Deciphering Disease Genomes in a Network Context

Borislav Hrisimirov Hristov

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Computer Science
Adviser: Professor Mona Singh

September 2019

Abstract

Despite the incredible influx of sequencing data, pinpointing the gene variants responsible for the development of heterogeneous diseases remains a particularly hard task because the same phenotypic outcome (disease) can result from a myriad of combinations of different alterations across the genome. A promising avenue is to consider genome alterations within the context of pathways instead of genes because different alterations within any of several genes comprising the same pathway can have similar consequences with respect to disease development. Large-scale biological networks provide a helpful proxy for biological pathway knowledge as genes that participate in the same pathway tend to interact with each other and form modules within the larger network. In this dissertation, I introduce two novel methods that further our ability to computationally highlight potential disease-causing genes by examining disease genomes in the context of biological networks.

First, in Chapter 2, I present a novel network-based approach which tackles cancer mutational heterogeneity by utilizing per-individual mutational profiles. I provide an intuitive formulation relying on balancing the size of a connected subgraph within the larger network with covering many patients. I describe a machine learning-like schema for selecting the value of the single required parameter and both an integer linear programming framework and a fast heuristic for optimizing the objective function. I demonstrate the outstanding performance of my method in identifying cancer-relevant genes, especially those mutated at very low rates.

Next, in Chapter 3, I propose a general computational framework that uses prior knowledge of disease-associated genes to guide a network-based search for novel ones based upon newly acquired information. I use a graph diffusion kernel to spread the signal from the set of already known disease genes and then use it to bias a random walk originating from the newly implicated genes to move closer to the known ones. I demonstrate that integrating the two types of information is better than using

either one of them alone. I show, in the context of cancer, that my method readily outperforms other network-based methods. Finally, I apply my approach to several complex diseases, thereby demonstrating its versatility in a broad range of settings.

Acknowledgements

My journey to a Ph.D has been a long trek on the winding path from youthful, cocky ignorance to a mature, thoughtful uncertainty. As I have grown and transformed, both personally and academically, I have come to appreciate how very little I know about the beautiful world we live in. And among the very few things I know is that I succeeded in my journey only because of the exceptional and unwavering support I have received from my advisor, friends, and family.

First and foremost, I want to thank my advisor Professor Mona Singh for her incredible mentorship. From the first day I joined her lab, Mona has provided me with invaluable advice and guidance. She painstakingly read every document I sent her, showed patience and found ways to get my research moving when progress was halted. Mona taught me how to write as a scientist, how to conduct research that is up to the highest standards, and most importantly, how to ask the right questions. I have been lucky to have incredible lab-mates as well— Chaitanya Aluru, Judy Du, Anat Etzion-Fuchs, Dario Ghersi, Eric Glynn, Peng Jiang, Daniel Munro, Antonio Muscarella, Sean King, Shilpa Kobren, Anton Persikov, Yuri Pritykin, Pawel Przytycki, Joshua Wetzel and José Zamalloa. They helped me on countless occasions. I would also like to express gratitude to my committee members: Olga Troyanskaya, Barbara Engelhardt, Ben Raphael, and Bernard Chazelle for always making time in their busy schedules to provide me with feedback and valuable insights on this dissertation.

During my time at Princeton, I had the privilege to interact with and become close friends with graduate students from very different fields than mine. John Martin, Matthew DCI, CJ Verbeck, Darl Lewis, Farzan Baros, Felicity Hills, Ashley Linder, and Eric Hubble— your friendship and support have made my time at Princeton immensely more enjoyable. We ventured together through grad school, travelled far away on bold hiking trips (Arvind Pawan, Link Patrick, and Jack Matt), lived together (Peter Humphries and Nick DeLuca), and helped each other when needed. But what

I would always remember are our hours-long dinners at Proctor Hall filled with heated discussions on every imaginable topic. These conversations were truly intellectually invigorating; they challenged my perception of the world, provided me with a window to peek through into the beautiful minds of young and inspiring Princeton doctoral students, and enriched me tremendously as a person. I also want to thank my good friends from undergraduate years— Kevin Lee, John Kawamura, Shane Easter, Josh Pascual, Jose Loya and Nick Vargas— for visiting, for organizing trips and events, and reminding me of the existence of the outside world. I need to give a shout out to the basketball crew at Dillon gym as well— the guys were always there, ready to play basketball when I needed to get some exercise and relax after a tedious day in the lab.

I am extremely grateful to my parents and sister for their constant love and support throughout my life. Although they live across the Atlantic, I always felt them close by, I always benefited from their encouragement and silent, unshakable belief in me. I would like to especially thank my grandmather Hristina for her warm, kind words every time we spoke on the phone, but above all, for instilling in me during my early childhood the love for learning and the passion for science which ultimately put me onto the long, winding path of becoming a scientist.

Chapter 2 of this dissertation has been published in *Cell Systems* [35]. I thank my coauthor Mona Singh for allowing me to include our joint work here. Mona Singh is also a collaborator on Chapter 3. Materials included in this dissertation have been publicly presented at the *Research in Computational Molecular Biology Conference* (RECOMB, Hong Kong, HK, May 2017), *Society for Industrial and Applied Mathematics Annual Meeting* (SIAM, Denver, CO, June 2018), and *Computational Biology Seminar at NCBI* (Bethesda, MD, May 2019). My Ph.D. work was supported by Princeton University, Princeton's Computer Science Department, Princeton's Gor-

don Wu Fellowship (to BHH), the National Institutes of Health (R01-CA208148 to MS), and the Forese Family Fund for Innovation. These funding bodies played no role in study design, data collection and analysis, nor in the writing of this dissertation.

To my family.

Contents

List of Figures

1 Introduction

We are now in an era of large-scale genomics data. Sequencing technologies have become so cheap that it is feasible to even imagine a world where everyone who goes to a doctor has his or her genome sequenced. Large-scale efforts such as the 1000 Genomes Project [17] and The Cancer Genome Atlas [81], as well as many smaller projects, have already sequenced tens of thousands of genomes cataloging millions of variants. Nevertheless, despite all this abundant data, understanding the genetic basis behind complex human diseases remains an open question of active research [47]. In contrast to simple Mendelian diseases, for which a small set of commonly shared genetic variants are responsible for disease phenotypes, complex heterogeneous diseases such as autism and cancer are driven by a myriad of combinations of different alterations across the genome. Individuals exhibiting the same phenotypic outcome (disease) may share very few, if any, genetic variants. This makes unraveling the genetic underpinnings of heterogeneous diseases a particularly hard task.

1.1 Biological networks

Biological networks have proven to be a helpful framework through which scientists can approach this task and investigate disease genomes [61, 18]. Broadly speaking, biological networks represent patterns of interaction between different entities in the cell. In protein–protein interaction networks, used in this dissertation, nodes represent the product of the genes—proteins—and an edge between two nodes indicates that

the corresponding proteins interact with one another. The human protein–protein interaction networks available today are massive, consisting of thousands of nodes and edges, and encoding an incredible amount of biological information.

It has been shown that genes related to the same disease manifest a significantly high tendency to interact with one another in the network [63, 28] and that genes related to diseases with similar phenotypes also have a higher propensity to interact with one another [95, 26]. The overarching conclusion is that genes associated with a disease are not randomly positioned in the network but rather, they cluster together [37] and that if a few disease genes are identified, other disease-related genes are likely to be found proximal in the network. These insights have led to the development of numerous network-based methods for identifying disease genes, as discussed next.

Early "linkage" approaches consider only direct interactions between genes located in the linkage interval of a disease with known disease-related genes [51, 61]. Later methods expand the use of topological information encoded in the network by reasoning that genes belonging to a module containing already known disease-related genes have a higher likelihood of being involved in the same disease [22, 52, 32]. These algorithms rely on diffusion startegies to "spread" or "release" disease signal from known disease genes [48, 89]. Genes that may not directly interact with any disease genes but are in close network vicinity to them would still receive relatively larger disease signal. Recently, with the rapid advancement of sequencing technologies and the widespread availability of exome sequencing data, the source of where information is released or diffused from has shifted from known disease genes to newly discovered variants that may be causal for disease [56, 39, 1].

With respect to heterogeneous diseases, and particularly cancer, a prominent viewpoint is to examine genomic alterations in the context of pathways. A biological pathway is typically described as a set of molecules and molecular interactions that

collectively mediate a specific biological activity within a cell [8]. The main insight is that even though different individuals may contain alterations in different genes, if these genes are part of the same pathway, disturbing one of them will have the same effect: perturbing the pathway and contributing to the development of the disease. Hence, analyzing known pathways for enrichment of mutations [41, 10] and pinpointing those that are significantly mutated across patients [93, 90] illuminates some of the mechanisms behind complex diseases. However, as our knowledge of pathways is incomplete, the power of these studies is somewhat limited. Thus, *de novo* discovery of disease-relevant pathways has been the focus of several new methods [88, 74, 4, 12]. Again, relying on the observation that genes that participate in the same biological pathway tend to interact with each other and form small modules within the larger network, studies successfully diffuse mutational signal from large-scale sequencing data to uncover disease genes [88, 34]. A different class of methods are based on the prize-collecting Steiner tree algorithm as they aim to identify modules that contain the most "prized" nodes with the minimal number of edges [84, 80].

1.2 Cancer

Most of the work in this dissertation aims to uncover genes that are causal in cancer. This is an exceptionally hard task because of the immense mutational heterogeneity of the disease. Although tumors from the same cancer type (or even from different cancer types originating in different tissues) display strong phenotypic similarities as all cancer cells exhibit certain hallmark behaviors such as uncontrolled growth and proliferation and resistance to cell death [29], they often contain very few genetic alterations in common. It has been shown that even cells from the same tumor can contain different sets of genetic variants [60]. Further, cancer cells harbor numerous, oftentimes hundreds of, somatic mutations [92], with the vast majority of these muta-

tions thought to play no role in cancer initiation or progression [68, 27]. Distinguishing between the numerous so-called "passenger" mutations and the "driver" mutations important for the oncogenic process has been a central goal of cancer genomics.

To aid in this goal, large-scale cancer genome sequencing efforts, such as The Cancer Genome Atlas (TCGA) [81] and the International Cancer Genome Consortium (ICGC) [83] have sequenced thousands of tumor samples across tens of different cancer types, revealing millions of somatic mutations. While these massive studies have provided researchers with an incredible influx of readily available data, untangling the genetic roots of cancer remains an open problem. Numerous analyses of the data, though, have confirmed that "driver" genes preferentially target specific signaling and regulatory pathways [82, 59], underscoring the power of network-based methods to provide a valuable context within which to study cancer genomes [64, 2]. Both of the approaches developed in this dissertation use biological networks and, while generally pertinent to complex diseases, are applied predominantly to cancer.

1.3 Contribution of this dissertation

In this dissertation, I propose new methods to uncover disease-relevant genes. I rely on the fact that biological networks provide a helpful proxy to biological knowledge of pathways and function and that this can be leveraged to tackle disease heterogeneity. The fundamental insight underlying my work is that the modularity of biological networks can be better exploited if information such as what the individual mutational profiles of patients are or what some of the already known disease-relevant genes are, is incorporated. My work consists of two main algorithms designed to help decipher the complexity of disease genomes.

In *Chapter 2*, I focus my attention on cancer because this is a disease that affects millions of people and which exhibits a vast mutational heterogeneity. I discuss

my work on using a network-based strategy to explain the mutational profiles of cancer patients. The main intuition behind this work is that, instead of aggregating mutational data, considering the per-individual mutational profiles of cancer patients helps uncover rare genomic variants (i.e., present only in a handful of individuals) that nevertheless play an important role in tumorigenesis and/or cancer progression. I present `nCOP`, a novel network-based method which aims to find genes with variants across large number of patients that also form a small connected subcomponent within the larger biological network. I describe two algorithms to solve the problem and a machine learning-like schema to automatically select the value of a single required parameter. I demonstrate that `nCOP` is more effective in discovering cancer genes than both a state of the art frequency-based method and other network-based methods, and that it excels at zooming in on infrequently mutated but cancer-relevant genes.

In *Chapter 3*, I broaden my scope to several complex diseases and discuss my work on incorporating prior knowledge of disease-associated genes to better interpret various types of new incoming data. The intuition behind this work is that existing knowledge can inform the way the new information is examined in a network context. I present `uKIN`, a method which first uses a graph diffusion kernel to spread a signal from the set of already known disease genes and then uses it to bias a random walk originating from the newly implicated genes to move closer to the known ones. First, I show how this approach successfully integrates existing knowledge and new information in the context of cancer and how it outperforms other methods in uncovering cancer-relevant genes. Then, I demonstrate the versatility of my method and integrate new information from genome-wide association studies in order to uncover disease genes for several other complex diseases.

In *Chapter 4*, I conclude by summarizing my findings and discussing the implications of and future directions for the original work presented in this dissertation.

2 Network-based Coverage of Mutational Profiles Reveals Cancer Genes

2.1 Introduction

Large-scale cancer genome sequencing consortia, such as The Cancer Genome Atlas (TCGA) [81], the International Cancer Genome Consortium (ICGC) [83] and other smaller, cancer-specific studies have sequenced the protein-coding regions of thousands of tumor samples across tens of different cancer types. Initial analyses of these data have revealed that while there may be numerous somatic mutations in a tumor that result in altered protein sequences, very few are likely to play a role in cancer development [5, 91, 27]. Therefore, a major challenge in cancer genomics is to develop methods that can distinguish the so-called "driver" mutations important for cancer initiation and progression from numerous other "passenger" mutations.

Early statistical approaches have identified cancer-driving genes by highlighting those genes that are mutated more frequently in a cohort of patients than expected by chance according to some background model [97, 20, 53]. However, the genetic underpinnings of cancer are highly heterogeneous: even when considering a single cancer type, very few genes are found to be somatically mutated across large numbers

of individuals [36]. Further, genes altered only in a few individuals may also be important for tumorigenesis and cancer progression [78]. Clearly, these rarely mutated but cancer-relevant genes cannot be detected by purely frequency-based approaches.

A promising alternative viewpoint is to consider somatic mutations in the context of pathways instead of genes. In particular, it has been proposed that alterations within any of several genes comprising the same pathway can have similar consequences with respect to cancer development, and that this contributes to the mutational heterogeneity evident across cancers. Consistent with this, numerous analyses of TCGA data have shown that certain known pathways are frequently altered across tumor samples of a particular cancer via mutations in different genes [82, 59]. Early studies have leveraged this observation by analyzing known pathways for enrichment of somatic mutations [41, 10] and pinpointing those that are significantly mutated across patients [93, 90]. The power of these studies is somewhat limited, however, as our knowledge of pathways is incomplete and new pathways cannot be identified by these approaches.

De novo discovery of cancer-relevant pathways using large-scale protein interaction networks has thus been the focus of several newer methods (e.g., [88, 10, 15, 65, 74, 4, 12]). In particular, since protein-protein interaction networks have a modular organization [30, 76], proteins taking part in the same pathways and processes tend to be close to each other in the network. One prominent class of techniques leverages this modular structure by propagating mutational information through protein interaction networks and deriving pathways from the induced subnetworks [88, 56, 39, 1]. For instance, Vandin et al. [88] diffuse a "heat" signal arising from the frequency with which proteins are somatically mutated across a cohort of samples to uncover cancer-relevant modules while Hofree et al. [34] approach the problem from a different angle, using biological network information to stratify cancer subtypes. A recent pan-cancer network analysis [56] affirms the power of diffusing mutational data across protein

interaction networks, especially for uncovering rarely mutated cancer genes. However, such diffusion approaches can be highly influenced by frequently mutated genes [56], and further, these methods do not consider whether most patients have mutations in any of the identified pathways. On the other hand, the tendency of most cancer pathways to have mutations in no more than a single component gene within an individual have led to the development of separate set of methods for identifying cancer-related genes based on mutational exclusivity [87, 55, 44, 46]. These methods, however, have limited power in detecting rarely mutated *cancer*-relevant genes.

Here we present a novel network-based approach to tackle cancer mutational heterogeneity by utilizing per-individual mutational profiles. Our method is based on the expectation that if a pathway is relevant for cancer, then (1) many individuals will have a somatic mutation within one of the genes comprising the pathway and (2) the genes comprising the pathway will interact with each other and together form a small connected subcomponent within the larger network. Therefore, given a biological network as well as patient sample data consisting of somatic point mutations, the goal of our approach is to find a set of candidate genes that both "cover" the most patients (i.e., individuals have mutations in one or more of these genes) and are connected in the network (i.e., these genes are likely to participate in the same cellular pathway or process). In contrast to network diffusion approaches, our framework focuses on per-individual mutational profiles and as a result, the "influence" of frequently mutated genes is not spread through the network. We note that network-based coverage approaches have been previously introduced to uncover pathways that are dysregulated [85, 13, 45] or mutated [19, 46] across cohort of samples. However, either patients were required to be covered by these approaches [85, 13, 45, 46], in some cases multiple times (which is especially relevant for dysregulated genes, since there are many of them), or these approaches were designed for data sets with significantly fewer mutations [19]; both cases lead to very different optimizations and

algorithms that are not effective for the task at hand. Alternatively, other approaches have attempted to discover sets of mutated genes that cover not patients but instead genes dysregulated in cancers, with coverage defined by short paths in interaction networks [3, 74, 4].

We devise a simple yet intuitive objective function that balances identifying a small subset of genes with covering a large fraction of individuals. Our objective has just a single parameter that is automatically set using a series of cross-validation tests, eliminating the need of many previous approaches to manually select values for various thresholds and parameters. We develop an integer linear programming formulation to solve this problem and also give a fast heuristic algorithm. We apply our method—**n**etwork-based **co**verage of **p**atients (`nCOP`)—to 24 cancer types from TCGA and uncover both well-known cancer driver genes as well as new potential cancer-related genes. We compare `nCOP` to previous methods that do not use network information, including a state-of-the-art frequency-based method [53] and a "set cover" version of our approach that attempts to find a set of genes that covers cancer samples without considering network connectivity, and demonstrate `nCOP`'s superior power in detecting known cancer genes and in zooming in on rarely mutated ones. Finally, we compare `nCOP` to recent network-based methods that aggregate mutational information and show that our per-patient approach readily outperforms them.

2.2 Results

2.2.1 Algorithm Overview

We begin by giving a brief summary of our method (Figure 2.1); each part is described in more detail in the Methods section 2.3. The biological network is modeled as an undirected graph where each vertex represents a gene, and there is an edge between two vertices if an interaction has been found between the corresponding proteins. We

annotate each node in the network with the IDs of the individuals having one or more mutations in the corresponding gene (Figure 2.1a). We aim to find a relatively small connected component such that most patients have mutations in one of the genes within it. A small subgraph is more likely to consist of functionally related genes and is less likely to be the result of overfitting to the set of individuals whose diseases we are analyzing. However, we would also like our model to have the greatest possible explanatory power—that is, to account for, or cover, as many patients as possible by including genes that are mutated within their cancers. We formulate our problem to balance these two competing objectives with a parameter α that controls the trade-off between keeping the subgraph small and covering more patients.

For a fixed value of α, we have developed two approaches to solve the underlying optimization problem. One is based on linear programming and the other is a fast greedy heuristic (see Methods 2.3.2 and 2.3.3 respectively). We use the greedy heuristic in the context of a carefully designed cross-validation procedure to select a value for α that results in good coverage of patients but avoids overfitting to them (Figure 2.1b). Once α is selected, this value is used within our objective function and we next analyze the entire patient cohort. In particular, multiple independent trials using α are run on randomly chosen subsets of the patient data (Figure 2.1c), as we have found that introducing a little bit of randomness helps increase performance as compared to a single run on the full data set. Each trial outputs a subgraph, and our final aggregated output is an ordered list of candidate genes ranked by how frequently each has been selected over the trials (Figure 2.1d).

We run nCOP, using the greedy heuristic algorithm, on somatic point mutation data from 24 different TCGA cancer types. Results in the main paper use the *HPRD* network [69] for all analysis and highlight kidney renal clear cell carcinoma (KIRC) with 416 samples as an examplar.

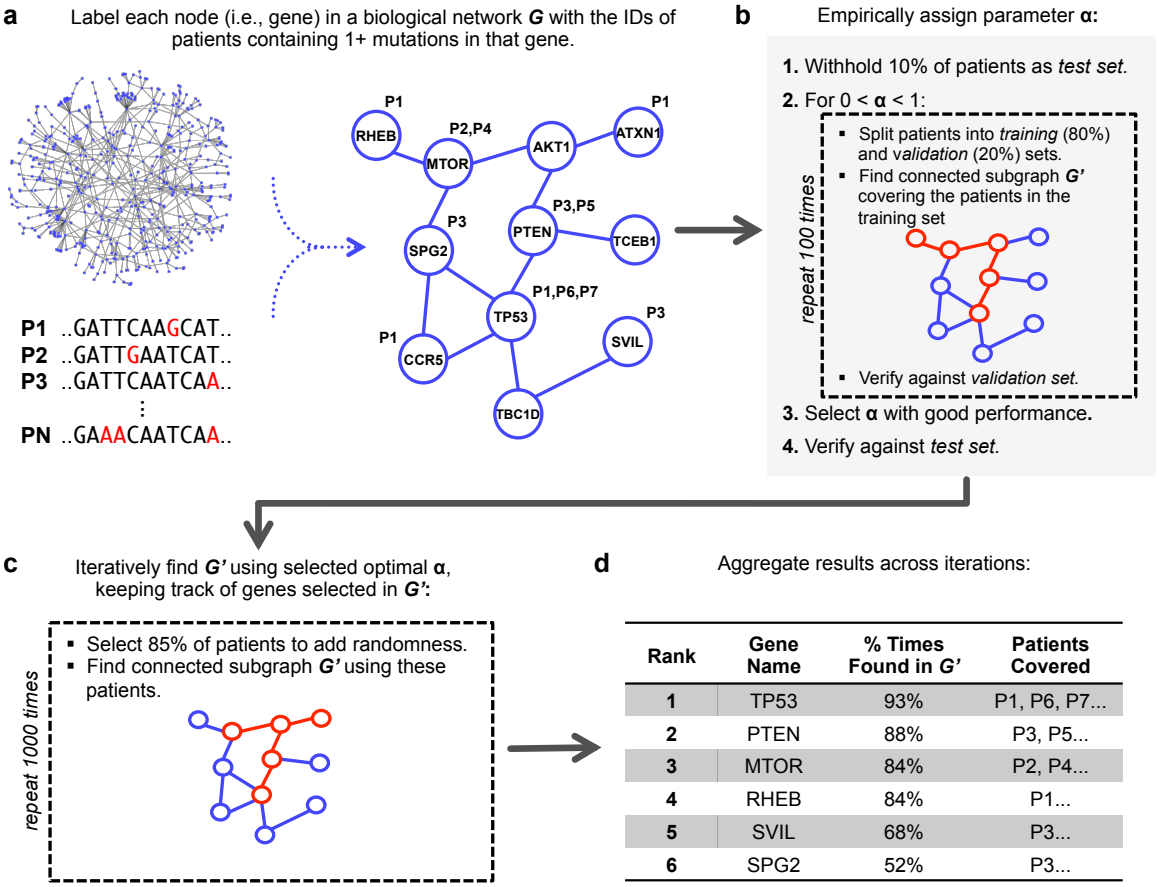


Figure 2.1: **Overview of our approach.** **(a)** Somatic mutations are mapped onto a protein-protein interaction network. Each node is associated with the set of individuals whose cancers have mutations in the corresponding gene. The overall goal is to select a small connected subnetwork such that most individuals in the cohort have mutations in one of the corresponding genes (i.e., are "covered"). **(b)** nCOP automatically selects a value for the parameter α by performing a series of cross-validation tests. First, 10% of the individuals are withheld as a test set. Next, the remaining individuals are repeatedly and randomly split into into two groups, a training set (80%) and a validation set (20%). For each split, the nCOP search heuristic is run for a range of α values ($0 < \alpha < 1$) using the individuals comprising the training set. The parameter α is selected to obtain high coverage of the individuals in the validation sets while maintaining similar coverage on the training sets (i.e., not overfitting to the training set). Coverage of individuals in the initially withheld test set is also calculated and confirmed to be similar to the validation sets. **(c)** Once α is selected, to avoid overfitting on the entire dataset, nCOP is run 1000 times using random subsets of 85% of the individuals. **(d)** Finally, the subnetworks output across the runs are aggregated and candidate genes are ranked by the number of the times they appear in the outputted subnetworks.

2.2.2 Automatic parameter selection reveals generalizability of uncovered subnetworks

Our optimization function for uncovering a subnetwork of mutated genes that covers many patients has one parameter, α. Large values of α result in a larger number of selected genes that cover more patients, yet may contain more irrelevant genes; this may especially be a factor if there are many samples where missense mutations are not the driving event. To choose an appropriate value for α for a set of cancer samples, we split our samples into training, validation and test sets [31], run our greedy heuristic using samples in the training set, and then choose an α where patient coverage deviates between the training and validation sets (see Methods 2.3.4). We note that this framework differs from a traditional machine learning cross-validation setting in that we are not training using a set of trusted examples; instead, our intuition is that cancer-relevant genes that are uncovered using the training samples should also cover samples outside of this set.

We demonstrate that, across the 24 cancer types, our cross-validation framework is a highly effective approach for choosing an α that balances patient coverage with subnetwork size. For all cancers, as α increases, the total number of genes in the chosen subnetwork G' increases (as expected), as does the fraction of patients in the training set that are covered by these genes (Figure 2.2a and Supplementary Figure A.1). For smaller values of α, coverage on the validation sets closely matches that obtained on the training sets; that is, the sets of genes chosen using patients in the training sets are also effective in covering patients in the corresponding validation sets. For KIRC, when $\alpha = 0.5$, genes chosen using the training sets cover on average nearly 70% of patients in the corresponding validation sets, with coverage on the completely withheld test set within 5% of this. The fact that a small subnetwork can be found that covers a large fraction of previously unseen patients is consistent with

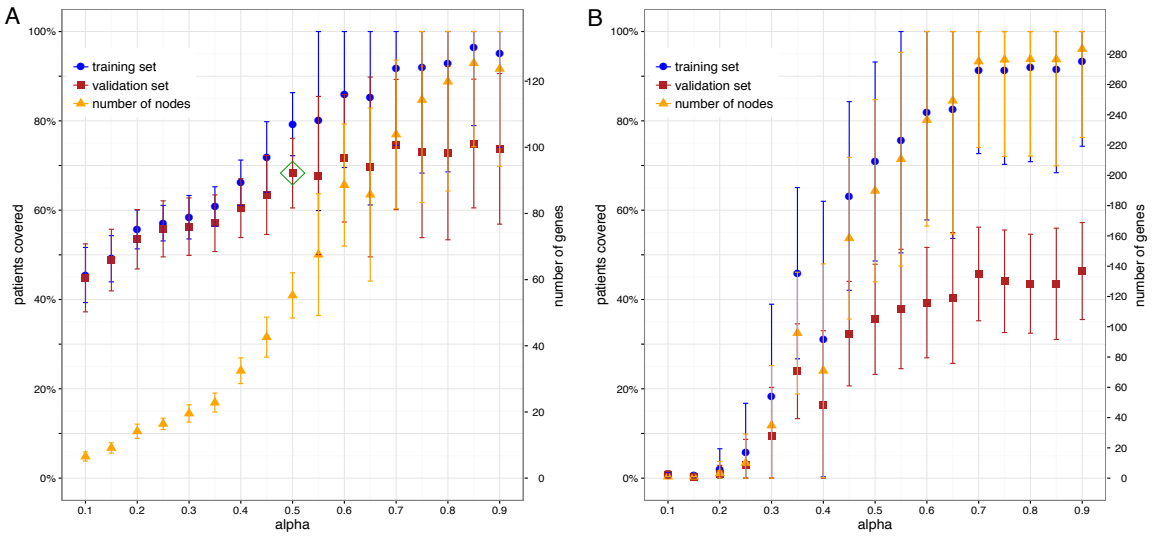


Figure 2.2: We illustrate our cross-validation procedure for parameter selection using the KIRC data set and the HPRD protein-protein interaction network. For each random split of the individuals, we run our algorithm on the training set for different values of α, and next plot the fraction of covered individuals in the training (blue) and validation (red) sets. We also give the number of proteins in the uncovered subgraphs (orange). For each plotted value, the mean and standard deviation over 100 random splits are shown. **(a)** When using somatic missense mutations, at higher values of α, overfitting occurs as the coverage on the validation set levels while coverage on the training set continues to increase. The parameter α is selected using an automated heuristic procedure (green rhombus) so that coverage on the validation set is good while overfitting on the training set is not extreme. **(b)** When using somatic synonymous mutations, there is poor coverage on the validation set regardless of coverage on the training set. Further, as compared to using missense mutation data, significantly more genes are required to cover the same fraction of individuals.

the hypothesis that a shared pathway or process plays a role in most (but not all) of these patients' cancers.

For larger values of α (> 0.6 for KIRC), however, coverage on the validation sets lags behind that observed on the training sets. For even larger values of α (> 0.85 for KIRC), the algorithm selects many genes, and eventually increases the coverage for most cancers on the training sets to nearly 100%. However, larger values of α do not substantially increase coverage of the withheld patients. This difference between the training and validation curves captures the overfitting of the model and also illustrates the trade-off between covering more patients and keeping the solution parsimonious.

We note that the eventual plateau of the validation curve is consistent across cancer types (Supplementary Figure A.1). For each cancer type, values of α are selected by our automated procedure (see Methods 2.3.4); this value is $\alpha = 0.5$ for the KIRC dataset shown in Figure 2.2.

As a control, we repeat the same procedure using only synonymous mutations (Figure 2.2b). We observe that the coverage on the validation sets is much poorer. Though coverage of course increases as more nodes are added, it never exceeds 50% even when α is increased to 1 or when we have nearly perfect coverage on the training set, despite adding many more nodes. This poor performance is consistent with the expectation that synonymous mutations do not result in altered protein sequences and do not disturb cellular pathways. Hence, given the differences observed between using missense versus silent mutation data when varying settings for α and comparing training and validation sets, our formulation appears to be well-suited for investigating mutational profiles in the context of interaction networks.

2.2.3 nCOP effectively uses network information to uncover known cancer genes

Having shown in the previous section how to select a value for the only parameter in the model, we next evaluate nCOP's performance in uncovering known cancer genes (CGCs) [24].

We first consider the KIRC data set, and find that our top predictions include a high fraction of CGC genes (Figure 2.3a). To illustrate the power of our network-based method, we compare its performance to approaches that do not consider any network information. In particular, we consider a set cover version of our approach that does not use network information at all, as well as a state-of-the-art frequency-based approach, `MutSigCV 2.0` [53]. For the same number of predicted genes, our approach

consistently has a larger fraction of CGCs than either approach, demonstrating the advantage of using network information.

We next compare nCOP to these two non-network approaches across all 24 cancer types. In particular, we compute the log ratio of the area under the precision-recall curve (AUPRC) of our approach versus each of the other approaches on each cancer type (Figure 2.3b). We outperform MutSigCV 2.0 in 22 of the 24 cancers and the set cover approach in all cancers, demonstrating the clear advantage of using network information; the performance improvement of nCOP over the set cover approach is particularly notable as the main difference between these approaches is the additional use of network information by nCOP. In several cancers, the performance improvements of nCOP are substantial. For example, nCOP shows a four-fold improvement over MutSigCV 2.0 in predicting cancer genes for liver hepatocellular carcinoma (LIHC) and an eight-fold improvement over MutSigCV 2.0 on pheochromocytoma and paraganglioma (PCPG). The overall results are consistent across different lists of known cancer genes (Supplementary Figure A.2a and b), numbers of predictions considered (Supplementary Figure A.2c), and networks (Supplementary Figure A.2d). The superior performance of nCOP as compared to these non-network based approaches on the vast majority of cancers demonstrates its considerable power.

Having shown that nCOP better identifies cancer-relevant genes than two approaches that do not use network information, we next consider whether the specific way in which nCOP uses network information is beneficial. Towards this end, we compare the effectiveness of nCOP in uncovering cancer genes to Muffinn [12], a method published last year that considers mutations found in interacting genes, and to DriverNet [3], a method that finds driver genes by uncovering sets of somatically mutated genes that are linked to dysregulated genes. We find that nCOP outperforms Muffinn on 20 and DriverNet on 21 of the 24 cancer types (Figure 2.3c). We also compare nCOP to Hotnet2 [56], a cutting-edge network diffusion method. As Hotnet2

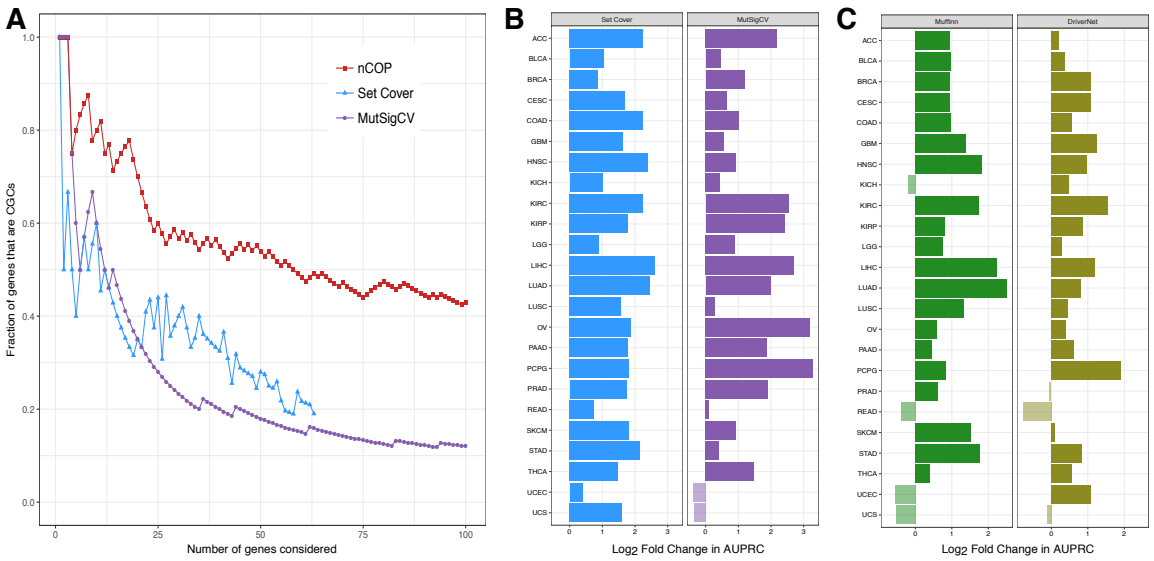


Figure 2.3: `nCOP` **is more successful than other methods in identifying known cancer genes.** **(a)** Our network-based algorithm `nCOP`, a set cover version of our algorithm that ignores network information, and `MutSigCV 2.0`, a frequency-based approach, are compared on the KIRC dataset. `nCOP` ranks genes based on how frequently they are output, and `MutSigCV 2.0` ranks genes by q-values. The set cover approach is run for increasing values of k until all patients are covered. For each method, as an increasing number of genes are considered, we compute the fraction that are CGCs. Over a range of thresholds, our algorithm `nCOP` outputs a larger fraction of CGC genes than the other two approaches. **(b)** Comparison of `nCOP` to two network-agnostic methods across 24 cancer types. For each cancer type, we compute AUPRCs for `nCOP`, the set cover approach, and `MutSigCV 2.0`, using their top 100 predictions. We give the log_2 ratios of `nCOP`'s AUPRCs to the other methods' AUPRCs. Our approach `nCOP` outperforms the set cover approach on all 24 cancers, and `MutSigCV 2.0` on 22 of the 24 cancer types. **(c)** Comparison of `nCOP` to two network-based methods, `Muffinn` and `DriverNet`, across 24 cancer types. For each cancer type, we compute the log_2 ratio of `nCOP`'s AUPRC to the other methods' AUPRCs. Our approach `nCOP` outperforms `Muffinn` and `DriverNet` on 20 and 21, respectively, of the 24 cancer types.

does not output a ranked list of genes, we could not compute an AUPRC. Instead, examining the complete list of genes highlighted by both methods, we observe that `nCOP` exhibits significantly better precision while trailing slightly in recall (Supplementary Figure A.3).

Robustness tests. We briefly describe some additional tests we performed to show that `nCOP` is robust and well-behaved. First, to confirm the importance of network

structure to nCOP, we have run nCOP on two types of randomized networks, degree-preserving and label shuffling, and have shown that (as expected) overall performance deteriorates across the cancer types (Supplementary Figure A.2e); we note that these randomized networks maintain the relationships between genes and the cancers they are found to be mutated in, and thus retain significant cancer-relevant information. Second, to make sure that the novel genes we uncover are not driven by patients with large numbers of passenger mutations (i.e., that the novel genes are not likely to be passenger genes), we have compared the overall number of mutations for patients having missense mutations only in CGC genes but not in any non-CGC (or novel) genes to the total number of such mutations for patients having missense mutations only in novel genes but not in any CGC genes (Supplementary Figure A.4), and have found that patients with only mutations in novel genes do not harbor more mutations. Finally, to make sure that genes are not more likely to be picked because they have higher degree, we have confirmed that newly predicted genes do not tend to exhibit higher degree than known cancer genes; indeed, among all novel genes found across all cancer types, most have degree less than 15, and there are only a couple with high degree (≥ 50).

2.2.4 nCOP newly predicts rarely mutated cancer genes

We next demonstrate that nCOP highlights genes with a range of mutation rates. When considering genes that are output by nCOP in at least 50% of the trials on the KIRC samples, we see many well-known cancer players: some are highly mutated, such as *VHL*, *BAP1* and *TP53*, while others, such as *ERBB2* and *RUNX1T1*, are each mutated only in a handful ($< 1\%$) of samples. While the former set of genes can be uncovered by any frequency-based technique, the latter have missense mutation rates that are similar to those of genes not relevant for cancer (Figure 2.4a) and are thus hard to uncover by frequency-based methods. Indeed, of the 4818 genes

that have any missense mutations across the KIRC samples, `nCOP` identifies 47 as cancer relevant, with 24 of those in the bottom 90% of mutated genes with respect to their missense mutation rates. Among these 24 genes, 12 are CGCs ($p < 10^{-8}$, hypergeometric test). The statistically significant enrichment of CGC genes in the rarely mutated genes found by nCOP is true across all cancers except for UCS where nCOP predicts only six genes. Thus, `nCOP` provides a means for pulling out cancer genes from the "long tail" [27] of infrequently mutated genes.

In addition to ranking known cancer genes highly, `nCOP` also gives high ranks to several non-CGC genes that may or may not be implicated in cancer, as our knowledge of cancer-related genes is incomplete. Among these novel predictions for KIRC are *HIF1A*, *NR5A2*, and *SALL1*, which have all recently been suggested to play a role in cancers [73, 94, 58] and are each mutated in less than 3% of the samples. *SALL1* is a zinc-finger transcription factor which is shown to play a role in kidney development [11] and mutations within it have been linked to Townes-Brocks syndrome, a rare genetic disease associated with kidney abnormalities and malformation [49]. Among the individuals in the KIRC dataset covered by the *SALL1* gene, one has no mutations affecting protein coding in any known cancer gene. Thus, while this particular individual's tumor is not driven by mutations in known cancer genes, `nCOP` pinpoints a role for *SALL1*.

Several of the genes uncovered by `nCOP` with low missense mutation rates in KIRC are part of the *PI3K-AKT signaling pathway*, a prominent cancer pathway that promotes cell survival and growth. When considering the 28 genes output by `nCOP` with missense mutation rates lower than that of *AKT2*, a key component of this pathway, we find that 18 of them form a small connected component (Figure 2.4b) and together are mutated in ∼14% of the samples. Three of our novel predictions, *STAT1*, *CDKN1A* and *HSP90AA1*, interact with *AKT1*. Existing literature [66, 50, 9, 14] supports a possible role of these genes in tumor progression. Notably, *STAT1*, a gene

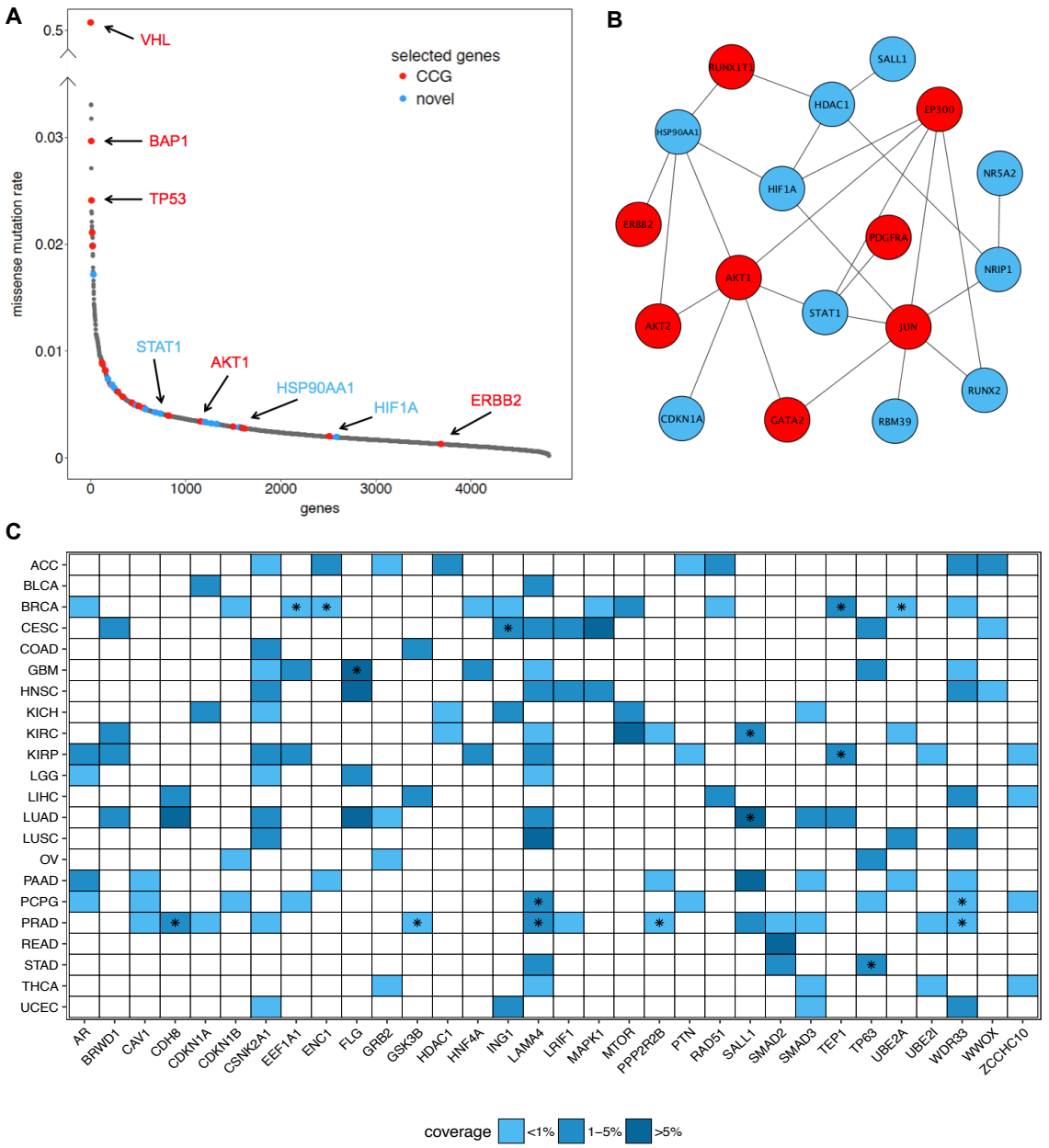


Figure 2.4: **nCOP identifies rarely mutated genes. (a)** The missense mutation rates, computed for each gene as the total number of missense mutations observed within it divided by the product of the number the samples and the length of the gene in nucleotides per 10^3 bases, are sorted from high to low and are shown for all mutated genes in the KIRC dataset. Genes that are output by nCOP in at least half the trials are shown in red for known cancer genes, and in blue for new predictions. All other genes are shown in grey. Well known cancer genes output by nCOP, such as *VHL* and *TP53*, are at the peak of the distribution. nCOP is also able to uncover known cancer genes with very low mutational rates lying at the tail of the distribution. **(b)** Several of the infrequently mutated genes selected by nCOP form a module with five genes

that belong to the prominent cancer PI3K-AKT signaling pathway. Red nodes denote CGC genes and blue nodes denote novel predictions. **(c)** Shown are all newly predicted, non-CGC genes that are uncovered by `nCOP` in more than 3 cancers. The majority of these predictions are mutated in less than 5% of the samples in the corresponding cancers in which they are implicated. A star indicates that the gene covers an individual of a particular cancer type who does not have any protein coding affecting variant in any CGC gene.

which modulates diverse cellular processes, such as proliferation, differentiation and cell death, also covers an individual with no variants in any known cancer gene.

When we consider the full ranked list of genes output by our procedure for KIRC and perform a rank-based gene set enrichment analysis using the Broad Institute GSEA tool [79], four pathways from the KEGG database, all cancer-relevant, are enriched at $p < 0.05$ (*microRNAs in cancer*, *pathways in cancer*, *jak stat signaling pathway*, and *choline metabolism in cancer*). Interestingly, the *thyroid hormone signaling pathway* is also enriched. It has been shown that thyroid hormones play a role in kidney growth and development [43] and four of our non-CGC predictions are part of that pathway, together with four known cancer genes.

When run individually on all 24 cancer types, `nCOP` newly implicates 32 genes as relevant in at least in at least three cancer types (Figure 4c). These genes typically are infrequently mutated, with 93% of them mutated in fewer than 5% of the samples in each of the cancers in which they are predicted to play a functional role. Several of the novel genes unveiled by `nCOP` are found in individuals whose cancers do not harbor somatic mutations in any known cancer gene; thus, somatic mutations within these novel genes are promising as candidate driver events within these cancers. Across all cancer types, there are 285 patients who do not have mutations affecting protein coding in any known cancer gene, and `nCOP` covers 114 of them (40%) by selecting 100 genes. The selection of these novel genes is not driven by samples with hypermutator phenotypes (Supplementary Figure A.4) and 13 appear in more than 3 cancers (Figure 2.4c). While some newly uncovered genes may be false positives, others (like

SALL1 and *STAT1*) are strong candidate genes for further investigation. This illustrates the power of `nCOP` to zoom in on rarely mutated genes and to help uncover the genetic underpinnings of the studied tumor samples.

2.3 Methods

2.3.1 General formulation

We model the biological network, as usual, as an undirected graph $G = (V, E)$ where each vertex represents a gene, and there is an edge between two vertices if an interaction has been found between the corresponding protein products. Each vertex v_j is associated with a set C_j containing the IDs of the individuals who have somatic mutations in the corresponding gene. We formulate our problem as that of finding a connected subgraph G' of G so as to minimize

$$\alpha X + (1 - \alpha) Size(G'),$$

where X is the fraction of patients that do not have an alteration in a gene included in G' (i.e., they are uncovered), $Size(G')$ is the size of the subgraph, and $0 \leq \alpha \leq 1$ is a fixed parameter controlling the trade-off between keeping the subgraph G' small and covering more patients. A patient with ID i is covered if $i \in \bigcup_{v_j \in G'} C_j$, and uncovered otherwise. We note that our problem is similar, though not identical, to the Minimum Connected Set Cover Problem [75], a NP-hard problem.

A simple and natural measure for the size of a subnetwork is its number of nodes (i.e., $Size(G') = |G'|$). However, longer genes may tend to acquire more mutations simply by chance. We correct for that by associating with each node v_j a weight w_j that is equal to the ratio of the length of the gene to the total number of mutations it has. The size of the subcomponent is then defined as $Size(G') = \sum_{v_j \in G'} w_j$. This

way, genes having longer length will be weighted more, correcting for a possible bias towards selecting longer genes. We note that since our objective function balances the fraction of uncovered patients with the size of the graph, we would like the size of the graph to be between 0 and 1; thus, we normalize each node weight by dividing by the unnormalized size of what we call a fully covering subgraph G^f—a connected subgraph of G that covers all patients. (In practice, we compute G^f using the greedy heuristic described below, with $\alpha = 1$).

2.3.2 Integer linear programming formulation

The problem of finding a minimum connected subgraph that covers as many patients as possible can be solved using constraint optimization. Let n be the number of patients in our sample. For each patient i, we define a binary variable p_i that is set to 1 if patient i is covered by the chosen subgraph G', and 0 otherwise. For each vertex (or gene) v_j, we define a binary variable x_j that is set to 1 if the vertex is included in the chosen subgraph G', and 0 otherwise. It is straightforward to set up constraints to ensure that a patient is considered uncovered if none of its mutated genes are part of G', and covered if at least one of its mutated genes is selected as part of G' (see Equations (1) and (2) below).

The challenging part of the ILP is setting up constraints to ensure that the chosen nodes form a connected subgraph G'. For this task, we employ a flow of commodity technique [23], which we now briefly describe. We inject $|G'|$ units of flow into G' (i.e., we inject $\sum x_i$ units of "flow" into a vertex that is included in the chosen subnetwork). Flow can move from one vertex to any of its neighbors in the network, and each vertex removes exactly one unit of flow as the flow passes through it. All flow must be removed from the subnetwork, and we set the constraints so that this is possible only if the subnetwork G' is connected. For the source of the flow we use an artificial external node v_{extr}. The main issue is that we do not know which node v_{extr}

should be connected to, as we do not know the nodes of G' in advance. To resolve this, we decide that v_{extr} connects to the node that covers the largest number of patients v_{max}; this is equivalent to determining in advance that $v_{max} \in G'$, though as an alternate approach we could also decide to choose this node probabilistically and run the ILP several times. Finally, to handle the flow constraints, for each edge $(i, j) \in E$, we introduce integer variables $y_{i,j}$ and $y_{j,i}$ to represent the amount of flow from node i to node j and from node j to node i, respectively. The full integer linear program is:

minimize $\alpha(n - \sum_i p_i)/n + (1 - \alpha) \sum_j x_j w_j$
subject to

$$p_i \geq x_j \qquad \forall i, j \text{ s. t. } i \in C_j \qquad (2.1)$$

$$p_i \leq \sum_{j:i \in C_j} x_j \qquad \text{for each patient } i \qquad (2.2)$$

$$\sum_{i:(i,j) \in E} y_{i,j} = x_j + \sum_{i:(i,j) \in E} y_{j,i} \qquad \text{for each vertex } v_j \qquad (2.3)$$

$$\sum_{j:(i,j) \in E} y_{i,j} \leq |V| x_i \qquad \text{for each vertex } v_i \qquad (2.4)$$

$$\sum_i x_i = y_{extr,max} \qquad (2.5)$$

$$p_i, x_i, y_{i,j} \in \{0, 1\} \qquad \text{for all such variables} \qquad (2.6)$$

Equation (1) ensures that a patient is considered covered if one of his or her somatically mutated genes is included in G'. Equation (2) ensures that a patient is not considered covered if none of his or her somatically mutated genes is chosen to be part of the subgraph. Equations (3), (4) and (5) enforce the connectivity requirement. Equation (3) requires that the flow going out of each vertex in the chosen subnetwork is 1 less than the flow coming in. Equation (4) requires that if a vertex is not part of

the chosen subgraph, the flow going through it is 0. Equation (5) sets the amount of flow injected into the subgraph to be equal to the number of chosen nodes.

2.3.3 Greedy heuristic

Solving the ILP yields an exact solution but is computationally difficult. Thus, we have also developed an efficient greedy heuristic. Our heuristic procedure initializes G' by randomly choosing the first gene from among the five most mutated genes, with probability proportional to the number of patients it is found mutated in. It then expands the subgraph G' iteratively as follows. At each iteration, all vertices that are at most distance 2 from a vertex in G' are examined and the one that improves the objective function the most is chosen; any ties are broken uniformly at random. If this vertex is not directly adjacent to the nodes in the subnetwork, the intermediary node is also added. The heuristic terminates when no improvement to the objective is possible. We repeat this heuristic multiple times, as it is probabilistic.

In practice, the greedy heuristic finds a solution that is on average $\sim$90% of the best value for the objective function as determined by the ILP formulation using `CPLEX` [38]. For example, on the glioblastoma dataset of 277 individuals, the ILP finds 61 genes covering 90% of the patients when using α=0.5. In comparison, for this value of α, the greedy heuristic finds on average 66 genes covering 88% of the patients with 39 genes in common. In the rest of the paper, we use the greedy optimization as it has comparable performance to the ILP, while being much faster.

2.3.4 Parameter selection and solution aggregation

We split our samples into training, validation and test sets [31]. A test set of (10%) of the patients is completely withheld. While varying α in small increments in the interval $(0; 1)$, the remaining data is repeatedly split (100 times for each value of α) into training (80%) and validation (20%) sets. For each split, the greedy heuristic

algorithm is run on the training set to find G'. The fractions of patients covered (by the selected G') in the training and validation sets are compared. The parameter α is selected where performance on the validation sets deviates as compared to the training sets. While this can be done visually, for all results reported here we do this automatically using a simple two-rule procedure that selects the smallest α for which the difference in average coverage between the training and validation set exceeds 5% and for which average performance on the validation set is within 10% from the maximum observed one for any α. Finally, the coverage of patients on the (completely withheld) test set is computed to ensure it is similar to the one on the validation set.

Once α is chosen for a set of cancer samples, we repeatedly (1000 times) run the algorithm on this set, each time withholding a fraction (15%) of the patients in order to introduce some randomness in the process. Genes are then ranked by the number of times they appear in G'. In practice, we have found that this improves performance as compared to running the algorithm once on the full data set.

2.3.5 Data sources and pre-processing

We downloaded all level 3 cancer somatic mutation data from The Cancer Genome Atlas (TCGA) [81] that was available as of October 1, 2014. This data consists of a total of 19,460 genes with somatic point mutations across 24 cancer types. For each cancer, samples that are obvious outliers with respect to their total number of mutated genes are excluded. See Supplementary Table 1 for a list of the cancer types, the cancer-specific thresholds to determine outlier samples, the number of patient samples considered for each cancer type, and other statistics about the TCGA somatic mutation dataset.

We use two different biological networks in our analysis: *HPRD* [69] (Release 9_041310) and *BioGrid* (Release 3.2.99, physical interactions only) [77]. Biological networks can exhibit several nodes with very high connectivity, often due to study

bias. As such high connectivity destroys the usefulness of the network information, we remove all nodes whose degrees are clear outliers with unusually high degree (degree > 900 and more than 10 standard deviations away from the mean). For *BioGrid*, this removes *UBC, APP, ELAVL1, SUMO2, CUL3*. For *HPRD*, we remove no nodes. For both networks, we exclude the nine longest genes (*TTN, MUC16, SYNE1, NEB, MUC19, CCDC168, FSIP2, OBSCN, GPR98*) as they tend to acquire numerous mutations by chance while covering many patients.

To further handle the connectivity arising within the networks due to high-degree nodes, we filter edges using the diffusion state distance (DSD) metric introduced in [7]; the DSD metric captures the intuition that edges between nodes that also share interactions with low degree nodes are more likely to be functionally meaningful than edges that do not (and thus are assigned closer distances). For each edge, the DSD scores (as computed by the software of [7]) between the corresponding nodes are Z-score normalized, and edges with Z-scores > 0.3 are removed. We note that the overall performance of our approach improves when performing this filtering (data not shown), supporting the claim of [7] that preprocessing a biological network in this manner is an important step. The final number of nodes and edges, respectively, for the filtered networks are 9,379 and 36,638 for *HPRD*; and 14,326 and 102,552 for *BioGrid*.

2.3.6 Performance evaluation

To evaluate the gene rankings of all the tested methods, we use the curated list of 517 cancer census genes (CGCs) available from COSMIC [25]. All genes in this list are considered as positives, and all other genes are considered as negatives. Though we expect that there are genes other than those already on the CGC list that play a role in cancer, this is a standard approach to judge performance (e.g., see [39]) and gives us an idea of how methods are performing as cancer genes should be highly

ranked by methods that perform well. To avoid potential biases due to using a single list of positives, we additionally tested using two different sets of cancer genes (Supplementary Figure A.2). Since only the top predictions by any method are relevant for cancer gene discovery, we judge performance by computing the area under the precision-recall curve (AUPRC) using the top 100 genes predicted by each method (without thresholding the output of any method by score or level of significance). If a method returns less than 100 genes total, we extend the precision-recall curve to 100 genes assuming that it performs as a random classifier. We note that reasonable changes to the number of predictions considered does not change our overall conclusions (Supplementary Figure A.2).

Other approaches. To ascertain the contribution of network information, we compare `nCOP` to two approaches that do not use network information: (1) `MutSigCV 2.0` [53], a state-of-the-art method that identifies genes that are mutated more frequently than expected according to a background model, and (2) a set cover approach that tries to find mutated genes that simply cover as many patients as possible. We formulate the set cover approach as an ILP that tries to find a good cover consisting of k vertices. Using the same notation as for `nCOP`, the set cover objective is to *maximize* $\sum_i p_i$, subject to Equations (1) and (2) of `nCOP`, and with the additional constraints that $\sum_j x_j \leq k$ and $\sum_j x_j \geq k$. We also compare `nCOP` to HOTNET2 [56], `Muffinn` [12], and `DriverNet` [3], three recent network-based approaches. To ensure fair comparisons, all methods are run on exactly the same cancer mutation data. Similarly, `Hotnet2`, `Muffinn` and `nCOP` are run on the same network. `DriverNet` instead uses an influence (i.e., functional interaction) graph and transcriptomic data; we use their default influence graph and provide as input TCGA normalized expression data. `MutSigCV 2.0`, `Hotnet2`, `Muffinn`, and `DriverNet` are run with default parameters (for `Hotnet2`, this is 100 permuted networks, and $\beta = 0.2$ for the restart probability for the insulated heat diffusion process).

2.4 Discussion

In this paper, we have shown that *nCOP*, a method that incorporates individual mutational profiles with protein–protein interaction networks, is a powerful approach for uncovering cancer genes. Our method is based on an intuitive mathematical formulation and demonstrates higher precision than other state-of-the-art methods in detecting known cancer genes. Further, our approach is particularly beneficial in highlighting infrequently mutated genes that are nevertheless relevant for cancer. Our approach therefore complements existing frequency-based methods (e.g., [53]) that generally rely on comparisons to background mutational models and lack the statistical power to detect genes mutated in fewer individuals.

In the future, *nCOP* can be extended in a number of natural ways. First, while *nCOP* currently analyzes only mutations within genes that affect protein coding, other alterations are also commonly observed in cancers. For example, copy number variants (CNVs) are found frequently in cancers and can play critical functional roles [98]. Although `nCOP` does not currently use CNV information, our framework can be extended to incorporate this data. Indeed, as the numbers of CNVs and point mutations found within each cancer genome appear to be inversely related [16], considering both types of alterations will increase the power of our approach. Second, *nCOP* may also benefit from incorporating gene weights that reflect likelihood to play a role in cancer; in our current work, we consider a gene's length but no other gene-specific attributes are considered. Such gene weights may be derived from existing approaches to detect frequency of mutation or to assess the functional impact of mutations. Finally, while `nCOP` can output groups of genes that are not part of a single connected component due to our randomized aggregation procedure, extending `nCOP`'s core algorithms to explicitly consider multiple subnetworks corresponding to distinct pathways may be a particularly promising avenue for future work.

We have applied nCOP across 24 different cancer types, and have shown that it is broadly effective in identifying cancer genes in each of them. However, cancers affecting the same tissue can often be grouped into distinct subtypes; breast cancer, for example, is broadly subtyped based on receptor status and expression patterns [67, 86]. In future applications, nCOP could be used to study how different known subtypes of a given type of cancer yield overlapping or differing perturbed pathways. Even more interesting, and with immediate clinical relevance, would be to develop additional techniques to stratify patients into different cancer subtypes based upon the differently perturbed modules that *nCOP* uncovers.

We conclude by noting that researchers can use our framework to rapidly and easily prioritize cancer genes, as *nCOP* requires only straightforward inputs and runs on a desktop machine. Indeed, nCOP's efficiency, robustness, and ease of use make it an excellent choice to investigate cancer as well as possibly other complex diseases. As sequencing costs plummet and cancer and other disease sequencing mutational data become more abundant, the predictive power of our method should only increase (Supplementary Figure A.5). In sum, we expect that our method nCOP will be of broad utility, and will represent a valuable resource for the cancer community.

3 Use of prior knowledge in networks

3.1 Introduction

Genetic variants have been identified in thousands of individuals with various acquired and inherited diseases, including cancer, autism, and Alzheimer's, among others. Despite this incredible influx of mutation data, pinpointing the gene variants responsible for the development of complex diseases remains a daunting task as the same phenotypic outcome (disease) can result from a myriad of combinations of different alterations across the genome. Therefore, a major challenge in computational biology is to develop methods that can decipher large genomic datasets and hone in on those genes that are causal for a particular disease.

Protein-protein networks provide a powerful framework within which to identify disease genes [37]. In particular, genes that take part in the same pathway or cellular process tend to be close to each other in the network [30, 76], and since genes relevant for a given disease typically target a relatively small number of biological pathways, they are not randomly positioned in the network but instead tend to interact with one another and cluster together in the network [63, 28, 26]. Consequently, if known disease genes are mapped to the network, other disease-relevant genes are likely to be found in their vicinity [64]. Indeed, biological networks have proven to be instru-

mental in identifying disease genes [18, 2]. While early methods consider only direct interactions between genes [51], later approaches exploit the full topological information in the network by "propagating" or "diffusing" signal from known disease genes [61, 48, 89]. In this manner, genes that do not directly interact with any known disease genes but are proximal to them in the network may still receive a relatively large amount of signal from disease genes, and thus be implicated as disease causing. With the widespread availability of exome sequencing data and genome-wide association studies (GWAS), the source of where information is propagated from has shifted from known disease genes to those that are newly identified as perhaps playing a role in disease [10, 88, 1, 65, 39, 56]. For instance, cancer genes and pathways have been identified based on diffusing a "heat" signal arising from the frequency with which genes are somatically mutated across tumors from a cohort of patients [56]. Thus, there are two dominant paradigms for uncovering disease genes using biological networks: spreading signal either from well-established, annotated disease genes or from genes that have been newly implicated as putatively causal.

Here, we argue that both sources of information should be utilized, and that existing knowledge of disease-genes should inform the way new data is examined within networks. In particular, while our prior knowledge of causal genes for a given disease may be incomplete, this information nevertheless is a valuable source of information about the biological processes underlying the disease. Towards this end, we introduce a guided random walk approach to uncover disease genes, where signal is propagated from the new data such that the signal tends to move towards genes that are closer to known disease genes. In contrast, other methods perform diffusion or random walks uniformly [88, 39], or where the diffusion is scaled by weights on network edges that reflect their estimated reliabilities [1]. Our guided random walk formulation relies on a single parameter that balances how much emphasis is placed on the new information versus the prior knowledge. We numerically solve for the stationary distribution that

the walk converges to and use how frequently each node is visited to rank the genes with respect to disease relevance.

We demonstrate the efficacy of our method `uKIN`—**u**sing **K**nowledge **I**n **N**etworks—by first applying it to discover genes causal for cancer. Here, new information consists of genes that are found to be somatically mutated in tumors—only a small subset of which are thought to be relevant for cancer initiation or progression—and prior information consists of "driver" genes annotated already to be cance-relevantr [25]. We demonstrate, across 24 cancer types, that propagating signal by integrating both sources of information performs substantially better in uncovering known cancer genes than propagating signal from either source alone. Next, we show that `uKIN` readily outperforms state-of-the-art network-based methods, and that `uKIN` can incorporate cancer-specific prior knowledge to better uncover causal genes for specific cancer types. Finally, we demonstrate `uKIN`'s versatility by applying it to three other complex diseases, where the genes comprising the new information arise from GWAS studies.

3.2 Results

3.2.1 Algorithm Overview

We first give a brief summary of our method `uKIN` (Figure 3.1). At a high level, our approach propagates new information across a network, while using prior information to guide this propagation. While our approach is generally applicable, here we focus on the case of propagating information across biological networks in order to find disease genes. We assume that prior knowledge about a disease is given by a set of genes already implicated as causal for that disease, and new information consists of genes that are potentially disease-relevant. In the scenario of uncovering cancer genes, prior information comes from the set of known cancer genes, and new information

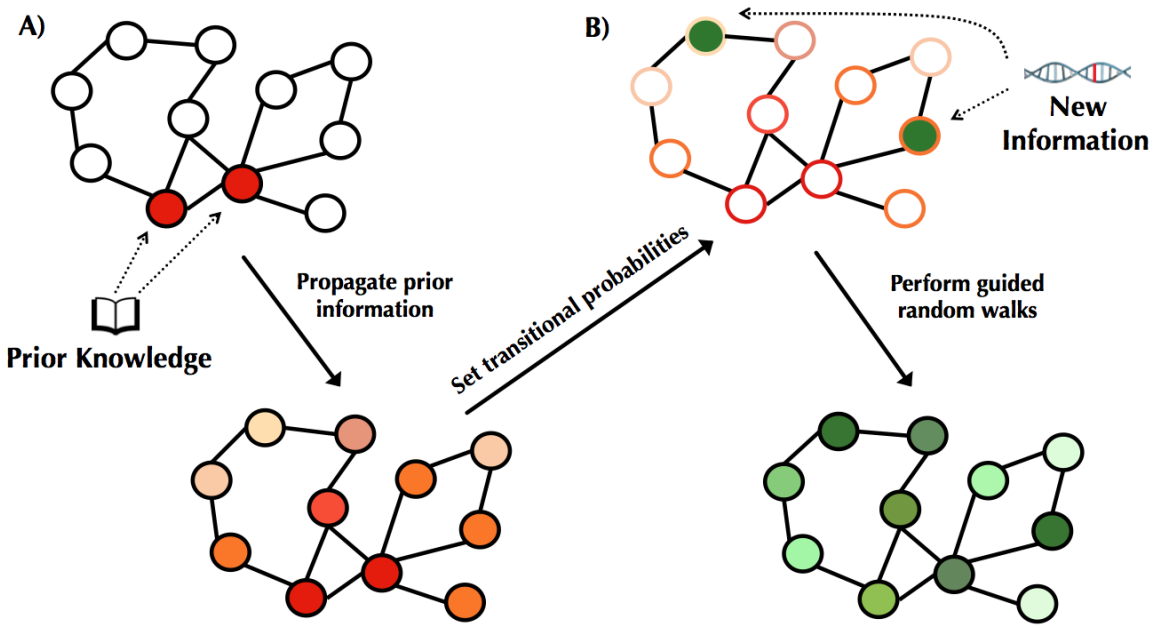


Figure 3.1: **Overview of our approach. (a)** Known disease-relevant genes (prior knowledge) are mapped onto a gene-gene interaction network (shown in red, top). Signal from this prior knowledge is propagated through the network via a network flow approach [71], resulting in each gene in the network being associated with a score such that higher scores (visualized in darker shades of red, bottom) correspond to genes closer to the set of known disease genes. These scores are used to set transition probabilities between genes such that a neighboring gene that is closer to the set of prior knowledge genes is more likely to be chosen. **(b)** Genes putatively associated with the disease—corresponding to the new information—are mapped onto the network (shown in green, top). To integrate both sources of information, random walks with restarts are initiated from the set of putatively associated genes, and at each step, the walk either restarts or moves to a neighboring gene according to the transition probabilities (i.e., walks tend to move towards genes outlined in darker shades of red). These prior-knowledge "guided" random walks with restarts have a stationary distribution corresponding to how frequently each gene is visited, and this distribution is used to order the genes. Higher scores correspond to more frequently visited genes (depicted in darker greens, bottom).

corresponds to those genes that are found to be somatically mutated across patient tumors. For other complex diseases, new information may arise from (say) genes weakly associated with a disease via GWAS studies or found to have *de novo* or rare mutations in a patient population of interest.

The first step of our approach is to compute for each gene a measure that captures how close it is in the network to the prior knowledge set of genes $\mathcal{K}$ (Figure 3.1a). To

accomplish this, we spread the signal from the genes in $\mathcal{K}$ using a diffusion kernel [71]. Next, we consider new information consisting of genes $\mathcal{M}$ that have been identified as potentially being associated with the disease. As we expect those that are actually disease-relevant to be proximal to each other and to the previously known set of disease genes, we spread the signal from these newly implicated genes $\mathcal{M}$, biasing the signal to move towards genes that are closer to the known disease genes $\mathcal{K}$ (Figure 3.1b). We accomplish this by performing random walks with restarts, where with probability α, the walk jumps back to one of the genes in $\mathcal{M}$. That is, α controls the extent to which we use new versus prior information, where higher values of α weigh the new information more heavily. With probability $1 - \alpha$, the walk moves to a neighboring node, but instead of moving from one gene to one of its neighbors uniformly at random as is typically done, the probability instead is higher for neighbors that are closer to the prior knowledge set of genes $\mathcal{K}$. Genes that are visited more frequently in these random walks are more likely to be relevant for the disease because they are more likely to be part of important pathways around $\mathcal{K}$ that are also close to $\mathcal{M}$. We thus numerically compute the probability with which each gene is visited in these random walks, and then use these probabilities to rank the genes. See Methods 3.3 for details.

We apply our method `uKIN` to uncover cancer genes as well as genes associated with three rare heterogeneous disorders. To uncover cancer genes, we use somatic point mutation data from 24 different TCGA cancer types. Genes that have missense and nonsense somatic mutations comprise the new information, and random walks start from these genes with probability proportional to their mutation rates. We use the curated list of 499 cancer census genes (CGCs) available from COSMIC [25] to derive both our prior knowledge $\mathcal{K}$ of cancer driver genes as well as the hidden set of true positivies which we will use for evaluation. We test our approach for all 24 cancer types, but showcase results for glioblastoma multiforme (GBM). To uncover

genes associated with each of the three rare diseases, we obtain our prior knowledge from the Online Mendelian Inheritance in Man (OMIM), and genes that have been implicated via GWAS studies provide our new information. All results in the main paper use the *HPRD* protein-protein interaction network [69], with results shown for *BioGrid* [77] in the Supplement.

3.2.2 uKIN successfully integrates prior knowledge and new information

We first demonstrate that our method successfully combines prior disease knowledge and new information by evaluating its performance on the GBM dataset. Briefly, we use 20 randomly drawn CGCs to represent the prior knowledge $\mathcal{K}$ and another 400 randomly drawn CGCs to be the hidden set H of unknown cancer-relevant genes that we aim to uncover (see Performance evaluation 3.3.5 for details). We analyze the ranked list of genes output by `uKIN` as we consider an increasing number of output genes, and compute what fraction are members of the hidden set $\mathcal{H}$ consisting of cancer-driver genes. We compare `uKIN`'s performance when using both prior and new knowledge with $\alpha = 0.5$, to versions of `uKIN` using either only new information ($\alpha = 1$) or only prior information ($\alpha = 0$). For all three versions, we average performance over 100 randomized runs. For $\alpha = 0.5$, we observe that a large fraction of the top predicted genes are part of the hidden set of known cancer genes (Figure 3.2a). Among the top 100 predictions, 24 are CGCs ($p < 10^{-10}$, hypergeometric test).

At $\alpha = 1$, our method completely ignores both the network and the prior information $\mathcal{K}$ and is equivalent to ordering the genes by their mutational frequencies. That is because the random walk restarts at each step with probability 1 with the starting locations chosen probabilistically according to their mutational frequencies. The very top of the list output by `uKIN` when $\alpha = 1$ consists of the most frequently mutated genes (in the case of GBM, this includes *TP53* and *PTEN*). As we con-

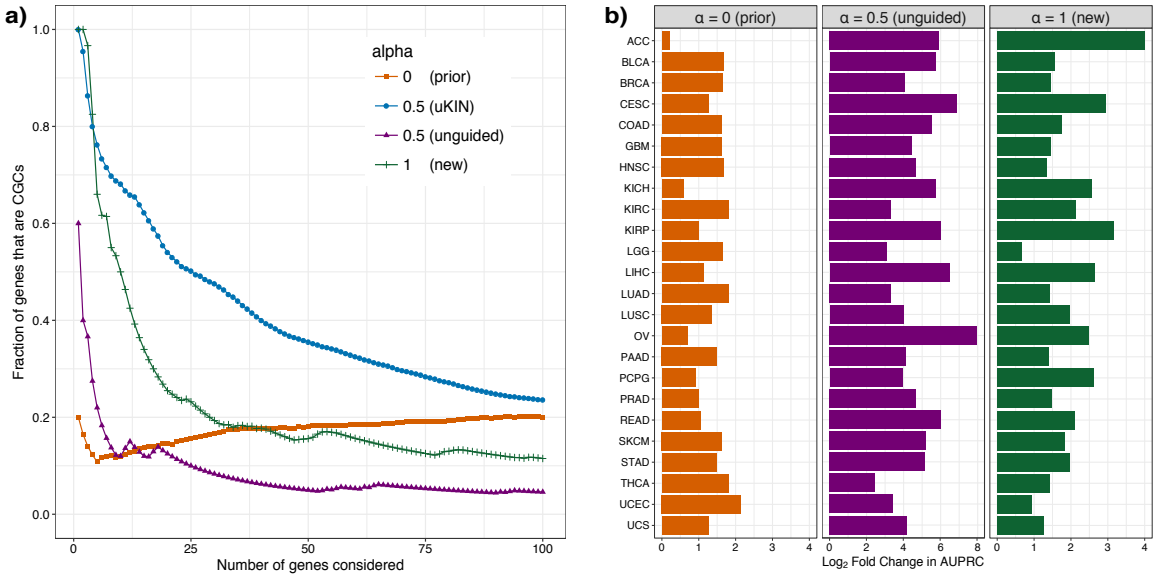


Figure 3.2: `uKIN` **successfully integrates new information and prior knowledge. (a)** We illustrate the effectiveness of our approach on the GBM data set and the HPRD protein-protein interaction network using 20 randomly drawn CGCs to represent the prior knowledge. We combine prior and new knowledge using a restart probability of $\alpha = 0.5$ (blue line). As we consider an increasing number of high scoring genes, we plot the fraction of these that are part of the hidden set of CGCs. As baseline comparisons, we also consider versions of our approach where we only use the new information ($\alpha = 1$) and order genes by their mutational frequency (green line), use the new information to perform *unguided* random walks with $\alpha = 0.5$ and order genes by their probabilities in the stationary distribution of the walk which depends on the network structure but not on the prior information (purple line), and where we only use prior information ($\alpha = 0$) and order genes based on propagating information from the set of genes comprising our prior knowledge (orange line). Integrating both prior and new sources of information results in better performance. **(b)** The performance of our network-based algorithm `uKIN` when integrating information at $\alpha = 0.5$ is compared to the three baseline cases where either only prior information is used ($\alpha = 0$, left) or only new information is used ($\alpha = 1$, right; $\alpha = 0.5$ (*unguided*), middle). In all three panels, for each cancer type, we compute the $\log_2$ ratio of `uKIN`'s AUPRC to the other approach's AUPRC. Across all 24 cancer types, using both sources of information outperforms using just one source of information.

sider an increasing number of genes, ordering them by mutational frequency is clearly outperformed by `uKIN` with $\alpha = 0.5$.

At the other extreme with $\alpha = 0$, the starting locations and their mutational frequencies are ignored as the random walk is memoryless and the stationary distribution depends only upon the propagated prior information Q. As expected, performance is

considerably worse than when running uKIN with $\alpha = 0.5$. Nevertheless, we observe that several CCGs are found for $\alpha = 0$; this is due to the fact that known cancer genes tend to cluster together in the network [10] and our propagation technique ranks highly the genes close to the genes in $\mathcal{K}$.

Another important basecase to consider is an *unguided* walk with the same restart probability $\alpha = 0.5$. In that case, the walk selects a neighboring node to move to uniformly at random. The stationary distribution that the walk converges to depends upon the starting locations and the network topology but is independent of the prior information. Such a walk provides a good baseline to judge the impact the propagated prior information Q has on the performance of our algorithm. As evident in Figure 3.2a, an *unguided* walk performs very poorly (purple line), highlighting the importance of Q in *guiding* the walk.

Noteworthily, the trends we observe on GBM hold across all 24 cancers (Figure 3.2b). For all cancer 24 cancers, the version of uKIN that uses both prior and new information with $\alpha = 0.5$ oupterforms using only prior information (Figure 3.2b, left) or only new information (Figure 3.2b, middle and right). Further, we have observe this improvement with using both prior and new information across all cancers for a wide range of α ($0.2 < \alpha < 0.8$), clearly demonstrating that using both sources of information is beneficial.

3.2.3 uKIN is effective in uncovering cancer relevant genes

Having shown in the previous section that our formulation is successful in integrating prior knowledge and new information, we next evaluate uKIN's performance in uncovering cancer relevant genes as compared to several previously published methods. In particular, for each of the 24 cancer types, we compute the $\log_2$ ratio of the area under the precision-recall curve (AUPRC) of uKIN with $\alpha = 0.5$ to the AUPRC for each of the other approaches. uKIN is run 100 times with 20 randomly sampled

genes comprising the prior knowledge, and evaluation is performed with respect to the 400 genes in the hidden set. First, we compare uKIN to MutSigCV 2.0 [53], a state-of-the-art frequency-based approach (Figure 3.3a). Our approach outperforms MutSigCV 2.0 on 22 of 24 cancer types. Second, we compare to three network-based approaches: Muffinn [12], a method that considers mutations found in interacting genes, DriverNet [3], a method that finds driver genes by uncovering sets of somatically mutated genes that are linked to dysregulated genes, and nCOP [35], a recent method that examines the per-individual mutational profiles of cancer patients in a biological network (Figure 3.3b). uKIN exhibits superior performance across all cancer types when compared to DriverNet, outperforms Muffinn in 23 out of 24 cancer types and nCOP in 17 of the 24 cancer types.

In several cancers, the performance improvements of uKIN are substantial. For example, uKIN has a four-fold improvement over MutSigCV 2.0 in predicting cancer genes for ovarian cancer (OV) and pancreas adenocarcinoma (PAAD), and a four-fold improvement over DriverNet for uterine corpus endometrial carcinoma (UCEC) and lung squamous cell carcinoma (LUSC). The limited number of patient samples available for uterine carcinosarcoma (UCS) limits nCOP's perfomance [35] whereas uKIN is able to leverage the prior knowledge available, resulting in uKIN's two fold improvement over nCOP; this highlights the benefits from incorporating existing knowledge of disease-relevant genes, especially when the new data is sparse. We also compare to Hotnet2 [56], a cutting-edge network diffusion method. As Hotnet2 does not output a ranked list of genes, we could not compute an AUPRC. Instead, examining the complete list of genes highlighted by both methods, we observe that uKIN exhibits both significantly better precision and recall (Supplementary Figure A.7). Overall, the clear advantage of uKIN over previous network-based approaches illustrates the benefits of using prior information in identifying cancer-relevant genes.

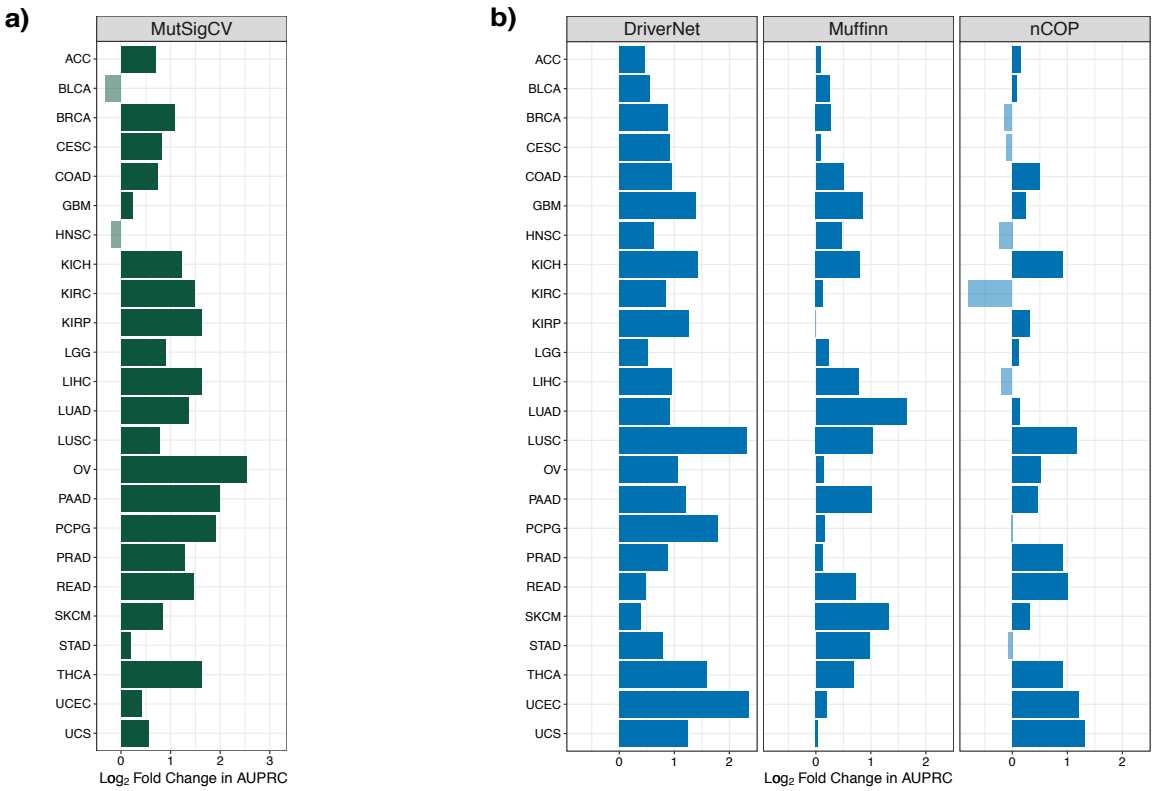


Figure 3.3: `uKIN` **is more effective than other methods in identifying known cancer genes.** For each method, for each cancer type, we compute the $\log_2$ ratio of `uKIN`'s AUPRC to its AUPRC. **(a)** Comparison of `uKIN` to `MutSigCV 2.0`, a state-of-the-art frequency-based approach. `uKIN` outperforms `MutSigCV 2.0` on 22 of the 24 cancer types. **(c)** Comparison of `uKIN` to `DriverNet` (left), `Muffinn` (middle), and `nCOP` (right). Our approach `uKIN` outperforms `DriverNet` on all cancer types, `Muffinn` on all but one cancer type and `nCOP` on 17 out of 24 cancer types.

Robustness tests. The overall results shown hold when we use different lists of known cancer genes used as a gold standard (Supplementary Figure A.8a), different numbers of predictions considered (Supplementary Figure A.8b), and different networks (Supplementary Figure A.8c). Further, we confirm the importance of network structure to `uKIN`, by running `uKIN` on two types of randomized networks, degree-preserving and label shuffling, and show that, as expected, overall performance deteriorates across the cancer types (Supplementary Figure A.8d); we note that while network structure is destroyed by these randomizations, per-gene mutational information is preserved, and thus highly mutated genes are still output.

3.2.4 Cancer-specific prior knowledge yields better performance

While many well-known cancer genes play a common role in the the development of multiple cancers (e.g., *TP53* and *PTEN*), others have been implicated in only a single or handful of cancer types. We next test how `uKIN`'s performance changes when using such highly specific prior knowledge. When filtering the set of CGC genes to those annotated to be drivers for a specific type of cancer, four cancer types, GBM, breast invasive carcinoma (BRCA), skin cutaneous carcinoma (SKCM), and thyroid carcinoma (THCA), have enough genes (33, 32, 42, 29, respectively) to split them in half to form the set of prior knowledge $\mathcal{K}$ and the hidden set $\mathcal{H}$.

We first use the genes specific to a cancer type of interest in $\mathcal{K}$ together with the TCGA data $\mathcal{M}$ for that cancer to uncover the genes in $\mathcal{H}$. Given the small number of genes in $\mathcal{H}$, we assess performance by measuring the average ranking over 100 splits of the data that `uKIN` assigns the genes in $\mathcal{H}$. Next, for the *same* cancer type, we use a set $\mathcal{K}$ corresponding to a *different* cancer type as prior knowledge (excluding any genes co-corresponding to the *original* cancer type) while still trying to uncover the genes in the *original* cancer of interest (i.e., using $\mathcal{M}$ and $\mathcal{H}$ belonging to the *original* cancer type). That is, we are testing the performance of `uKIN` when using knowledge corresponding to a different cancer type. For all four cancer types, we find that performance deteriorates when `uKIN` uses prior knowledge for another cancer type (Figure 3.4a), as genes in $\mathcal{H}$ appear further down in the list of genes output by `uKIN`. This suggests that `uKIN` can utilize cancer-type specific knowledge and highlights the benefits of having accurate prior information.

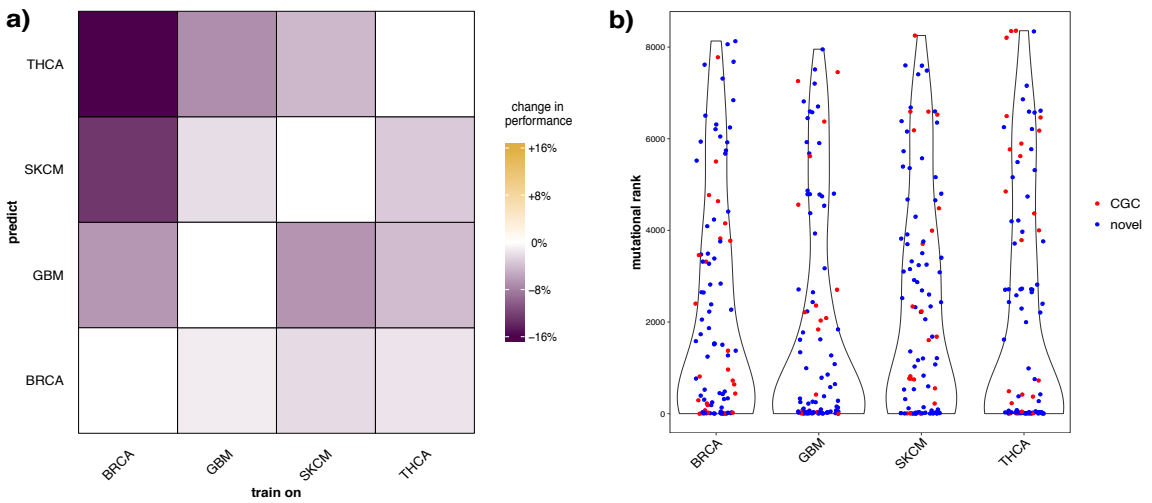


Figure 3.4: **(a) Use of cancer-type specific knowledge improves performance.** To assess the ability of our method to discern between knowledge specific to different cancer types, we split the genes from CGC annotated to be drivers only for a particular cancer type in two sets. We use the first one as prior knowledge while trying to uncover the genes in the second. This process is repeated 100 times. Next, we use the set of genes belonging to a different cancer type as prior knowledge while still trying to uncover the genes in the original cancer of interest. This leads to a decrease in perfomance (as measured by the increase in the average `uKIN`'s ranking of genes we aimed to uncover) across all possible combinations. **(b) `uKIN` identifies rarely mutated genes.** To illustrate `uKIN`'s ability to pull genes from the long tail of the mutational distribution, we run `uKIN` with $\alpha = 0.5$ and with 20 genes as prior knowledge 100 times. For each gene, its final score is averaged across the runs. For each of the top 100 genes, we consider the rank of its mutational rate (y-axis). Known CGC genes are in red and novel predictions in blue. The top predictions consist of many heavily mutated genes (i.e., those with low ranks), but `uKIN` is also able to uncover known cancer genes with very low mutational ranks (red dots towards the top).

3.2.5 `uKIN` highlights infrequently mutated cancer-relevant genes

We next demonstrate that `uKIN` highlights genes with a broad range of mutational rates. When we run `uKIN` on each of the 24 cancer types using prior knowledge consisting of 20 genes sampled 100 times, and consider the top 100 predictions (averaged across the runs), we observe that these genes have vastly diverse mutational rates

(Figure 3.4b for GBM, BRCA, SKCM and THCA and Supplemental Figure A.6 for all cancer types).

Naturally, because the starting locations of the random walk are chosen probabilistically proportionally to the genes' mutational frequencies, highly mutated genes are ranked among the top prediction. This makes the presence of many genes with very low mutational rates somewhat unexpected. In the case of GBM, among those rarely mutated genes are *LAND1A* and *SMAD4*, which are two well known cancer players. These genes have mutational rates that are similar to those of genes not relevant for cancer and are therefore hard to detect with frequency-based approaches. Among the 23 genes with mutational rank at the bottom half, 5 are CGCs ($p < 10^{-2}$, hypergeometric test). This statistically significant enrichment of CGC genes with low mutational rank found by `uKIN` is true across all cancers. Thus, `uKIN` provides a means for pulling out cancer genes from the "long tail" [27] of infrequently mutated genes.

In addition to highlighting known cancer genes, `uKIN` also ranks highly several non-CGC genes that may or may not play role in the initiation and progression of cancer, as our knowledge of cancer-related genes is incomplete. Among these novel predictions for GBM are *ATXN1*, *SMURF1*, and *CCR3* which have all recently been suggested to play a role in cancers [42, 57, 54] and are each mutated in less than 5% of the samples. *ATXN1* is a chromatin-binding factor that plays a critical role in the development of spinocerebellar ataxia, a neurodegenerative disorder [72], and mutants of *ATXN1* have been found to stimulate the proliferation of cerebellar stem cells in mice [21]. This is a promising gene for further investigation because glioblastoma is a cancer that usually starts in the cerebrum and the potential role of *ATXN1* in tumorigenesis has only recently been suggested [42]. *SMURF1* and its highly ranked by `uKIN` network-interactor *SMAD1* have already been implicated in the development

of several cancers [96]. *SMURF1* also interacts with the nuclear receptor *TLX* whose inhibitory role in glioblastoma has been revealed [40].

We further find that the genes `uKIN` highlights are enriched in many KEGG pathways and GO terms relevant for cancer, including *microRNAs in cancer*, *cell proliferation*, *choline metabolism in cancer* and *apoptosis* (Bonferroni-corrected $p < 0.001$, hypergeometric test).

3.2.6 Larger and more accurate prior knowledge improves performance

As our method relies on the use of prior knowledge, we examine the effect of the amount and accuracy of such knowledge on `uKIN`'s performance. To probe how much the amount of knowledge affects performance, we consider 10 randomly sampled hidden sets, which are held fixed as we sample 10 times per hidden set different sizes of already implicated disease genes $\mathcal{K}$ ($|\mathcal{K}| = 5, 10, 20, 40, \ldots, 100$). We run our framework on the kidney renal cell carcinoma dataset for three different values of α and compute the $\log_2$ ratio of the respective AUPRCs versus the AUPRC for $\alpha = 1$, as when $\alpha = 1$ the results do not depend on $\mathcal{K}$ at all (i.e., the AUPRC for $\alpha = 1$ is constant).

For $\alpha = 0.3$, `uKIN`'s performance in recapitulating the hidden set of known cancer genes steadily improves as a larger amount of prior knowledge is utilized (Figure 3.3a). For small $|\mathcal{K}| < 20$ `uKIN` with $\alpha = 0.5$ performs better than $\alpha = 0.3$ which is as expected, since at $\alpha = 0.5$ `uKIN` relies more on the new information $\mathcal{M}$ than on the limited prior knowledge $\mathcal{K}$. However, when $\mathcal{K}$ consists of a larger number of genes ($|\mathcal{K}| > 30$), $\alpha = 0.3$ overtakes $\alpha = 0.5$, suggesting that when substantial prior knowledge is available, `uKIN` can leverage it and a smaller α is preferred. On the other hand, when knowledge is sparse, a larger α allows `uKIN` to focus on the new information. Of course, as the number of genes comprising the set of prior

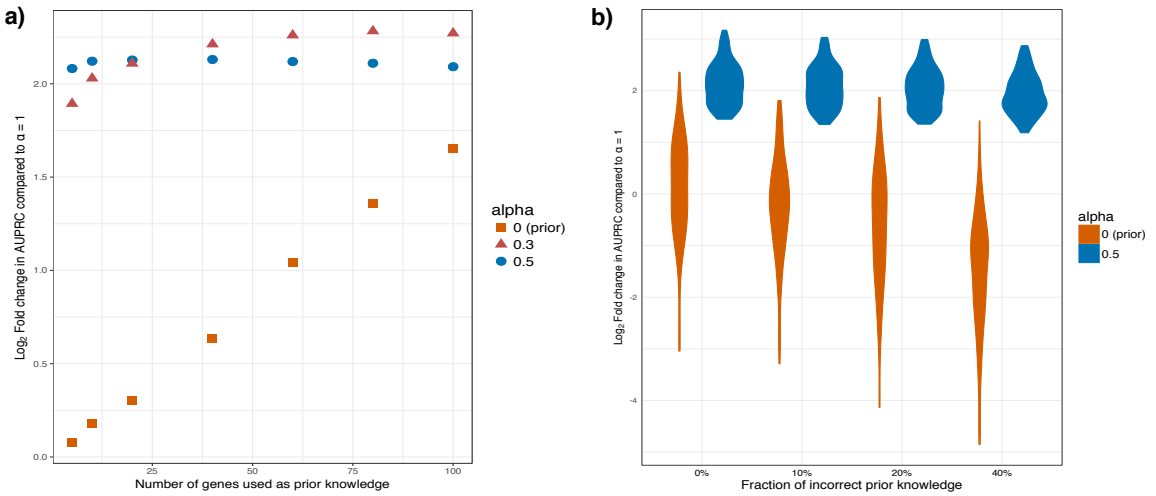


Figure 3.5: **(a) `uKIN` benefits from more knowledge.** As we consider larger numbers of genes comprising the set of prior knowledge ($|\mathcal{K}| = 5, 10, 20, 40, \ldots, 100$), we examine the ability of `uKIN` to uncover CGC genes in the same fixed set $\mathcal{H}$ when using $\alpha = 0.5$ (blue triangles), $\alpha = 0.3$ (pink circles) or $\alpha = 0$ (red squares). We show the $\log_2$ ratio, averaged over 100 runs, of the AUPRC of each version of `uKIN` to the AUPRC for $\alpha = 1$ which is constant across all possible $\mathcal{K}$. For small $\mathcal{K}$, $\alpha = 0$ performs poorly as is expected; as the prior knowledge available increases so does the performance. For both $\alpha = 0.3$ and $\alpha = 0.5$, an increase in the size of $\mathcal{K}$ leads to an initial increase in the performance but eventually performance plateaus. When limited prior knowledge is available ($|\mathcal{K}| < 20$), $\alpha = 0.5$, which uses more of the new information, does better then $\alpha = 0.3$, which relies more on using prior knowledge. When prior knowledge is abundant ($|\mathcal{K}| > 40$), `uKIN` with $\alpha = 0.3$ outperforms $\alpha = 0.5$. **(b) `uKIN` is robust to small amounts of erroneous knowledge.** We replace a fraction of the CGCs in the set of prior knowledge $\mathcal{K}$ with non-cancerous genes chosen uniformly at random from the set of non-CGC genes in the network. `uKIN` remains robust to some incorrect knowledge (10%); it's performance decreases significantly when $> 30\%$ of the prior knowledge becomes incorrect. As expected, for $\alpha = 0$, the decrease is more notable because in that case `uKIN` uses only prior knowledge.

knowledge increases, spreading information just from those genes ($\alpha = 0$), the better the propagated knowledge does as a stand alone predictor. This is consistent with the observed clustering of CGC genes within biological networks [10]. However, even when propagating information from 100 known cancer genes, the performance is worse than that when integrating it with new information (with either $\alpha = 0.3$ or $\alpha = 0.5$, Figure 3.3a).

We next investigate the effect of having some incorrect prior knowledge (a plausible real world scenario). We simulate the presence of erroneous knowledge by replacing a fraction (10%, 20%, 30%, and 40%) of the CGCs in the set $\mathcal{K}$ with non-cancerous genes chosen uniformly at random from the set of non-CGC genes in the network and rerunning `uKIN`. While only a small drop in performance is observed when 10% of the genes in $\mathcal{K}$ are replaced, the decrease becomes significant when $> 30\%$ are replaced (Figure 3.3b)). Further, as expected, this decrease is more notable for $\alpha = 0$ then for $\alpha = 0.5$ as the former relies entirely on prior knowledge than the latter. Overall, our results suggest that `uKIN` is robust to some noise in the prior knowledge $\mathcal{K}$ ($\leq 10\%$) but that if there is uncertainty about the knowledge in $\mathcal{K}$, a larger α should be used.

3.2.7 Application to identify disease genes for complex inherited disorders

A major advantage of our method is that it can be easily applied to other scenarios, using a wide variety of different types of information. To demonstrate this versatility, we applied `uKIN` to detect disease genes for three complex diseases: *macular degeneration*, *amyotrophic lateral sclerosis (ALS)* and *epilepsy*. For each disease, we randomly split in half the OMIM database's [62] list of genes associated with the disease 100 times to form the set of prior knowledge $\mathcal{K}$ and the hidden set $\mathcal{H}$. We use the GWAS catalogue list of genes with their corresponding $p-$values to form the set $\mathcal{M}$. For all three diseases, we find that spreading the signal using only knowledge from OMIM ($\alpha = 0$) performs worse than combining both sources of information ($\alpha = 0.5$). For each of these diseases, there is virtually no overlap between the GWAS hits $\mathcal{M}$ and a set of OMIM genes $\mathcal{H}$; simply sorting genes by their significance in GWAS studies (i.e., `uKIN` with $\alpha = 1$) results in AUPRC of 0. Instead, we spread information from the set of GWAS genes $\mathcal{M}$ in the same fashion as from OMIM and observe again that

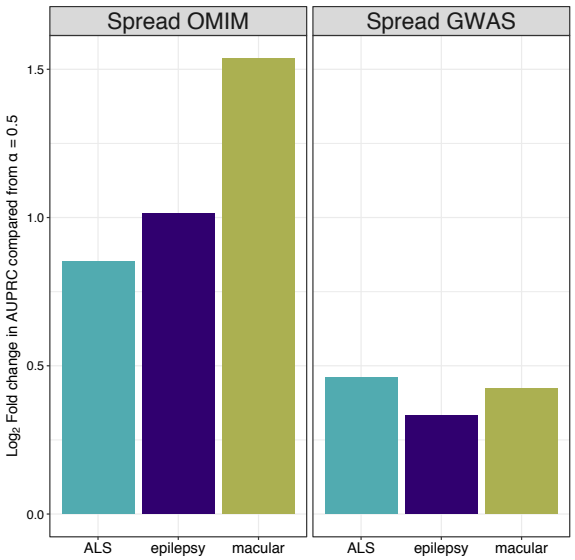


Figure 3.6: **uKIN is effective in identifying complex disease genes.** We demonstrate the versatility of the uKIN framework by integrating OMIM and GWAS data for three complex diseases, epilepsy, ALS and macular degeneration. For each disease, we compare uKIN's performance when using OMIM annotated genes as prior information and GWAS hits as new information with $\alpha = 0.5$, to baseline versions that propagate only information from OMIM ($\alpha = 0$, left) or GWAS studies (right). In all cases, we compute the $\log_2$ ratio of the AUPRC obtained by uKIN using both prior and new information to the baseline methods.

using this single source of information alone has worse performance then combining it with another (Figure 3.6, right panel).

3.3 Methods

3.3.1 Background and notation

The biological network is modeled, as usual, as an undirected graph $G = (V, E)$ where each vertex represents a gene, and there is an edge between two vertices if an interaction has been found between the corresponding protein products. We require G to be connected, restricting ourselves to the largest connected component if necessary. We explain our formulation with respect to cancer, but note that it is applicable in

other settings (both disease and otherwise). The set of genes already known to be cancer associated is denoted by $\mathcal{K} = \{k_1, k_2, ..., k_l\}$. The set of genes that have been found to be somatically mutated in a cohort of individuals with cancer is denoted by $\mathcal{M} = \{m_1, m_2, ..., m_p\}$, with $\mathcal{F} = \{f_{m_1}, f_{m_2}, ..., f_{m_p}\}$ corresponding to the rate with which each of these genes is mutated. We refer to $\mathcal{K}$ as the prior knowledge and $\mathcal{M}$ as the new information. We assume that $\mathcal{K} \subset V$ and $\mathcal{M} \subset V$; in practice, we remove genes not present in the network. The genes within $\mathcal{K}$ and $\mathcal{M}$ may overlap (i.e., it is not required that $\mathcal{K} \cap M = \emptyset$). Our goal is to integrate all three types of information, G, $\mathcal{K}$ and $\mathcal{M}$, in order to uncover new cancer genes. Our method is based on the intuition that genes close to $\mathcal{K}$ are more likely to be involved in the same cellular processes or pathways as genes in $\mathcal{K}$ and hence more likely to be relevant for disease. We thus perform random walks over the network G, starting from $\mathcal{M}$ but biased towards going closer to $\mathcal{K}$, and rank genes with respect to disease relevance by how frequently they are visited.

3.3.2 Algorithm

For each gene $v \in V$, assume that we have a measure Q_v that represents how close v is to the set of genes $\mathcal{K}$. We will use the measure Q, which we describe in the next section, to guide a random walk starting at the nodes in M and walking towards the nodes in $\mathcal{K}$. Each walk starts from a gene i in M, chosen with probability proportional to its mutational rate f_i. At each step, with probability α the walk can restart from a gene j in M, and with probability $1 - \alpha$ the walk moves to a neighboring gene picked probabilistically based upon Q. Specifically, if $\mathcal{N}(u)$ are the neighbors of node u, the walk goes from node u to node $v \in \mathcal{N}$ with probability proportional to $Q(v) / \sum_{w \in \mathcal{N}} Q(w)$. That is, if at time t the walk is at node u, the probability that it

transitions to node v at time $t+1$ is

$$p_{uv} = (1-\alpha)\delta_{uv} \cdot \frac{Q(v)}{\sum_{w \in N(u)} Q(w)} + \alpha \cdot \frac{f_v}{\sum_{i \in \mathcal{M}} f_i}$$

where $\delta_{uv} = 1$ if $v \in \mathcal{N}(u)$ and 0 otherwise. Hence, the guided random walk is fully described by a transitional matrix P with entries p_{uv}. This stochastic matrix is non-negative and by the Perron-Frobenius theorem it has a right eigenvector π corresponding to eigenvalue 1. Therefore, $\pi P^t = \pi$ and π is the stationary distribution the guided random walk converges to and this can be computed numerically. For each gene i, its score is given by the ith element of π. Those with high scores are most frequently visited and, therefore, are more likely relevant to cancer as they are close to both the mutated starting nodes as well as to known cancer genes.

3.3.3 Incorporating prior knowledge

For each gene in the network, we wish to compute how close it is to the set of cancer associated genes $\mathcal{K}$. While many approaches have been proposed to compute "distances" in networks, we use a network flow technique where each node $k \in K$ introduces a continuous unitary flow which diffuses uniformly across the edges of the graph and is lost from each node $v \in V$ in the graph at a constant first-order rate λ [71]. Briefly, let $A = \{a_{i,j}\}$ denote the adjacency matrix of G (i.e., $a_{ij} = 1$ if $(i,j) \in E$ and 0 otherwise) and let S be the diagonal matrix where s_{ii} is the degree of node $i \in V$. Then, the Laplacian of the graph G shifted by λ is defined as $L = -(A - S - \lambda I)$. The equilibrium distribution of fluid density on the graph is computed as $Q = L^{-1}b$ [71], where b is the elementary unit vector with 1 for the nodes introducing the flow and 0 for the rest (i.e., $b_i = 1$ if $v_i \in K$ and $b_i = 0$ if $v_i \notin K$ for $\forall v_i \in V$). Q can be efficiently computed numerically. Thus, at equilibrium, each

node v in the graph is associated with the score Q_v which reflects how close it is to the nodes already marked as causal for cancer.

3.3.4 Data sources and pre-processing

We use two different biological networks in our analysis: *HPRD* (Release 9_041310) [69] and *BioGrid* (Release 3.2.99, physical interactions only) [77]. Biological networks often contain spurious interactions as well as "hub" proteins with many interactions. Since both are problematic for network analysis, we pre-process the networks as in [35]. Briefly, we remove all proteins with an unusually high number of interactions (> 900 interactions and more than 10 standard deviations away from the mean number of interactions). For *BioGrid*, this removes *UBC*, *APP*, *ELAVL1*, *SUMO2* and *CUL3*. For *HPRD*, this removes no proteins. Additionally, to remove spurious interactions, we remove those that have a Z-score normalized diffusion state distance [7] > 0.3. This process leaves us with 9,379 proteins and 36,638 interactions for *HPRD* and 14,326 proteins and 102,552 interactions for *BioGrid*.

We download level 3 cancer somatic mutation data from The Cancer Genome Atlas (TCGA) [81] for 24 cancer types (Table A). For each cancer type, we process the data as previously described and exclude samples that are obvious outliers with respect to their total number of mutated genes [35]. Our set of prior knowledge comes from the 719 CGC genes that are labeled by COSMIC (version August 2018) as being causally implicated in cancer [25]. For each cancer type, our new information consists of genes that have somatic missense mutations, and we compute the mutational frequency of a gene as the number of observed somatic missense mutations across tumors, divided by the number of amino acids in the encoded protein.

We obtain 24, 28, and 63 genes associated with three complex diseases, macular degeneration, ALS and epilepsy, respectively, from the OMIM database [62]. These genes are used to construct the set of prior knowledge. For each disease, we form

the set M by querying from the GWAS database [6] the genes implicated for the disease and using the corresponding p-values to compute the starting frequencies f. Specifically, for each disease, for each study GWAS i, if a gene j's p-value is $p_{i,j}$, we set its frequency to $-\log(p_{i,j})/\sum_k -\log(p_{i,k})$ and then for each gene average these frequencies over the studies.

3.3.5 Performance evaluation

To evaluate our method in the context of cancer, we subdivide the CGC genes that appear in our network into two subsets. One subset will serve as our prior knowledge $\mathcal{K}$ and we will test how well our approach scores the genes in the other subset. In particular, we randomly draw from the CGCs 400 genes to form a set $\mathcal{H}$ of positives that we aim to uncover. From the remaining 199 CGCs present in the network, we randomly draw a fixed number l to represent the prior knowledge $\mathcal{K}$ and run our framework. As we consider an increasing number of most highly ranked genes, we compute the fraction that are in the set $\mathcal{H}$ of positives. All CGC genes not in $\mathcal{H}$ are ignored in these calculations; this allows us to compare performance when varying the number l of genes that comprise our prior knowledge.

We also compute area under the precision-recall curves (AUPRCs). In this case, all CGC genes in $\mathcal{H}$ are considered positives, all CGC genes not in $\mathcal{H}$ are neutral (ignored), and all other genes are negatives. Though we expect that there are genes other than those already in the CGC that play a role in cancer, this is a standard approach to judge performance (e.g., see [39]) as cancer genes should be highly ranked. We compute AUPRCs using the top 100 predicted genes. To account for the randomness in sampling, we repeatedly draw (10 times) the set $\mathcal{H}$ and for each draw we sample the genes comprising the prior knowledge $\mathcal{K}$ 10 times. The final AUPRC results from averaging the AUPRCs across all 100 runs.

We compare uKIN on the cancer datasets to `MutSigCV 2.0` [53], `nCOP`[35], `Muffinn` [12], and `DriverNet` [3]. To ensure fair comparisons, all methods are run on exactly the same cancer mutation data and the same network if applicable. All methods' AUPRCs are computed against the same randomly sampled test sets $\mathcal{H}$ and averaged at the end. All methods are run with their default parameters.

To evaluate our method in the context of the three complex diseases, we subdivide evenly the set of OMIM genes associated with each disease into the prior knowledge set $\mathcal{K}$ and the set of positives $\mathcal{H}$. Similarly to our cancer evaluation, this is done repeatedly (100 times) and respective AUPRCs are averaged at the end.

3.4 Discussion

Here, we have shown that `uKIN`, a method that incorporates both existing knowledge as well as new information, is an effective and versatile approach for uncovering disease genes. Our method is based upon the intuition that prior knowledge of disease-relevant genes can be used to guide the way information from new data is spread and interpreted in the context of biological networks. Our approach demonstrates higher precision than other state-of-the-art methods in detecting known cancer genes and excels at highlighting infrequently mutated genes that are nevertheless relevant for cancer.

The framework presented here can be extended in a number of natural ways. First, in addition to positive knowledge of known disease genes, we also have "negative" knowledge of genes that are not involved in the development of a given disease. These genes can propagate their "negative" information, thereby biasing the random walk to move away from their respective modules and perhaps further enhancing the performance of our method. Second, `uKIN` may also benefit from incorporating edge weights that reflect the reliability of interactions between proteins (e.g., inter-

actions between proteins that are co-expressed are more likely to be reliable, as are interactions that are seen in multiple experiments); these weights will have an impact on both the propagation of prior knowledge as well as the guided random walks. Third, since a recent study [70] has shown that contrasting cancer mutation data with natural germline variation data helps boost the true disease signal by downgrading genes that mutate frequently in nature, `uKIN`'s performance may benefit from scaling the starting probabilities of the new putatively implicated genes to account for their variation in healthy populations. More interestingly, additional techniques could examine if pathways or modules accommodate large natural variation and hence, the guided walks should move away from them. Fourth, while here we have demonstrated how `uKIN` can use cancer-type specific knowledge, cancers of the same type can often be grouped into distinct subtypes, and such highly-detailed knowledge may improve `uKIN`'s performance even further.

In conclusion, `uKIN` is a flexible method that handles diverse types of new information, is robust, fast, runs on a desktop machine, and is freely available online. As our knowledge of disease-associated genes continues to grow and be refined, and as new experimental data becomes more abundant, we expect that `uKIN` will prove to be a powerful and broadly applicable framework for accurately, rapidly, and easily prioritizing disease genes.

4 Conclusions

In this dissertation, I introduced two new approaches for deciphering disease genomes in the context of large biological networks. In *Chapter 2*, I developed a novel method that examines per-individual mutational profiles of cancer patients. I showed that my approach readily outperforms other state-of-the-art approaches in discovering cancer genes. In *Chapter 3*, I described a general framework for incorporating prior knowledge and new information. I showed how the signal from an already known set of disease-associated genes can be used to guide the way newly acquired experimental data is interpreted. My approach led to identifying disease genes with higher precision than using either source of information alone. Both of my methods successfully tackle the overarching problem of disease heterogeneity and are able to uncover rarely mutated but highly relevant disease genes. Another important underlying commonality between my two approaches is that they are both well-positioned to take advantage of the rapidly increasing amount of available biological data. In the case of `nCOP`, as the number of individuals for which we have mutational data increases, the method's power in detecting cancer genes will also increase. Similarly, in the case of `uKIN`, as either our set of known disease genes or the amount of diverse new disease information increases, its ability to uncover additional disease genes will also increase.

My dissertation contributes to the growing body of network-based methods designed to tackle the Herculean task of understanding how genetic changes lead to a disease. My methods will serve as a valuable resource to the scientific community as it

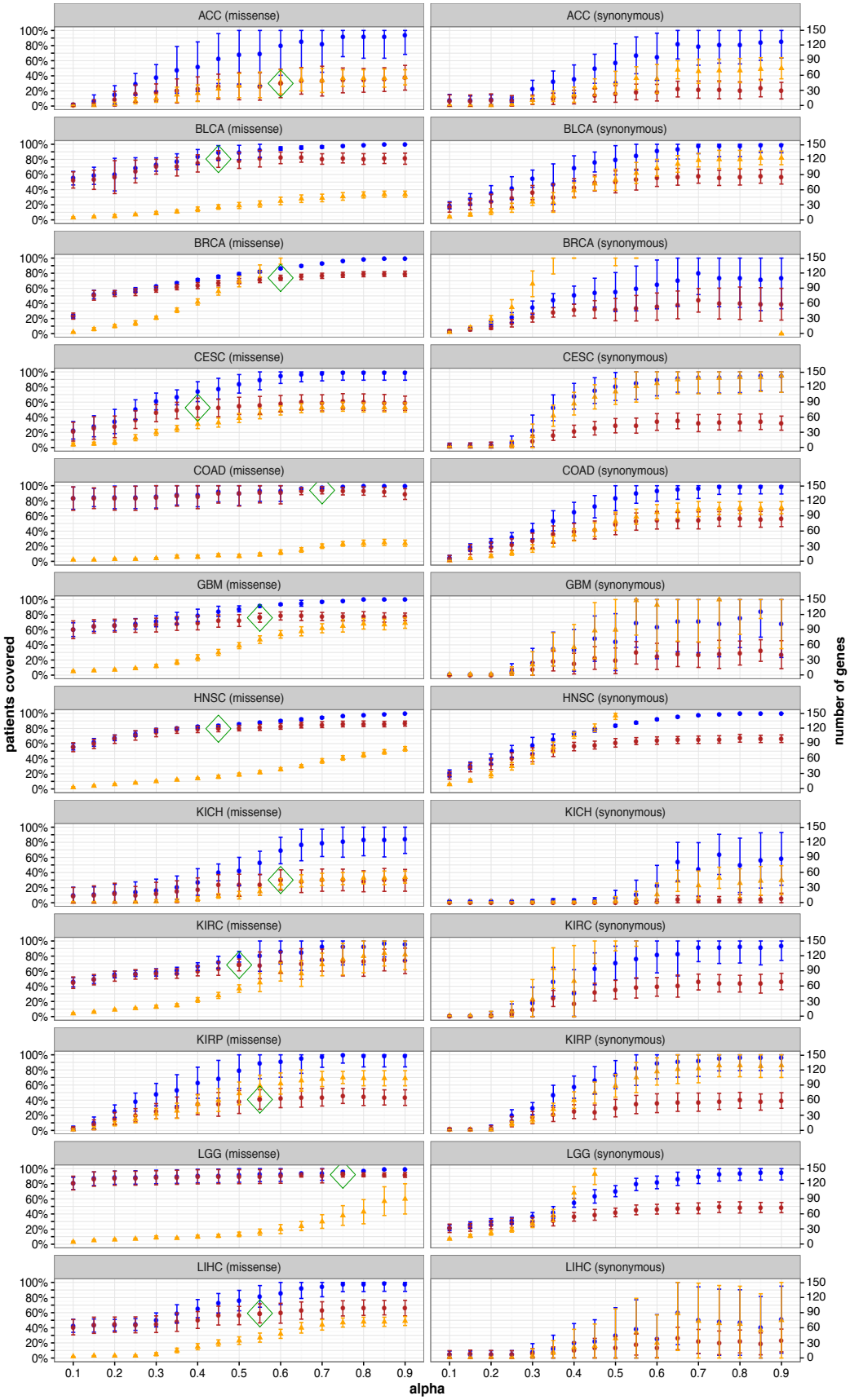
continues its quest to make exciting discoveries in both basic science and biomedical research.

A Supplementary Figures

The following appendix contains a table summarizing the TCGA data I use along with 8 supplementary figures that support the findings in the *Chapters 2* and *3*.

Cancer Symbol	Cancer Type	Number of Patients	Number of Mutated Genes		
			Total	Average	Cut off
ACC	Adrenocortical carcinoma	76	2068	32.1	80
BLCA	Bladder Urothelial Carcinoma	196	11407	135.7	300
BRCA	Breast invasive carcinoma	882	10813	27	80
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	173	6907	63	200
COAD	Colon adenocarcinoma	153	6521	74.4	150
GBM	Glioblastoma multiforme	278	7250	46.8	80
HNSC	Head and Neck squamous cell carcinoma	435	13048	87.9	200
KICH	Kidney Chromophobe	64	661	11	50
KIRC	Kidney renal clear cell carcinoma	416	9212	40.9	100
KIRP	Kidney renal papillary cell carcinoma	166	5687	47.7	100
LGG	Brain Lower Grade Glioma	451	7130	28.8	60
LIHC	Liver hepatocellular carcinoma	196	7705	67.3	200
LUAD	Lung adenocarcinoma	487	15481	172.8	500
LUSC	Lung squamous cell carcinoma	167	12264	212	500
OV	Ovarian serous cystadenocarcinoma	138	3390	30.7	80
PAAD	Pancreatic adenocarcinoma	124	3228	36.8	100
PCPG	Pheochromocytoma and Paraganglioma	183	1819	11.7	30
PRAD	Prostate adenocarcinoma	238	4792	28.1	50
READ	Rectum adenocarcinoma	34	1214	40.7	150
SKCM	Skin Cutaneous Melanoma	329	14748	240.1	1000
STAD	Stomach adenocarcinoma	242	10595	103.5	500
THCA	Thyroid carcinoma	401	2268	7.4	30
UCEC	Uterine Corpus Endometrial Carcinoma	155	4282	38.8	100
UCS	Uterine Carcinosarcoma	54	1787	38.9	80

Table A.1: **TCGA dataset and statistics.** We list the 24 cancer types studied along with their abbreviations. For each cancer type, we give the total number of patient samples considered after highly mutated samples are filtered out, the total number of mutated genes across these samples, the average number of mutated genes across all samples, and the cutoff on the number of mutated genes within a sample that was used to filter samples.


ACC (missense)
ACC (synonymous)
BLCA (missense)
BLCA (synonymous)
BRCA (missense)
BRCA (synonymous)
CESC (missense)
CESC (synonymous)
COAD (missense)
COAD (synonymous)
GBM (missense)
GBM (synonymous)
HNSC (missense)
HNSC (synonymous)
KICH (missense)
KICH (synonymous)
KIRC (missense)
KIRC (synonymous)
KIRP (missense)
KIRP (synonymous)
LGG (missense)
LGG (synonymous)
LIHC (missense)
LIHC (synonymous)
patients covered
number of genes
alpha
0%
20%
40%
60%
80%
100%
0
30
60
90
120
150
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

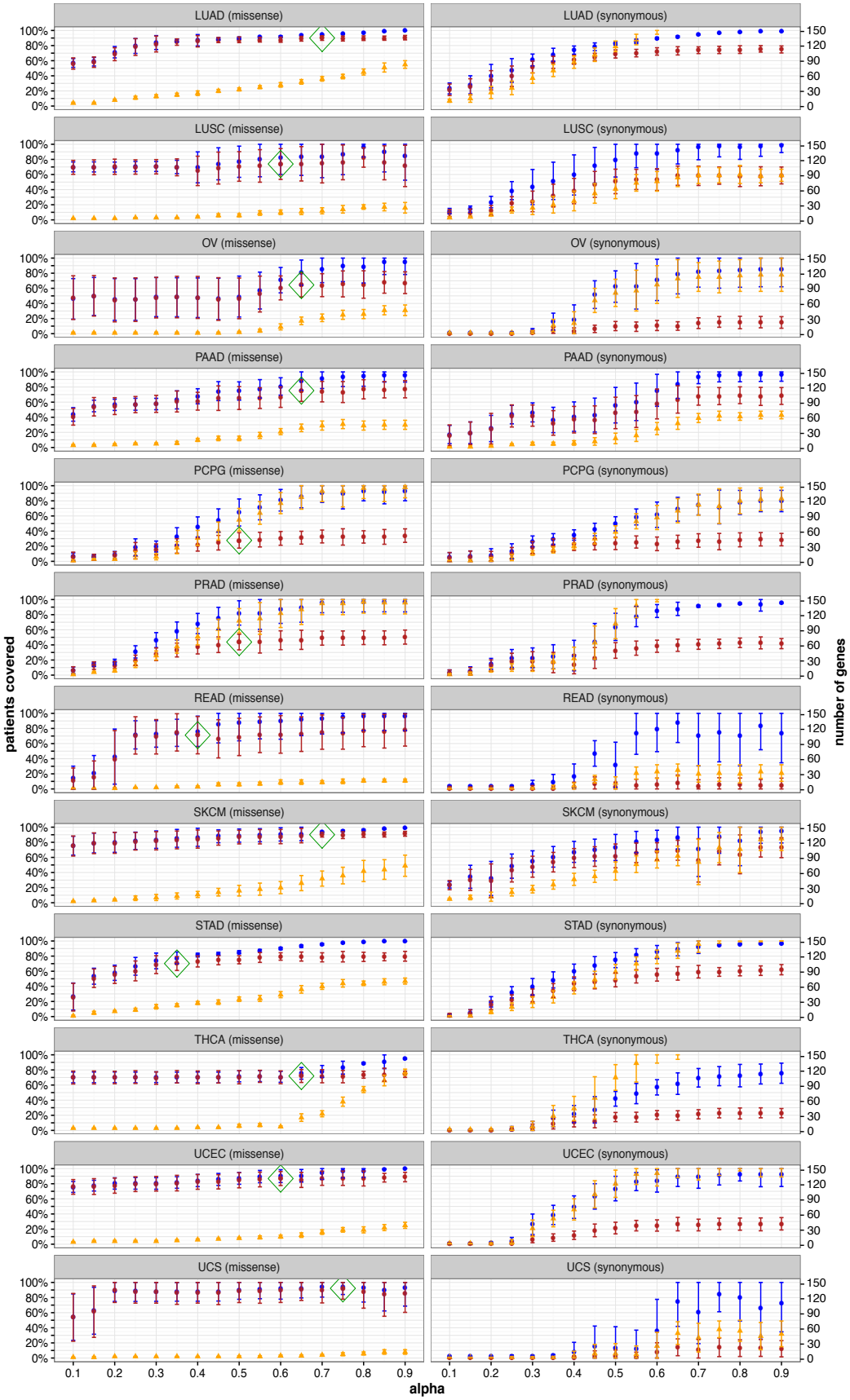

LUAD (missense)
LUAD (synonymous)
LUSC (missense)
LUSC (synonymous)
OV (missense)
OV (synonymous)
PAAD (missense)
PAAD (synonymous)
PCPG (missense)
PCPG (synonymous)
PRAD (missense)
PRAD (synonymous)
READ (missense)
READ (synonymous)
SKCM (missense)
SKCM (synonymous)
STAD (missense)
STAD (synonymous)
THCA (missense)
THCA (synonymous)
UCEC (missense)
UCEC (synonymous)
UCS (missense)
UCS (synonymous)
patients covered
number of genes
alpha
0%
20%
40%
60%
80%
100%
0
30
60
90
120
150
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

Figure A.1: **Fraction of individuals covered as α varies across all cancers.** For each random split of the individuals, we run our algorithm on the training sets for different values of α, and plot the fraction of covered individuals in the training (blue) and validation (red) sets. We also give the number of proteins in the uncovered subgraphs G' (orange). For each plotted value, the mean and standard deviation over 100 random splits are shown and the automatically selected α for the missense mutation data is indicated by a green rhombus. The performances on both the training and validation sets are much worse when using synonymous mutations compared to when using missense mutations. Coverage on the validation set for synonymous mutations is consistently lower for the same values of α across respective cancer types than that for missense mutations, with maximum possible coverage on the validation set not exceeding 50% in many cases. Further, it takes significantly more nodes to cover the same fraction of patients when using synonymous mutations.

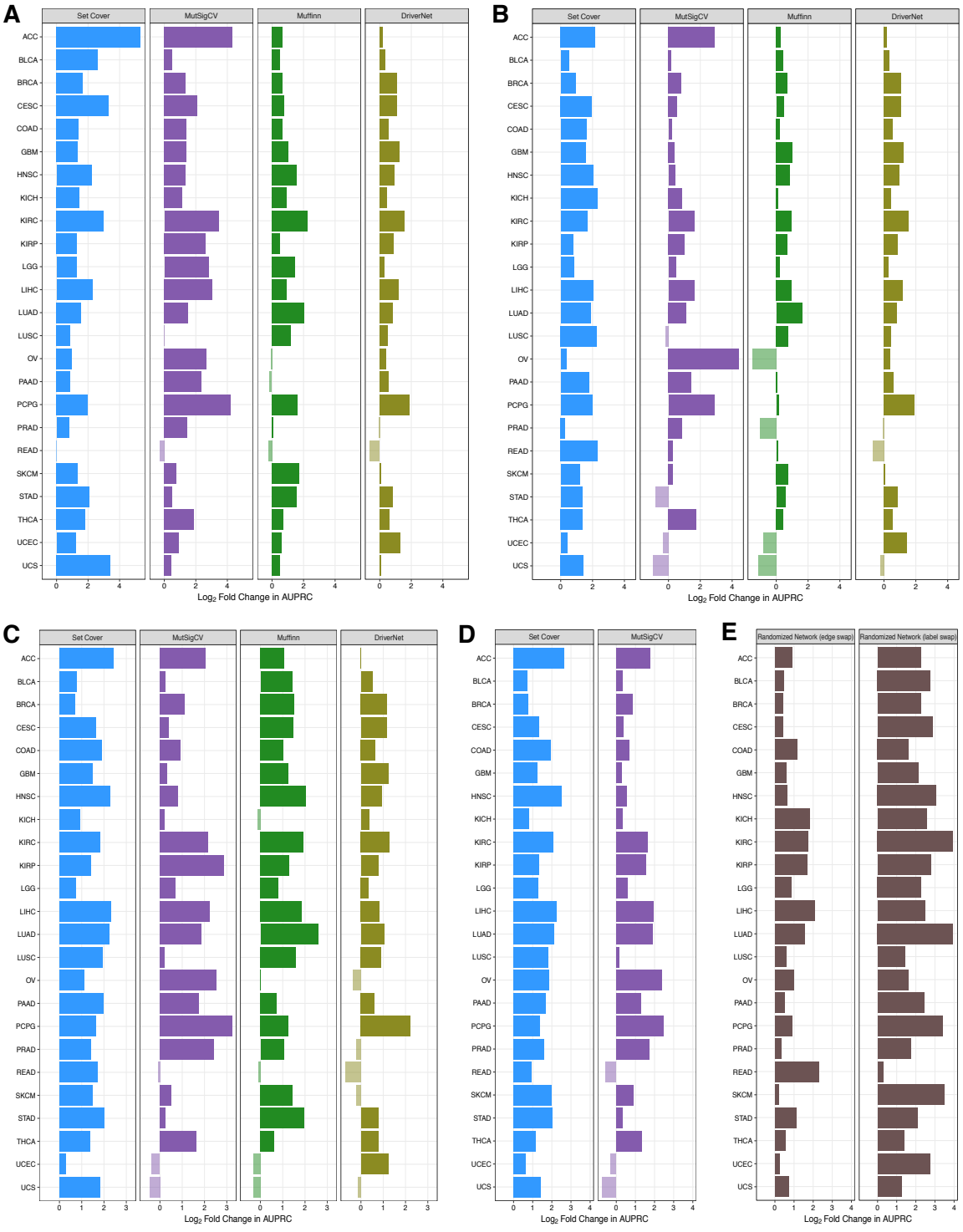

A
B
C
D
E
Set Cover
MutSigCV
Muffinn
DriverNet
Randomized Network (edge swap)
Randomized Network (label swap)
ACC
BLCA
BRCA
CESC
COAD
GBM
HNSC
KICH
KIRC
KIRP
LGG
LIHC
LUAD
LUSC
OV
PAAD
PCPG
PRAD
READ
SKCM
STAD
THCA
UCEC
UCS
Log2 Fold Change in AUPRC

Figure A.2: **Robustness of nCOP. (a)** To make sure that our method is robust with respect to the set of labelled cancer genes, instead of the Cancer Gene Census (CGC) list, we use the list of 413 genes provided by Hofree et al. in [33] which they obtained by querying the UniprotKB database for the keyword-terms 'proto-oncogene,' 'oncogene' and 'tumoursuppressor' gene. Log-fold AUPRCs are computed as described in the main text. The results are consistent with those shown in Figure 3 based on the CGC list and show the superior performance of nCOP as compared to the other methods in recapitulating known cancer genes. **(b)** Results using the Vogelstein et al. list of cancer genes [91]. **(c)** To assess the robustness of our evaluation, we compute AUPRCs using the top 50 predicted genes. The results are consistent with those shown in Figure 3 which use the top 100 predicted genes and show the superior performance of nCOP as compared to the other methods in deriving known cancer genes. The results are also consistent when computing AUPRC's using 150 genes (data not shown). **(d)** To make sure that our method is robust with respect to the specific network utilized, we repeat our entire analysis procedure using the Biogrid network. Our approach nCOP outperforms the network-agnostic methods in 21 out of 24 of the cancer types. **(e)** Comparison to randomized networks. In the left panel, we use a classic degree-preserving randomization (edge swapping) and in the right panel, we use a node label shuffling randomization where the network structure is maintained but gene names are swapped (thereby genes can have very different numbers of interactions in the randomizations). For each of the 24 cancers, we compute the log_2 ratio of the area under the precision recall curve using nCOP on the real network and on the randomized network and show the average over 10 different randomizations. Performance, as expected, is worse for both randomizations across all cancers. We note that significant cancer-relevant information is retained in these randomized networks. In particular, in both types of network randomizations, we maintain the relationships between genes and the patients that they are found to be somatically mutated in. Thus, some highly mutated CGC genes may still be output by nCOP when running on randomized networks.

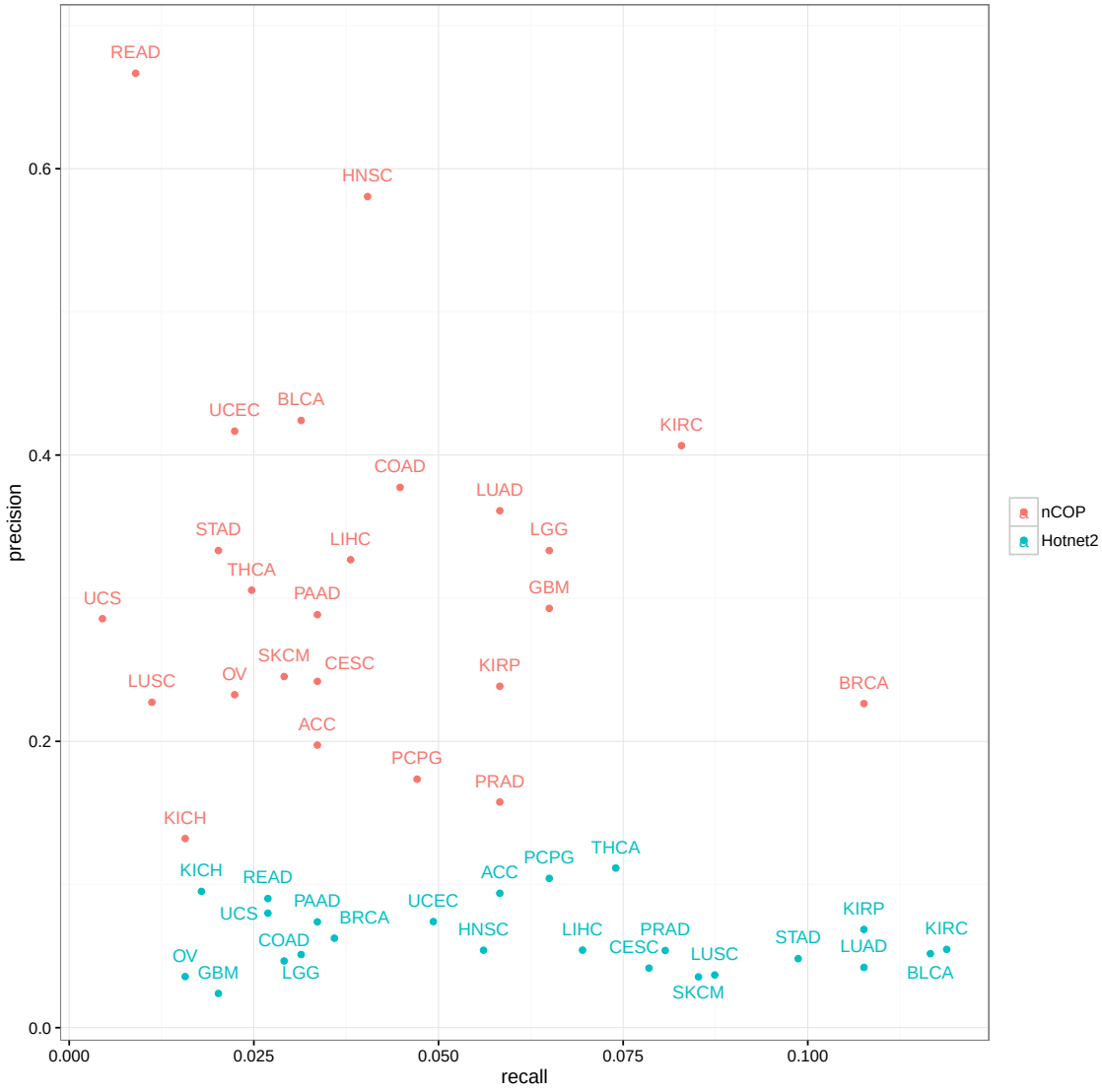


Figure A.3: **Comparison between `nCOP` and `Hotnet2`.** For each cancer type, we compute the precision and recall of the genes returned by `nCOP` and `Hotnet2`. For `nCOP`, we choose a single threshold to select predicted cancer genes, corresponding to those genes that occur in at least 25% of the runs. While `Hotnet2` achieves slightly greater recall due to the larger number of genes it highlights, `nCOP`'s precision using this threshold is superior. `nCOP` uncovers fewer but potentially more relevant cancer genes.

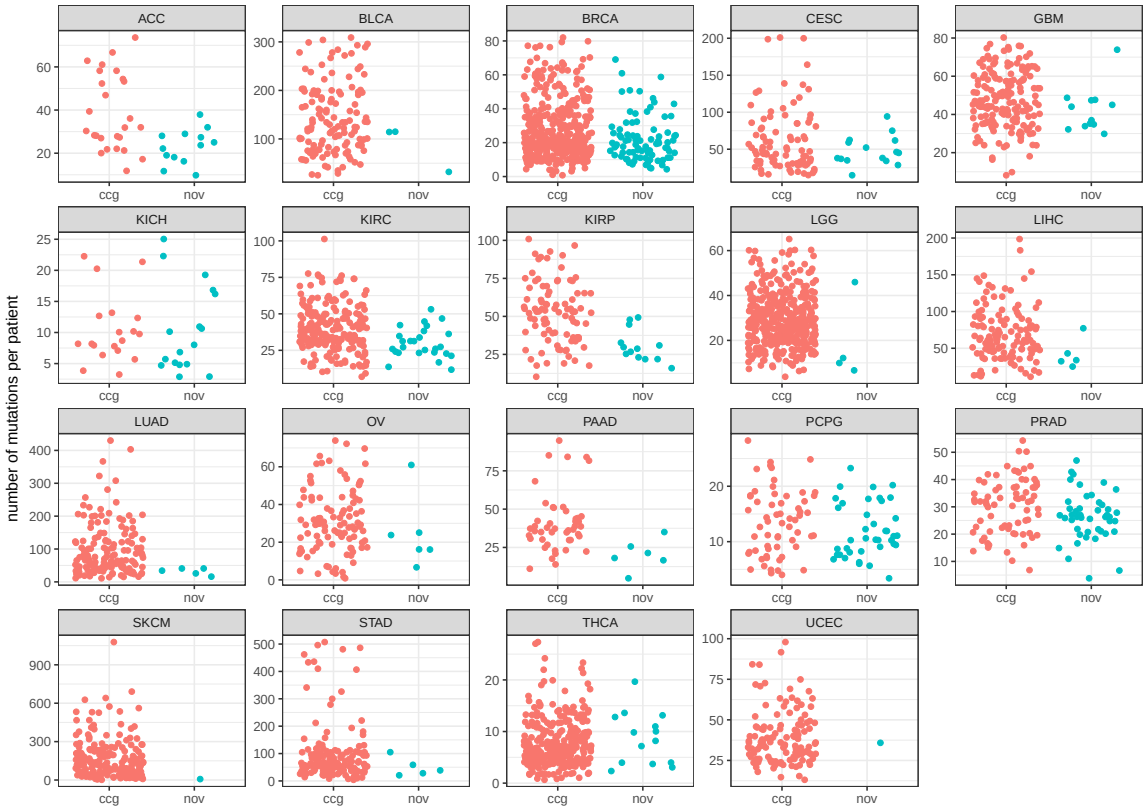


Figure A.4: **Novel genes uncovered by `nCOP` are not due to patients with many mutations.** Plotted for each cancer type are the total number of missense mutations for patients having missense mutations only in known CGC genes and not in novel genes (left) and the total number of missense mutations for patients having missense mutations only in novel genes and not in CGC genes (right). The novel genes uncovered by `nCOP` are not due to patients with large numbers of mutations.

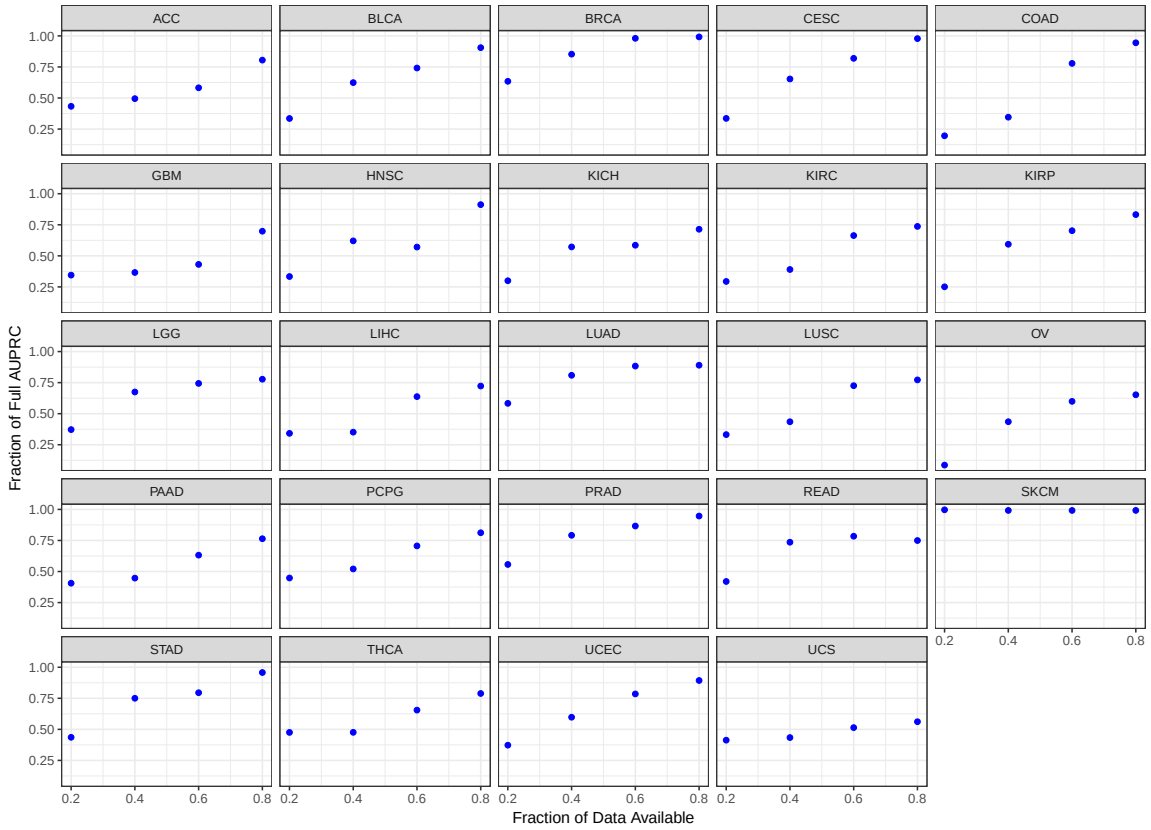


Figure A.5: **The predictive power of nCOP increases with more data.** For each cancer type, we repeatedly sample a fraction of the patients (20%, 40%, 60%, and 80%), rerun our method on the reduced data set, compute the ratio between the AUPRC using the sampled data set and the full data set, and plot the median ratio across 50 samples per fraction. As nCOP uses more data, its predictive power increases and becomes similar to the one on the full data set.

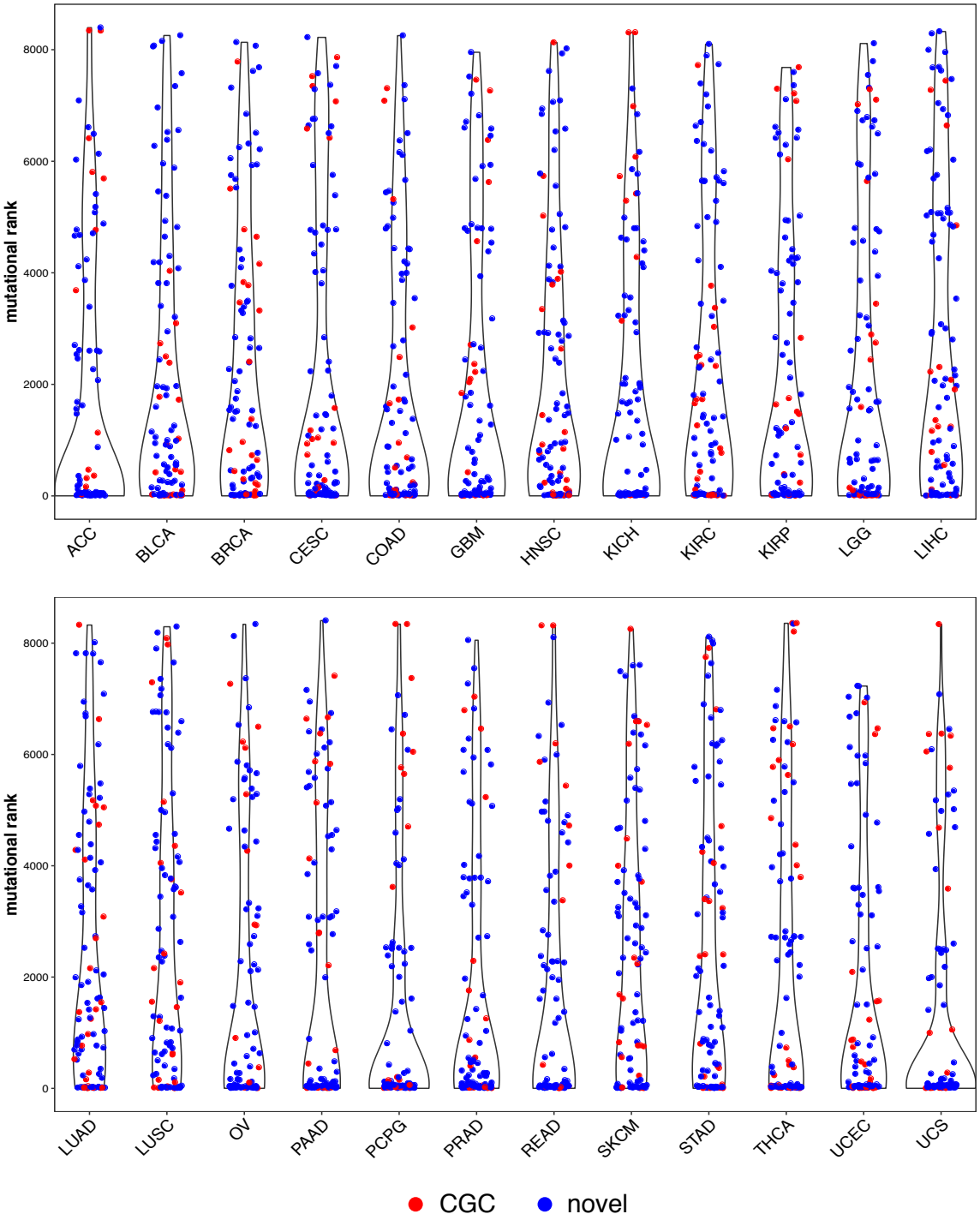

mutational rank
8000
6000
4000
2000
0
ACC
BLCA
BRCA
CESC
COAD
GBM
HNSC
KICH
KIRC
KIRP
LGG
LIHC
mutational rank
8000
6000
4000
2000
0
LUAD
LUSC
OV
PAAD
PCPG
PRAD
READ
SKCM
STAD
THCA
UCEC
UCS
CGC
novel

Figure A.6: `uKIN` **identifies rarely mutated genes.** To illustrate `uKIN`'s ability to pull genes from the long tail of the mutational distribution, we run `uKIN` with $\alpha = 0.5$ and with 20 genes as prior knowledge 100 times. For each gene, its final score is averaged across the runs. For each of the top 100 genes, we consider the rank of its mutational rate (y-axis). Known CGC genes are in red and novel predictions in blue. The top predictions consist of many heavily mutated genes (i.e., those with low ranks), but `uKIN` is also able to uncover known cancer genes with very low mutational ranks (red dots towards the top).

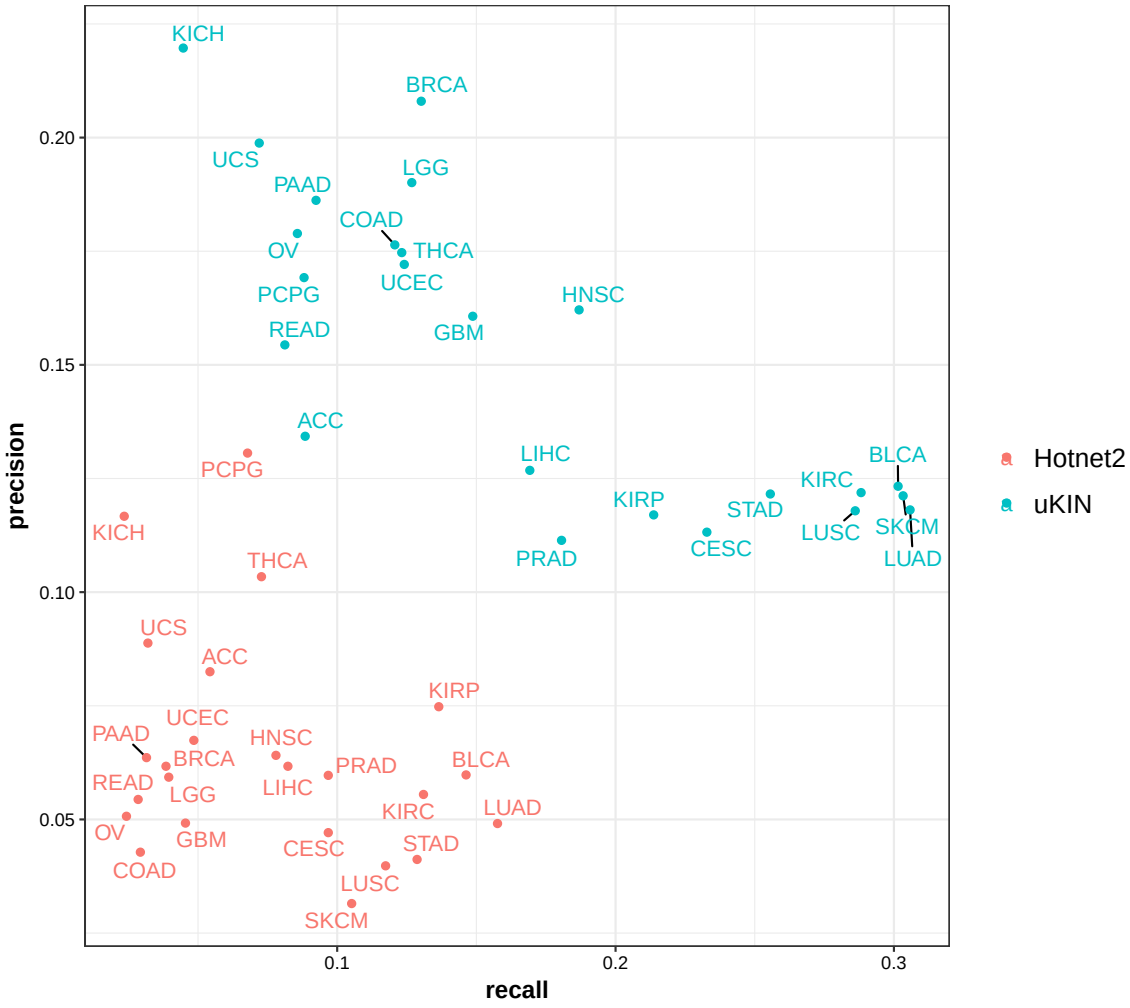


Figure A.7: **Comparison between `uKIN` and `Hotnet2`.** For each cancer type, we compute the precision and recall of the genes returned by `uKIN` and `Hotnet2`. For `uKIN`, we choose the same number of genes as highlighted by `Hotnet2`. `uKIN` clearly demonstrates both higher precision and recall than `Hotnet2` across all 24 cancer types.

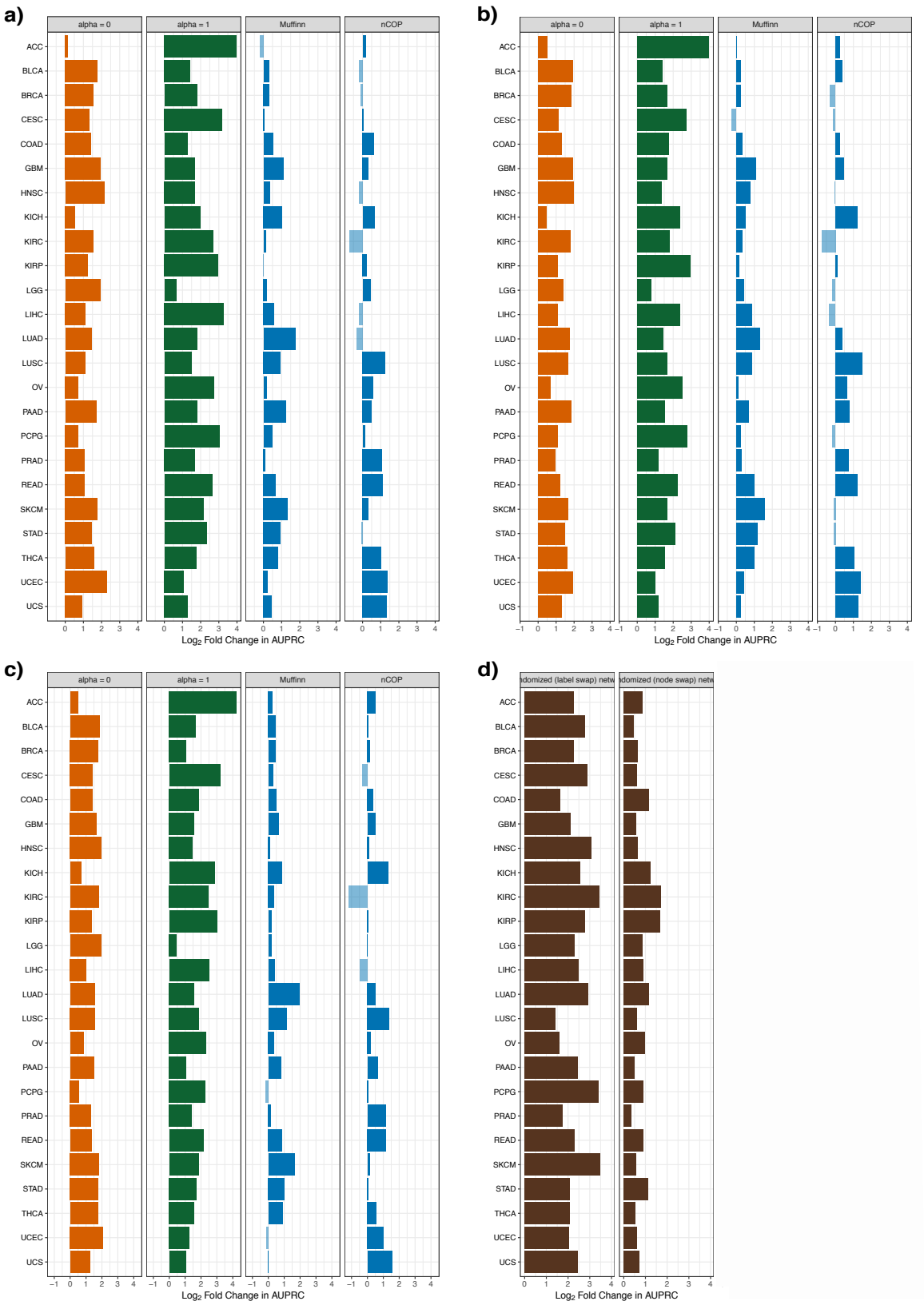

a)
b)
c)
d)
alpha = 0
alpha = 1
Muffinn
nCOP
ACC
BLCA
BRCA
CESC
COAD
GBM
HNSC
KICH
KIRC
KIRP
LGG
LIHC
LUAD
LUSC
OV
PAAD
PCPG
PRAD
READ
SKCM
STAD
THCA
UCEC
UCS
Log2 Fold Change in AUPRC

Figure A.8: **Robustness of uKIN. (a)** To make sure that uKIN is robust with respect to the set of labelled cancer genes $\mathcal{H}$, instead of randomly sampling 400 genes from the Cancer Gene Census (CGC) list, we form $\mathcal{H}$ using genes from other sources. Specifically, we aggregate the cancer genes provided by Hofree et al. in [33] (which they obtained by querying the UniprotKB database for the keyword-terms 'proto-oncogene,' 'oncogene' and 'tumoursuppressor' gene) and Vogelstein et al. [91], excluding any genes present in the set of prior knowledge $\mathcal{K}$. Log-fold AUPRCs are computed as described in the main text. The results are consistent with those shown in Figures 3.2 and 3.3 based on the CGC list and show the superior performance of uKIN as compared to the other methods in recapitulating known cancer genes. **(b)** To assess the robustness of our evaluation, we compute AUPRCs using the top 50 predicted genes. The results are consistent with those shown in Figures 3.2 and 3.3 which use the top 100 predicted genes and show the superior performance of uKIN as compared to the baselines and other methods in deriving known cancer genes. The results are also consistent when computing AUPRC's using 150 genes (data not shown). **(c)** To make sure that our method is robust with respect to the specific network utilized, we repeat our entire analysis procedure using the Biogrid network. The results are consistent with those shown in Figures 3.2 and 3.3, based on the HPRD network. **(d)** Comparison to randomized networks. In the left panel, we use a classic degree-preserving randomization (edge swapping) and in the right panel, we use a node label shuffling randomization where the network structure is maintained but gene names are swapped (thereby genes can have very different numbers of interactions in the randomizations). For each of the 24 cancers, we compute the log_2 ratio of the area under the precision recall curve using uKIN on the real network and on the randomized network and show the average over 10 different randomizations. Performance, as expected, is worse for both randomizations across all cancers. We note that significant cancer-relevant information is retained in these randomized networks. In particular, in both types of network randomizations, we maintain the relationships between genes and the samples that they are found to be somatically mutated in. Thus, some highly mutated CGC genes may still be output by uKIN when running on randomized networks.

Bibliography

[1] Sepideh Babaei, Marc Hulsman, Marcel Reinders, and Jeroen de Ridder. Detecting recurrent gene mutation in interaction network context using multi-scale graph diffusion. *BMC Bioinformatics*, 14:29, 2013.

[2] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature reviews genetics*, 12(1):56, 2011.

[3] Ali Bashashati, Gholamreza Haffari, Jiarui Ding, Gavin Ha, Kenneth Lui, Jamie Rosner, David G Huntsman, Carlos Caldas, Samuel A Aparicio, and Sohrab P Shah. Drivernet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome biology*, 13(12):1, 2012.

[4] Denis Bertrand, Kern Rei Chng, Faranak Ghazi Sherbaf, Anja Kiesel, Burton KH Chia, Yee Yen Sia, Sharon K Huang, Dave SB Hoon, Edison T Liu, Axel Hillmer, et al. Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. *Nucleic acids research*, 43(7):e44–e44, 2015.

[5] I. Bozic, T. Antal, H. Ohtsuki, H. Carter, D. Kim, S. Chen, R. Karchin, K. W. Kinzler, B. Vogelstein, and M. A. Nowak. Accumulation of driver and passenger mutations during tumor progression. *Proc Natl Acad Sci U S A*, 107(43):18545–50, 2010.

[6] Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, et al. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012, 2018.

[7] Mengfei Cao, Hao Zhang, Jisoo Park, Noah M Daniels, Mark E Crovella, Lenore J Cowen, and Benjamin Hescott. Going the distance for protein function prediction: a new distance metric for protein interaction networks. *PloS one*, 8(10):e76339, 2013.

[8] Michael P Cary, Gary D Bader, and Chris Sander. Pathway information for systems biology. *FEBS letters*, 579(8):1815–1820, 2005.

[9] J-B Cazier, SR Rao, CM McLean, AK Walker, BJ Wright, EEM Jaeger, C Kartsonaki, L Marsden, C Yau, C Camps, et al. Whole-genome sequencing of bladder cancers reveals somatic CDKN1A mutations and clinicopathological associations with mutation burden. *Nature communications*, 5, 2014.

[10] E. Cerami, E. Demir, N. Schultz, B. S. Taylor, and C. Sander. Automated network analysis identifies core pathways in glioblastoma. *PLoS ONE*, 5(2):e8918, 2010.

[11] Li Chai, Jianchang Yang, Chunhui Di, Wei Cui, Kiyoshi Kawakami, Raymond Lai, and Yupo Ma. Transcriptional activation of the sall1 by the human six1 homeodomain during kidney development. *Journal of Biological Chemistry*, 281(28):18918–18926, 2006.

[12] Ara Cho, Jung Eun Shim, Eiru Kim, Fran Supek, Ben Lehner, and Insuk Lee. Muffinn: cancer gene discovery via network analysis of somatic mutation data. *Genome Biology*, 17(1):129, 2016.

[13] SA Chowdhury and M Koyuturk. Identification of coordinately dysregulated subnetworks in complex phenotypes. In *Pac Symp Biocomput*, pages 133—-144, 2010.

[14] Shuhua Chu, Yuewang Liu, Li Zhang, Bei Liu, Li Li, and Jun-zhen Shi. Regulation of survival and chemoresistance by HSP90AA1 in ovarian cancer SKOV3 cells. *Molecular biology reports*, 40(1):1–6, 2013.

[15] G. Ciriello, E. Cerami, C. Sander, and N. Schultz. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.*, 22(2):398–406, Feb 2012.

[16] G Ciriello, M Miller, B. Aksoy, Y. Senbabaoglu, N. Schultz, and C. Sander. Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics*, 45:1127–1133, 2013.

[17] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[18] Lenore Cowen, Trey Ideker, Benjamin J Raphael, and Roded Sharan. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, 18(9):551, 2017.

[19] Nick Dand, Frauke Sprengel, Volker Ahlers, and Thomas Schlitt. BioGranatIG: a network analysis tool to suggest mechanisms of genetic heterogeneity from exome-sequencing data. *Bioinformatics*, 29(6):733–741, 2013.

[20] N. Dees, Q. Zhang, C. Kandoth, M. Wendl, W. Schierding, D. Koboldt, et al. MuSiC: Identifying mutational significance in cancer genomes. *Genome Res.*, 22:1589—-1598, 2012.

[21] Chandrakanth Reddy Edamakanti, Jeehaeh Do, Alessandro Didonna, Marco Martina, and Puneet Opal. Mutant ataxin1 disrupts cerebellar development in spinocerebellar ataxia type 1. *The Journal of clinical investigation*, 128(6):2252–2265, 2018.

[22] Ayla Ergün, Carolyn A Lawrence, Michael A Kohanski, Timothy A Brennan, and James J Collins. A network biology approach to prostate cancer. *Molecular systems biology*, 3(1), 2007.

[23] Shimon Even and R Endre Tarjan. Network flow and testing graph connectivity. *SIAM journal on computing*, 4(4):507–518, 1975.

[24] Simon A Forbes, Nidhi Bindal, Sally Bamford, Charlotte Cole, Chai Yin Kok, David Beare, Mingming Jia, Rebecca Shepherd, Kenric Leung, Andrew Menzies, et al. COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic acids research*, page gkq929, 2010.

[25] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton. A census of human cancer genes. *Nat Rev Cancer*, 4(3):177–83, 2004.

[26] TKB Gandhi, Jun Zhong, Suresh Mathivanan, L Karthick, KN Chandrika, S Sujatha Mohan, Salil Sharma, Stefan Pinkert, Shilpa Nagaraju, Balamurugan Periaswamy, et al. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature genetics*, 38(3):285, 2006.

[27] L. A. Garraway and E. S. Lander. Lessons from the cancer genome. *Cell*, 153(1):17–37, 2013.

[28] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[29] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[30] L. Hartwell, J. Hopfield, S. Leibler, and A. Murray. From molecular to modular cell biology. *Nature*, 402:C47–52, 1999.

[31] Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani. *The elements of statistical learning*, volume 2. Springer, 2009.

[32] Laura M Heiser, Nicholas J Wang, Carolyn L Talcott, Keith R Laderoute, Merrill Knapp, Yinghui Guan, Zhi Hu, Safiyyah Ziyad, Barbara L Weber, Sylvie Laquerre, et al. Integrated analysis of breast cancer cell lines reveals unique signaling pathways. *Genome biology*, 10(3):R31, 2009.

[33] Matan Hofree, Hannah Carter, Jason F Kreisberg, Sourav Bandyopadhyay, Paul S Mischel, Stephen Friend, and Trey Ideker. Challenges in identifying cancer genes by analysis of exome sequencing data. *Nature Communications*, 7, 2016.

[34] Matan Hofree, John P Shen, Hannah Carter, Andrew Gross, and Trey Ideker. Network-based stratification of tumor mutations. *Nature methods*, 2013.

[35] Borislav H Hristov and Mona Singh. Network-based coverage of mutational profiles reveals cancer genes. *Cell systems*, 5(3):221–229, 2017.

[36] Thomas J Hudson, Warwick Anderson, Axel Aretz, Anna D Barker, Cindy Bell, Rosa R Bernabé, MK Bhan, Fabien Calvo, Iiro Eerola, Daniela S Gerhard, et al. International network of cancer genome projects. *Nature*, 464(7291):993–998, 2010.

[37] Trey Ideker and Roded Sharan. Protein networks in disease. *Genome research*, 18(4):644–652, 2008.

[38] ILOG CPLEX 7.1, 2016. http://www.ilog.com/products/cplex/.

[39] Peilin Jia and Zhongming Zhao. Varwalker: personalized mutation network analysis of putative cancer genes from next-generation sequencing data. *PLoS Comput Biol*, 10(2):e1003460, 2014.

[40] Erik Johansson, Qiwei Zhai, Zhao-jun Zeng, Takeshi Yoshida, and Keiko Funa. Nuclear receptor tlx inhibits tgf-β signaling in glioblastoma. *Experimental cell research*, 343(2):118–125, 2016.

[41] Siân Jones, Xiaosong Zhang, D Williams Parsons, Jimmy Cheng-Ho Lin, Rebecca J Leary, Philipp Angenendt, Parminder Mankoo, Hannah Carter, Hirohiko Kamiyama, Antonio Jimeno, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science Signaling*, 321(5897):1801, 2008.

[42] A-Ram Kang, Hyoung-Tae An, Jesang Ko, Eui-Ju Choi, and Seongman Kang. Ataxin-1 is involved in tumorigenesis of cervical cancer cells via the egfr–ras–mapk signaling pathway. *Oncotarget*, 8(55):94606, 2017.

[43] AI Katz, DS Emmanouel, and MD Lindheimer. Thyroid hormone and the kidney. *Nephron*, 15(3-5):223–249, 1975.

[44] YA Kim, Raheleh Salari, Stefan Wuchty, and TERESA M Przytycka. Module cover–a new approach to genotypephenotype studies. *Pacyfic Synposium on Biocomputing*, 18:103–110, 2013.

[45] YA Kim, S Wuchty, and T Przytycka. Identifying causal genes and dysregulated pathways in complex diseases. *PLoS Comput Biol*, 7:e1001095, 2011.

[46] 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*, 31(12):i284–i292, 2015.

[47] Yoo-Ah Kim and Teresa M Przytycka. Bridging the gap between genotype and phenotype via network approaches. *Frontiers in genetics*, 3:227, 2013.

[48] Sebastian Köhler, Sebastian Bauer, Denise Horn, and Peter N Robinson. Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics*, 82(4):949–958, 2008.

[49] Jürgen Kohlhase, Annegret Wischermann, Herbert Reichenbach, Ursula Froster, and Wolfgang Engel. Mutations in the SALL1 putative transcription factor gene cause Townes-Brocks syndrome. *Nature genetics*, 18(1):81–83, 1998.

[50] Antonis E Koromilas and Veronika Sexl. The tumor suppressor function of STAT1 in breast cancer. *Jak-Stat*, 2(2):e23353, 2013.

[51] Michael Krauthammer, Charles A Kaufmann, T Conrad Gilliam, and Andrey Rzhetsky. Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in alzheimer's disease. *Proceedings of the National Academy of Sciences*, 101(42):15148–15153, 2004.

[52] Kasper Lage, E Olof Karlberg, Zenia M Størling, Páll I Olason, Anders G Pedersen, Olga Rigina, Anders M Hinsby, Zeynep Tümer, Flemming Pociot, Niels Tommerup, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature biotechnology*, 25(3):309, 2007.

[53] Michael S Lawrence, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, Chip Stewart, Craig H Mermel, Steven A Roberts, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, 2013.

[54] Yeo Song Lee, So-Young Kim, Su Jeong Song, Hye Kyung Hong, Yura Lee, Bo Young Oh, Woo Yong Lee, and Yong Beom Cho. Crosstalk between ccl7 and ccr3 promotes metastasis of colon cancer cells via erk-jnk signaling pathways. *Oncotarget*, 7(24):36842, 2016.

[55] M. D. Leiserson, D. Blokh, R. Sharan, and B. J. Raphael. Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput. Biol.*, 9(5):e1003054, 2013.

[56] Mark D M Leiserson, Fabio Vandin, Hsin-Ta Wu, Jason R Dobson, Jonathan V Eldridge, Jacob L Thomas, Alexandra Papoutsaki, Younhun Kim, Beifang Niu, Michael McLellan, Michael S Lawrence, Abel Gonzalez-Perez, David Tamborero, Yuwei Cheng, Gregory A Ryslik, Nuria Lopez-Bigas, Gad Getz, Li Ding, and Benjamin J Raphael. Pan-cancer network analysis identifies combinations of

rare somatic mutations across pathways and protein complexes. *Nature Genetics*, 47:106–114, 2015.

[57] H Li, N Xiao, Y Wang, R Wang, Y Chen, W Pan, D Liu, S Li, J Sun, K Zhang, et al. Smurf1 regulates lung cancer cell growth and migration through interaction with and ubiquitination of pipkiγ. *Oncogene*, 36(41):5668, 2017.

[58] Qiushi Lin, Arihiro Aihara, Waihong Chung, Yu Li, Zheping Huang, Xuesong Chen, Shaofan Weng, Rolf I Carlson, Jack R Wands, and Xiaoqun Dong. LRH1 as a driving factor in pancreatic cancer growth. *Cancer letters*, 345(1):85–90, 2014.

[59] Roger McLendon, Allan Friedman, Darrell Bigner, Erwin G Van Meir, Daniel J Brat, Gena M Mastrogianakis, Jeffrey J Olson, Tom Mikkelsen, Norman Lehman, Ken Aldape, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, 2008.

[60] Corbin E Meacham and Sean J Morrison. Tumour heterogeneity and cancer cell plasticity. *Nature*, 501(7467):328, 2013.

[61] Saket Navlakha and Carl Kingsford. The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, 26(8):1057–1063, 2010.

[62] OMIM® Online Mendelian Inheritance in Man. Mckusick-nathans institute of genetic medicine, 2000.

[63] Martin Oti and Han G Brunner. The modular nature of genetic diseases. *Clinical genetics*, 71(1):1–11, 2007.

[64] Kivilcim Ozturk, Michelle Dow, Daniel E Carlin, Rafael Bejar, and Hannah Carter. The emerging potential for network analysis to inform precision cancer medicine. *Journal of molecular biology*, 430(18):2875–2899, 2018.

[65] E. O. Paull, D. E. Carlin, M. Niepel, P. K. Sorger, D. Haussler, and J. M. Stuart. Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics*, 29(21):2757–2764, Nov 2013.

[66] Sara Pensa, Gabriella Regis, Daniela Boselli, Francesco Novelli, and Valeria Poli. STAT1 and STAT3 in tumorigenesis: two sides of the same coin? 2013.

[67] C. M. Perou, T. Sorlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, et al. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, Aug 2000.

[68] Erin D Pleasance, R Keira Cheetham, Philip J Stephens, David J McBride, Sean J Humphray, Chris D Greenman, Ignacio Varela, Meng-Lay Lin, Gonzalo R Ordóñez, Graham R Bignell, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463(7278):191, 2010.

[69] TS Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, et al. Human protein reference database——2009 update. *Nucleic acids research*, 37(suppl 1):D767–D772, 2009.

[70] Pawel F Przytycki and Mona Singh. Differential analysis between somatic mutation and germline variation profiles reveals cancer-related genes. *Genome medicine*, 9(1):79, 2017.

[71] Yan Qi, Yasir Suhail, Yu-yi Lin, Jef D Boeke, and Joel S Bader. Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. *Genome research*, pages gr–077693, 2008.

[72] Maxime WC Rousseaux, Tyler Tschumperlin, Hsiang-Chih Lu, Elizabeth P Lackey, Vitaliy V Bondar, Ying-Wooi Wan, Qiumin Tan, Carolyn J Adamski, Jillian Friedrich, Kirk Twaroski, et al. Atxn1-cic complex is the primary driver of cerebellar pathology in spinocerebellar ataxia type 1 through a gain-of-function mechanism. *Neuron*, 97(6):1235–1243, 2018.

[73] Luciana P Schwab, Danielle L Peacock, Debeshi Majumdar, Jesse F Ingels, Laura C Jensen, Keisha D Smith, Richard C Cushing, and Tiffany N Seagroves. Hypoxia-inducible factor 1α promotes primary tumor growth and tumor-initiating cell activity in breast cancer. *Breast Cancer Research*, 14(1):R6, 2012.

[74] Raunak Shrestha, Ermin Hodzic, Jake Yeung, Kendric Wang, Thomas Sauerwald, Phuong Dao, Shawn Anderson, Himisha Beltran, Mark A Rubin, Colin C Collins, et al. Hit'ndrive: multi-driver gene prioritization based on hitting time. In *International Conference on Research in Computational Molecular Biology*, pages 293–306. Springer, 2014.

[75] Tian-Ping Shuai and Xiao-Dong Hu. Connected set cover problem and its applications. In *International Conference on Algorithmic Applications in Management*, pages 243–254. Springer, 2006.

[76] V. Spirin and L. A. Mirny. Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. USA.*, 100:12123–12128, 2003.

[77] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. BioGRID: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl 1):D535–D539, 2006.

[78] M. R. Stratton, P. J. Campbell, and P. A. Futreal. The cancer genome. *Nature*, 458(7239):719–24, 2009.

[79] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based

approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[80] Yahui Sun, Chenkai Ma, and Saman Halgamuge. The node-weighted steiner tree approach to identify elements of cancer-related signaling pathways. *BMC bioinformatics*, 18(16):551, 2017.

[81] TCGA Research Network:. http://cancergenome.nih.gov/.

[82] The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487:330–337, 2012.

[83] The International Cancer Genome Consortium. International network of cancer genome projects. *Nature*, 464:993–998, 2010.

[84] Nurcan Tuncbag, Alfredo Braunstein, Andrea Pagnani, Shao-Shan Carol Huang, Jennifer Chayes, Christian Borgs, Riccardo Zecchina, and Ernest Fraenkel. Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem. *Journal of computational biology*, 20(2):124–136, 2013.

[85] Igor Ulitsky, Akshay Krishnamurthy, Richard M Karp, and Ron Shamir. DEGAS: de novo discovery of dysregulated pathways in human diseases. *PLoS one*, 5(10):e13367, 2010.

[86] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, Jan 2002.

[87] F. Vandin, E. Upfal, and B. J. Raphael. De novo discovery of mutated driver pathways in cancer. *Genome Res*, 22(2):375–85, 2012.

[88] Fabio Vandin, Eli Upfal, and Benjamin J Raphael. Algorithms for detecting significantly mutated pathways in cancer. *Journal of Computational Biology*, 18(3):507–522, 2011.

[89] Oron Vanunu, Oded Magger, Eytan Ruppin, Tomer Shlomi, and Roded Sharan. Associating genes and protein complexes with disease via network propagation. *PLoS computational biology*, 6(1):e1000641, 2010.

[90] C. J. Vaske, S. C. Benz, J. Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler, and J. M. Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, 26(12):i237–245, Jun 2010.

[91] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, Jr. Diaz, L. A., and K. W. Kinzler. Cancer genome landscapes. *Science*, 339(6127):1546–58, 2013.

[92] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, and Kenneth W Kinzler. Cancer genome landscapes. *science*, 339(6127):1546–1558, 2013.

[93] M. C. Wendl, J. W. Wallis, L. Lin, C. Kandoth, E. R. Mardis, R. K. Wilson, and L. Ding. PathScan: a tool for discerning mutational significance in groups of putative cancer genes. *Bioinformatics*, 27(12):1595–1602, Jun 2011.

[94] J Wolf, K Müller-Decker, C Flechtenmacher, F Zhang, M Shahmoradgoli, GB Mills, JD Hoheisel, and M Boettcher. An in vivo RNAi screen identifies sall1 as a tumor suppressor in human breast cancer with a role in CDH1 regulation. *Oncogene*, 33(33):4273–4278, 2014.

[95] Jianzhen Xu and Yongjin Li. Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics*, 22(22):2800–2805, 2006.

[96] Daming Yang, Tieying Hou, Lei Li, Yimin Chu, Fengli Zhou, Ying Xu, Xinyu Hou, Huan Song, Kai Zhu, Zhaoyuan Hou, et al. Smad1 promotes colorectal cancer cell migration through ajuba transactivation. *Oncotarget*, 8(66):110415, 2017.

[97] A. Youn and R. Simon. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics*, 27(2), 2011.

[98] T. I. Zack, S. E. Schumacher, S. L. Carter, A. D. Cherniack, G. Saksena, B. Tabak, et al. Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.*, 45(10):1134–1140, Oct 2013.