initialized with two mutation datasets:

1. TCGA Pan-Cancer project [52]: 18,526 mutated genes in 3110 tumor samples from twelve different cancer types. Copy number aberrations were ex-

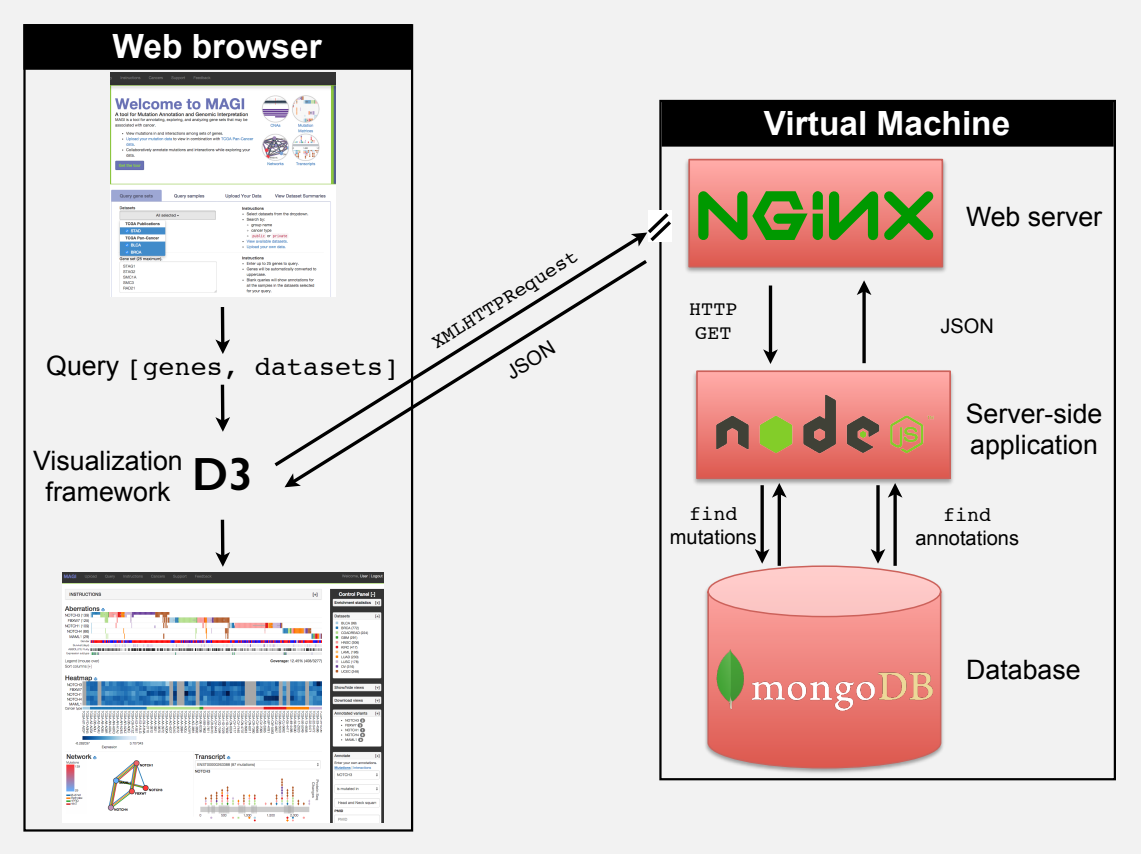



**Figure 5.9:** Schematic of the software technologies used in MAGI. MAGI uses D3 client-side to load the mutations and annotations from a web server running Nginx. The web server constructs a JSON object payload using the Node.js framework, which performs the query and collates the resulting data from MongoDB.

tracted from GISTIC2 [23] output via FireHose, using recurrently aberrant regions from the GISTIC2 maxpeaks. We restricted aberrations to those within 50kb of regions with significantly recurrent aberrations in one or more cancer types, or across cancer types. MAGI also includes expression data from the TCGA Pan-Cancer project downloaded from syn1715753. Samples from the TCGA Pan-Cancer dataset are annotated with clinical data (survival time and gender) downloaded from Firehose (`http://gdac.broadinstitute.org/runs/stddata__2014_02_15/data/PANCAN12/20140215`), and purity estimates from the ABSOLUTE algorithm [177] downloaded from syn1710466. TCGA breast cancer (BRCA) samples are annotated with gene expression subtypes from [8], downloaded from `http://tcga-data.nci.nih.gov/docs/publications/brca_2012/BRCA.547.PAM50.SigClust.Subtypes.txt`.

2. TCGA Gastric Cancer Project [55]: 9,316 mutated genes in 215 tumor samples. Mutations and copy number aberrations were downloaded from syn1725886. CNAs were extracted from GISTIC2 output. We excluded 74 tumor samples classified as hypermutators in [55]. We also downloaded clinical data (survival time and gender) and tumor purity estimates from the ABSOLUTE algorithm [177] from syn1725886.

We provide these datasets online on the MAGI website (see below).

MAGI maps copy number aberrations to gene locations on the hg19 reference genome. Mutations in each gene were mapped to a single ENSEMBL transcript using the vcf2maf software package (`https://github.com/ckandoth/vcf2maf`). MAGI includes 44 cancer type abbreviations from TCGA (`https://tcga-data.nci.nih.gov/datareports/codeTablesReport.htm`) and ICGC (`https://dcc.icgc.org/projects`), such that samples of the same cancer type are colored consistently across views. Because MAGI uses standard file types (MAF files and GISTIC output) and cancer types used in TCGA, it is easy to update the TCGA data in MAGI directly from FireHose output.

MAGI is pre-loaded with ∼40,000 annotations of known cancer variants. Each of these annotations maps a PubMed or PubMed Central (PMC) ID to a variant (protein sequence change). MAGI includes nearly 4,000 annotations from the Database of Curated Mutations (DoCM; `http://docm.genome.wustl.edu/`). We generated the remaining annotations by performing text searches using the PMC API for each of the protein sequence changes in the TCGA Pan-Cancer and TCGA STAD datasets.

MAGI also includes annotations – protein-protein interaction networks and protein domain datasets – that describe features, properties, or interactions between genes. MAGI is initialized with interactions from four different protein-protein interaction networks – HINT [139], HPRD [178], iRefIndex [141], and Multinet [142] – which comprise 157,026 interactions among 16,448 proteins. MAGI also includes

174,362 protein domains in 64,186 transcripts from the Conserved Domain [179], PFAM [180], and SMART [181, 182] protein domain databases. We describe the format of the data used by MAGI in Table 5.3.

### 5.2.2 Visualization components

Given a query gene set $G$ and a collection of mutation datasets, MAGI generates five types of visualizations, or views. The views are generated using our publicly available GD3 library (`https://github.com/raphael-group/gd3`), a library of genomic visualization elements written using the D3.js library [176]. These views are each displayed as part of a single-page web application (Figure 5.1). The views are interactive and linked (described further in Section 5.2.3). We describe the views in detail below.

- *Aberrations view.* The aberrations view is a matrix indexed by genes (rows) and samples (column) (Figure 5.1). The matrix displays binary aberration data; in each sample, a gene is marked as containing an aberration or not. By default, the aberrations view shows the SNVs and CNAs in the query gene set. For each mutation in each gene, the corresponding cell in the aberrations matrix is marked with a colored shape that indicates the type of mutation (SNV, inactivating SNV, amplification, deletion, or fusion). Cells may be colored by the sample type; the standard TCGA Pan-Cancer color scheme [52] is used for these cancer types and users may define colors for their own samples. Aberration matrices that include only a single mutation dataset color the shape based on whether or not the sample is mutated in one gene (mutually exclusive) or more than gene (co-occurring) in $G$ (Figure 5.10). The samples (columns) of the aberrations view are dynamically sortable based on attributes of the genes, the samples, and the mutations themselves. The aberrations view is interactive, and supports zooming and panning such that users can view the pattern of mutations in a gene set across thousands of tumors, or restrict their view to the mutations in individual tumors. Users can also dynamically filter the aberrations by toggling on and off different mutation types.

  The aberrations view offers researchers the opportunity to view mutations in any combination of genes to facilitate understanding of how the genes are mutated within individuals and/or across cancer types. We demonstrate this view using the mutations in eight genes of the SWI/SNF complex in the TCGA Pan-Cancer dataset (Figure 5.11). The aberrations view generated by MAGI demonstrates visually that this complex is enriched for mutations in kidney cancer (KIRC); in fact, 153/592 (25%) samples with mutations in this complex occur in KIRC (Figure 5.11). Furthermore, MAGI demonstrates that most of the inactivating SNVs or deletions occur in KIRC, bladder, or endometrial

| Visualization Component(s) | Data | Description | Example |
|---|---|---|---|
| Aberrations view and transcript view | SNVs and small indels | Gene symbol | *ARID1A* |
| | | Sample ID | TCGA-A6-2676 |
| | | Transcript ID | ENST00000324856 |
| | | Transcript length | 2285 |
| | | Position | 1262 |
| | | Mutation type | Missense |
| | | Original Amino Acid | R |
| | | New Amino Acid | H |
| Aberrations view and CNA view | CNAs | Gene symbol | *ARID1A* |
| | | Sample ID | TCGA-04-1367 |
| | | Type (amplification or deletion) | DEL |
| | | Left point | 19153764 |
| | | Right point | 33238714 |
| Transcript view | Protein domain annotations* | Dataset | PFAM |
| | | Transcript ID | ENST00000324856 |
| | | Domain name | DUF3518 |
| | | Domain start | 1975 |
| | | Domain end | 2231 |
| CNA view | Genome annotations* | Gene symbol | *ARID1A* |
| | | Chromosome | 1 |
| | | Start | 27022521 |
| | | End | 27108601 |
| Network view | Protein-protein interactions | Source (gene symbol) | *ARID1A* |
| | | Target (gene symbol) | *SMAD2* |
| | | Network name | iRefIndex |
| Heatmap view | Continuous-valued data (e.g. gene expression, DNA methylation) | Gene | *ARID1A* |
| | | Sample | TCGA-A6-2676 |
| | | Values | 0.40 |
| Aberrations view | Binary data | Gene | *ARID1A* |
| | | Sample | TCGA-A6-2676 |
| Sample annotations | Categorical (e.g. expression subtype) or continuous (e.g. tumor purity) data | Sample | TCGA-A6-2676 |
| | | Value (continuous) | 0.88 |
| | | Value (categorical) | HER2-Enriched |

**Table 5.3:** Data types included in MAGI, organized by the visualization component in which the data is used. Users can upload all data types, except those indicated with an asterisk (*).

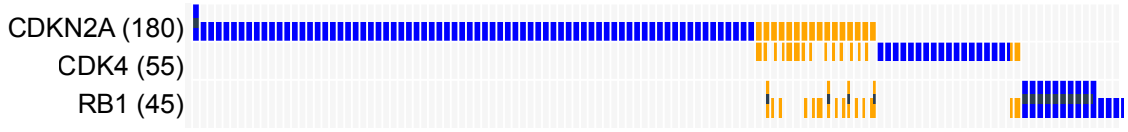



**Figure 5.10:** Aberrations view of the mutations in the CDKN2A, CDK4, and RB1 genes in the glioblastoma tumors from the TCGA Pan-Cancer dataset. Full ticks represent SNVs while black stripes represent inactivating SNVs, downticks represent deletions, and upticks represent amplifications. Exclusive mutations – those that occur in only one gene in the gene set – are colored blue, and co-occurring mutations are colored orange.

cancer (207/301 samples), which is easily accomplished by toggling off the other cancer types.

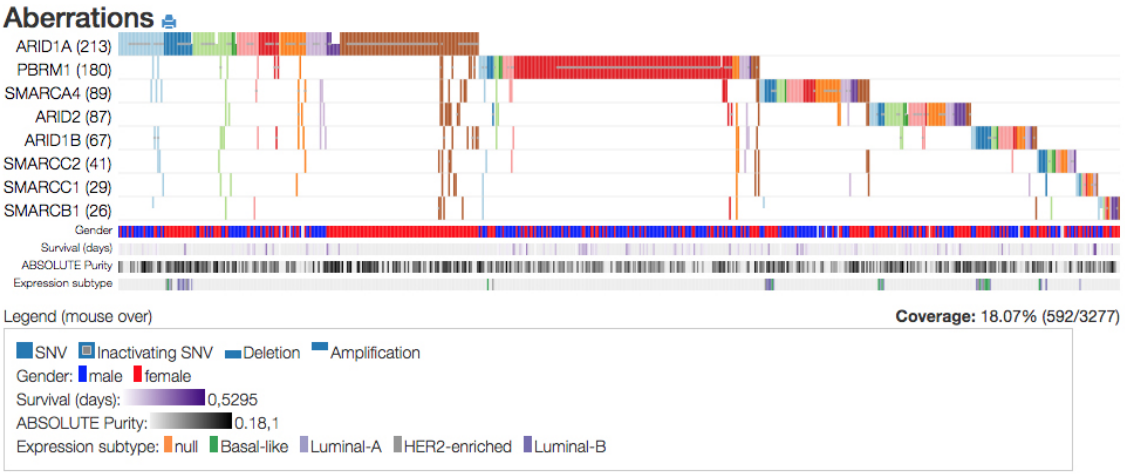



**Figure 5.11:** Aberrations view of genes in the SWI/SNF complex from the TCGA Pan-Cancer dataset. The genes are enriched for mutations in kidney cancer (KIRC, red) and endometrial cancer (UCEC, brown). Most of the inactivating SNVs and deletions in these genes occur in the BLCA, KIRC, and UCEC cancer types, which is easily seen in MAGI by toggling off the other cancer types in the sample type legend on the right (not shown).

An additional use case for the aberrations view is in examining the pattern of mutations in a gene set across a cohort of tumors (Figure 5.10). The aberrations view makes it simple to determine visually if a set of mutations are mutually exclusive or co-occurring. Mutually exclusive mutations may indicate that a set of genes is in the same functional pathway [38, 39]. For example, in Figure 5.10, mutations across 159 glioblastoma samples in three genes (CDKN2A, CDK4, RB1) in the Rb signaling pathway are shown. Only 3/159 samples have mutations in more than one gene, and indeed, these genes have been previously shown to have significant mutually exclusive mutations in glioblastoma [1]. An alternative explanation for mutual exclusivity is that some genes are targeted more often in different cancer (sub)types. This is easy to observe using MAGI as mutations are color-coded by (sub)type, such as in Figure 5.11, where

PBRM1 is mutated predominantly in kidney cancers (131/180 mutations).

- *Heatmap view.* MAGI generates a heatmap view for visualizing continuous-valued data in the query gene set across a cohort of samples (Figure 5.1). By default, the heatmap shows gene expression data in the samples in which the genes are mutated. The heatmap is not restricted to displaying gene expression data; users can upload any continuous-valued data to MAGI as a matrix (e.g. DNA methylation data).

- *Transcript view.* MAGI generates transcript views for each transcript in each gene in the query set $G$ (Figure 5.1). The transcript plot shows the locations of mutations for a given gene in the protein sequence and constituent protein domains. The protein sequence and domains are shown in the middle of the plot, with in-frame and missense mutations shown above the sequence and inactivating mutations (nonsense, nonstop, splice-site, and frameshift insertions/deletions) shown below. Mutations are represented by different colored symbols, where each symbol represents a different mutation type (e.g. missense or nonsense) and is colored by the dataset in which the mutation occurred. The transcript plot is interactive: users can dynamically switch the protein domain database used to annotate the protein sequence, toggle on and off different mutation types, and zoom and pan along the protein sequence to view mutations at the individual base-pair level, or across hundreds or thousands of base-pairs.

  The transcript view is especially useful for addressing a common challenge in cancer genomics: determining the functional impact of mutations. Methods to determine the functional impact of non-synonymous SNVs fall into three categories, and involve determining (1) whether the mutations are inactivating; (2) if there is a cluster of these mutations at a particular position, which suggests that these mutations are activating [16]; or (3) if the mutation affected a known protein domain. This latter method is especially useful when examining sets of genes with the transcript view, as it is possible to determine if mutations are targeting an (multiple) interaction domain(s) between pairs of proteins.

  To demonstrate the transcript plot, we examined the mutations in BRAF in glioblastoma tumors (GBM) from the TCGA Pan-Cancer dataset (Figure 5.12). BRAF is a well-known oncogene and the V600E mutation, a valine to glutamic acid substitution at residue 600, is a driver mutation in multiple cancer types, including 60% of melanomas [183] as well as in colon cancer [184]. In GBM, we find that BRAF mutations are rare with only 7/290 (2.4%) samples having a BRAF mutation. However, 5/7 (71%) of the mutations in these samples are V600E mutations. This demonstrates the benefits of interactive, exploratory analysis of TCGA data.

- *Network view.* The network view (Figure 5.1) shows the interactions among the protein products of the genes in the query gene set $G$. The subnetwork plot is a multigraph, where each gene (node) is connected by one or more edges, each edge indicating a known protein-protein interaction from one of the four

interaction networks11-14 included in MAGI. Genes in the graph are colored by the number of samples in which they are mutated. The networks used in the plot can be toggled on or off by the user. The positions of nodes are initially determined by the force-directed layout from D3 [176] (layout based off of [185]), but nodes can be moved and/or fixed to create custom static layouts. Each edge is annotated with PubMed IDs (PMIDs) of publications that support the underlying interaction.

- *Copy number aberration (CNA) view.* MAGI generates a copy number aberration view (Figure 5.1) for each gene in $G$. The CNA view shows the amplified or deleted segments adjacent to a particular gene in each sample. The middle of the view shows all genes within 500kb of the query gene. The top (respectively bottom) of the view shows the amplified (respectively deleted) segments in each sample with at least one CNA in the region. Each segment is a rectangle that spans the genomic region that is amplified or deleted, and is colored by the dataset of the sample. The CNA view allows for zooming and panning, such that users can determine if the amplified or deleted regions are focused around the particular gene, or span many genes.

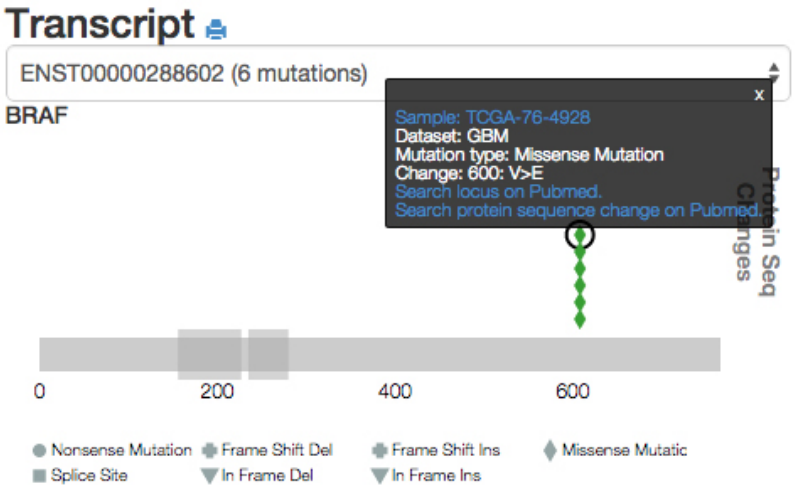



**Figure 5.12:** Mutations in *BRAF* in glioblastoma tumors in the TCGA Pan-Cancer dataset. Shown is the screenshot of the transcript plot of *BRAF* mutations in transcript ENST00000288602. Each diamond indicates a missense mutation in an individual sample – five of six are clustered at position 600.

Since somatic copy number aberrations often include multiple genes and typically vary greatly in size and position across different samples, it is difficult to determine which, if any, of the genes in a copy number aberration is the target of the aberrations. Methods such as GISTIC2 [23] identify recurrently amplified or deleted segments of the genome, and in some cases predict the genes that are the targets of these aberrations. The MAGI CNA view can be useful in identifying likely targets of recurrent aberrations and rare aberrations missed by these computational approaches:

1. Missed targets of recurrent copy number aberrations. The MAGI CNA view allows the user to see the entire genomic region in which a set of CNAs occur, including the genomic neighbors of the target identified by a computational approach. This can help users quickly identify likely mistakes in the computational approaches method for assigning targets to recurrent CNAs. For example, we viewed the aberrations the FGF19 gene from the TCGA STAD dataset (Figure 5.13a). GISTIC2 identified FGF19 as the target of amplifications in 30/215 (14%) of STAD samples, potentially a novel discovery, as amplifications in FGF19 have not been reported in gastric cancer before. However, FGF19 is located just 44kb from CCND1 on chromosome 11 (Figure 5.13b). CCND1 is a well-studied cancer gene and is frequently amplified in cancer [186], and thus seems likely to be a target of these amplifications. Furthermore, all 12 amplifications that span FGF19 also span CCND1. MAGI makes it easy for users to quickly identify potential missed targets of CNAs by computational approaches.
2. Rare copy number aberrations. The MAGI CNA view allows the user to see the size and position of copy number aberrations in each sample. This aids in the identification of CNAs that may be significant, particularly in combination with other types of mutation in the same or different genes, shown in the other view. We viewed the CNAs in the PDGFRA gene in the TCGA Pan-Cancer dataset (Figure 5.14). GISTIC2 identified PDGFRA as a significantly recurrently amplified gene only in GBM tumors, but Figure 5.14 shows that PDGFRA contains less frequent amplifications in other cancers as well. For example, in the lung squamous cancer (LUSC) samples in the TCGA Pan-Cancer data [52], PDGFRA has focal aberrations in 9/178 tumors (5%). While the amplifications of PDGFRA in LUSC are not significantly recurrent [9], they may be worthy of additional investigation since similar amplifications in PDGFRA are significantly recurrent in GBM.

### 5.2.3 Interactive linking of data views

The views shown in MAGI are interactively linked in two ways. First, users can dynamically filter the data shown in all views (e.g. by toggling on or off individual datasets). Second, when a user interacts with (i.e. moves the cursor over) a sample in one view, all the data for that sample is highlighted in each other view (Figure 5.15). This improves user exploration of the data in a number of ways. For example, when a user moves the cursor over a sample (column) in the aberrations view, all of the mutations in that sample for the transcript shown in the transcript view are highlighted. This shows the user immediately if a sample has multiple mutations

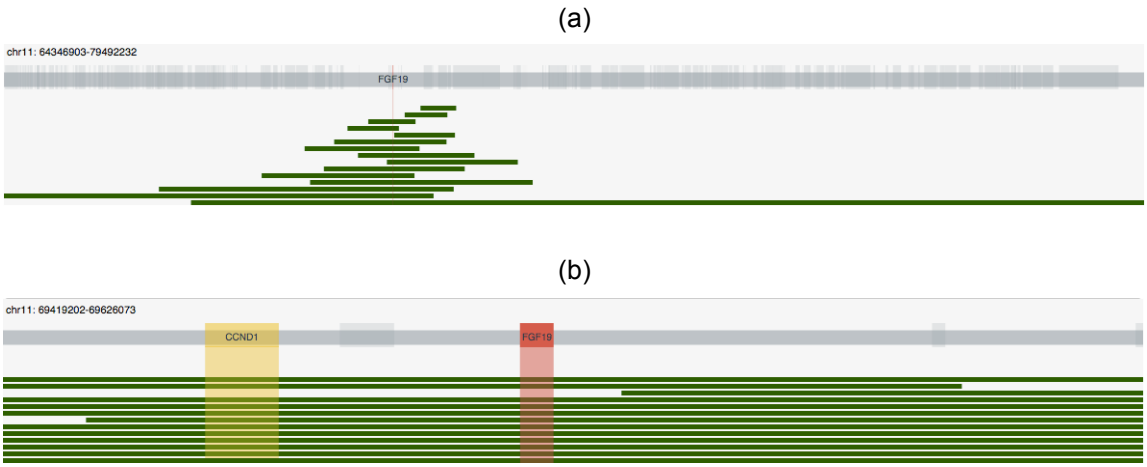



**Figure 5.13:** An example of a potential missed target of recurrent CNAs by GISTIC2 in TCGA STAD. Shown are amplifications assigned by GISTIC2 to the $FGF19$ gene in the TCGA STAD dataset. (a) View of all the amplifications in $FGF19$. (b) Zoomed in view of the amplifications, where $CCND1$ – a well-studied gene frequently amplified in cancer [186] – is visible. $CCND1$ is only 44kb from $FGF19$, and thus may also be a target of the CNAs in $FGF19$.

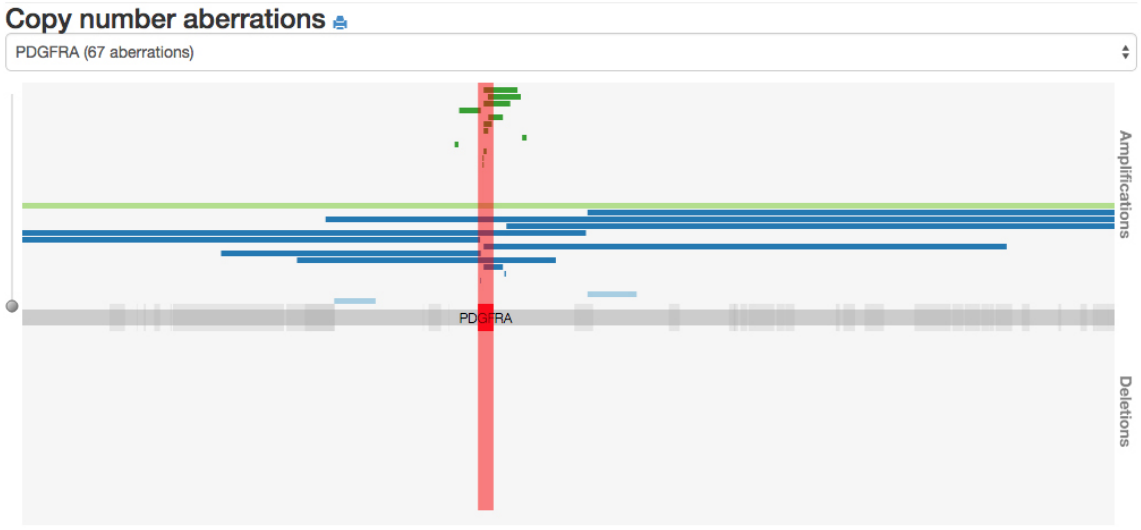



**Figure 5.14:** Copy number aberrations in $PDGFRA$ in the TCGA Pan-Cancer dataset. Amplifications in $PDGFRA$ were identified by GISTIC2 to be significantly recurrent in GBM tumors (green bars), although $PDGFRA$ is also amplified in many other cancers. Vertical bar indicates the genomic coordinates of $PDGFRA$ gene.

in a single sample. Additional examples of tasks enabled by linked views include the correspondence between gene expression and mutation types (does a gene with a deletion or inactivating mutation in a sample also have lower gene expression?), or showing whether a sample has two hits in a gene, e.g. a deletion and inactivating mutation.

### 5.2.4 Uploading private mutation datasets

With the declining costs of whole-genome/-exome sequencing, many researchers are now sequencing their own tumor cohorts. MAGI provides a platform for these researchers to analyze and explore their mutation data in concert with larger datasets of public mutation data; e.g. from TCGA. By combining public and private mutation data in this manner researchers may interactively explore hypotheses about individual mutations or genes, and combinations thereof. For example, researchers can quickly determine if the mutations in their dataset are present in the larger public datasets. MAGI is not intended to replace sophisticated algorithms to predict driver mutations or genes in new datasets, but rather complements these approaches by allowing researchers to quickly determine the nature of mutations in a gene or set of genes in a particular cancer type.

MAGI offers a tool for users to upload their own data using simple plain text tab-separated value (TSV) formats. Users may upload files directly from their personal machine or provide a public URL to the file (e.g. from Synapse, or a TCGA Publication Page). Users can upload SNVs using a TSV format, or the standard mutation annotation (MAF) format used by TCGA for SNV data. Similarly, users can upload CNAs using a TSV format, or a TAR file (`http://www.gnu.org/software/tar/`) of the output of the GISTIC2 algorithm [23]. Users can also upload TSV files of "generic" aberration data that lists the mutated genes in a set of samples to MAGI, which can include for example gene expression or methylation data. MAGI also offers users the ability to upload continuous-valued data (e.g. expression or methylation data) which will be drawn as a heatmap. Users can annotate the samples in each of the views by uploading a matrix of sample annotations – such as clinical data or the output of an algorithm analyzing the tumor – containing continuous (e.g. survival time) or categorical data (e.g. expression subtype). MAGI supports GZIP compression (`http://www.gzip.org/`) and will decompress any provided GZIP files. The file formats and specifications are available online at `http://magi.cs.brown.edu/upload`.

Alternatively, users can upload a single manifest file that includes the URLs for all the data files in a given dataset. As example, we have posted the data files for the TCGA Pan-Cancer and gastric cancer datasets that are pre-loaded into MAGI (`http://compbio-research.cs.brown.edu/software/magi/data`), and created man-

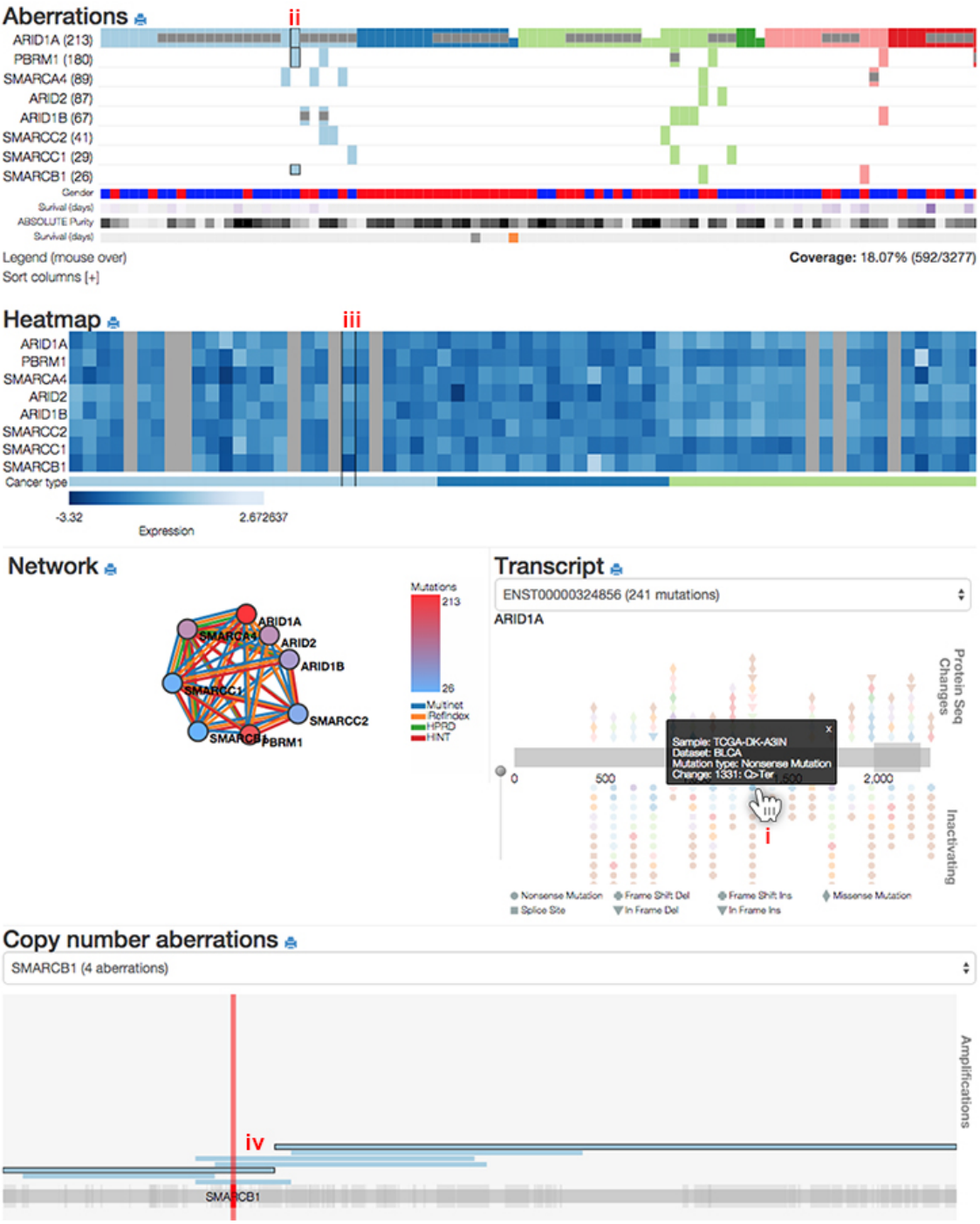



**Figure 5.15:** Example of sample linking. As users scroll over mutations in the transcript view (i), the corresponding samples are highlighted in the aberration matrix (ii), heatmap (iii), and copy number views (iv).

ifest files for each of these datasets (`http://magi.brown.edu/manifests`). This feature also makes it easy to pull the latest TCGA mutation data from Firehose (`http://gdac.broadinstitute.org/`).

After the data upload is complete, MAGI annotates the mutation data using the annotations described above, and then makes the data available to be queried by the user. Thus, users can upload and query their own mutation data on MAGI in seconds.

MAGI automatically generates a dataset summary page that allows users to sort or search for mutated genes or functionally-related gene sets (from KEGG [28] and PINdb [29]) based on their mutations in order to aid users in creating their query gene set(s) (Figure 5.4). Both the genes and gene sets are displayed in sortable and searchable tables that list the number of SNVs, CNAs, and mutated samples per gene or gene set. The rows of these tables include a link to an automatically generated MAGI query, which allows users to quickly view the mutation data of genes and gene sets with possible driver mutations in their data. Users can also dynamically create plots of two attributes of the mutations in each gene in a dataset (by default the number of SNVs and number of CNAs). Users can zoom and pan these plots to quickly identify outlier genes that may have interesting patterns of mutations.

No local software installation or browser plugins are required for a user to upload their own data. The only requirement is that a user must login to MAGI using an Open ID provider (`http://openid.net/`), including Google, Facebook, or Twitter. Mutation data uploaded with MAGI is available exclusively to the user who uploaded, but can be queried in combination with public mutation data, which facilitates comparing the mutations in a gene set from two studies (e.g. of the same cancer type). MAGI allows users to share the results of individual queries with colleagues and collaborators without making their mutation data publicly available by hashing each query to make a unique, bookmarkable link, such that only individuals with the link can view the MAGI page generated from the query.

### 5.2.5 Collaborative annotation

Users who log into MAGI can record annotations for mutations and protein interactions. These annotations are available to all users, enabling collaborative annotation of public datasets. The main MAGI view includes a panel (Figure 5.5) that lists the number of annotations in the MAGI database (in the form of PubMed IDs) for the displayed genes. Each gene links to a page that lists all the variant annotations for each gene as well as links to the corresponding PubMed pages (Figure 5.6).

The interface for adding annotations is integrated with each of the views, thus reducing the effort required to add data. Clicking on a mutation or interaction in any view populates an annotation form (Figure 5.5) with information about the corresponding mutation or interaction (Figure 5.5). For example, clicking on an SNV in the transcript view records the position, protein domain, mutation type (e.g. missense), and cancer type of the mutation. Users may edit this default information and then enter one or more PubMed identifiers (PMIDs) and optionally a text comment. Users may also upvote or downvote existing annotations. By linking the information in the view directly with the annotation form, MAGI automates data entry thereby encouraging users to add annotations.

MAGI also links out to public databases to facilitate user addition of new annotations. For each mutation in the transcript view, we provide a link to a PMC search for the same protein sequence change. This simplifies the process of determining whether a variant has been previously reported or functionally validated. For example, examining the mutations in the SMAD4 gene in the TCGA Pan-Cancer dataset, we see four mutations at position 537, three in colorectal cancer and one in lung squamous cancer. The PMC searches reveal two publications for the lung squamous variant (D537E) and two publications for one of the colorectal variants (D537Y). These publications date to 1998, 2000, 2001, and 2004, and report functional studies of these variants. MAGI also links the gene names in the aberrations view to the corresponding NCBI pages to further facilitate the discovery of annotations which the user can then add to MAGI.

For protein-protein interactions, users can annotate existing interactions or add new interactions to a "Community" interaction database. Users may also upvote or downvote (Figure 5.5) the recorded annotations – including those that are pre-loaded from PPI databases – so that others will know whether the community believes the annotation to be appropriate and useful (See example below). This feature is intended to facilitate the annotation of protein-protein interactions and correct errors that are present in the curated databases [187]. One example of a possible error is in the iRefIndex 9 [141] protein interaction network. iRefIndex 9 includes interactions of NOTCH1 with PIK3CA, PIK3R1, PIK3CG. These interactions are curated from [188], which reports NOTCH1 interacting with the PI3K proteins as part of a NOTCH1-p56lck-PI3K complex, but only p85 (PIK3R1) is specifically listed in one of the co-immunoprecipitation experiments. MAGI is designed to help remedy such ambiguities through collaborative annotation, and we downvoted each of these interactions in MAGI.

By default, all annotations recorded for public data (e.g. TCGA data or protein-protein interaction data), will be available to other users. Mutation annotations are stored for a given gene and mutation type, with optional fields for protein sequence change and cancer type. The votes for each annotation are stored with the user ID of the voter, in order to ensure that the same voter cannot vote more than

once on a given annotation. Users who login will see all contributed annotations in the interactive views by clicking on the corresponding mutation (e.g. deletions in ARID1A). All users can view expert-sourced interactions as a "community" PPI network.

The annotation features in MAGI facilitate "expert-sourcing" of public datasets, further increasing the value of these datasets by linking them directly to the scientific literature or expert knowledge. To our knowledge, MAGI is the first web platform to offer expert-sourcing of annotations for cancer genomics.

### 5.2.6 Tumor sample view

A major issue in cancer genomics is to interpret the variants in a single sample. MAGI includes a sample view that shows the variants and sample annotation for a single tumor sample (Figure 5.7). The view lists all the mutations in a given tumor sample, ordering them by the annotation score of the variants in the MAGI database. Thus mutations with more annotations – presumably the drivers – are shown near the top of the list. As more researchers use MAGI and annotate variants, the mutations that are supported by literature will be further separated from likely passenger mutations.

Furthermore, since MAGI allows users to upload mutation data using a simple web form, researchers or clinicians can upload data from an individual tumor and have the mutations prioritized by MAGI using literature annotations.

MAGI shows the tumor sample view in a separate page from the gene set visualizations. For example, Figure 5.7 displays a bladder urothelial carcinoma (BLCA) tumor sample's mutations.

### 5.2.7 Statistical tests of enrichment

Users can load sample annotations (e.g. sample clusters or subtypes) into MAGI – either for public or private datasets – and test whether a group of annotated samples is enriched for mutations (Figure 5.16). MAGI computes the enrichment by cross-classifying samples into an contingency table, and then using Fisher's exact test (when $r = c = 2$) and the $\chi^2$ association test. MAGI displays the enrichments for a given query set of genes and datasets on a separate page.

We demonstrate the statistical tests using breast cancer gene expression subtypes from TCGA [8], computing the association between ERBB2 aberrations and

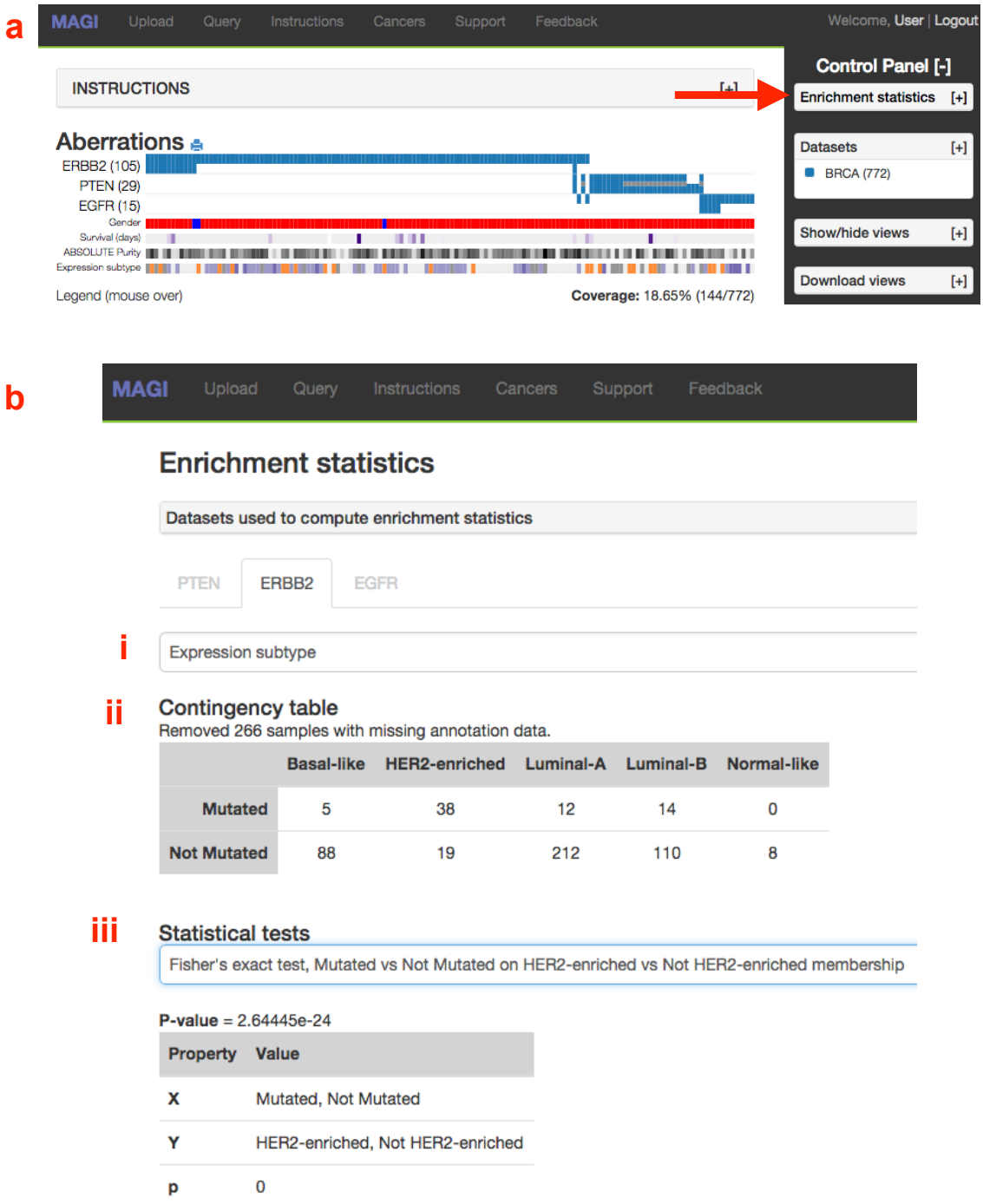



**Figure 5.16:** MAGI interface for computing statistical tests of association between mutation status and sample annotations. (a) Users view enrichments for their query by following the "Enrichment statistics" link on the view page. (b) The MAGI enrichment statistics page, shown here for the genes *PTEN*, *ERBB2*, and *EGFR* in BRCA samples annotated by the gene expression subtypes from [8]. i) Users choose the sample annotation they want to test from a dropdown of all discrete sample annotation categories. ii) MAGI then cross-classifies samples into a contingency table. iii) Users view enrichments by choosing from a dropdown of statistical tests.

the HER2-enriched subtype (Figure 5.16). ERBB2 amplifications largely define the HER2-enriched subtype [8], and MAGI indeed finds that the HER2-enriched subtype is enriched for aberrations ($P < 10^{-15}$). Identifying associations between aberrations and sample annotations is a very common task in cancer genomics research. MAGI allows the user to effortlessly perform these calculations, integrating visual exploration of the data with statistical analyses. To the best of our knowledge, MAGI is the only online tool that computes statistical tests of enrichment between mutations and sample annotations interactively across a combination of public and user-uploaded datasets.

### 5.2.8 Software availability

- *MAGI web server* (`https://github.com/raphael-group/magi`). Git repository including the Node.js web server used to run MAGI.
- *GD3 visualization library* (`https://github.com/raphael-group/gd3`). Git repository including the Javascript library (GD3) that uses D3 to generate the MAGI visualizations.
- *Amazon machine image* (AMI ID: ami-44e4442c). Preconfigured Amazon machine image named "MAGI Web Server" that includes the MAGI web server and all of its software dependencies. Users can clone the AMI to create their own private version of MAGI.

## 5.3 Related work

MAGI fulfills different needs from the current data portals, genome browsers and web servers that have been developed to view and analyze genomics data. Portals such as CBio [173, 174] allow users to query and explore cancer genomics data, but do not allow straightforward uploads of private data. Several systems [173, 174] allow direct queries or bulk downloads of TCGA and/or ICGC data. With these systems, joint analysis of public and private datasets typically requires complicated software installation and/or bioinformatics expertise. Recently, the UCSC Genome Browser has added data track hubs [189] for users to upload browser tracks, but the genome-centric analysis does not facilitate analysis of mutation combinations in pathways/networks. In contrast, genome browsers such as the Integrative Genomics Viewer [190], UCSC Cancer Genomics Browser [191], or Galaxy [192] allow researchers to view the genomic locations of aberrations from multiple tumor samples, but are not integrated with public datasets. Different goals are pursued by web

servers such as IntOGen-mutations [193, 194] and others [21], that facilitate analysis of the mutations in a cohort of tumors through a simple web interface.

Current cancer data portals and browsers [173, 174, 175, 195] (See [170] for a review) have limited features to facilitate a bidirectional interaction between large-scale public and smaller scale/private datasets. In all current systems, information flows in only one direction. There is no ability for researchers to record insights about the public datasets, and the results of follow-up studies remain scattered in the literature and disconnected from the datasets. These computational/software limitations impede the scientific process and lessen the impact of these large cancer datasets.

## 5.4 Case study

To demonstrate the power of using MAGI to analyze private mutation data in conjunction with TCGA data, we uploaded mutation data from the TCGA stomach adenocarcinoma (STAD) study, a cancer type that was not one of the 12 cancer types in the Pan-Cancer dataset [52]. The TCGA STAD dataset [55] consists of SNVS and CNAs in 9,326 genes in 215 samples (See Section 5.2.1). We queried the TCGA STAD dataset using a list of 22 known gastric cancer genes compiled from an earlier publication [196]. This list includes both canonical cancer genes (TP53, ERBB2, KRAS, MYC, RB1, and PTEN), as well as genes with roles in cancer more specific to STAD (ERBB3, EZH2, RUNX3, SMAD1, SMAD2, SMAD3, SMAD4, SMAD7, CDKN1A, CDKN1B, RELA, and TGFBR2). We found that two of the canonical cancer genes are mutated in ¿25 samples – TP53 (106 samples), MYC (57), ERBB2 (56), and KRAS (34) – while PTEN (10) and RB1 (4) were less frequently mutated. PTEN and RB1 each have multiple inactivating mutations consistent with their role as tumor suppressors.

Of the remaining gastric cancer genes, CDH1 (21), SMAD4 (48) ERBB3 (10), TGFBR2 (6), and SMAD2 (4) are the most mutated. SMAD4, TGFBR2, and SMAD2 each have patterns of mutations that indicate a strong functional impact (Figure 5.8).

- In SMAD4, all 19 mutations (including three frameshift insertions) are in the MH2 binding domain, including three missense mutations and two in-frame insertions at position 361 where mutations associated with polyposis have been previously reported [197]. We added this publication as an annotation of the SMAD4 mutation, thus making this information available to all users of MAGI. SMAD4 was also deleted in 3 samples.

- In SMAD2, three of the four mutations are inactivating nonsense mutations, and the fourth mutation is a missense mutation in the MH2 binding domain.
- In TGFBR2, 4 of the 5 missense mutations occur in the protein kinase domain, while the other mutation is an inactivating frameshift deletion.

The observed pattern of inactivating mutations in SMAD2, SMAD4, and TGFBR2 is particularly striking as TGFBR2 is a member of the transforming growth factor beta (TGF-$\beta$) signal transduction pathway. This pathway activates the SMAD family of proteins and is known to have a key role in many cancers [198], including gastric cancer [199, 200].

Combining the STAD mutation data with TCGA Pan-Cancer data in MAGI suggests a strong link between SMAD4 and SMAD2 mutations in gastric cancer and colorectal cancer (COADREAD).

- 12 of the 49 mutations in SMAD4 in the STAD and COADREAD datasets occur at position 361 in the MH2 binding domain. Three missense and two in-frame insertions occur at position 361 in STAD, and six missense mutations occur at position 361 in COADREAD. Furthermore, the MAGI database includes 37 annotations for variants at position 361 in SMAD4, including 21 annotations from DoCM specifically for COADREAD. This provides further support for a mutational hotspot at position 361.
- There are two R321Q missense mutations in SMAD2 in COADREAD, which match the position and residue change of the single SMAD2 missense mutation in STAD. While the single R321Q mutation in STAD would not stand out among the many other mutations measured in the STAD samples, the conservation of this mutation in COADREAD lends support for this mutation having a role in cancer. Moreover, there are five nonsense mutations at position 464 in SMAD2, including two mutations in STAD and three in COADREAD.

These results show that not only are the SMADs mutated in both STAD and COADREAD, but particular mutations in SMAD2 and SMAD4 are common to these two cancer types. In addition, mutations in TGFBR2 and more generally in the TGF-$\beta$ pathway are well-known in colon cancer [201, 202], and mutations in the TGF-$\beta$ pathway were discussed in the TCGA COADREAD publication [171]. However, none of these particular mutations were discussed in either the COADREAD [171] or STAD publications [55], nor were the similarities between the mutations in these cancer types reported. The interactive visualizations generated by MAGI led immediately to these discoveries.

## 5.5 Discussion

We demonstrate MAGI's functionality using cancer mutation data. However, the software is not limited to the display and annotation of cancer data. The aberrations view in MAGI could equally be used to display germline variants or other transcriptomic, genomic, or proteomic data across samples. Similarly, the network view can display other types of interactions between genes, proteins, protein domains, or individual residues. Any of these can then be annotated interactively, enabling collaborative annotation of a wide variety of datasets. Individual researchers, teams of researchers, or international collaborations may create custom MAGI installations to leverage other public and private datasets and engage different communities of researchers in annotation of these datasets.

# Chapter 6

# Conclusions

Cancer remains one of the most common and deadly diseases in the United States despite the significant resources invested in cancer research. According to the National Cancer Institute, an estimated 1.6 million people will be diagnosed with and nearly 600 thousand will die from cancer in 2016.[1] In addition to the untold human cost, 125 billion dollars will be spent on cancer care. Amid these grim numbers there are signs of progress, as cancer mortality dropped 1.4-2% from 2003-2012. Part of this progress is due to the billions of dollars invested in cancer research every year, which have led to better understanding of cancer biology and new therapeutics.

The precipitous decline in the cost of DNA sequencing offers a promising new path to therapies personalized for the mutations in a given patient's cancer cells. As the cost of sequencing has dropped to around one-thousand dollars per genome it has become routine to sequence tumor samples, and public consortia have assembled datasets with the sequence of tens of thousands of tumors. Researchers have begun analyzing this data and leveraging the large sample sizes in their private studies. However, these datasets are large (hundreds of gigabytes per genome) and cancer is a family of complex genetic diseases with many factors contributing to both onset and prognosis, and requires sophisticated computational methods to prioritize testable hypotheses.

In this thesis, we have presented methods for identifying combinations of driver (causal) mutations in large tumor cohorts. Three of these methods (COMEt, the weighted test for mutual exclusivity, and HotNet2) search for combinations of mutations targeting the genetic pathways that are perturbed in cancer. The output of these methods are prioritized and testable predictions of the mutations responsible for cancer in a cohort of tumors. The fourth method (MAGI) is a web application

[1] http://www.cancer.gov/about-cancer/what-is-cancer/statistics

biologists can use to search mutation data for combinations of driver mutations. We applied all of these approaches in collaboration with biologists on datasets of hundreds to thousands of tumors.

We have learned many lessons during the years of work represented in this thesis, but two general lessons stand out in particular. First, there is a key tradeoff in prior knowledge and the number of hypotheses being tested when searching for combinations of driver mutations. Methods that use no prior knowledge, such as our CoMEt algorithm for searching for mutually exclusive mutations, are less biased and easier to apply, but require more sophisticated approaches and assumptions to account for bias in the data. Second, biological experts are a valuable resource that can be leveraged through computational approaches. We introduced a web application to help biologists with no computational expertise identify driver mutations, and we anticipate that researchers will develop more "human-in-the-loop" methods for cancer genomics in the near future.

Much work remains to achieve personalized therapy for cancer. One pressing issue is to analyze somatic mutations in the non-coding regions that make up $\sim$99% of the genome. All the work in this thesis was restricted to mutations in the coding sequence of genes, but projects such as ENCODE [203] have demonstrated the functional and regulatory importance of elements in these non-coding regions. In that vein, several recent efforts have begun to shed light on non-coding mutations in cancer [204, 205, 206, 207]. Multiple studies have identified recurrent mutations in the *TERT* promoter [204, 205], and [205] validated the functional impact of these mutations in cancer cell lines. More recently, researchers have begun to systematically search for recurrent non-coding mutations in cancer [206, 207], which is an important step in analyzing whole cancer genomes. Despite these efforts, the significance of non-coding mutations in cancer remains relatively unexplored, and is an important direction for future research.

# Appendix A

# CoMEt

## A.1 Results

### A.1.1 Comparison to muex on real data

We compared CoMEt to muex [62] using two different versions of the TCGA glioblastoma (GBM) dataset: (1) the dataset from Leiserson *et al.* [1] containing 398 alterations and 261 samples; (2) the dataset from Szczurek *et al.* [62], containing 83 alterations and 236 samples (See Section 2.2.8). There are 184 samples in both the Multi-Dendrix GBM and muex GBM datasets. Besides the samples, the main difference between these two datasets is that the muex dataset is restricted to only 83 significantly recurrent alterations.

Since the muex score is for single alteration sets, we ran muex iteratively to identify collections of alteration sets. That is, we run muex to find the top scoring alteration set, remove those alterations, and repeat $t-1$ times. We ran muex with the parameters used in [62], restricting to alteration sets with coverage at least 0.3, impurity lower than 0.5, and a significance cutoff of 0.05. On the muex GBM dataset, we ran CoMEt and muex with $k=4$ and $t=3$ to match the parameters used in [62]. On the Multi-Dendrix GBM dataset, we ran CoMEt and muex with $k=3$ and $t=3$, since muex aborted with an out-of-memory error for $k=4$ on this dataset.

On both GBM datasets, CoMEt identifies collections with much more significant exclusivity. Moreover, more of the genes in the CoMEt collections are known cancer genes (according to the COSMIC Cancer Census [208]) compared to the genes in the muex collections (Table S13). On the Multi-Dendrix GBM dataset, CoMEt

identifies three collections that overlap the Rb (*CDK4*, *CDKN2A*, *RB1*), p53 (*TP53*, *MDM2*, *CDKN2A*), and PI(3)K (*PTEN*, *IDH1*) signaling pathways. Each of these sets include surprisingly exclusive alterations, with $\Phi(M)$ ranging from $10^{-8}$ to $10^{-19}$, and all the alterations are in cancer genes. In contrast, muex identifies sets with lower coverage and less surprising exclusivity, with $\Phi(M) > 10^{-3}$ for each set, and three of the alterations are not in known cancer genes.

On the muex GBM dataset, CoMEt again identifies more exclusive alteration sets that overlap more known cancer genes, while muex reports few known cancer genes with most having an uncertain association with cancer. In general this dataset seems to include more spurious alterations, as both algorithms identify less exclusive sets with fewer cancer genes than on the Multi-Dendrix GBM dataset. This might be a result of the different handling of copy number aberrations in the two papers (see [1] and [62]).

## A.2 Methods

### A.2.1 MCMC Algorithm

We define a Markov chain whose states $\Omega$ are possible collections $\mathbf{M}$ and where transitions between states (collections) are defined such that the chain is ergodic. Finite and ergodic Markov chains converge to a unique stationary distribution. In this case, because we want to sample from collections $\mathbf{M}$ in proportion to their weights

$$\Phi(\mathbf{M})^{-\alpha} = \prod_{M \in \mathbf{M}} \Phi(M)^{-\alpha},$$

our desired stationary distribution is

$$\pi_{\mathbf{M}} = \frac{\Phi(\mathbf{M})^{-\alpha}}{\sum_{\mathbf{M}' \in \Omega} \Phi(\mathbf{M}')^{-\alpha}}. \tag{A.1}$$

Note that we use $\Phi(\mathbf{M})^{-\alpha}$ so more exclusive collections have higher weights. The Metropolis-Hastings algorithm [75, 76] is a method for defining transition probabilities for an irreducible Markov chain such that the modified chain is ergodic and has a desired stationary distribution. A Metropolis-Hastings algorithm to sample collections $\mathbf{M}$ according to this stationary distribution is as follows:

**Initialization.** Choose $tk$ genes uniformly at random from $\mathcal{E}$, and assign $k$ genes at random to initialize $\mathbf{M} = M_1, \ldots, M_t$.

**Iteration.** For $N = 1, 2, \ldots,$ obtain $\mathbf{M}_{N+1}$ from $\mathbf{M}_N$ as follows:

1. Select a gene $g$ uniformly at random from $\mathcal{E}$.
2. Define the proposed collection $\mathbf{M}'_N$ as follows:
   i) If $g \notin \mathbf{M}_N$, then choose uniformly at random gene $g' \in M_i$, and replace $g'$ with $g$.
   ii) Else, choose uniformly at random gene $g' \in M_i$, and *swap* genes $g$ and $g'$. Note that if $g, g' \in M_i$, then $M_i$ will be unchanged.
3. Let $P(\mathbf{M}_N, \mathbf{M}'_N) = \min\{1, \frac{\Phi(\mathbf{M}_N)^\alpha}{\Phi(\mathbf{M}'_N)^\alpha}\}$.
4. With probability $P(\mathbf{M}_N, \mathbf{M}'_N), \mathbf{M}_{N+1} = \mathbf{M}'_N$, else $\mathbf{M}_{N+1} = \mathbf{M}_N$.

It is easy to see that this chain is ergodic (it is possible to reach any state (collection) from any other state (collection), it is finite, and it is not bipartite) and thus it converges to our desired stationary distribution. We use the parameter $\alpha$ to increase/decrease the difference between $\Phi(\mathbf{M}'_N)$ and $\Phi(\mathbf{M}_N)$ (we used $\alpha = 2$ except where noted). Also, in the second step of the algorithm, we ensure that the number of exclusive alterations is larger than the number of co-occurring by checking that the Dendrix weight $W(M) > 0$. This is to avoid examining sets alterations with high coverage (e.g. altered over 90% of samples) that may have significant exclusivity even though relatively few samples harbor exclusive alterations. We assess convergence of the MCMC algorithm by calculating *total variation distance* of the the sampling distributions from multiple chains with different initializations (details below).

**Convergence of MCMC from different initial gene sets** We assessed the convergence of the MCMC algorithm by comparing the sampling distributions from multiple chains initialized at different starting states. The rationale is that if the multiple chains have converged (i.e. are sampling from the posterior distribution) then the sampling distributions obtained from the different initializations should be very similar. Conversely, if the sampling distributions are different, then one or more of the chains has failed to converge. We computed the distance between the sampling distributions of chains $C$ with different initializations as follows. First, let $P_c(\mathbf{M})$ be the proportion of iterations in which collection $\mathbf{M} \in \Omega$ was sampled in chain $c \in C$. We compute the total variation distance [209] between the distribution $P_c$ for each chain $c$ to the distribution $P_u(\mathbf{M})$, where $u$ is a chain formed by concatenating the chains in $C$, defined as

$$||P_c - P_u||_{TV} = \max_{s \in u} \|P_c(s) - P_u(s)\|. \tag{A.2}$$

A small total variational distance implies that the chain $c$ has converged. We take the mean of the total variation distance across chains.

To implement the above procedure, we ran CoMEt with 5 to 10 different initializations. For one of these initializations, we used the collection output by MultiDendrix [1] (using the same values of the parameters $t$ and $k$ as in CoMEt). The remaining initializations were random collections. We start CoMEt with 100 million iterations for each of these initializations. If after the 100 million iterations the mean total variation distance is smaller than 0.005, then we consider the chains to have converged. Otherwise, we increase the number of iterations by a factor of 1.5, or stop the process if the number of iterations reaches 1 billion. The output of the MCMC algorithm is the the union of the sampling distributions from the different initializations. As an example, Figure A.1 shows the results of this procedure on AML mutation data for $t = 3$ and $k = 4$, after 1 million and 10 million iterations.

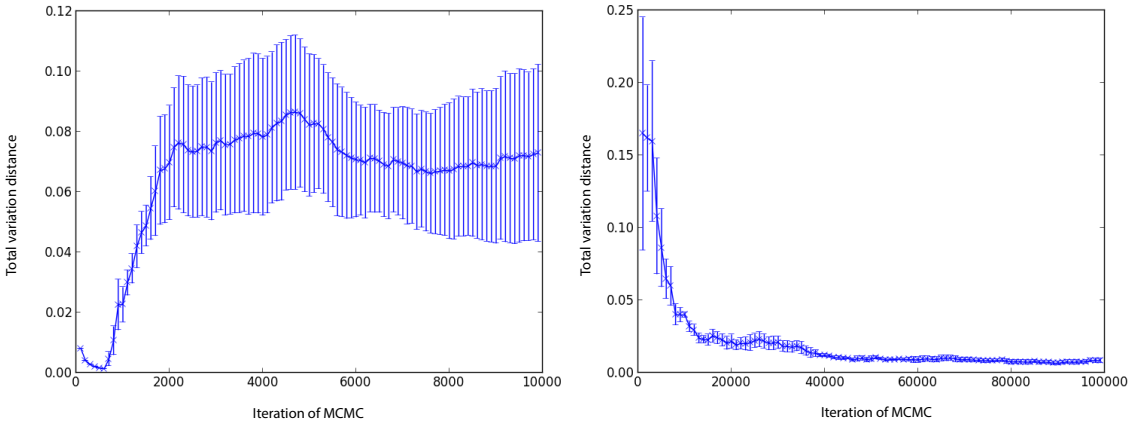



**Figure A.1:** Plots of the total variation distance distribution in each iteration for an MCMC run with $1M$ iterations (left) and an MCMC run with $10M$ iterations (right) on AML mutation data for $t = 3$ and $k = 4$.

### A.2.2 Parameter selection

We select $\delta$ with the following heuristic procedure. When we run CoMEt with $t$ sets in the collection, ideally we should obtain $t$ cliques in the marginal probability graph. To find the best $\delta$ that fulfills the expectation, we search for an "L-corner" in a graph of the number of edges in the marginal probability graph as a function of the edge weight.

More precisely, we first plot a log-log distribution with the number of edges in the marginal probability graph with edge weight $\geq p$ against edge weight $p$ (Figure A.2). We choose $\delta$ starting from the minimum edge weight $p_{min}$ that contains at least $t \times \binom{k}{2}$ edges in the marginal probability graph. e.g. the yellow horizontal line in Figure A.2 shows the number of edges in GBM with $k = 3$ and $t = 3$. We identify a value $\delta$ where the number of edges increases dramatically after this value as the

probability threshold decreases. To find this value, for each value $x$ we perform a linear regression of two best-fit lines (using root mean squared error) before and after this value. We the first $p > p_{min}$ that forms a "L-corner", i.e. the slope of the two best-fit lines changes from a smaller negative value to a larger negative value as the value $x$ decreases (e.g. moving leftward in Figure A.2).

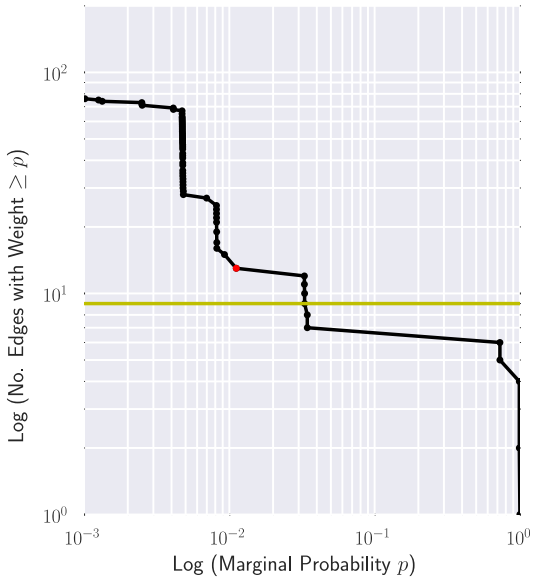



**Figure A.2:** The distribution of the number of edges with weight $\geq p$ in GBM with $k = 3$ and $t = 3$ in log-log scale. The red dot indicates the first hitting edge weight where the change in slope is negative (when moving leftward) such that the number of edges in the subgraph is at least $t \times \binom{k}{2} = 9$ (as the horizontal yellow line).

For each TCGA dataset, we ran CoMEt with $\alpha = 2$, $k = 4$ and 100 million iterations using 5 to 10 random initializations. We used $t = 3$ for BRCA and $t = 4$ for AML, GBM, and STAD. For BRCA and STAD with subtypes, we ran CoMEt with $k = 4$ and $t$ equal to the number of pre-defined subtypes (4 and 3, respectively), and 100 million iterations using 10 random initializations. See Appendix 2.2.8 and Appendix A.2.1 for additional details.

## A.3 Data

### A.3.1 Simulated data

We generated simulated datasets using the following approach. Recall $\mathbf{C}$ is a set of highly altered genes whose alterations are not necessarily exclusive.

1. Select $k$ genes to form an "implanted pathway" $P$.
2. Let $\gamma_P$ be the fraction of mutated samples in $P$. Select $\gamma_P \times n$ samples to be exclusively mutated in $P$, where the proportion of mutations in each gene in $P$ is given by the tuple $\mu_P = (c_1, \ldots, c_k)$.
3. Randomly select samples to be mutated in each gene in $\mathbf{C}$, where the fraction of mutated samples per gene is given by $\gamma_{\mathbf{C}}$.
4. For each of the $n$ samples $s$ in each of the $m$ genes $g$ (including the implanted and cancer genes), mutate $g$ in $s$ with fixed probability $q$. This step introduces noise into the dataset.

We used $m = 100$, $n = 500$, $k = 3$, $\mu_P = (0.5, 0.35, 0.15)$, $|\mathbf{C}| = 5$, $\gamma_{\mathbf{C}} = (0.67, 0.49, 0.29, 0.29, 0.2)$, and $q = 0.0027538462$.[1] We removed alterations that occurred in fewer than 5 alterations (resulting in the average number of genes of 276.44). We ran CoMEt 100 million iterations from 3 random initial starts.

[1] We chose values for $\mathbf{C}$ and $q$ using values calculated from real data. We choose $C$ to match the mutation frequencies of the five most mutated genes in the TCGA glioblastoma dataset. We calculated $q$ empirically from the TCGA breast cancer mutation matrix.

# Appendix B

# HotNet2

## B.1 HotNet2 algorithm

### B.1.1 Motivation

We developed the HotNet2 (HotNet diffusion oriented subnetworks) algorithm to identify subnetworks of a genome-scale interaction network that are mutated more than expected by chance (Figure 4.4). Standard computational approaches to analyze mutations on pathways and protein complexes are severely limited by the statistical requirement of defining a priori a reasonable number of gene sets, or combinations of genes to evaluate. This imposes multiple undesirable constraints. First, gene sets often contain dozens of genes, reducing the power to identify a smaller subset containing a few directly interacting proteins. Second, there is typically extensive overlap between annotated gene sets, complicating the interpretation of analyses that report tens to hundreds of "significant," but overlapping, pathways. Third, by ignoring the topology of interactions, gene-set analyses have reduced ability to identify crosstalk between pathways. Finally, *a priori* definition of gene sets prevents the discovery of novel combinations of mutations.

While interaction networks have proven useful in analyzing various types of genomic data [34, 115], statistically robust identification of significantly mutated subnetworks is a difficult problem with three major challenges. First, the number of subnetworks is too large to test exhaustively in a computationally efficient and statistically rigorous manner; e.g. $> 10^{14}$ subnetworks of $\geq 8$ proteins in a medium-size human interaction network. Second, the topology of biological interaction networks is heterogeneous; many proteins, and in particular many cancer-related proteins,

have a large number of neighbors. Third, both the frequency of somatic mutations in individual genes/proteins and the topology of the interactions between proteins determine the significance of a subnetwork. While approaches have been introduced to find network modules [34, 115], rank gene sets, or prioritize disease-related genes [210] these approaches consider only network topology and not also the scores of individual genes.

### B.1.2 Insulated heat diffusion as a random walk

HotNet2 uses an insulated heat process (See Section 4.4) that can also be described in terms of a random walk with restart. The random walk starts from a protein $g$, and at each time step $t$ moves to one of the neighbors (chosen uniformly at random) of the current node $g_t$ with probability $1 - \beta$ ($0 \leq \beta \leq 1$), while the walk restarts from $g$ with probability $\beta$. This process is defined by a transition matrix $W$:

$$W_{ij} = \begin{cases} \frac{1}{\deg(j)} & \text{if node } i \text{ interacts with node } j, \\ 0 & \text{otherwise.} \end{cases} \tag{B.1}$$

where $\deg(i)$ is the number of neighbors (i.e., the degree) of protein $g_i$ in the interaction network. The locality of the walk – and therefore the insulated heat diffusion process – is governed by the parameter $\beta$, representing the probability with which the walk starting at $g_i$ is forced to restart from $g_i$. Assuming the graph is connected (in practice we restrict attention to the largest connected component) the Ergodic Theorem guarantees that the random walk starting at protein $g_i$ reaches a stationary distribution described by the vector $\vec{s}_i$:

$$\vec{s}_i = \beta(I - (1 - \beta)W)^{-1}\vec{e}_i, \tag{B.2}$$

where $\vec{e}_i$ is the vector with a 1 in position $i$ and 0's in the remaining positions. The $j$-th entry $\vec{s}_i(j)$ of $\vec{s}_i$ gives the probability that, in the limit, the random walk starting at node $g_i$ is at node $g_j$. We define the diffusion matrix $F$ for a graph $G$ as

$$F = \beta(I - (1 - \beta)W)^{-1}. \tag{B.3}$$

Note that $\vec{s}_i$ is the $i^{\text{th}}$ column of $F$.

#### Asymmetry of heat diffusion

Any clustering algorithm – whether on a graph or not – depends on a notion of distance, or oppositely similarity, between points. Distances are by definition symmetric; however similarity might not be symmetric. For example, some models of

protein sequence similarity are non-symmetric [211]. In the case of HotNet2, similarity is non-symmetric for two reasons: first, the local topology of a pair of nodes $u$ and $v$ in the network which is encoded in the heat diffusion process – is not symmetric. A simple example is shown in Figure 4.5: a node $u$ of degree 1 sends all its heat to its neighbor $v$, but $v$ sends heat to many nodes, including $u$. Thus, $u$ might be "closer" to $v$ than $v$ is to $u$. Second, the score on node $u$ and the score on node $v$ are typically different.

### B.1.3 Statistical significance

We evaluated the statistical significance of HotNet2 subnetworks and the HotNet2 consensus using the two-stage statistical test introduced in the original HotNet algorithm [34, 115]. In brief, the first stage is to compute a $P$-value for the statistic $X_k$, the number of subnetworks of size $\geq k$ reported by HotNet2. We compute the empirical distribution of $X_k$ by running HotNet2 on random data where we permute the heat scores on genes, restricting the permutation to the genes that are in the network and not removed by the expression filter. Although this permutation does not preserve any correlation between a gene's heat and its network topology, we found that a more restricted permutation test did not substantially change the empirical distribution (Appendix B.5.3 and Figure 4.24). For the HotNet2 consensus, we permuted data preserving the relationship between the mutation frequency and MutSigCV score (See Section 4.4.2).

The second stage computes the False Discovery Rate for the set of significant subnetworks identified, as described in HotNet [34, 115]. In this analysis we used

$$\gamma_i = \frac{1}{2^{i-1}},$$

for the set of subnetworks of size $i$, for $i = 2, \ldots, 8$, and

$$\gamma_9 = 1 - \sum_{i=2} \frac{1}{2^{i-1}},$$

since we considered only subnetworks of size between 2 and 9. (We note that the parameter $\gamma$ was called $\beta$ in the HotNet publication [34, 115].)

### B.1.4 Parameter selection

**Insulated heat-diffusion:** $\beta$

The HotNet2 parameter $\beta$ is chosen for a given protein-protein interaction network, and remains fixed for different heat datasets.We chose $\beta$ to balance the amount of heat that diffuses from a protein to its immediate neighbors and to the rest of the network. This was done by computing the amount of heat retained in the neighbors of vertices ("source proteins") with different network centrality. In particular, we computed the *betweenness centrality* for each protein $v$, the number of shortest paths between all pairs of proteins that pass through $v$. We then picked five source proteins from the interaction network; those with the minimum, 25% quantile, median, 75% quantile, and maximum betweenness centrality. We examine vertices with different network centrality in order to choose diffusion parameter $\beta$ such that all proteins retain most of their heat in their immediate neighbors. To determine whether the heat from each source protein $v$ was retained in $v$'s neighbors $N(v)$ or spread throughout the network for a given $\beta$, we compared the influence each source protein had on three sets of proteins: $N(v)$, all nodes distance 2 from $v$, and all other nodes in the network. We compared the three sets by counting the number of proteins in each set on which $v$ had at least $\theta$ influence. We considered the 20 distributions with $\beta = 0.05, 0.10, \ldots, 1$. From these distributions we chose $\beta$ such that each source protein had most of its heat concentrated in its neighbors. As $\beta$ decreases from 1 (the maximum restart probability), the amount of heat on $N(v)$ increases as less and less heat is retained in $v$. However, there is an inflection point at which the amount of heat on $N(v)$ will decrease, as more and more heat will reach the neighbors of $N(v)$. We choose $\beta$ as the minimum $\beta$ before this inflection point. For example, in Figure 4.28, we can see the inflection point near $\beta = 0.4$ for HINT+HI2012. We used $\beta = 0.4$ for the HINT+HI2012 network, $\beta = 0.45$ for the iRefIndex network, and $\beta = 0.50$ for the Multinet network (see below).

We evaluated the sensitivity of the HotNet2 results to the value of parameter $\beta$ by comparing subnetworks reported by HotNet2 on the HINT+HI2012 network and mutation frequency scores (See below), varying $\beta$ $\pm 10\%$ from the value 0.40 determined by the automated procedure. We found that changing $\beta$ has only a minor effect on the results, with at most 7 genes (3.8% of total) added/removed from the subnetworks (See Supplementary Table 29).

**Minimum edge weight:** $\delta$

We choose the edge weight parameter $\delta$ such that HotNet2 will not find large subnetworks using random data. Specifically, for each mutation dataset and interaction

network, we generated 100 random networks with the same degree distribution as the original network by performing $Q \times |E|$ connected edge swaps ($E$ is the set of edges in the interaction network), ensuring that each node retains the same degree as in the original network and that the resulting network is connected, setting $Q = 100$ [212]. For each permuted network, we identified the minimum $\delta$ such that all strongly connected components found by HotNet2 have size $\leq L_{\max}$, for $L_{\max} = 5, 10, 15, 20$. For each $L_{\max}$, we report the median of these values of $\delta$ across the 100 permuted networks (Figure 4.29). For each run, we selected the smallest $\delta$ with the most significant ($P < 0.05$) subnetwork sizes $k$.

We evaluated the sensitivity of the HotNet2 results to the value of the parameter $\delta$ by comparing subnetworks reported by HotNet2 on the HINT+HI2012 network and mutation frequency scores (See below) varying $\delta \pm 5\%$ from the value 0.000496 determined by the automated procedure. We found that varying $\delta$ changed at most 35 genes (12.3% of total) in the subnetworks (See Supplementary Table 30).

### B.1.5 Consensus over multiple networks and gene scores

We derive "consensus" subnetworks from HotNet2 runs on different networks and/or gene scores using the following iterative procedure on a weighted graph. We built a complete, weighted graph $G$ with genes on the vertices, and an edge between a pair of genes $(g, g')$ weighted by the number of networks in which HotNet2 reports $g, g'$ in the same subnetwork (independent of the scores used). We then identified the consensus subnetworks from $G$ as follows. First, we initialized the consensus subnetworks $C$ as the connected components of $G$ when restricting attention to edges with weight $= 3$ (i.e. pairs of genes found in the same subnetwork by HotNet2 in all three interaction networks). Then, we extended each consensus subnetwork $S \in C$ by adding each gene $g$ such that all $g$'s weight two edges end in $S$. We then extended the consensus subnetworks once more using the same procedure but restricting to only weight one edges. We call genes "linkers" if they have weight two edges in multiple consensus subnetworks.

## B.2 Data

### B.2.1 Somatic aberration data

The somatic aberration data downloaded from Synapse and Firehose consists of SNVs in 20,472 genes from 3,281 samples, and CNAs in 720 genes from 4,334 sam-

ples. We describe the sources and our pipeline for processing somatic mutation data in Section 4.4.1 and Figure 4.2. The dataset after processing consists of somatic aberrations in 19,459 genes from 3,110 samples, and after applying the gene expression filter includes somatic aberrations in 11,565 genes in 3,110 samples.

### B.2.2 Gene scores

Many factors determine the significance of mutations in individual genes, including the frequency of recurrence across samples, gene length, mutation context, regional variation in mutation rates, etc. Several approaches have been introduced to account for these factors [17, 25, 143, 26]. We performed HotNet2 analysis using two approaches to assign heat to individual genes according to recurrence and predicted functional impact.

**Mutation frequency** Also used by Vandin *et al.* in HotNet [34]. We computed the mutation frequency for each gene as the number of samples with at least one SNV or CNA in that gene. After processing the dataset contains somatic aberrations in 19,459 genes from 3110 samples. See Figure 4.2.

**MutSigCV $q$-values** MutSigCV [17] computes the statistical significance of mutations in genes across a cohort of samples. MutSigCV first estimates a per-gene BMR, and then computes the probability of observing the given number of mutations in each gene given its BMR. We used the $-\log_{10}$ of the MutSigCV $q$-values for all 18,371 genes.

We use these two scoring schemes because they are at opposite extremes in terms of stringency. The mutation frequency scores provide an opportunity for higher sensitivity, particularly for rarely mutated genes or genes primarily mutated by copy number aberrations. MutSigCV assesses the statistical significance of individual genes, and identifies only 175 genes with $q < 0.05$ in the dataset, which we expect to result in high specificity.

### B.2.3 Expression filtering

We downloaded a list of expressed genes from syn1734155. This list includes 12,081 genes with at least 3 RNA-Seq reads per sample in at least 70% of samples, as used in [164]. We restricted HotNet2 scores to these expressed genes plus a list of 18 well-known cancer genes (*AR, CDH4, EGFR, EPHA3, ERBB4, FGFR2, FLT3,*

*FOXA1, FOXA2, MECOM, MIR142, MSH4, PDGFRA, SOX1, SOX9, SOX17, TBX3, WT1*) that have low transcript detection levels. After expression filtering, the input datasets to HotNet2 consisted of the following:

- somatic aberrations in 11,565 genes from 3,110 samples;
- MutSigCV $q$-values for 11,215 genes.

When we restrict these genes to those also in one of the three protein interaction networks, the somatic aberration data includes 10,208 genes and the MutSigCV $q$-values include 10,215 genes.

### B.2.4 Copy number calling and target selection

We determined targets of copy number alterations and performed copy number calling by the following two steps. First, we downloaded GISTIC2 results for individual cancer types and Pan-Cancer (Pan12) from Firehose. For the Pan-Cancer GISTIC2 dataset, we selected target genes in each maxpeak by checking whether the peak contains genes from [165]. For the individual cancer type GISTIC2 datasets, we selected target genes in each maxpeak by checking whether the peak contains cancer genes defined by The Sanger Institute Gene Census [208]. For max peaks that did not have target genes, we picked a max peak as a target gene if the peak contains only one gene. Second, to determine if each sample had a copy number aberration in a given target gene, we extracted discretized copy number amplifications and deletions from focal copy number binary table provided by GISTIC2 with $\log_2$ ratio value above 0.9 and below $-0.9$, respectively (Figure 4.2).

### B.2.5 Protein-protein interaction networks

To date there is no single database of human protein-protein interactions with high sensitivity and specificity. We used three interaction networks with varying numbers of interactions to allow for different false positive and false negative interactions in the analysis: (1) HINT+HI2012, a combination of high-quality protein-protein interactions from HINT [139] and the recent HI-2012 [140] (See URLs) set of protein-protein interactions; (2) MultiNet [142] a network that integrates multiple types of interactions from different databases; (3) iRefIndex [141], an integrated network from multiple data sources. These three interaction networks have different trade-offs in sensitivity vs. specificity that would not be represented by simply merging the three networks. Figure 4.33 shows the overlap between these networks. Unfortunately,

a simple merge of networks will create a highly-connected network of 16,843 nodes and 180,271 edges. While such a network might have higher sensitivity for pairwise protein-protein interactions than individual networks, it will have very low specificity: the merged network will contain *all* false positive interactions from individual networks. Such a network is not expected to be a reasonable representative of the underlying biology. Moreover, a merged interaction network will have different network properties than the the individual interaction networks (e.g. the average pairwise distances between genes are 3.302 on the merged network compared to 4.087, 3.644, and 3.394 on HINT+HI2012, iRefIndex, and Multinet, respectively). Changes in topology may have large effects for HotNet2 which analyzes network topology (not just nearest neighbors of a protein). Rather than creating a merged network, we run HotNet2 on the individual network individually and then form a consensus across the three networks from the HotNet2 runs (See Section 4.4.4). This procedure is in effect giving greater weight to the interactions that are common among the three networks, while also accounting for the topology of each individual network.

We provide details of our processing of each of the three networks below.

**iRefIndex.** We downloaded iRefIndex version 9.0[1], and constructed an interaction network using all interactions except colocalizations and genetic interactions, including disulfide bond, ubiquitination, palmitoylation, deacetylation, sumoylation, physical interaction, methylation, hydroxylation, phosphorylation, aggregation, adp ribosylation, physical association, enzymatic reaction, covalent binding, phosphotransfer, deubiquitination, cleavage, acetylation, deneddylation, genetic inequality, demethylation, protein cleavage, glycosylation, dephosphorylation, neddylation, and direct interaction. The iRefIndex network consists of 91,872 interactions among 12,338 proteins.

**HINT+HI2012.** We created HINT+HI2012 protein interaction network by considering two interactome databases: HI-2012 prepublication data in human HI2 Interactome database[2] (HI2012) and high-quality interactomes database[3] (HINT). We merged HI-2012 and the binary and co-complex interactomes in HINT to create the HINT+HI2012 protein-protein interaction network. The HINT+HI2012 network consists of 40,783 interactions among 10,008 proteins.

**MultiNet.** We downloaded MultiNet[4], and constructed a interaction network using the entire set of interactions, including protein-protein, phosphorylation, metabolic, signaling, genetic, and regulatory interactions from multiple databases. The MultiNet network consists of 109,597 interactions among 14,445 proteins.

[1] `http://irefindex.uio.no/wiki/Sources_iRefIndex_9.0`
[2] `http://interactome.dfci.harvard.edu`
[3] `http://hint.yulab.org/batch.html`
[4] `http://homes.gersteinlab.org/`

We removed self-loop interactions and multiple edges between interactions in all three networks. We restricted our HotNet2 analysis to the largest connected component consisting of the vast majority of genes and interactions in each network: 91,808 interactions among 12,128 proteins in iRefIndex; 40,704 interactions among 9,858 proteins in HINT+HI2012; and 109,569 interactions among 14,398 proteins in MultiNet.

### B.2.6 Germline variants

To identify potential germline mutations, we compared the mutations used in our HotNet2 analysis to the ones identified using more stringent germline filtering (See syn1729383). This more stringent germline filtering removes variants that are common in NHLBI exomes or 1000 Genomes data (global minor allele frequency $> 0.1\%$). However, this more stringent filtering removes some *bonafide* somatic mutations.

We ran HotNet2 on this strictly filtered set of SNVs. We found 13 subnetworks in the strictly filtered consensus, 12 of which were also found in the original HotNet2 consensus (Supplementary Table 39). The only subnetworks that were missing were the RTK subnetwork and the *BOC-CDON* subnetworks. The strictly filtered consensus included an additional subnetwork not included in the HotNet2 consensus that consists of the *ANK1*, *SPTB*, and *ITPR3* genes. We note that we still identify the MHC Class I subnetwork using the strictly filtered dataset, emphasizing that there are *bonafide* somatic mutations driving the subnetwork even though many of its genes are known to have germline mutations.

## B.3 Statistical test to evaluate HotNet2 Pan-Cancer results

### B.3.1 Finding positional and structural clusterings: NMC and iPAC algorithms

We tested whether genes in subnetworks identified by HotNet2 had either significant clusters of missense mutations or at least 20% inactivating mutations (defined as nonsense, nonstop, splice-site, and frameshift insertions and deletions). These two mutational signals are the "20/20 Rule" that is hypothesized to distinguish cancer genes [16]. We use the NMC [213] algorithm to test for non-random clustering of missense mutations in the protein sequence, and the iPAC [144] algorithm to test for

non-random clustering in the protein structure. We ran the NMC Algorithm [213] on genes with missense mutations. We matched the ENSEMBL transcript ID to the corresponding protein sequence. We used only those mutations for which there was concordance between the missense mutation and the amino acid at the appropriate position in the FASTA-format protein sequence. We applied the NMC algorithm with Bonferroni correction and reported the genes with at least one mutated amino acid cluster that had a $P$-value smaller than the significance threshold of $\alpha < 0.05$. Similarly, we applied the iPAC [144] algorithm on a subset of the above data to identify non-random somatic mutational clusters while taking into account the protein tertiary structure via Multidimensional Scaling (MDS). We only used genes with a defined protein structure in the PDB.

### B.3.2 Cancer type specificity test

We annotated the subnetworks output by HotNet2 on each interaction network by determining which subnetworks are enriched for mutations from a specific cancer type. For a subnetwork $S$ and cancer type $\tau$, we calculated the following statistic. Let $N_\tau$ be the number of samples of type $\tau$, and let $\Gamma(S)$ be the set of samples with at least one mutation in a gene in $S$. Let $C_\tau(S) = |\{p | p \in \Gamma(S), \text{type}(p) = \tau\}|$, i.e. the number of mutated samples from $S$ of type $\tau$. We calculate the enrichment of subnetwork $S$ for mutations of cancer type $\tau$ conditioning on the number of mutations in the subnetwork, and the number of samples of type $\tau$. The enrichment statistic is calculated using Fisher's exact test of the $2 \times 2$ contingency table having as categories the mutation status of a sample ('Mutated' if the sample has a mutation in the subnetworks, 'Not mutated' otherwise) and the cancer type of the sample ('In $\tau$' if the sample is in type $\tau$, 'Not in $\tau$' otherwise). The entries of the contingency table are (in row-wise order): $C_\tau(S)$; $N_\tau - C_\tau(S)$; $\sum_{\tau' \neq \tau} C_{\tau'}(S)$ ; $\sum_{\tau' \neq \tau} N_{\tau'} - C_{\tau'}(S)$. We correct the $P$-value from Fisher's exact test using the Bonferroni correction, where the number of hypotheses is the product of the number of cancer types and the number of genes in the consensus, and only report the corrected $P$-values.

We also tested each gene in each subnetwork identified by HotNet2 for enrichment for mutations from a particular cancer type. We test this by conditioning on the number of observed mutations for the gene, regardless of the number of mutations in the subnetwork, by using Fisher's exact test for the $2 \times 2$ contingency table with categories the mutations status of the ('Mutated' if the sample has a mutation in the gene, 'Not mutated' otherwise) and the cancer type of the sample ('In $\tau$' if the sample is in type $\tau$, 'Not in $\tau$' otherwise). Again, we correct the $P$-value from Fisher's exact test using the Bonferroni-correction, where the number of hypotheses is the number of cancer types, and only report the corrected $P$-values.

### B.3.3 Statistical approach to evaluate exclusivity and co-occurrence of mutations

We assessed the statistical significance of the mutual exclusivity or co-occurrence of mutations in subnetworks identified by Hotnet2 using a one-sided Fisher's exact test on a $2 \times 2$ contingency table. In particular, let $x$ be a $2 \times 2$ contingency table, each cell $x_{i_1 i_2}$ denotes the count of samples corresponding to the mutation status (not mutated (0) or mutated (1)) of the pair of subnetworks, e.g. $x_{10}$ is the number of samples with mutations in one subnetwork($i_1 = 1$) and without mutations in another subnetwork($i_2 = 0$). To test the exclusivity of a contingency table, we use as a test statistic $T_x = x_{10} + x_{01}$.

Since $P$-values from exact tests like Fisher's exact test are typically overconservative, we use the mid $P$-value instead [66]. The mid $P$-value is the average of the probability of a value at least as extreme as the observed value and the probability of a value more extreme than observed. In our case, the mid $P$-value is defined as

$$\frac{1}{2}(P(T \geq T_x) + P(T > T_x)) = \frac{1}{2}P(T = T_x) + P(T > T_x).$$

We assess the statistical significance of the mutual exclusivity or co-occurrence on each cancer type using the above method. To assess on the whole Pan-Cancer data, we performed the Cochran-Mantel-Haenszel test to test for independence across the contingency table of each cancer type.

## B.4 Website

By default, the HotNet2 software package outputs an interactive website that includes visualizations of each component identified by HotNet2. These visualizations incorporate protein-protein interaction network and mutation data (See example in Figure 4.7).

The Pan-Cancer HotNet2 results are available online at `http://compbio.cs.brown.edu/pancancer/hotnet2/`. This website includes visualizations for each subnetwork which incorporate protein-protein interaction network, mutation, gene transcript, and protein domain data (See example in Figure 4.7).

## B.5 HotNet2 Pan-Cancer results

We ran HotNet2 for each combination of interaction network and score. HotNet2 identified a number of significant subnetworks (Supplementary Tables 1-2) in all six runs. Below we describe some of the subnetworks identified using the consensus procedure (Supplementary Table 3-5), as well as other subnetworks (Supplementary Table 6-18). Supplementary Tables 6-18 report the cancer type enrichments $P$-value and the NMC and iPAC clustering $P$-values for the subnetworks in the main text. Most of the subnetworks identified by HotNet2 are not identified by standard pathways enrichment analysis (Supplementary Table 35) and also contain genes not reported as significantly or recurrently mutated by MuSiC [26] or MutSigCV [17] or Oncodrive [25, 143] (Supplementary Table 20). The subnetworks identified by HotNet2 show cancer type enrichments in mutations and genes in the subnetworks show enrichments for mutations in different cancer types (Supplementary Tables 6-18). Genes in the subnetworks also show significant position or structure clustering.

### B.5.1 Genes with positional and structural clusters

We found that high scoring genes in our data were enriched for both positional and structural clusters of mutations ($P < 10^{-20}$, details below). We find sequence and structural clustering for 100 and 16 genes, respectively, in the subnetworks (Supplementary Table 6-18). In addition, we find 25 genes mutated in $\geq 1\%$ of samples that have at least 20% inactivating mutations. Overall, genes in the HotNet2 consensus subnetworks are enriched for NMC clusters ($P < 0.0001$, Appendix B.5.1) or 20% inactivating mutations ($P < 0.0001$, Appendix B.5.1) compared to genes not in subnetworks.

#### Contingency tests for genes with NMC clusters

One method to distinguish driver mutations from passenger mutations is to look for positional clustering of mutations in genes across patient samples [16]. We used an implementation of the NMC algorithm to look for clusters of missense mutations [213] in genes that have mutation heat in any of the three PPI networks used for Hotnet2 runs. We then classified these 10,478 genes based on whether they have a significant clustering of missense mutations (NMC corrected $P < 0.05$), are found in a consensus subnetwork, are neither, or are both. Using a $\chi^2$ test with Yate's correction, we find that genes with significant clusters of missense mutations were significantly ($P < 0.0001, \chi^2 = 42.23$) overrepresented in consensus subnetworks for Hotnet2 (See Supplementary Table 21). Specifically, we found: there were 5,467 genes that do

not have either an NMC cluster or were in a Hotnet2 consensus subnetwork; there were 4,886 genes that have an NMC cluster and were not in a consensus subnetwork; there were 29 genes which were in a Hotnet2 consensus subnetwork and did not have an NMC cluster; and there were 96 genes that had an NMC cluster and were found in a Hotnet2 consensus subnetwork.

### Contingency tests for genes with inactivating mutations

One method for discovering driver mutations in genes is to look at the frequency at which inactivating mutations occur in a gene across patient samples. Inactivating mutations are those which would cause aberrant transcription (splice site mutations), translational frameshift (frameshift indels), or premature translational stop (nonsense mutations). A proposed threshold for distinguishing driver mutations from passenger mutations is to look for genes whose inactivation frequency across patients samples is $\geq 20\%$ [16]. To classify a gene as being inactivated by the proposed 20/20 rule, we only considered genes that were mutated in $\geq 1\%$ of patients to reduce false positives from the high number of genes that have an inactivating mutation in only 1 or 2 samples. Using a $\chi^2$ test with Yate's correction, we find that genes with inactivating mutations were significantly ($P < 0.0001, \chi^2 = 280.86$) overrepresented in consensus subnetworks for Hotnet2 (See Supplementary Table 22). Specifically, we found: there were 10,221 genes that were not inactivated in $\geq 20\%$ of samples or were in a Hotnet2 consensus subnetwork; there were 132 genes that were inactivated in $\geq 20\%$ of patient samples and were not in a consensus subnetwork; there were 100 genes which were in a Hotnet2 consensus subnetwork and were not inactivated in $\geq 20\%$ of patient samples; and there were 25 genes that were inactivated in $\geq 20\%$ of patient samples and were found in a Hotnet2 consensus subnetwork.

### Correlation between heat and mutation clusters and inactivating mutations

We evaluated whether hotter genes had more clustering of missense mutations or $\geq 20\%$ of mutations as inactivating. We indeed find that hotter genes (according to mutation frequency) are enriched for these two properties ($P < 10^{-20}$ by Mann-Whitney U test). We performed a similar analysis for MutSigCV scores finding $P = 3.09 \times 10^{-7}$. See Figure 4.15.

### B.5.2 Sample mutation rates in novel genes

We compared the mutation rates of samples that have a mutation in at least one of the 80 novel genes found by HotNet2 but no mutations in the 139 known cancer genes (reported in [16, 164, 165], or by MutSigCV or Oncodrive) to the mutation rates of samples with at a mutation in at least one of the 139 known cancer genes, but no mutations in the 80 novel genes. We found no significant difference in these rates (mean of 32.26 mutations per sample for novel genes, versus 32.30 mutations per sample for known genes, $P$-value = 0.4 by Mann-Whitney U test), showing that mutations in the novel genes are not restricted to those samples with large numbers of passenger mutations (See Supplementary Figure 4.16).

### B.5.3 Heat scores in the network

The null hypothesis tested by HotNet2 is that the heat scores on genes are independent of the network topology. Two issues that confound this analysis are:

1. The distribution of scores for genes in the network is different from the distribution of heat scores not in the network. We find that this is the case. We performed a two-sample z-test to determine if the heat scores in genes in or out of the PPI network were significant. We repeated this experiment for the mutation frequency and MutSigCV datasets for each of the HINT+HI2012, iRefIndex, and Multinet PPI networks. For each dataset and network combination, the distributions were statistically significantly different, with genes in the network having higher overall heat. We show the CDFs of each distribution, including Z-scores and $P$-values, in Figure 4.25.

   We address this concern by using the score distribution for genes in the network; i.e. our permutation test is restricted to genes in the network.

2. There is a correlation between gene score and degree in the network. We computed the correlation between degree and heat scores. We find the correlation ranges between $\rho = 0.09$ and $\rho = 0.15$ across the three interaction networks and two heat scores (MutSigCV and mutation frequency). However, we found that this correlation was primarily due to the top 100 highest scoring genes in the network. Supplementary Figure 4.26 shows a plot of the correlation as we remove the $N$ highest-scoring genes from each network. After $N = 100$, the correlation has dropped below $\rho = 0.05$ for each network-score pair except Multinet-MutSigCV.

   To assess the effect of the degree-heat correlation on the HotNet2 permutation test, we computed the number of subnetworks found by HotNet2 on the

HINT+HI2012 network, using a permutation test that preserves the correlation between degree and heat. Specifically, we permute the scores of 100 highest scoring genes amongst themselves, and permute the scores of the remaining genes amongst themselves. We generated 1000 such datasets. This permutation test maintains the correlation between degree and heat (Figure 4.24c). The sizes of subnetworks found by HotNet2 were very consistent with those from the original permutation test that permuted heat scores across the nodes uniformly at random (See Figure 4.24a-b). We compared the distributions of the number of components of size at least 3 for both permutation methods, and found that while the permutation method that does not preserve degree produces significantly more components using a t-test ($P = 0.02$), the difference between the distributions is tiny (permuted mean: 25.375, degree permuted mean: 25.294).

### B.5.4 *TP53*, PI(3)K, and NOTCH subnetworks

We show mutation matrices for the PI(3)K, and NOTCH subnetworks in Figures 4.18 and 4.19. The mutation matrix for the *TP53* subnetwork is too large to fit in one page, but is viewable in the online browser of the HotNet2 Pan-Cancer results (Appendix B.4). These subnetworks include genes that were *not* marked as significant by individual gene scores. For example, in the *TP53* subnetwork, mutations in *EPHA3* have been reported in several cancers; we find a cluster of 6 mutations ($P = 0.001$) in a domain that has been shown to significantly reduce the catalytic activity of *EPHA3* [214].

### B.5.5 RTK subnetwork

HotNet2 Pan-Cancer analysis found a subnetwork containing several receptor tyrosine kinases (RTK) that have important roles in the development of multiple cancer types, which includes *EGFR*, *ERBB2* and *ERBB4* (Figure 4.9 and Supplementary Table 11). The RTK subnetwork also contains *ELF3*, a gene involved in cell growth that was recently identified by [65] as a promising candidate cancer gene. *ELF3* is mutated in 19 samples, 11 (58%) of which have inactivating mutations, and is enriched for mutations in BLCA ($P < 3 \times 10^{-5}$) and COADREAD ($P < 0.02$).

### B.5.6 ASCOM complex and interactors

The fifth most mutated consensus subnetwork (16.9% of all samples) contained the ASCOM complex (*MLL2* and *MLL3*), the putative ASCOM-interacting protein encoded by *KDM6A*, as well as *E2F3*, *N4BP2* and *PROSER1* (Figure 4.13 and Supplementary Table 16). The interaction between these latter two genes and *KDM6A* was reported in a mass spectrometry screen [215], and thus is less characterized than the other interactions. The ASCOM complex and *KDM6A* are involved in coordinate deposition of histone modifications that promote transcription. This subnetwork is mutated in at least 5 samples from each cancer type, and is enriched for mutations in BLCA, HNSC, and LUSC. At the individual gene level, we identified a number of enrichment for mutations in different cancer types, including the association of *KDM6A* mutations with BLCA ($P < 10^{-15}$), and the association of *MLL3* mutations with BLCA ($P = 5 \times 10^{-5}$)). Both *MLL3* and *KDM6A* have been previously reported to be associated with transitional cell carcinoma of the bladder[216]. *MLL2* is enriched for mutations in BLCA ($P = 7.5 \times 10^{-7}$), HNSC ($P = 5 \times 10^{-14}$), and LUSC ($P = 1.8 \times 10^{-9}$). *MLL3* and *PROSER1* show significant ($P < 10^{-14}$ and $P < 3\text{x}10^{-4}$, respectively) clustering of missense mutations. *MLL3* mutations are clustered in the zinc-finger domain.Thus HotNet2 identifies novel cancer type specificities for mutations in the ASCOM complex and interactors, as well as some of the recently reported associations for their mutations. *KDM6A* showed significant clustering of mutations ($FDR < 0.1$ from iPAC) in its protein structure.

### B.5.7 SWI/SNF complex

The SWI/SNF subnetworks consist of two consensus subnetworks. One includes the core members *SMARCA4* and *SMARCB1* of SWI/SNF complex, while the other includes SWI/SNF members *ARID1A*, *ARID1B*, *ARID2*, *PBRM1* plus *ADNP*. We jointly analyze these subnetworks as many individual HotNet2 runs report them together and they are members of the same protein complex. Supplementary Table 12 shows the enrichments for mutations for the subnetworks as a whole as well as for the single genes in the SWI/SNF complex subnetwork. See main text for discussion of this subnetwork.

### B.5.8 BAP1 complex

Supplementary Table 13 show the enrichments for mutations for the subnetworks as a whole as well as for the single genes in the BAP1 complex subnetwork (Figure 3b). The subnetwork is enriched for mutations in BLCA ($P = 0.01$), KIRC ($P = 0.00023$),

and LUSC ($P = 0.006$). KIRC enrichment was driven mostly by *BAP1*, while LUSC enrichment was due mostly to mutations in *ASXL1* and *ASXL2*. *BAP1* showed significant ($P < 9 \times 10^{-6}$) clustering of missense mutations in the N-terminal region where mutations have been observed to be associated with mesothelioma [217] and have been shown to inhibit the nuclear localization and deubiquitinase functions of BAP1 [218]). *FOXK1*, *FOXK2*, and *KDM1B* were the genes in the subnetwork with lowest mutation frequencies. *FOXK2*, a putative member of the BAP1 core complex [153], is a transcription factor whose phosphorylation results in apoptosis [219]. None of the mutations in *FOXK2*, were within these critical phosphorylation residues [219], suggesting that they are not perturbing this interaction; however most of the missense mutations in *FOXK2* (6/13) were found in the forkhead transcription factor domain and forkhead associated domain, which may be inactivating the DNA-binding properties of *FOXK2*. With a few exceptions [220, 221, 222], many of these associations are novel.

#### Mutual exclusivity between BAP1 and SWI/SNF complexes

We find that the mutations in the BAP1 and SWI/SNF subnetworks display strong mutual exclusivity ($P = 9.4 \times 10^{-5}$) due to 12 mutually exclusive mutations in the BAP1 complex and 21 mutations (20/21 mutually exclusive) in SWI/SNF complex. These include rare mutations in *ASXL1*, *ASXL2*, *ANKRD17*, *FOXK1*, and *FOXK2* in BAP1 complex (2 of which are inactivating) and *ARID1A*, *SMARCA4*, *ARID2*, *ARID1B* and *SMARCB1* in the SWI/SNF complex (8 of which are inactivating), consistent with the inactivation of *BAP1* or *PBRM1*. We observe that 26 of the 42 KIRC samples with *BAP1* mutations have inactivating mutations. Moreover, 12 of the 16 KIRC samples with missense mutations have these mutations in the Peptidase_C12 domain ($P < 3 \times 10^{-4}$), and thus we classify such missense mutations as inactivating. Using this classification, we observe that 5/10 of mutated samples in BRCA, GBM, HNSC and LUAD are inactivating, suggesting that BAP1 inactivation also occurs in these samples. *ASXL1* mutations are common in LAML [82] and we observe that 5/5 mutations in LAML are inactivating mutations. In addition 6/9 *ASXL1* mutations in HNSC are inactivating, while 3/7 *ASXL2* mutations in BRCA are inactivating. These *ASXL1* and *ASXL2* mutations are exclusive of *BAP1* inactivation, demonstrating alternative strategies for inactivation of the BAP1 complex.

### B.5.9 Core-binding factors

HotNet2 identifies a consensus subnetwork containing *RUNX1*, *CBFB*, and *ELF4* mutated in 2.8% of samples (Figure 4.12 and Supplementary Table 14). The subnetwork is enriched for mutations in LAML ($P = 4.6 \times 10^{-6}$) and BRCA ($P =$

$9.9 \times 10^{-5}$), which is largely driven by *RUNX1* and *CBFB* mutations that are enriched in BRCA (respectively $P = 2 \times 10^{-4}$, $P = 9 \times 10^{-7}$). We observed 16/17 of missense mutations clustered in the Runt-homology domain (RHD) of *RUNX1* ($P = 2 \times 10^{-6}$), which is required for DNA binding as well as interaction with *CBFB* [223], and with the exception of one missense mutation outside of the RHD, all other mutations in *RUNX1* were inactivating. RHD mutations have been associated with development of LAML [224] and our analysis suggests that abrogation of DNA-binding capabilities of *RUNX1* through mutation of *RUNX1*-RHD or *CBFB* may be a critical step in breast cancer progression [225]. Further, the mutations in *RUNX1* that occur in BLCA, COADREAD and HNSC patients were located in the RHD. In addition, each of the mutations in *CBFB* that occur in BLCA, LAML, and UCEC samples were located in the domain that is critical for interaction between *CBFB* and *RUNX1* or were inactivating. These observations suggest that members of the core-binding complex are inactivated by somatic mutations in cancers other than BRCA and LAML, although perhaps only in rare cases. Thus, the HotNet2 Pan-Cancer analysis further characterizes the pattern of mutation in core binding factors across BRCA, AML, and other cancer types.

### B.5.10 Cohesin complex

Supplementary Table 15 show the enrichments for mutations for the subnetworks as a whole as well as for the single genes in the cohesin complex subnetwork. We identified gene-specific enrichments for *STAG2* in BLCA ($P = 0.005$). In STAG1, we found three significant clusters of missense mutations that together encompass 28 total mutations. All of the genes in the subnetwork have $> 79\%$ of their mutations as missense, with the exception of *STAG2* and *RAD21* (57% and 69% respectively). Interestingly, all of the 8 mutations in these two genes in LAML were not missense (3 nonsense for *STAG2*; 2 nonsense, 2 frame shift insertions, and 1 frame shift deletion for *RAD21*), suggesting that these genes are inactivated by mutation in LAML and other cancer types.

### B.5.11 Condensin complex

HotNet2 Pan-Cancer analyses on the MultiNet and iRefIndex networks identified two subnetworks containing six proteins in the condensin complex, which together were mutated in 4.2% of samples. Supplementary Table 6 show the enrichments for mutations for the subnetworks as a whole as well as for the single genes in the condensin complex subnetwork. See main text for discussion of this subnetwork.

### B.5.12 *KEAP1*, *NFE2L2*, and interactors

The seventh most mutated consensus subnetwork (8.5% of all samples) includes *KEAP1*, *NFE2L2*, *CHD6*, *WAC*, and *PTMA* (Figure 4.10 and Supplementary Table 18). This subnetwork is mutated in at least 10 samples of each cancer type with the exception of GBM (3 mutated samples) and LAML (3 mutated samples). *NFE2L2* is a transcription factor that modulates the response to oxidative stress and is normally sequestered in the cytoplasm by *KEAP1* [226]. The subnetwork was enriched for mutations in BLCA ($P = 0.04$), HNSC ($P = 0.0004$), LUAD ($P = 4 \times 10^{-8}$), and LUSC ($P < 10^{-15}$). We find enrichment for mutations in *KEAP1* in LUAD ($P < 10^{-15}$), and in LUSC ($P = 8 \times 10^{-6}$), and enrichment for *NFE2L2* mutations in HNSC ($P = 2 \times 10^{-5}$) and LUSC ($P = 9 \times 10^{-13}$), consistent with previous reports [9, 227, 228]. Missense mutations in *KEAP1* and *NFE2L2* were clustered in their respective protein structures ($P < 4 \times 10^{-5}$ and $P < 10^{-15}$ respectively from NMC analyses, and $Q < 0.025$ for *KEAP1* from iPAC analysis). The 4 clustered mutations for *KEAP1* were all in the position of the NFE2L2 binding Kelch domain [229], and were in different cancer types (2 in LUSC, 1 in LUAD, and 1 in HNSC), suggesting mutation of this domain is important as it is targeted in multiple types. The 36 significantly clustered mutations of *NFE2L2* are in the N-terminal region of the protein.

### B.5.13 MHC Class I proteins

HotNet2 identified a significantly mutated subnetwork (3.1% of samples) containing five genes – *HLA-A*, *CD1D*, *HLA-B*, *B2M*, and *PH4TM* – with *HLA-A* and *HLA-B* of which are from the major histocompatibility complex (MHC) Class I (Figure 4.14 and Supplementary Table 17). Two of these genes were previously reported to harbor somatic mutations: *HLA-A* in squamous lung cancer [9] and *B2M* in colon cancer [230]. However, we find that the mutations in LUSC are distributed throughout the subnetwork, while BLCA mutations are concentrated in *HLA-A* ($P = 0.001$). We find that the genes in this subnetwork harbor many inactivating mutations (41/97 samples), with *B2M* (7/17), *HLA-A* (19/34), and *HLA-B* (10/20) harboring the most. We also find a significant cluster of missense mutations in *B2M*'s first amino acid residue ($P = 1.53 \times 10^{-9}$), suggesting that different cancers utilize different mechanisms for inactivation of *B2M*. Finally, since *HLA-A* and *HLA-B* are highly polymorphic [230], we used more stringent filtering (See `https://www.synapse.org/#!Synapse:syn1729383`) to remove potential germline variants. This procedure classified 14/34 samples with *HLA-A* mutations and 11/21 samples with *HLA-B* mutations as probable germline variants. These results demonstrate that different somatic aberrations, and possibly a combination of germline and somatic aberrations, perturb this subnetwork in different cancer types.

### B.5.14 Telomerase complex and interactors

HotNet2 identified a significantly mutated subnetwork (8.5% of samples) containing two members of the telomerase complex, telomerase reverse transcriptase (*TERT*) and telomerase-associated protein 1 (*TEP1*), as well as *SMG1*, *SMG5*, *SMG6*, and *SMG7* (Figure 4.34 and Supplementary Table 38). HotNet2 identified this subnetwork in the consensus of the runs where no expression filter was applied to the mutation frequency or MutSigCV genes (Appendix B.2.3). The subnetwork is enriched for mutations in LUAD ($P = 0.004$) and LUSC ($P = 0.0009$), and *SMG7* is itself enriched for mutations in LUSC ($P = 0.03$).

### B.5.15 CLASP and CLIP proteins

HotNet2 identified a significantly mutated subnetwork (2% of samples) containing the linker protein CLIP2 and the CLIP associated proteins CLASP1 and CLASP2 (Figure 4.6 and Supplementary Table 7). HotNet2 identified this subnetwork on the Multinet network using mutation frequency gene scores. The CLASP1/2 proteins have been shown to be involved in cellular migration [231], and LOH in the chromosomal region of CLASP2 is associated with non-small cell lung cancer [232]. The association with lung cancer is supported by the mutations in this subnetwork, as the subnetwork ($P = 0.0001$) and both *CLASP1* ($P = 0.01$) and *CLASP2* ($P = 0.008$) are enriched for mutations in LUSC. Furthermore, 7/12 LUSC samples with a mutation in this subnetwork have an inactivating mutation.

## B.6 Mutation validation

To provide further support for a subset of novel cancer genes identified in the HotNet2 subnetworks, we examined RNA-Seq and whole-genome sequencing (WGS) data from the same TCGA samples, data that was generated independently from the whole-exome sequencing data that was used to call the somatic mutations. We focused our analysis on rare mutations in the condensin complex, which was one of the novel discoveries of our analysis (Figure 4.22), as well as genes that were not identified by other methods (marked by '**' in the figures).

Examination of reference and variant allele counts in the RNA-Seq and WGS data validated 39 somatic mutations. These include a total of 12 non-silent mutations in condensin genes using RNA-Seq, 9 non-silent mutations using WGS, and one mutation using both data types. These mutations occur in condensin genes

*SMC2*, *SMC4*, *NCAPD2*, *NCAPD3*, *NCAPH2*, and *NCAPG2*. In the larger set of novel genes, we have validated 8 non-silent mutations using RNA-Seq, 14 non-silent mutations using WGS, and 3 non-silent mutations using both types of data. These include mutations in the cohesin subunit *STAG1*, and the *BAP1* complex subunit *ASXL2*, among others.

Supplementary Table 23 provides the read counts of validated mutations (green rows for RNA-Seq, yellow for WGS, and orange for both), with missing data indicated by -1. This analysis demonstrates that several of the novel genes identified by the HotNet2 analysis contain *bonafide* somatic mutations, and provide promising targets for further functional characterization.

## B.7 Comparison of HotNet2 to HotNet

### B.7.1 Stars and spider graphs in Pan-Cancer data

We compared the proportion of subnetworks returned by HotNet and HotNet2 on the Pan-Cancer mutation data that are star/spider graphs dominated by a hot, central node. We say that a subnetwork $S$ is a hot spider/star graph if the center (root) node $u \in S$ contains $> 50\%$ of the heat in $S$. A star graph of $n > 2$ nodes is a tree where the root has degree $n-1$ and all other nodes have degree one. A spider graph of $n > 3$ nodes is a tree where the root has degree of at least three and all other nodes have degree of at most 2. Note that any star graph of at least 4 nodes is also a spider graph, and that we restrict our analysis to subnetworks of at least 3 nodes as star/spider graphs are defined for subnetworks of at least 3 nodes.

Because HotNet2 uses non-symmetric edge weights to identify strongly connected components, we hypothesized that HotNet2 would identify fewer hot star/spider subnetworks than HotNet. Indeed, HotNet2 returned only 12/90 hot spider/star graphs (13%) compared to 67/94 hot spider/star graphs (71%) returned by HotNet across the three PPI networks and two gene scores (Supplementary Table 31). Thus, HotNet2 returns $> 80\%$ fewer hot stars/spiders than HotNet. This is a major difference between the algorithms and is one of the reasons why HotNet fails to find statistically significant results ($P \leq 0.01$ for any subnetwork size $k$) on three of six runs (Supplementary Table 32,33), while HotNet2 finds statistically significant results on all six runs.

The difference is explained by the undirected vs. directed heat similarity measures used in HotNet vs. HotNet2. In HotNet (undirected heat), placing a high heat score on a node with relatively low-degree is likely to result in a hot star/spider.

Thus, we find hot stars/spiders frequently in randomly permuted data, meaning that the number and size of subnetworks found in real data will not be significant. In contrast, in HotNet2 (directed heat), a single high heat score placed on a random node will not result in a hot star spider. Consequently, by accounting for the direction of heat flow, HotNet2 returns fewer hot stars/spiders on both real and random data, and thus the number and size of subnetworks found in real data is significant.

We note that the goal of HotNet2 is not to eliminate hot stars/spiders, but rather to reduce the number of such subnetworks that are false positives. Due to the incomplete and noisy PPI networks, there are true positive subnetworks that are stars/spiders; e.g. the three most mutated members of the BAP1 complex – *BAP1*, *ASXL1*, *ASXL2* – are a hot star in the iRefIndex PPI network. Restricting attention to the statistically significant subnetworks identified by each algorithm: only 2/36 (5%) of the significant subnetworks reported by HotNet2 are hot stars/spiders, compared to 15/35 (43%) of the significant subnetworks reported by HotNet. Furthermore, only 4/399 (1%) of protein-protein interactions within the HotNet2 subnetworks occur in hot spiders/stars, while 45/245 (18%) of interactions within the HotNet subnetworks occur in hot spiders/stars. It is difficult to determine which, if any, of these stars/spiders are false positives. However, this analysis demonstrates that hot stars/spiders are a much smaller proportion of HotNet2s subnetworks. This results in HotNet2 having higher statistical power and returning subnetworks that have a higher fraction of interactions with proteins other than a hot central node.

### B.7.2 Simulated data

We performed three different simulations comparing HotNet2 and HotNet with: (1) randomized datasets; (2) highly connected subnetworks; (3) implanting highly connected groups of known cancer genes from [16]. The results of (1) show that neither algorithm identifies significantly mutated subnetworks in randomized datasets. The results of (2) and (3) show that HotNet2 achieves high sensitivity and specificity, and outperforms HotNet.

#### Randomized datasets

We evaluated whether HotNet2 and HotNet identify significant subnetworks on randomized datasets by permuting the MutSigCV heat scores 50 times and running both algorithms on the HINT+HI2012 network. For each run, we chose the smallest $\delta$ value for the run with lowest $P$-value across any subnetwork size $k$. On these simulations, we find that the $P$-values for different subnetwork size $k$ are approximately uniformly distributed (Figure 4.27a and Supplementary Table 28), and even appear

to be overly conservative for larger subnetwork sizes. On 450 total subnetwork sizes (9 sizes $\times$ 50 runs), HotNet2 found only 3 subnetwork sizes with $P < 0.01$ (compared to 4 for HotNet). The $P$-values for HotNet and HotNet2 are deflated (more conservative) only at larger $P$-values and for $k \geq 6$, and thus would not result in false negatives. Moreover, we note a larger deflation for HotNet (Figure 4.27a bottom) compared to HotNet2 (Figure 4.27(a) top), demonstrating one advantage of the newer algorithm.

The observed deflation of larger $P$-values is a result of the discreteness of the distribution of our test statistic $X_k$(= number of subnetworks of size at least $k$) for large values of $k$ (Figure 4.27b-c). In particular, on the 50 randomized datasets, we have that for large values of $k$ (e.g., $k \geq 6$) the test statistic $X_k$ assumes an integer value $x*$ within a limited range (e.g. for HotNet2, $X_k$ takes values between 0 and 3 for $k = 7$ (Figure 4.27c) and $k = 8$, and between 0 and 2 for $k = 9$ and $k = 10$ - see Supplementary Table 37). This implies that if $x*$ is the number of subnetworks of size $\geq k$ on a randomly selected dataset, then for large $k$, $x* = 0$ with high probability; e.g. for $k = 7$ we have $Pr[x* = 0] \approx 0.75$ for HotNet. On such a dataset, it follows that the tail probability $Pr[X_k \geq x*] = 1$, resulting in the deflated $p$-values seen on the Q-Q plot for larger $P$-values and large $k$.

**Highly connected subnetworks**

We evaluated the performance of HotNet2 and HotNet in the ideal case where the highest heat scores are placed on genes that each algorithm considers "most connected" in the HINT+HI2012 interaction network. We identified the components to implant by running HotNet and HotNet2 with uniform heat scores on all genes in the network that passed the gene expression filter. We then selected the resulting components of size at least 10, giving us 16 components to implant for each algorithm. These components consisted of 203 genes for HotNet and 214 genes for HotNet2. We then randomly assigned the top mutation frequency scores to the genes in these selected components and randomly assigned the remaining scores to the rest of the genes in the network that had mutation frequency scores and passed expression filtering. HotNet2 returns 16 components of size at least 10 with a $P < 0.01$. These 16 components consist of 223 genes and contain all 214 implanted genes, giving a sensitivity of 100% and a specificity of 96.0%. Furthermore, all genes that are in the same implanted component are in the same recovered component. HotNet, by contrast, performs poorly. It recovers only 12 components of size at least 10 with $P < 0.01$. These 12 components consist of 206 genes of which only 151 are implanted genes, giving a sensitivity of 74.4% and a specificity of 73.3%. HotNet does recover some additional implanted genes in smaller components, but many of these components are of a size that does not achieve statistical significance. Even if we consider the best case for HotNet without regard to statistical significance, it only achieves

sensitivity of 92.6% and specificity of 76.7%.

### Highly connected cancer genes

We compared HotNet2 and HotNet on simulated mutation data where subnetworks of known cancer genes from [16] were mutated randomly in samples according to a specific driver mutation probability and other genes were mutated randomly in samples according to a passenger mutation probability. The passenger mutation probability was lower than the driver mutation probability, but these probabilities were chosen so that the distributions of the number of mutations in cancer genes and passenger genes overlapped (Figure 4.30a). We selected subnetworks to implant by greedily adding the highest weight edges from the diffusion matrix $F$ of the HINT+HI2012 interaction network to identify all non-overlapping subnetworks of size at most 15 that include multiple cancer genes from [16]. We used the subnetworks with at least four genes, resulting in implanted networks of 50 genes in six subnetworks. We also chose 5 known cancer genes that were not part of the implanted networks to have scores from the driver distribution with the hypothesis that HotNet would identify these subnetworks as hot stars/spiders. We then generated 30 simulated datasets by assigning heat scores from overlapping normal distributions for the 50 implanted (mean: 0.05, std: 0.01) and 2945 passenger (mean: 0.01, std: 0.01) genes (Figure 4.30a). We then compared HotNet2 and HotNet on these datasets.[5]

Across the 30 datasets, both HotNet2 and HotNet identify statistically significant subnetworks that overlap the implanted subnetworks with high sensitivity and specificity (Figure 4.30b-c). HotNet2 achieves a better balance in the tradeoff of sensitivity vs. specificity, as summarized by the Youden's index [233] which has a mean value of 0.86 for HotNet2 versus 0.76 for HotNet (difference is significant: $P = 4.2 \times 10^{-11}$ by t-test).

In order to determine if the relative advantage of HotNet2 compared to HotNet came from its selection of the parameter $\delta$, we created ROC curves for each of these datasets varying $\delta$. The ROC curves show that the performance of HotNet2 dominates HotNet with average area under the curve (AUC) = 0.998 for HotNet2 and 0.947 for HotNet. These results demonstrate that HotNet2 has consistently better performance than HotNet over a range of $\delta$ values and hence p-value thresholds.

We note that in practice HotNet2 chooses the value of the $\delta$ parameter auto-

[5]Running HotNet2 and HotNet on simulated datasets requires an automated procedure for choosing which subnetworks to report, which is dependent both on the minimum edge weight $\delta$ and the minimum component size $k$. For each run, we chose the delta with the largest number of $k$ with $P < 0.05$, and chose the $k$ with $P < 0.05$ with the maximum Youden's index [233] ($J = sensitivity + specificity - 1$).

matically using randomized networks (See Section 4.4 and Appendix B.1.4). This is to avoid testing numerous $\delta$ values which can result in overfitting to the data. The ROC curves in Figure 4.30d show that the $\delta$ values automatically chosen by HotNet2 (dots) provide a reasonable balance between sensitivity and specificity.

### B.7.3 Cross-validation

We compared the performance and stability of HotNet2 and HotNet using two-fold cross-validation. We repeatedly split the Pan-Cancer mutation dataset in two halves, by randomly selecting half of the samples from each cancer type. We ran HotNet2 (respectively HotNet) on the first half, finding the value of the parameter $\delta$ and the subnetwork size $k$ that gave the smallest $P$-value under the HotNet2 (respectively HotNet) statistical test. HotNet2 returned significant subnetworks ($P < 0.05$, after multi-hypothesis correction for $\delta$ and $k$ values) for all 15 cross-validation datasets, while HotNet had returned significant subnetworks for only 3 out of the 15 datasets. This result demonstrates that HotNet2 has higher power to detect mutated subnetworks with fewer samples.

Next, we compared the performance of HotNet2 and HotNet in detecting putative cancer genes (genes satisfying the "20/20 Rule"; See Appendix B.8.1) using only 50% of the samples. We summarized the tradeoff in sensitivity vs. specificity of each method using the diagnostic odds ratio (DOR):

$$DOR = \frac{TP/FN}{FP/TN} = \frac{sensitivity \times specificity}{(1 - sensitivity) \times (1 - specificity)}$$

We found that the DORs for the subnetworks found by HotNet2 were significantly higher than those found by HotNet: $14.93 \pm 2.81$ for HotNet2 vs. $7.09 \pm 1.08$ for HotNet (Figure 4.32a), demonstrating that HotNet2 maintains a performance advantage using only 50% of the samples. Moreover, for HotNet2 the DOR was typically much lower when using 50% of the samples than using 100% of samples, demonstrating HotNet2's improved performance with more data. In contrast, the DOR's for HotNet on 50% of samples (including those DORs from runs that were not significant according to HotNet's statistical test) fluctuate around the DOR for 100% of samples, showing that HotNet does not show consistent gains in performance with more data. We suspect that the reason is because the benefits of additional data are outweighed in HotNet (See Section 4.4.3).

Finally, we performed two-fold cross validation on the DOR. We repeatedly computed the values of the parameters $\delta$ and the subnetwork size $k$ from one half of the data (training set) and then used this values of $\delta$ and $k$ to compute subnetworks in the other half of the data (test set). We found that the relative change in DOR

(Figure 4.32b) was much smaller for HotNet2 (-0.0078 ± 0.257) compared to HotNet (-0.481 ± 0.814) demonstrating that HotNet2 is more stable than HotNet.

## B.8 Comparison of HotNet2 to other approaches

### B.8.1 Pathway and gene set analysis

We performed a comparison of the HotNet2 results with standard pathway-based enrichment analyses. Specifically, we used the DAVID [32, 33] and GSEA [31, 30] tools to analyze the genes input to HotNet2 for enrichment in known pathways and gene sets.

These comparisons show that pathway and gene set enrichment approaches produce results that are both qualitatively and quantitatively very different from HotNet2. HotNet2 reports a small number of subnetworks of interacting genes, while gene set methods report large numbers of overlapping and redundant gene sets. The resulting gene sets are difficult to interpret and also do not contain some of the well-known and novel protein complexes identified by HotNet2. The pathway permutation test results also show that many of the subnetworks identified by HotNet2 are significantly enriched for overlap with known pathways/complexes more often than expected by chance. This suggests that HotNet2 provides new insights and is a useful complement (or arguably a replacement for) other pathway tests. We provide details of the pathway and known gene set analysis below.

#### DAVID analysis

We first analyzed all 11,565 mutated genes in the mutation frequency dataset to HotNet2 using DAVID to search for functional enrichment in pathways from the BBID[6], BioCARTA[7], and KEGG [27, 234] databases (Supplementary Table 35(a)). DAVID reported 31 of the tested pathways were enriched with FDR $< 0.05$, 5 of which include "cancer" in their name. These pathways overlapped considerably: the 31 pathways contained 1484 distinct genes, with each gene being contained in an average of 1.6 pathways. We repeated this analysis with the top 200 most mutated genes and DAVID reported 20 of the tested pathways were enriched. These 20 pathways contained only 46 distinct genes, with each gene being contained in an average of 6.4 pathways. Furthermore, none of the pathways found with either

[6] `http://bbid.grc.nia.nih.gov/`

[7] `http://www.biocarta.com/genes/index.asp`

gene list overlap some of the subnetworks (complexes and pathways) discovered by HotNet2 with known (or posited) roles in cancer. For example, the pathways include no genes from the BAP1, condensin, or SWI/SNF complexes.

### GSEA analysis

We also used the GSEA algorithm [31, 30] to identify known gene sets with an over-representation of highly mutated genes. We applied GSEA to the list of 11,565 genes in the mutation frequency dataset to HotNet2 ranked by their mutation frequency, and searched the MSigDB curated gene sets [30]. We restricted the analysis to the 9,270 gene sets of sizes 5-200 that included genes in the input list. GSEA identified 190 enriched gene sets with FDR $< 0.05$, 17 of which included "cancer" in their name (Supplementary Table 35(b)). The significant gene sets overlap considerably, as they include 2028 distinct genes with an average of 2.5 pathways per gene. While the significant gene sets are more comprehensive than the pathways identified by DAVID in terms of the subnetworks (complexes and pathways) identified by HotNet2, they still do not include well-known cancer genes such as *ARID1A*, *MDM4*, or *MLL2*, and completely miss HotNet2 subnetworks such as the ASCOM complex.

### Pathway permutation test

We compared the functional enrichment of HotNet2 subnetworks to that of random subnetworks as measured by overlap with known pathways and protein complexes. We computed pathway/complex enrichment using KEGG pathways and PINdb [29] complexes. We call a subnetwork "enriched" if it has a corrected hypergeometric $P$-value less than 0.05 for any pathway/complex, where we Bonferronni-correct by the product of the number of tested pathways and number of tested subnetworks. We tested each subnetwork identified by HotNet2 for enrichment against each complex and pathway. We find that 15/30 subnetworks are enriched (Supplementary Table 36). The 15 subnetworks without enrichment were mostly small subnetworks (13/15 had size $\leq 7$) where enrichment with large pathways is difficult. We then computed an empirical $P$-value for each of these enriched subnetworks by comparing the minimum corrected $P$-value (over all tested pathways) for a subnetwork against the minimum corrected $P$-value (over all tested pathways) of 1000 random subnetworks of the same size drawn from the same protein-protein interaction network in which HotNet2 identified the subnetwork. We find that 14/15 of the enriched subnetworks have an empirical $P$-value of less than 0.05.

**Identifying putative cancer genes**

We compared the performance of HotNet2 in identifying putative cancer genes to the performance of HotNet and two pathway-based approaches, GSEA and DAVID. Because there is not a comprehensive list of cancer genes to use as a gold-standard, we constructed a list of putative cancer genes following the "20/20 Rule" recently proposed by Vogelstein, *et al.* [16]. The 20/20 rule posits that cancer genes contain either a surprising cluster of missense mutations or enrichment for inactivating mutations. We constructed a list of 278 such "20/20" genes from the genes in the mutation frequency dataset that were mutated in $\geq 1\%$ of TCGA samples. The list was constructed in two steps. First, we identified 155 genes with significant ($P < 5 \times 10^{-5}$) clusters of missense mutations using the NMC algorithm [213]. Next, we identified 123 additional genes that have $\geq 20\%$ of their mutations as inactivating (nonsense, frame shift indels, nonstop, or splice site mutations).

We compared the sensitivity and specificity of HotNet2, HotNet, DAVID, and GSEA in identifying these 278 20/20 genes. Note that we did not include gene-centric methods such as MuSiC, MutSigCV, or Oncodrive in this comparison, because these methods use clustering of missense mutations and/or enrichment of inactivating mutations in deriving their gene scores. In contrast, HotNet2, HotNet, and pathway-based approaches do not consider these signals, and thus the 20/20 gene list provides an independent evaluation of these methods.

We compared the output of HotNet2 and HotNet on the mutation frequency dataset and HINT+HI2012 network to the results of the GSEA and DAVID algorithms applied to the 11,565 genes in that dataset. We found that HotNet2 had a false positive rate (FPR) that was 52%, 61%, 83% lower than HotNet, GSEA, and DAVID, respectively, at equal sensitivity (Supplementary Table 34(a)). To evaluate the performance of each method over a range of sensitivities/specificities, we computed the receiver operator characteristic (ROC) curves for each method by varying the parameter $\delta$ for HotNet and HotNet2, and the FDR for DAVID and GSEA (Figure 4.31a). The ROC curves demonstrate that HotNet2 dominates the other approaches in the range of low FPR (FPR $< 0.1$) that is typically used in the identification of cancer genes [18]. At higher FPR, the HotNet algorithm begins to outperform the other approaches (Figure 4.31b). However, this result should be viewed with caution for two reasons: (1) FPR $> 0.1$ corresponds to more than 1100 false positive genes; (2) these results will not be significant according to HotNet's statistical test.

The ROC curves provide a comprehensive comparison of the tradeoff in sensitivity vs. specificity of the different methods; summarizing this tradeoff in a single number entails some loss of information [235]. Nevertheless, to provide such a number, we computed the partial area under the curve (pAUC), a standard measure for

summarizing the ROC curve in a reduced FPR range. We computed the pAUC up to FPR = 0.1 and standardized the pAUC so that it is 1 for a perfect predictor and 0.5 for a random predictor as in [236]. HotNet2 has an 4%, 4%, and 13% higher standardized pAUC (Supplementary Table 34(b)) than HotNet, GSEA, and DAVID, respectively, demonstrating the advantages of HotNet2 over HotNet and pathway based methods for predicting cancer genes *and* their combinations.

We also test whether HotNet2's advantage over DAVID and GSEA was due solely to HotNet2's (and HotNet's) restriction to genes only in the HINT+HI2012 PPI network. We compared HotNet2, HotNet, DAVID, and GSEA on the reduced set of 6930 genes in HINT+HI2012 that mutated in at least one sample in the mutation dataset. HotNet2 again outperformed the other algorithms, with a false positive rate (FPR) that was 52%, 46%, and 55% lower than HotNet, GSEA, and DAVID, respectively, at equal sensitivity (Supplementary Table 38 (a)). The ROC curve for HotNet2 dominated other methods, both at FPR $< 0.1$ (Figure 4.31c) and FPR $< 0.3$ (Figure 4.31d). For FPR $< 0.1$, HotNet2's pAUC was 4%, 2%, and 8% higher than HotNet, GSEA, and DAVID, respectively (Supplementary Table 34 (b)).

### B.8.2 Comparison to MEMo

We attempted to run MEMo [39] on the 3110 samples in the Pan-Cancer dataset using SNVs and CNAs from the mutation frequency dataset. However, MEMo was not able to run on a dataset this large. In particular, MEMo reported that it was using 11,179 genes after filtering, but terminated with a heap error after 6 minutes on a large memory machine. We note that the challenge of running MEMo on large datasets is reported by the authors who recommend "that the list of recurrently altered genes be kept below 100 [for optimal results]" [61].

### B.8.3 Comparision to Ciriello *et al.* and Lawrence *et al.*

We compare our analysis to the Pan-Cancer study from Ciriello *et al.* [166]. We emphasize that our study and Ciriello *et al.* have very different goals. Ciriello *et al.* focus on the clustering of samples in the Pan-Cancer dataset according to the presence/absence of individually significant mutations. While Ciriello *et al.* manually annotate a few known pathways, they do not assess the clustering of mutations in these pathways, nor do they attempt to identify novel combinations. Specifically, Ciriello *et al.* limit attention to 479 significant mutation events. Only 209 of these events are single gene mutations (or promoter methylation) with the remainder being copy number aberrations that affect many genes. In addition, 36 of the 479 events do not occur in any of the 3299 samples analyzed in Ciriello *et al.*, according to

their mutation matrix. Finally, the clustering of samples further restricts attention to only those events that appear in $\geq 1\%$ of the samples; this implies that only 164 events are used in the first stage of the clustering when all samples are considered. While the mutations considered in Ciriello *et al.* are useful for clustering samples, Ciriello *et al.* do not claim that they form a complete list of driver mutations in these samples.

We also note that Ciriello *et al.* analyze a slightly different dataset than the dataset analyzed by us and several other TCGA Pan-Cancer projects. In particular, compared to the official TCGA Pan-Cancer data freeze [52], Ciriello *et al.* analyze:

- 284 fewer BRCA samples;
- 268 more COADREAD samples; and,
- 131 more OV samples

with other cancers having smaller discrepancies. 2639 samples are common to the datasets.

We also compare our study to Lawrence *et al.* [65], who also recently performed a Pan-Cancer study. Lawrence *et al.* predict cancer genes from somatic SNVs – not including any CNAs – in a larger cohort of samples from 21 cancer types, containing more than 2,000 tumors outside of the TCGA Pan-Cancer cohort. Similar to our analysis, Lawrence *et al.* identify both known and novel cancer genes, and there is overlap between the novel genes identified by both approaches (e.g. *ELF3*, *HLA-A*, *ARID2*, and *ASXL2*). Notably, since HotNet2 analyzes combinations of mutations, we identify these genes as significant using fewer samples. We also identify additional candidate cancer genes not reported in Lawrence *et al.*, but with supporting evidence of their function in cancer from protein interactions and mutation clustering.

# Bibliography


[1] Mark D M Leiserson, Dima Blokh, Roded Sharan, and Benjamin J Raphael. Simultaneous identification of multiple driver pathways in cancer. *PLoS computational biology*, 9(5):e1003054, May 2013.

[2] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. The cancer genome. *Nature*, 458(7239):719–724, April 2009.

[3] Funda Meric-Bernstam and Gordon B Mills. Overcoming implementation challenges of personalized cancer therapy. *Nature reviews. Clinical oncology*, 9(9):542–548, September 2012.

[4] Guillaume Assié, Eric Letouzé, Martin Fassnacht, Anne Jouinot, Windy Luscap, et al. Integrated genomic characterization of adrenocortical carcinoma. *Nature genetics*, 46(6):607–612, June 2014.

[5] The Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *The New England journal of medicine*, 368(22):2059–74, May 2013.

[6] The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–8, October 2008.

[7] Cameron W Brennan, Roel G W Verhaak, Aaron McKenna, Benito Campos, Houtan Noushmehr, and *et al.* The somatic genomic landscape of glioblastoma. *Cell*, 155(2):462–77, October 2013.

[8] The Cancer Genome Atlas Research Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, October 2012.

[9] Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–525, September 2012.

[10] Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, 499(7456):43–49, July 2013.

[11] The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–15, June 2011.

[12] The Cancer Genome Atlas Research Network, W Marston Linehan, Paul T Spellman, Christopher J Ricketts, Chad J Creighton, et al. Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. *The New England journal of medicine*, 374(2):135–145, January 2016.

[13] Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. *Cell*, 159(3):676–690, October 2014.

[14] Thomas J Hudson, Warwick Anderson, Axel Artez, Anna D Barker, Cindy Bell, and *et al.* International network of cancer genome projects. *Nature*, 464(7291):993–8, April 2010.

[15] Benjamin J Raphael, Jason R Dobson, Layla Oesper, and Fabio Vandin. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome medicine*, 6(1):5, January 2014.

[16] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis a Diaz, and Kenneth W Kinzler. Cancer genome landscapes. *Science (New York, N.Y.)*, 339(6127):1546–58, March 2013.

[17] Michael S Lawrence, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, and *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–8, July 2013.

[18] Douglas Hanahan and Robert a Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–74, March 2011.

[19] Pauline C Ng and Steven Henikoff. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–3814, July 2003.

[20] Ivan A Adzhubei, Steffen Scmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimoa, et al. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–249, 2010.

[21] Boris Reva, Yevgeniy Antipin, and Chris Sander. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research*, 39(17):e118, September 2011.

[22] Levi A Garraway and Eric S Lander. Lessons from the cancer genome. *Cell*, 153(1):17–37, March 2013.

[23] Craig H Mermel, Steven E Schumacher, Barbara Hill, Matthew L Meyerson, Rameen Beroukhim, and Gad Getz. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome biology*, 12(4):R41, January 2011.

[24] Hsin-Ta Wu, Iman Hajirasouliha, and Benjamin J Raphael. Detecting independent and recurrent copy number aberrations using interval graphs. *Bioinformatics (Oxford, England)*, 30(12):i195–203, June 2014.

[25] Abel Gonzalez-Perez and Nuria Lopez-Bigas. Functional impact bias reveals cancer drivers. *Nucleic acids research*, pages 1–10, August 2012.

[26] Nathan D Dees, Qunyuan Zhang, Cyriac Kandoth, Michael C Wendl, William Schierding, et al. MuSiC: identifying mutational significance in cancer genomes. *Genome research*, 22(8):1589–98, August 2012.

[27] M Kanehisa and S Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, Jan 2000.

[28] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, 40(Database issue):D109–114, 2012.

[29] Phuong-Van Luc and Paul Tempst. PINdb: a database of nuclear protein complexes from human and yeast. *Bioinformatics (Oxford, England)*, 20(9):1413–1415, June 2004.

[30] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, October 2005.

[31] Vamsi K Mootha, Cecilia M Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics*, 34(3):267–273, July 2003.

[32] D W Huang, B T Sherman, and R A Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, 4:44–57, 2008.

[33] D W Huang, B T Sherman, and R A Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37:1–13, 2009.

[34] Fabio Vandin, Eli Upfal, and Benjamin J Raphael. Algorithms for detecting significantly mutated pathways in cancer. *Journal of Computational Biology*, 18(3):507–22, March 2011.

[35] Enrico Glaab, Anaïs Baudot, Natalio Krasnogor, Reinhard Schneider, and Alfonso Valencia. EnrichNet: network-based gene set enrichment analysis. *Bioinformatics (Oxford, England)*, 28(18):i451–i457, September 2012.

[36] Roman K Thomas, Alissa C Baker, Ralph M Debiasi, Wendy Winckler, Thomas Laframboise, et al. High-throughput oncogene mutation profiling in human cancer. *Nature genetics*, 39(3):347–351, March 2007.

[37] Chen-Hsiang Yeang, Frank McCormick, and Arnold Levine. Combinatorial patterns of somatic gene mutations in cancer. *The FASEB journal*, 22(8):2605–22, August 2008.

[38] Fabio Vandin, Eli Upfal, and Benjamin J Raphael. De novo discovery of mutated driver pathways in cancer. *Genome research*, 22(2):375–85, February 2012.

[39] Giovanni Ciriello, Ethan Cerami, Chris Sander, and Nikolaus Schultz. Mutual exclusivity analysis identifies oncogenic network modules. *Genome research*, 22(2):398–406, February 2012.

[40] Christopher a Miller, Stephen H Settle, Erik P Sulman, Kenneth D Aldape, and Aleksandar Milosavljevic. Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. *BMC medical genomics*, 4(1):34, January 2011.

[41] Daniel C Koboldt, Qunyuan Zhang, David E Larson, Dong Shen, Michael D McLellan, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, 22(3):568–576, March 2012.

[42] David E Larson, Christopher C Harris, Ken Chen, Daniel C Koboldt, Travis E Abbott, David J Dooling, Timothy J Ley, Elaine R Mardis, Richard K Wilson, and Li Ding. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, 28(3):311–317, 2012.

[43] Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*, 31(3):213–219, March 2013.

[44] Ruibin Xi, Angela G Hadjipanayis, Lovelace J Luquette, Tae-Min Kim, Eunjung Lee, et al. Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proceedings of the National Academy of Sciences of the United States of America*, 108(46):E1128–36, November 2011.

[45] Derek Y Chiang, Gad Getz, David B Jaffe, Michael J T O'Kelly, Xiaojun Zhao, et al. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature methods*, 6(1):99–103, January 2009.

[46] K Lindblad-Toh, D M Tanenbaum, M J Daly, E Winchester, W O Lui, A Villapakkam, et al. Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *Nature biotechnology*, 18(9):1001–1005, September 2000.

[47] R Mei, P C Galipeau, C Prass, A Berno, G Ghandour, et al. Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. *Genome research*, 10(8):1126–1137, August 2000.

[48] Mark D M Leiserson, Hsin-ta Wu, Fabio Vandin, and Benjamin J Raphael B. Research in Computational Molecular Biology. In Teresa M. Przytycka, editor, *RECOMB 2015: The 19th Annual International Conference on Research in Computational Molecular Biology*, volume 9029 of *Lecture Notes in Computer Science*, pages 202–204, Warsaw, 2015. Springer International Publishing.

[49] Mark Dm Leiserson, Hsin-Ta Wu, Fabio Vandin, and Benjamin J Raphael. CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome biology*, 16:160, 2015.

[50] Mark DM Leiserson, Matthew A Reyna, and Benjamin J Raphael. 2016. In submission: *European Conference on Computational Biology*.

[51] Mark D M Leiserson, Fabio Vandin, Hsin-Ta Wu, Jason R Dobson, Jonathan V Eldridge, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature genetics*, 47(2):106–114, February 2015.

[52] Kyle Chang, Chad J Creighton, Caleb Davis, Lawrence Donehower, Jennifer Drummond, and *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120, September 2013.

[53] Mark D M Leiserson, Connor C Gramazio, Jason Hu, Hsin-Ta Wu, David H Laidlaw, and Benjamin J Raphael. MAGI: visualization and collaborative annotation of genomic aberrations. *Nature methods*, 12(6):483–484, June 2015.

[54] Charles Lu, Mingchao Xie, Michael C Wendl, Jiayin Wang, Michael D McLellan, Mark D M Leiserson, Kuan-Lin Huang, Matthew A Wyczalkowski, Reyka Jayasinghe, Tapahsama Banerjee, Jie Ning, Piyush Tripathi, Qunyuan Zhang, Beifang Niu, Kai Ye, Heather K Schmidt, Robert S Fulton, Joshua F McMichael, Prag Batra, Cyriac Kandoth, Maheetha Bharadwaj, Daniel C Koboldt, Christopher A Miller, Krishna L Kanchi, James M Eldred, David E Larson, John S Welch, Ming You, Bradley A Ozenberger, Ramaswamy Govindan, Matthew J Walter, Matthew J Ellis, Elaine R Mardis, Timothy A Graubert, John F Dipersio, Timothy J Ley, Richard K Wilson, Paul J Goodfellow, Benjamin J Raphael, Feng Chen, Kimberly J Johnson, Jeffrey D Parvin, and Li Ding. Patterns and functional implications of rare germline variants across 12 cancer types. *Nature communications*, 6:10086, 2015.

[55] Adam J. Bass, Vesteinn Thorsson, Ilya Shmulevich, Sheila M. Reynolds, Michael Miller, and *et al.* Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, 513(7517):202–209, July 2014.

[56] Junjun Zhang, Joachim Baran, a Cros, Jonathan M Guberman, Syed Haider, and *et al.* International Cancer Genome Consortium Data Portal–a one-stop shop for cancer genomics data. *Database : the journal of biological databases and curation*, 2011:bar026, January 2011.

[57] Jesse J Salk, Edward J Fox, and Lawrence a Loeb. Mutational heterogeneity in human cancers: origin and consequences. *Annual review of pathology*, 5:51–75, January 2010.

[58] Li Ding, Benjamin J Raphael, Feng Chen, and Michael C Wendl. Advances for studying clonal evolution in cancer. *Cancer letters*, 340(2):212–9, November 2013.

[59] Junfei Zhao, Shihua Zhang, Ling-Yun Wu, and Xiang-Sun Zhang. Efficient methods for identifying mutated driver pathways in cancer. *Bioinformatics (Oxford, England)*, 28(22):2940–7, November 2012.

[60] Hai-Tao Li, Yu-Lang Zhang, Chun-Hou Zheng, and Hong-Qiang Wang. Simulated annealing based algorithm for identifying mutated driver pathways in cancer. *BioMed research international*, 2014:375980, January 2014.

[61] Giovanni Ciriello, Ethan Cerami, Bulent Arman Aksoy, Chris Sander, and Nikolaus Schultz. Using MEMo to discover mutual exclusivity modules in cancer. *Current protocols in bioinformatics*, 41(8.17.1):8.17.2, March 2013.

[62] Ewa Szczurek and Niko Beerenwinkel. Modeling mutual exclusivity of cancer mutations. *PLoS computational biology*, 10(3):e1003503, March 2014.

[63] Özgün Babur, Mithat Gönen, Bülent Arman Aksoy, Nikolaus Schultz, Giovanni Ciriello, Chris Sander, and Emek Demir. Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome biology*, 16(1):1–10, 2015.

[64] Mark D M Leiserson, Hsin-ta Wu, Fabio Vandin, and Benjamin J Raphael B. Research in Computational Molecular Biology. In Teresa M. Przytycka, editor, *RECOMB 2015: The 19th Annual International Conference on Research in Computational Molecular Biology*, volume 9029 of *Lecture Notes in Computer Science*, pages 202–204, Warsaw, 2015. Springer International Publishing.

[65] Michael S Lawrence, Petar Stojanov, Craig H Mermel, James T Robinson, Levi a Garraway, and *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495–501, January 2014.

[66] HO Lancaster. Significance tests in discrete distributions. *Journal of the American Statistical Association*, 56(294):223–234, 1961.

[67] Martin Dyer, Ravi Kannan, and John Mount. Sampling Contingency Tables. *Random Structures and Algorithms*, 10(4):487–506, 1997.

[68] Cyrus R Mehta and Nitin R Patel. A network algorithm for performing fisher's exact test in r× c contingency tables. *Journal of the American Statistical Association*, 78(382):427–434, 1983.

[69] F Requena and N Martín Ciudad. A major improvement to the network algorithm for fisher's exact test in 2× c contingency tables. *Computational statistics & data analysis*, 51(2):490–498, 2006.

[70] A Berrington de González and DR Cox. Additive and multiplicative models for the joint effect of two risk factors. *Biostatistics*, 6(1):1–9, 2005.

[71] Gill Bejerano, Nir Friedman, and Naftali Tishby. Efficient exact p-value computation for small sample, sparse, and surprising categorical data. *Journal of Computational Biology*, 11(5):867–886, 2004.

[72] Daniel Zelterman, ISF Chan, and PW Mielke Jr. Exact tests of significance in higher dimensional tables. *The American Statistician*, 49(4):357–361, 1995.

[73] Alexander Barvinok, Zur Luria, and Alexander Yong. An approximation algorithm for counting contingency tables. *Random Structures and Algorithms*, 37(1):25–66, 2010.

[74] Jeffrey W. Miller and Matthew T. Harrison. Exact sampling and counting for fixed-margin matrices. *The Annals of Statistics*, 41(3):1569–1592, June 2013.

[75] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087, 1953.

[76] WK Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[77] Andrea Gobbi, Francesco Iorio, Kevin J Dawson, David C Wedge, David Tamborero, and *et al.* Fast randomization of large genomic datasets while preserving alteration counts. *Bioinformatics (Oxford, England)*, 30(17):i617–i623, September 2014.

[78] Roel G W Verhaak, Katherine a Hoadley, Elizabeth Purdom, Victoria Wang, Yuan Qi, and *et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer cell*, 17(1):98–110, January 2010.

[79] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[80] Linghua Wang, Shigeru Yamaguchi, Matthew D Burstein, Keita Terashima, Kyle Chang, Ho-Keung Ng, Hideo Nakamura, Zongxiao He, Harshavardhan Doddapaneni, Lora Lewis, Mark Wang, Tomonari Suzuki, Ryo Nishikawa, Atsushi Natsume, Shunsuke Terasaka, Robert Dauser, William Whitehead, Adesina Adekunle, Jiayi Sun, Yi Qiao, Gábor Marth, Donna M Muzny, Richard a Gibbs, Suzanne M Leal, David a Wheeler, and Ching C Lau. Novel somatic and germline mutations in intracranial germ cell tumours. *Nature*, pages 860–862, 2014.

[81] Maria E Figueroa, Omar Abdel-Wahab, Chao Lu, Patrick S Ward, Jay Patel, Alan Shih, Yushan Li, Neha Bhagwat, Aparna Vasanthakumar, Hugo F Fernandez, et al. Leukemic idh1 and idh2 mutations result in a hypermethylation phenotype, disrupt tet2 function, and impair hematopoietic differentiation. *Cancer cell*, 18(6):553–567, 2010.

[82] Omar Abdel-Wahab and Ross L Levine. Mutations in epigenetic modifiers in the pathogenesis and therapy of acute myeloid leukemia. *Blood*, 121(18):3563–3572, 2013.

[83] Klaus H Metzeler, Kati Maharry, Michael D Radmacher, Krzysztof Mrózek, Dean Margeson, Heiko Becker, John Curfman, Kelsi B Holland, Sebastian Schwind, Susan P Whitman, et al. Tet2 mutations improve the new european leukemianet risk classification of acute myeloid leukemia: a cancer and leukemia group b study. *Journal of clinical oncology*, 29(10):1373–1381, 2011.

[84] D Kamnasaran, N Ajewung, M Rana, and P Gould. 393 npas3 is a novel late-stage acting progression factor in gliomas with tumour suppressive functions. *European Journal of Cancer Supplements*, 8(5):100, 2010.

[85] Frederico Moreira, Tim-Rasmus Kiehl, Kelvin So, Norbert F Ajeawung, Carmelita Honculada, Peter Gould, Russell O Pieper, and Deepak Kamnasaran. Npas3 demonstrates features of a tumor suppressive role in driving the

progression of astrocytomas. *The American journal of pathology*, 179(1):462–476, 2011.

[86] S Thalappilly, P Soubeyran, JL Iovanna, and NJ Dusetti. Vav2 regulates epidermal growth factor receptor endocytosis and degradation. *Oncogene*, 29(17):2528–2539, 2010.

[87] S Rea, G Xouri, and A Akhtar. Males absent on the first (mof): from flies to humans. *Oncogene*, 26(37):5385–5394, 2007.

[88] Jan Mollenhauer, Stefan Wiemann, Wolfram Scheurlen, Bernhard Korn, Yutaka Hayashi, Klaus K Wilgenbus, Andreas von Deimling, and Annemarie Poustka. Dmbt1, a new member of the srcr superfamily, on chromosome 10q25. 3–26.1 is deleted in malignant brain tumours. *Nature genetics*, 17(1):32–39, 1997.

[89] Kazuya Motomura, Michel Mittelbronn, Werner Paulus, Benjamin Brokinkel, Kathy Keyvani, Ulrich Sure, Karsten Wrede, Yoichi Nakazato, Yuko Tanaka, Daniela Pierscianek, et al. Dmbt1 homozygous deletion in diffuse astrocytomas is associated with unfavorable clinical outcome. *Journal of Neuropathology & Experimental Neurology*, 71(8):702–707, 2012.

[90] James S Hale, Balint Otvos, Maksim Sinyuk, Alvaro G Alvarado, Masahiro Hitomi, Kevin Stoltz, Qiulian Wu, William Flavahan, Bruce Levison, Mette L Johansen, et al. Cancer stem cell-specific scavenger receptor cd36 drives glioblastoma progression. *Stem Cells*, 32(7):1746–1758, 2014.

[91] Guiomar Solanas, Carme Cortina, Marta Sevillano, and Eduard Batlle. Cleavage of e-cadherin by adam10 mediates epithelial cell sorting downstream of ephb signalling. *Nature cell biology*, 13(9):1100–1107, 2011.

[92] Ju-Hee Lee, Eun-Goo Jeong, Moon-Chang Choi, Sung-Hak Kim, Jung-Hyun Park, Sang-Hyun Song, Jinah Park, Yung-Jue Bang, and Tae-You Kim. Inhibition of histone deacetylase 10 induces thioredoxin-interacting protein and causes accumulation of reactive oxygen species in snu-620 human gastric cancer cells. *Molecules and cells*, 30(2):107–112, 2010.

[93] Chen Shochat, Noa Tal, Obul R Bandapalli, Chiara Palmi, Ithamar Ganmore, Geertruy te Kronnie, Gunnar Cario, Giovanni Cazzaniga, Andreas E Kulozik, Martin Stanulla, et al. Gain-of-function mutations in interleukin-7 receptor-$\alpha$ (il7r) in childhood acute lymphoblastic leukemias. *The Journal of experimental medicine*, 208(5):901–908, 2011.

[94] Karolina Holm, Johan Staaf, Göran Jönsson, Johan Vallon-Christersson, Haukur Gunnarsson, Adalgeir Arason, Linda Magnusson, Rosa B Barkardottir, Cecilia Hegardt, Markus Ringnér, et al. Characterisation of amplification patterns and target genes at chromosome 11q13 in ccnd1-amplified sporadic and familial breast tumours. *Breast cancer research and treatment*, 133(2):583–594, 2012.

[95] F Graziano, B Humar, and P Guilford. The role of the E-cadherin gene (CDH1) in diffuse gastric cancer susceptibility: from the laboratory to clinical practice. *Annals of oncology*, pages 1705–1713, 2003.

[96] Shunsuke Hiraguri, Tony Godfrey, Haruhiko Nakamura, and Jeremy Graff. Mechanisms of inactivation of E-cadherin in breast cancer cell lines. *Cancer Research*, 1:1972–1978, 1998.

[97] Jerry Usary, Victor Llaca, Gamze Karaca, Shafaq Presswala, Mehmet Karaca, Xiaping He, Anita Langerød, Rolf Kåresen, Daniel S Oh, Lynn G Dressler, et al. Mutation of gata3 in human breast tumors. *Oncogene*, 23(46):7669–7678, 2004.

[98] Jonathan Chou, Jeffrey H Lin, Audrey Brenot, Jung-whan Kim, Sylvain Provot, and Zena Werb. Gata3 suppresses metastasis and modulates the tumour microenvironment by regulating microrna-29b expression. *Nature cell biology*, 15(2):201–213, 2013.

[99] Wei Yan, Qing Jackie Cao, Richard B Arenas, Brooke Bentley, and Rong Shao. GATA3 inhibits breast cancer metastasis through the reversal of epithelial-mesenchymal transition. *The Journal of biological chemistry*, 285(18):14042–51, April 2010.

[100] Pamela Cowin, Tracey M Rowlands, and Sarah J Hatsell. Cadherins and catenins in breast cancer. *Current opinion in cell biology*, 17(5):499–508, 2005.

[101] Andrew R Green, Sophie Krivinskas, Peter Young, Emad A Rakha, E Claire Paish, Desmond G Powe, and Ian O Ellis. Loss of expression of chromosome 16q genes dpep1 and ctcf in lobular carcinoma in situ of the breast. *Breast cancer research and treatment*, 113(1):59–66, 2009.

[102] Galina N Filippova, Chen-Feng Qi, Jonathan E Ulmer, James M Moore, Michael D Ward, Ying J Hu, Dmitri I Loukinov, Elena M Pugacheva, Elena M Klenova, Paul E Grundy, et al. Tumor-associated zinc finger mutations in the ctcf transcription factor selectively alter its dna-binding specificity. *Cancer research*, 62(1):48–52, 2002.

[103] Zhen Zhao, Chi-Chao Chen, Cory D Rillahan, Ronglai Shen, Thomas Kitzing, Megan E McNerney, Ernesto Diaz-Flores, Johannes Zuber, Kevin Shannon, Michelle M Le Beau, Mona S Spector, Scott C Kogan, and Scott W Lowe. Cooperative loss of RAS feedback regulation drives myeloid leukemogenesis. *Nature Genetics*, 47(5):539–543, 2015.

[104] Fabio Vandin, Benjamin J. Raphael, and Elil Upfal. Research in Computational Molecular Biology. In Teresa M. Przytycka, editor, *RECOMB 2015: The 19th Annual International Conference on Research in Computational Molecular Biology*, volume 9029 of *Lecture Notes in Computer Science*, pages 326–337, Warsaw, 2015. Springer International Publishing.

[105] The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, July 2012.

[106] The Cancer Genome Atlas Research Network, Cyriac Kandoth, Nikolaus Schultz, Andrew D Cherniack, Rehan Akbani, et al. Integrated genomic characterization of endometrial carcinoma. *Nature*, 497(7447):67–73, May 2013.

[107] The Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. *Cell*, 159(3):676–690, October 2014.

[108] Michael C Wendl, John W Wallis, Ling Lin, Cyriac Kandoth, Elaine R Mardis, Richard K Wilson, and Li Ding. PathScan: a tool for discerning mutational significance in groups of putative cancer genes. *Bioinformatics (Oxford, England)*, 27(12):1595–602, June 2011.

[109] Vamsi K Mootha, Cecilia M Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics*, 34(3):267–273, July 2003.

[110] Fabio Vandin, Eli Upfal, and Benjamin J Raphael. Algorithms for detecting significantly mutated pathways in cancer. *Journal of computational biology : a journal of computational molecular cell biology*, 18(3):507–522, March 2011.

[111] Giovanni Ciriello, Ethan Cerami, Chris Sander, and Nikolaus Schultz. Mutual exclusivity analysis identifies oncogenic network modules. *Genome research*, 22(2):398–406, February 2012.

[112] Matthew Ruffalo, Mehmet Koyutürk, and Roded Sharan. Network-Based Integration of Disparate Omic Data To Identify "Silent Players" in Cancer. *PLoS computational biology*, 11(12):e1004595, December 2015.

[113] Mark D M Leiserson, Fabio Vandin, Hsin-Ta Wu, Jason R Dobson, Jonathan V Eldridge, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature genetics*, 47(2):106–114, February 2015.

[114] Christopher A Miller, Stephen H Settle, Erik P Sulman, Kenneth D Aldape, and Aleksandar Milosavljevic. Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. *BMC medical genomics*, 4:34, 2011.

[115] Fabio Vandin, Patrick Clay, Eli Upfal, and Benjamin J Raphael. Discovery of mutated subnetworks associated with clinical data in cancer. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 55–66, 2012.

[116] Mark D M Leiserson, Dima Blokh, Roded Sharan, and Benjamin J Raphael. Simultaneous identification of multiple driver pathways in cancer. *PLoS Computational Biology*, 9(5):e1003054, 2013.

[117] Ewa Szczurek and Niko Beerenwinkel. Modeling mutual exclusivity of cancer mutations. *PLoS computational biology*, 10(3):e1003503, March 2014.

[118] Simona Constantinescu, Ewa Szczurek, Pejman Mohammadi, Jörg Rahnenführer, and Niko Beerenwinkel. TiMEx: a waiting time model for mutually exclusive cancer alterations. *Bioinformatics (Oxford, England)*, July 2015.

[119] Özgün Babur, Mithat Gönen, Bülent Arman Aksoy, Nikolaus Schultz, Giovanni Ciriello, et al. Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome biology*, 16:45, 2015.

[120] Mark D M Leiserson, Hsin-Ta Wu, Fabio Vandin, and Benjamin J Raphael. CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome biology*, 16(1):160, 2015.

[121] Yoo-Ah Kim, Dong-Yeon Cho, Phuong Dao, and Teresa M Przytycka. MEMCover: integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types. *Bioinformatics (Oxford, England)*, 31(12):i284–92, June 2015.

[122] Timothy J Ley, Christopher Miller, Li Ding, Benjamin J Raphael, Andrew J Mungall, et al. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *New England Journal of Medicine*, 368(22):2059–2074, 2013.

[123] Steven A Roberts and Dmitry A Gordenin. Hypermutation in human cancer genomes: footprints and mechanisms. *Nature Reviews Cancer*, 14(12):786–800, 2014.

[124] David Manescu and Uri Keich. A Symmetric Length-Aware Enrichment Test. *RECOMB*, pages 224–242, 2015.

[125] Yili Hong. On computing the distribution function for the Poisson binomial distribution. *Computational Statistics & Data Analysis*, pages 41–51, 2013.

[126] Ronald W Butler. *Saddlepoint approximations with applications*, volume 22. Cambridge University Press, 2007.

[127] Junhua Zhang, Ling-Yun Wu, Xiang-Sun Zhang, and Shihua Zhang. Discovery of co-occurring driver pathways in cancer. *BMC bioinformatics*, 15:271, 2014.

[128] Y Benjamini and Y Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*, 57:289–300, 1995.

[129] The Cancer Genome Atlas Research Network. Genomic Classification of Cutaneous Melanoma. *Cell*, 161(7):1681–1696, June 2015.

[130] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, et al.

[131] Annette Woiwode, Sandra A S Johnson, Shuping Zhong, Cheng Zhang, Robert G Roeder, et al. PTEN represses RNA polymerase III-dependent transcription by targeting the TFIIIB complex. *Molecular and cellular biology*, 28(12):4204–4214, June 2008.

[132] I Hlavata, B Mohelnikova-Duchonova, R Vaclavikova, V Liska, P Pitule, P Novak, et al. The role of ABC transporters in progression and clinical outcome of colorectal cancer. *Mutagenesis*, 27(2):187–196, March 2012.

[133] I Miklós and J Podani. Randomization of presence-absence matrices: comments and new algorithms. *Ecology*, 85(1):86–92, 2004.

[134] Katherine A Hoadley, Christina Yau, Denise M Wolf, Andrew D Cherniack, David Tamborero, Sam Ng, Max D M Leiserson, Beifang Niu, Michael D McLellan, Vladislav Uzunangelov, Jiashan Zhang, Cyriac Kandoth, Rehan Akbani, Hui Shen, Larsson Omberg, Andy Chu, Adam A Margolin, Laura J Van't Veer, Nuria Lopez-Bigas, Peter W Laird, Benjamin J Raphael, Li Ding, A Gordon Robertson, Lauren A Byers, Gordon B Mills, John N Weinstein, Carter Van Waes, Zhong Chen, Eric A Collisson, Cancer Genome Atlas Research Network, Christopher C Benz, Charles M Perou, and Joshua M Stuart. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4):929–944, August 2014.

[135] Krishna L Kanchi, Kimberly J Johnson, Charles Lu, Michael D McLellan, Mark D M Leiserson, et al. Integrated analysis of germline and somatic variants in ovarian cancer. *Nature communications*, 5:3156, 2014.

[136] Cancer Genome Atlas Research Network, Cyriac Kandoth, Nikolaus Schultz, Andrew D Cherniack, Rehan Akbani, Yuexin Liu, Hui Shen, A Gordon Robertson, Itai Pashtan, Ronglai Shen, Christopher C Benz, Christina Yau, Peter W Laird, Li Ding, Wei Zhang, Gordon B Mills, Raju Kucherlapati, Elaine R Mardis, and Douglas A Levine. Integrated genomic characterization of endometrial carcinoma. *Nature*, 497(7447):67–73, May 2013.

[137] Catherine S Grasso, Yi-Mi Wu, Dan R Robinson, Xuhong Cao, Saravana M Dhanasekaran, Amjad P Khan, Michael J Quist, Xiaojun Jing, Robert J Lonigro, J Chad Brenner, Irfan A Asangani, Bushra Ateeq, Sang Y Chun, Javed Siddiqui, Lee Sam, Matt Anstett, Rohit Mehra, John R Prensner, Nallasivam Palanisamy, Gregory A Ryslik, Fabio Vandin, Benjamin J Raphael, Lakshmi P Kunju, Daniel R Rhodes, Kenneth J Pienta, Arul M Chinnaiyan, and Scott A Tomlins. The mutational landscape of lethal castration-resistant prostate cancer. *Nature*, 487(7406):239–243, July 2012.

[138] Matan Hofree, John P Shen, Hannah Carter, Andrew Gross, and Trey Ideker. Network-based stratification of tumor mutations. *Nature methods*, 10(11):1108–1115, November 2013.

[139] Jishnu Das and Haiyuan Yu. HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC systems biology*, 6:92, 2012.

[140] Haiyuan Yu, Leah Tardivo, Stanley Tam, Evan Weiner, Fana Gebreab, Changyu Fan, Nenad Svrzikapa, Tomoko Hirozane-Kishikawa, Edward Rietman, Xinping Yang, Julie Sahalie, Kourosh Salehi-Ashtiani, Tong Hao, Michael E Cusick, David E Hill, Frederick P Roth, Pascal Braun, and Marc Vidal. Next-generation sequencing to generate interactome datasets. *Nature methods*, 8(6):478–480, June 2011.

[141] Sabry Razick, George Magklaras, and Ian M Donaldson. iRefIndex: a consolidated protein interaction database with provenance. *BMC bioinformatics*, 9:405, 2008.

[142] Ekta Khurana, Yao Fu, Jieming Chen, and Mark Gerstein. Interpretation of genomic variants using a unified biological network approach. *PLoS computational biology*, 9(3):e1002886, 2013.

[143] David Tamborero, Nuria Lopez-Bigas, and Abel Gonzalez-Perez. Oncodrive-CIS: a method to reveal likely driver genes based on the impact of their copy number changes on expression. *PloS one*, 8(2):e55489, 2013.

[144] Gregory A Ryslik, Yuwei Cheng, Kei-Hoi Cheung, Yorgo Modis, and Hongyu Zhao. Utilizing protein structure to identify non-random somatic mutations. *BMC Bioinformatics*, 14:190, 2013.

[145] Annemarie Greife, Silvia Jankowiak, Jochen Steinbring, Parvaneh Nikpour, Günter Niegisch, Michèle J Hoffmann, and Wolfgang A Schulz. Canonical Notch signalling is inactive in urothelial carcinoma. *BMC cancer*, 14:628, 2014.

[146] Boris G Wilson and Charles W M Roberts. SWI/SNF nucleosome remodellers and cancer. *Nature reviews. Cancer*, 11(7):481–492, July 2011.

[147] Ignacio Varela, Patrick Tarpey, Keiran Raine, Dachuan Huang, Choon Kiat Ong, Philip Stephens, Helen Davies, David Jones, Meng-Lay Lin, Jon Teague, Graham Bignell, Adam Butler, Juok Cho, Gillian L Dalgliesh, Danushka Galappaththige, Chris Greenman, Claire Hardy, Mingming Jia, Calli Latimer, King Wai Lau, John Marshall, Stuart McLaren, Andrew Menzies, Laura Mudie, Lucy Stebbings, David A Largaespada, L F A Wessels, Stephane Richard, Richard J Kahnoski, John Anema, David A Tuveson, Pedro A Perez-Mancera, Ville Mustonen, Andrej Fischer, David J Adams, Alistair Rust, Waraporn Chan-on, Chutima Subimerb, Karl Dykema, Kyle Furge, Peter J Campbell, Bin Tean Teh, Michael R Stratton, and P Andrew Futreal. Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature*, 469(7331):539–542, January 2011.

[148] Cigall Kadoch, Diana C Hargreaves, Courtney Hodges, Laura Elias, Lena Ho, Jeff Ranish, and Gerald R Crabtree. Proteomic and bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human malignancy. *Nature genetics*, 45(6):592–601, June 2013.

[149] Mark Sausen, Rebecca J Leary, Siân Jones, Jian Wu, C Patrick Reynolds, Xueyuan Liu, Amanda Blackford, Giovanni Parmigiani, Luis A Diaz, Nickolas Papadopoulos, Bert Vogelstein, Kenneth W Kinzler, Victor E Velculescu, and Michael D Hogarty. Integrated genomic analyses identify ARID1A and ARID1B alterations in the childhood cancer neuroblastoma. *Nature genetics*, 45(1):12–17, January 2013.

[150] Yoshinori Tsurusaki, Nobuhiko Okamoto, Hirofumi Ohashi, Tomoki Kosho, Yoko Imai, Yumiko Hibi-Ko, Tadashi Kaname, Kenji Naritomi, Hiroshi Kawame, Keiko Wakui, Yoshimitsu Fukushima, Tomomi Homma, Mitsuhiro Kato, Yoko Hiraki, Takanori Yamagata, Shoji Yano, Seiji Mizuno, Satoru Sakazume, Takuma Ishii, Toshiro Nagai, Masaaki Shiina, Kazuhiro Ogata, Tohru Ohta, Norio Niikawa, Satoko Miyatake, Ippei Okada, Takeshi

Mizuguchi, Hiroshi Doi, Hirotomo Saitsu, Noriko Miyake, and Naomichi Matsumoto. Mutations affecting components of the SWI/SNF complex cause Coffin-Siris syndrome. *Nature genetics*, 44(4):376–378, April 2012.

[151] Shmuel Mandel and Illana Gozes. Activity-dependent neuroprotective protein constitutes a novel element in the SWI/SNF chromatin remodeling complex. *The Journal of biological chemistry*, 282(47):34448–34456, November 2007.

[152] R A Steingart and I Gozes. Recombinant activity-dependent neuroprotective protein protects cells against oxidative stress. *Molecular and cellular endocrinology*, 252(1-2):148–153, June 2006.

[153] Michele Carbone, Haining Yang, Harvey I Pass, Thomas Krausz, Joseph R Testa, and Giovanni Gaudino. BAP1 and cancer. *Nature reviews. Cancer*, 13(3):153–159, March 2013.

[154] Samuel Peña-Llopis, Silvia Vega-Rubín-de Celis, Arnold Liao, Nan Leng, Andrea Pavía-Jiménez, Shanshan Wang, Toshinari Yamasaki, Leah Zhrebker, Sharanya Sivanand, Patrick Spence, Lisa Kinch, Tina Hambuch, Suneer Jain, Yair Lotan, Vitaly Margulis, Arthur I Sagalowsky, Pia Banerji Summerour, Wareef Kabbani, S W Wendy Wong, Nick Grishin, Marc Laurent, Xian-Jin Xie, Christian D Haudenschild, Mark T Ross, David R Bentley, Payal Kapur, and James Brugarolas. BAP1 loss defines a new class of renal cell carcinoma. *Nature genetics*, 44(7):751–759, July 2012.

[155] Rui Fang, Andrew J Barbera, Yufei Xu, Michael Rutenberg, Thiago Leonor, Qing Bi, Fei Lan, Pinchao Mei, Guo-Cheng Yuan, Christine Lian, Junmin Peng, Dongmei Cheng, Guangchao Sui, Ursula B Kaiser, Yang Shi, and Yujiang Geno Shi. Human LSD2/KDM1b/AOF1 regulates gene transcription by modulating intragenic H3K4me2 methylation. *Molecular cell*, 39(2):222–233, July 2010.

[156] Yujiang Shi, Fei Lan, Caitlin Matson, Peter Mulligan, Johnathan R Whetstine, Philip A Cole, Robert A Casero, and Yang Shi. Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. *Cell*, 119(7):941–953, December 2004.

[157] Huiling Xu, Jonathan M Tomaszewski, and Michael J McKay. Can corruption of chromosome cohesion create a conduit to cancer? *Nature reviews. Cancer*, 11(3):199–210, March 2011.

[158] Eric D Rubio, David J Reiss, Piri L Welcsh, Christine M Disteche, Galina N Filippova, Nitin S Baliga, Ruedi Aebersold, Jeffrey A Ranish, and Anton Krumm. CTCF physically links cohesin to chromatin. *Proceedings of the National Academy of Sciences of the United States of America*, 105(24):8309–8314, June 2008.

[159] Dominic Schmidt, Petra C Schwalie, Caryn S Ross-Innes, Antoni Hurtado, Gordon D Brown, Jason S Carroll, Paul Flicek, and Duncan T Odom. A CTCF-independent role for cohesin in tissue-specific transcription. *Genome research*, 20(5):578–588, May 2010.

[160] Ayana Kon, Lee-Yung Shih, Masashi Minamino, Masashi Sanada, Yuichi Shiraishi, Yasunobu Nagata, Kenichi Yoshida, Yusuke Okuno, Masashige Bando, Ryuichiro Nakato, Shumpei Ishikawa, Aiko Sato-Otsubo, Genta Nagae, Aiko Nishimoto, Claudia Haferlach, Daniel Nowak, Yusuke Sato, Tamara Alpermann, Masao Nagasaki, Teppei Shimamura, Hiroko Tanaka, Kenichi Chiba, Ryo Yamamoto, Tomoyuki Yamaguchi, Makoto Otsu, Naoshi Obara, Mamiko Sakata-Yanagimoto, Tsuyoshi Nakamaki, Ken Ishiyama, Florian Nolte, Wolf-Karsten Hofmann, Shuichi Miyawaki, Shigeru Chiba, Hiraku Mori, Hiromitsu Nakauchi, H Phillip Koeffler, Hiroyuki Aburatani, Torsten Haferlach, Katsuhiko Shirahige, Satoru Miyano, and Seishi Ogawa. Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. *Nature genetics*, 45(10):1232–1237, October 2013.

[161] Andrew J Wood, Aaron F Severson, and Barbara J Meyer. Condensin and cohesin complexity: the expanding repertoire of functions. *Nature reviews. Genetics*, 11(6):391–404, June 2010.

[162] Tatsuya Hirano. Condensins: universal organizers of chromosomes with diverse functions. *Genes & development*, 26(15):1659–1678, August 2012.

[163] Jacques Lapointe, Sameer Malhotra, John P Higgins, Eric Bair, Maxwell Thompson, Keyan Salari, Craig P Giacomini, Michelle Ferrari, Kelli Montgomery, Robert Tibshirani, Matt van de Rijn, James D Brooks, and Jonathan R Pollack. hCAP-D3 expression marks a prostate cancer subtype with favorable clinical behavior and androgen signaling signature. *The American journal of surgical pathology*, 32(2):205–209, February 2008.

[164] Cyriac Kandoth, Michael D McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, Qunyuan Zhang, Joshua F McMichael, Matthew A Wyczalkowski, Mark D M Leiserson, Christopher A Miller, John S Welch, Matthew J Walter, Michael C Wendl, Timothy J Ley, Richard K Wilson, Benjamin J Raphael, and Li Ding. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471):333–339, October 2013.

[165] Travis I Zack, Stephen E Schumacher, Scott L Carter, Andre D Cherniack, Gordon Saksena, Barbara Tabak, Michael S Lawrence, Cheng-Zhong Zhsng, Jeremiah Wala, Craig H Mermel, Carrie Sougnez, Stacey B Gabriel, Bryan Hernandez, Hui Shen, Peter W Laird, Gad Getz, Matthew Meyerson, and Rameen Beroukhim. Pan-cancer patterns of somatic copy number alteration. *Nature genetics*, 45(10):1134–1140, October 2013.

[166] Giovanni Ciriello, Martin L Miller, Bülent Arman Aksoy, Yasin Senbabaoglu, Nikolaus Schultz, and Chris Sander. Emerging landscape of oncogenic signatures across human cancers. *Nature genetics*, 45(10):1127–1133, October 2013.

[167] Koyel Mitra, Anne-Ruxandra Carvunis, Sanath Kumar Ramesh, and Trey Ideker. Integrative approaches for finding modular structure in biological networks. *Nature reviews. Genetics*, 14(10):719–732, October 2013.

[168] *The heat kernel as the pagerank of a graph*, volume 104, 2007.

[169] P Berkhin. Bookmark-coloring algorithm for personalized pagerank computing. *Internet Mathematics*, 3:41–62, 2006.

[170] Michael P Schroeder, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. Visualizing multidimensional cancer genomics data. *Genome medicine*, 5(1):9, 2013.

[171] The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–7, July 2012.

[172] Li Ding, Michael C Wendl, Joshua F McMichael, and Benjamin J Raphael. Expanding the computational toolbox for mining cancer genomes. *Nature reviews. Genetics*, 15(8):556–570, August 2014.

[173] Jianjiong Gao, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S Onur Sumer, Yichao Sun, Anders Jacobsen, Rileen Sinha, Erik Larsson, Ethan Cerami, Chris Sander, and Nikolaus Schultz. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling*, 6(269):pl1, April 2013.

[174] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J Byrne, Michael L Heuer, Erik Larsson, Yevgeniy Antipin, Boris Reva, Arthur P Goldberg, Chris Sander, and Nikolaus Schultz. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery*, 2(5):401–404, May 2012.

[175] Melissa S Cline, Brian Craft, Teresa Swatloski, Mary Goldman, Singer Ma, David Haussler, and Jingchun Zhu. Exploring TCGA Pan-Cancer data at the UCSC Cancer Genomics Browser. *Scientific reports*, 3:2652, 2013.

[176] M Bostock, V Ogievetsky, and J Heer. D3: data-driven documents. *Visualization and Computer ...*, 2011.

[177] Scott L Carter, Kristian Cibulskis, Elena Helman, Aaron McKenna, Hui Shen, Travis Zack, Peter W Laird, Robert C Onofrio, Wendy Winckler, Barbara A Weir, Rameen Beroukhim, David Pellman, Douglas A Levine, Eric S Lander, Matthew Meyerson, and Gad Getz. Absolute quantification of somatic DNA alterations in human cancer. *Nature biotechnology*, 30(5):413–421, May 2012.

[178] T S Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, Lavanya Balakrishnan, Arivusudar Marimuthu, Sutopa Banerjee, Devi S Somanathan, Aimy Sebastian, Sandhya Rani, Somak Ray, C J Harrys Kishore, Sashi Kanth, Mukhtar Ahmed, Manoj K Kashyap, Riaz Mohmood, Y L Ramachandra, V Krishna, B Abdul Rahiman, Sujatha Mohan, Prathibha Ranganathan, Subhashri Ramabadran, Raghothama Chaerkady, and Akhilesh Pandey. Human Protein Reference Database–2009 update. *Nucleic acids research*, 37(Database issue):D767–72, January 2009.

[179] Aron Marchler-Bauer, Chanjuan Zheng, Farideh Chitsaz, Myra K Derbyshire, Lewis Y Geer, Renata C Geer, Noreen R Gonzales, Marc Gwadz, David I Hurwitz, Christopher J Lanczycki, Fu Lu, Shennan Lu, Gabriele H Marchler, James S Song, Narmada Thanki, Roxanne A Yamashita, Dachuan Zhang, and Stephen H Bryant. CDD: conserved domains and protein three-dimensional structure. *Nucleic acids research*, 41(Database issue):D348–52, January 2013.

[180] Marco Punta, Penny C Coggill, Ruth Y Eberhardt, Jaina Mistry, John Tate, Chris Boursnell, Ningze Pang, Kristoffer Forslund, Goran Ceric, Jody Clements, Andreas Heger, Liisa Holm, Erik L L Sonnhammer, Sean R Eddy, Alex Bateman, and Robert D Finn. The Pfam protein families database. *Nucleic acids research*, 40(Database issue):D290–301, January 2012.

[181] *SMART, a simple modular architecture research tool: identification of signaling domains*, volume 95, 1998.

[182] Ivica Letunic, Tobias Doerks, and Peer Bork. SMART 7: recent updates to the protein domain annotation resource. *Nucleic acids research*, 40(Database issue):D302–5, January 2012.

[183] Ramin Nazarian, Hubing Shi, Qi Wang, Xiangju Kong, Richard C Koya, Hane Lee, Zugen Chen, Mi-Kyung Lee, Narsis Attar, Hooman Sazegar, Thinle Chodon, Stanley F Nelson, Grant McArthur, Jeffrey A Sosman, Antoni Ribas, and Roger S Lo. Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK or N-RAS upregulation. *Nature*, 468(7326):973–977, December 2010.

[184] Wade S Samowitz, Carol Sweeney, Jennifer Herrick, Hans Albertsen, Theodore R Levin, Maureen A Murtaugh, Roger K Wolff, and Martha L Slattery. Poor survival associated with the BRAF V600E mutation in microsatellite-stable colon cancers. *Cancer research*, 65(14):6063–6069, July 2005.

[185] Tim Dwyer. Scalable, Versatile and Simple Constrained Graph Layout. *Comput. Graph. Forum ()*, 28(3):991–998, 2009.

[186] Elizabeth A Musgrove, C Elizabeth Caldon, Jane Barraclough, Andrew Stone, and Robert L Sutherland. Cyclin D as a therapeutic target in cancer. *Nature reviews. Cancer*, 11(8):558–572, August 2011.

[187] Michael E Cusick, Haiyuan Yu, Alex Smolyar, Kavitha Venkatesan, Anne-Ruxandra Carvunis, Nicolas Simonis, Jean-François Rual, Heather Borick, Pascal Braun, Matija Dreze, Jean Vandenhaute, Mary Galli, Junshi Yazaki, David E Hill, Joseph R Ecker, Frederick P Roth, and Marc Vidal. Literature-curated protein interaction datasets. *Nature methods*, 6(1):39–46, January 2009.

[188] Hadassah Sade, Sudhir Krishna, and Apurva Sarin. The anti-apoptotic effect of Notch-1 requires p56lck-dependent, Akt/PKB-mediated signaling in T cells. *The Journal of biological chemistry*, 279(4):2937–2944, January 2004.

[189] Brian J Raney, Timothy R Dreszer, Galt P Barber, Hiram Clawson, Pauline A Fujita, Ting Wang, Ngan Nguyen, Benedict Paten, Ann S Zweig, Donna

Karolchik, and W James Kent. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics (Oxford, England)*, 30(7):1003–1005, April 2014.

[190] Helga Thorvaldsdóttir, James T Robinson, and Jill P Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2):178–192, March 2013.

[191] Mary Goldman, Brian Craft, Teresa Swatloski, Kyle Ellrott, Melissa Cline, Mark Diekhans, Singer Ma, Chris Wilks, Josh Stuart, David Haussler, and Jingchun Zhu. The UCSC Cancer Genomics Browser: update 2013. *Nucleic acids research*, 41(Database issue):D949–54, January 2013.

[192] Jeremy Goecks, Carl Eberhard, Tomithy Too, Galaxy Team, Anton Nekrutenko, and James Taylor. Web-based visual analysis for high-throughput genomics. *BMC genomics*, 14:397, 2013.

[193] Gunes Gundem, Christian Perez-Llamas, Alba Jene-Sanz, Anna Kedzierska, Abul Islam, Jordi Deu-Pons, Simon J Furney, and Nuria Lopez-Bigas. IntOGen: integration and data mining of multidimensional oncogenomic data. *Nature methods*, 7(2):92–93, February 2010.

[194] Abel Gonzalez-Perez, Christian Perez-Llamas, Jordi Deu-Pons, David Tamborero, Michael P Schroeder, Alba Jene-Sanz, Alberto Santos, and Nuria Lopez-Bigas. IntOGen-mutations identifies cancer drivers across tumor types. *Nature methods*, 10(11):1081–1082, November 2013.

[195] Junjun Zhang, Joachim Baran, A Cros, Jonathan M Guberman, Syed Haider, Jack Hsu, Yong Liang, Elena Rivkin, Jianxin Wang, Brett Whitty, Marie Wong-Erasmus, Long Yao, and Arek Kasprzyk. International Cancer Genome Consortium Data Portal–a one-stop shop for cancer genomics data. *Database : the journal of biological databases and curation*, 2011:bar026, 2011.

[196] Siddavaram Nagini. Carcinoma of the stomach: A review of epidemiology, pathogenesis, molecular genetics and chemoprevention. *World journal of gastrointestinal oncology*, 4(7):156–169, July 2012.

[197] R Houlston, S Bevan, A Williams, J Young, M Dunlop, P Rozen, C Eng, D Markie, K Woodford-Richens, M A Rodriguez-Bigas, B Leggett, K Neale, R Phillips, E Sheridan, S Hodgson, T Iwama, D Eccles, W Bodmer, and I Tomlinson. Mutations in DPC4 (SMAD4) cause juvenile polyposis syndrome, but only account for a minority of cases. *Human molecular genetics*, 7(12):1907–1912, November 1998.

[198] Joan Massagué. TGFbeta in Cancer. *Cell*, 134(2):215–230, July 2008.

[199] C Oliveira, R Seruca, M Seixas, and M Sobrinho-Simões. The clinicopathological features of gastric carcinomas with microsatellite instability may be mediated by mutations of different "target genes": a study of the TGFbeta RII, IGFII R, and BAX genes. *The American journal of pathology*, 153(4):1211–1219, October 1998.

[200] Seok-Hyung Kim, Seung-Hyun Lee, Yoon-La Choi, Li-Hui Wang, Cheol Keun Park, and Young Kee Shin. Extensive alteration in the expression profiles of TGFB pathway signaling components and TP53 is observed along the gastric dysplasia-carcinoma sequence. *Histology and histopathology*, 23(12):1439–1452, December 2008.

[201] Swati Biswas, Patricia Trobridge, Judith Romero-Gallo, Dean Billheimer, Lois L Myeroff, James K V Willson, Sanford D Markowitz, and William M Grady. Mutational inactivation of TGFBR2 in microsatellite unstable colon cancer arises from the cooperation of genomic instability and the clonal outgrowth of transforming growth factor beta resistant cells. *Genes, chromosomes & cancer*, 47(2):95–106, February 2008.

[202] Naresh Bellam and Boris Pasche. Tgf-beta signaling alterations and colon cancer. *Cancer treatment and research*, 155:85–103, 2010.

[203] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012.

[204] João Vinagre, Ana Almeida, Helena Pópulo, Rui Batista, Joana Lyra, Vasco Pinto, Ricardo Coelho, Ricardo Celestino, Hugo Prazeres, Luis Lima, Miguel Melo, Adriana Gaspar da Rocha, Ana Preto, Patrícia Castro, Ligia Castro, Fernando Pardal, José Manuel Lopes, Lúcio Lara Santos, Rui Manuel Reis, José Cameselle-Teijeiro, Manuel Sobrinho-Simões, Jorge Lima, Valdemar Máximo, and Paula Soares. Frequency of TERT promoter mutations in human cancers. *Nature communications*, 4:2185, 2013.

[205] Franklin W Huang, Eran Hodis, Mary Jue Xu, Gregory V Kryukov, Lynda Chin, and Levi A Garraway. Highly recurrent TERT promoter mutations in human melanoma. *Science (New York, N.Y.)*, 339(6122):957–959, February 2013.

[206] Nils Weinhold, Anders Jacobsen, Nikolaus Schultz, Chris Sander, and William Lee. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nature genetics*, 46(11):1160–1165, November 2014.

[207] Collin Melton, Jason A Reuter, Damek V Spacek, and Michael Snyder. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nature genetics*, 47(7):710–716, July 2015.

[208] Simon a Forbes, Nidhi Bindal, Sally Bamford, Charlotte Cole, Chai Yin Kok, and *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic acids research*, 39(Database issue):D945–50, January 2011.

[209] L Tierney. Markov-Chains for Exploring Posterior Distributions. *Annals of Statistics*, 22(4):1701–1728, December 1994.

[210] Sebastian Köhler, Sebastian Bauer, Denise Horn, and Peter N Robinson. Walking the interactome for prioritization of candidate disease genes. *American journal of human genetics*, 82(4):949–58, April 2008.

[211] T Müller, S Rahmann, and M Rehmsmeier. Non-symmetric score matrices and the detection of homologous transmembrane proteins. *Bioinformatics (Oxford, England)*, 17 Suppl 1:S182–9, January 2001.

[212] R. Milo, N. Kashtan, S. Itzkovitz, M. E. J. Newman, and U. Alon. On the uniform generation of random graphs with prescribed degree sequences. *arXiv*, 2003.

[213] Jingjing Ye, Adam Pavlicek, Elizabeth A Lunney, Paul A Rejto, and Chi-Hse Teng. Statistical method on nonrandom clustering with application to somatic mutations in cancer. *BMC bioinformatics*, 11(1):11, January 2010.

[214] Tara L Davis, John R Walker, Peter Loppnau, Christine Butler-Cole, Abdellah Allali-Hassani, and Sirano Dhe-Paganon. Autoregulation by the juxtamembrane region of the human ephrin receptor tyrosine kinase A3 (EphA3). *Structure*, 16(6):873–84, June 2008.

[215] A A Zada, J A Pulikkan, D Bararia, M Geletu, A K Trivedi, M Y Balkhi, W D Hiddemann, D G Tenen, H M Behre, and G Behre. Proteomic discovery of max as a novel interacting partner of c/ebpalpha: a myc/max/mad link. *Leukemia*, 20(12):2137–46, Dec 2006.

[216] Yaoting Gui, Guangwu Guo, Yi Huang, Xueda Hu, Aifa Tang, Shengjie Gao, Renhua Wu, Chao Chen, Xianxin Li, Liang Zhou, Minghui He, Zesong Li, Xiaojuan Sun, Wenlong Jia, Jinnong Chen, Shangming Yang, Fangjian Zhou, Xiaokun Zhao, Shengqing Wan, Rui Ye, Chaozhao Liang, Zhisheng Liu, Peide Huang, Chunxiao Liu, Hui Jiang, Yong Wang, Hancheng Zheng, Liang Sun, Xingwang Liu, Zhimao Jiang, Dafei Feng, Jing Chen, Song Wu, Jing Zou, Zhongfu Zhang, Ruilin Yang, Jun Zhao, Congjie Xu, Weihua Yin, Zhichen Guan, Jiongxian Ye, Hong Zhang, Jingxiang Li, Karsten Kristiansen, Michael L Nickerson, Dan Theodorescu, Yingrui Li, Xiuqing Zhang, Songgang Li, Jian Wang, Huanming Yang, Jun Wang, and Zhiming Cai. Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder. *Nat Genet*, 43(9):875–8, Sep 2011.

[217] Matthew Bott, Marie Brevet, Barry S Taylor, Shigeki Shimizu, Tatsuo Ito, Lu Wang, Jenette Creaney, Richard A Lake, Maureen F Zakowski, Boris Reva, Chris Sander, Robert Delsite, Simon Powell, Qin Zhou, Ronglai Shen, Adam Olshen, Valerie Rusch, and Marc Ladanyi. The nuclear deubiquitinase bap1 is commonly inactivated by somatic mutations and 3p21.1 losses in malignant pleural mesothelioma. *Nat Genet*, 43(7):668–72, Jul 2011.

[218] Karen H Ventii, Narra S Devi, Kenneth L Friedrich, Tatiana A Chernova, Mourad Tighiouart, Erwin G Van Meir, and Keith D Wilkinson. Brca1-associated protein-1 is a tumor suppressor that requires deubiquitinating activity and nuclear localization. *Cancer Res*, 68(17):6953–62, Sep 2008.

[219] Anett Marais, Zongling Ji, Emma S Child, Eberhard Krause, David J Mann, and Andrew D Sharrocks. Cell cycle-dependent regulation of the forkhead transcription factor FOXK2 by CDKcyclin complexes. *The Journal of biological chemistry*, 285(46):35728–39, November 2010.

[220] M Katoh. Functional and cancer genomics of asxl family members. *Br J Cancer*, Jun 2013.

[221] Nivedita Sahu and Jennifer Rubin Grandis. New advances in molecular approaches to head and neck squamous cell carcinoma. *Anticancer Drugs*, 22(7):656–64, Aug 2011.

[222] Qianxing Mo, Sijian Wang, Venkatraman E Seshan, Adam B Olshen, Nikolaus Schultz, Chris Sander, R Scott Powers, Marc Ladanyi, and Ronglai Shen. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences of the United States of America*, 110(11):4245–50, March 2013.

[223] J Bravo, Z Li, N A Speck, and A J Warren. The leukemia-associated AML1 (Runx1)–CBF beta complex functions as a DNA-induced molecular clamp. *Nature structural biology*, 8(4):371–8, April 2001.

[224] Jörg Cammenga, Birte Niebuhr, Stefan Horn, Ulla Bergholz, Gabriele Putz, Frank Buchholz, Jürgen Löhler, and Carol Stocking. RUNX1 DNA-binding mutants, associated with minimally differentiated acute myelogenous leukemia, disrupt myeloid differentiation. *Cancer research*, 67(2):537–45, January 2007.

[225] Kevin a Janes. RUNX1 and its understudied role in breast cancer. *Cell cycle (Georgetown, Tex.)*, 10(20):3461–5, October 2011.

[226] H C Huang, T Nguyen, and C B Pickett. Regulation of the antioxidant response element by protein kinase c-mediated phosphorylation of nf-e2-related factor 2. *Proc Natl Acad Sci U S A*, 97(23):12475–80, Nov 2000.

[227] Michael B Sporn and Karen T Liby. Nrf2 and cancer: the good, the bad and the importance of context. *Nat Rev Cancer*, 12(8):564–71, Aug 2012.

[228] Tatsuhiro Shibata, Tsutomu Ohta, Kit I Tong, Akiko Kokubu, Reiko Odogawa, Koji Tsuta, Hisao Asamura, Masayuki Yamamoto, and Setsuo Hirohashi. Cancer related mutations in nrf2 impair its recognition by keap1-cul3 e3 ligase and promote malignancy. *Proc Natl Acad Sci U S A*, 105(36):13568–73, Sep 2008.

[229] K Itoh, N Wakabayashi, Y Katoh, T Ishii, K Igarashi, J D Engel, and M Yamamoto. Keap1 represses nuclear activation of antioxidant responsive elements by nrf2 through binding to the amino-terminal neh2 domain. *Genes Dev*, 13(1):76–86, Jan 1999.

[230] R Dawkins, C Leelayuwat, S Gaudieri, G Tay, J Hui, S Cattley, P Martinez, and J Kulski. Genomics of the major histocompatibility complex: haplotypes, duplication, retroviruses and disease. *Immunol Rev*, 167:275–304, Feb 1999.

[231] Ksenija Drabek, Marco van Ham, Tatiana Stepanova, Katharina Draegestein, Remco van Horssen, Carmen Laura Sayas, Anna Akhmanova, Timo Ten Hagen, Ron Smits, Riccardo Fodde, Frank Grosveld, and Niels Galjart. Role of CLASP2 in microtubule stabilization and the regulation of persistent motility. *Current biology : CB*, 16(22):2259–64, November 2006.

[232] a Maitra, I I Wistuba, C Washington, a K Virmani, R Ashfaq, S Milchgrub, a F Gazdar, and J D Minna. High-resolution chromosome 3p allelotyping of breast carcinomas and precursor lesions demonstrates frequent loss of heterozygosity and a discontinuous pattern of allele loss. *The American journal of pathology*, 159(1):119–30, July 2001.

[233] W.J. Youden. Index for rating diagnostic tests. *Cancer*, 3:32–35, 1950.

[234] Minoru Kanehisa. Molecular network analysis of diseases and drugs in kegg. *Methods Mol Biol*, 939:263–75, 2013.

[235] Lobo J.M., Jimnez-Valverde A., and Real R. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17(2):141–145, 2008.

[236] McClish D.K. Analyzing a portion of the ROC curve. *Medical Decision Making*, 9(3):190–195, 1989.