Title

Learning and Optimization for Personalized Cancer Treatment

Author

Sarto Basso, Rebecca

Publication Date

2020

Peer reviewed|Thesis/dissertation

Learning and Optimization for Personalized Cancer Treatment

by

Rebecca Sarto Basso

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Industrial Engineering and Operations Research

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Anil Aswani , Chair
Professor Philip Kaminsky
Professor Elizabeth Purdom

Spring 2020

Abstract

Learning and Optimization for Personalized Cancer Treatment

by

Rebecca Sarto Basso

Doctor of Philosophy in Engineering - Industrial Engineering and Operations Research

University of California, Berkeley

Professor Anil Aswani , Chair

Personalized cancer therapy is an emerging treatment strategy based on the ability to predict which patients are more likely to respond well to specific treatments. It involves the systematic use of genetic or other information about an individual patient to optimally select a course of treatment. This dissertation presents mathematical models and algorithms to predict drug response on a personalized level and to understand the causes of different responses.

Recent large cancer studies have measured somatic mutations in an unprecedented number of tumours. These large datasets finally allow the identification of cancer-related sets of genetic mutations, in particular we are interested in identifying groups of genetic mutations that are associated with positive or negative drug response. We propose a combinatorial formulation for the problem, and prove that it's computationally hard. We design two optimization algorithms to solve the problem and implement them in our tool UNCOVER. We provide analytic evidence of the effectiveness of UNCOVER in finding high-quality solutions and show experimentally that UNCOVER finds sets of alterations significantly associated with functional targets in a variety of scenarios. In particular, we show that our algorithms find sets which are better than the ones obtained by the state-of-the-art method. In addition, our algorithms are much faster than the state-of-the-art, allowing the analysis of large datasets of thousands of target profiles from cancer cell lines. While we formulate this as a more general computational problem, we use UNCOVER to analyze drug response data, identifying sets of mutations associated with drug sensitivity.

Our next contribution is a computational method, named NETPHIX (NETwork-to-PHenotype assocIation with eXclusivity), which aims to identify subnetworks of genes whose genetic alterations are associated with a continuous cancer phenotype. Leveraging the properties of cancer mutations and the interactions among genes, we formulate the problem as an integer linear program and solve it optimally to obtain a set of associated genes. Note that this algorithm solves a related but different mathematical problem than the one considered by UNCOVER since it also takes into account functional relationship among genes, which can be captured as an input network. Additionally NETPHIX, unlike UNCOVER, allows to pick up mixed sensitivity modules. Applied to a large-scale drug screening dataset, NETPHIX uncovered gene modules significantly associated with drug response, and many of the modules are also validated in another independent dataset. Utilizing interaction information, NETPHIX modules are functionally coherent, and can thus provide important insights into drug action.

We also include a case study that provides novel biological insight obtained from NETPHIX and expression correlation analysis to investigate the genetic mutations associated with mutational signatures. Specifically, our analysis aims to answer the following two complementary questions: (i) what are functional pathways whose gene expression activities correlate with the strengths of mutational signatures, and (ii) are there pathways whose genetic alterations might have led to specific mutational signatures. Analyzing a breast

cancer dataset, we identified pathways associated with mutational signatures on both expression and mutation levels, elucidating differences between related signatures .

UNCOVER and NETPHIX can be used directly for personalized cancer treatment by looking at the genomic alterations of patients and checking if these are correlated with drug sensitivity for any candidate treatment. While these methods can lead to useful insight towards personalized drug treatment our last contribution attempt to solve the more practical problem of predicting drug response based on all the information oncologists have available about the patient, this can include genetic information but is more often based on demographics, histology report, baseline labs and medical history recorded in the patient's Electronic Health Record. We propose a framework to simultaneously predict multiple outcomes for each treatment i.e. we are not only concerned with the expected survival time of the patient, other relevant factors such as quality of life and side effects are also considered as important quantifiable outcomes. These outcomes are heavily correlated to each other and one can leverage this property to improve prediction performance over predicting each outcome separately. Furthermore to the authors best knowledge there is no current published work or package that is able to handle a mix of survival, continuous and categorical outcomes. We provide a unified framework for prediction of heterogeneous outcomes in a clinical setting, leveraging an ensemble learning method known as random forests. We propose an updated node splitting rule that captures the heterogeneity of clinical outcomes.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my advisor, Professor Anil Aswani, for the helpful encouragement and guidance in completing this path and my committee members, Professor Phil Kaminski and Professor Elizabeth Purdom for their time and service. I would like to also extend my gratitude to Professor Rhonda Righter and Professor Max Shen for their advice during this journey. Besides the fantastic staff and faculty in the department, I was lucky to also have some great collegues, in particular I would like to thank Quico and Mark for their support and friendship.

Over the course of my PhD I had the wonderful opportunity to collaborate with many individuals outside of UC Berkeley, including Professor Fabio Vandin who's experience and advice were essential in helping me begin to understand the complexities of computational biology research. I had the privilege to spend 8 months at the National Institute of Health in Teresa Przytycka's lab, where I worked closely with Dr Yoo-Ah Kim and Teresa herself. The work on Chapter 3 and 4 would not have been possible without their contribution and guidance. Lastly I had the pleasure to join the Biostatistics team at Project Ronin for the last year of my PhD program. In Mike Elashoff's team I learned to look at my research from a more practical point of view and identify areas where my knowledge could make a bigger impact. Mike, Clara, Sarah and Lu were all wonderful mentors and collaborators.

I could have not completed my PhD without the support of my loving family in Italy and friends around the world. While there are too many names to mention, special gratitute goes to Alex, Nereo, Marinella and Gianni for being there for me when I most needed it, Nicola, Carlotta, Silvia e Arianna that offered a constant support from afar and Carlos, Adam, Dan, Kim, Maria, Hadley, Genya, Miranda, Tenzin, Edna, Xabi, Emily, Mostafa, Sam, Johnny, Mike, Kate, Ignasi, Kasmir and Joni for providing distraction and adventures throughout this journey. I will also always be grateful to my mum for raising me as a strong independent woman and for making me ambitious and hard-working from a very early age. This past year would not have been nearly as fun, interesting and exciting without a very special person in my life, thank you Evan for being a constant light in my life and for giving me the motivation I needed even through the rough patches.

Fianlly, I would not be where I am today without the mentorship and ecouragement of my high school math teacher Prof. Fiorenzo Cieri, he always believed in me and is responsible for my early love of mathematics, thank you for all you do.

Chapter 1

Introduction

Cancer is a devastating disease that takes the lives of hundreds of thousands of people every year. Due to disease heterogeneity, the effectiveness of any specific cancer therapy, such as chemotherapy or radiation, widely differs between individual patients. This inherent variability of cancer lends itself to the growing field of precision and personalized medicine. Personalized cancer therapy is an emerging treatment strategy based on the ability to predict which patients are more likely to respond well to specific treatments. It involves the systematic use of genetic or other information about an individual patient to optimally select a course of treatment.

Recent advances in sequencing technologies now allow the collection of genome-wide measurements in large cohorts of cancer patients. In particular, they allow the measurement of the entire complement of somatic (i.e., appearing during the lifetime of an individual) alterations in all samples from large tumor cohorts. The study of such alterations has lead to an unprecedented improvement in our understanding of how tumors arise and progress.

Several computational and statistical methods have been recently designed to identify driver alterations, associated to the disease, and to distinguish them from random, passenger alterations not related with the disease. The identification of genes associated with cancer is complicated by the extensive intertumour heterogeneity, with large (100-1000's) and different collections of alterations being present in tumors from different patients and no two tumours having the same collection of alterations. Two main reasons for such heterogeneity are that i) most mutations are passenger, random mutations, and, more importantly, ii) driver alterations target cancer pathways, groups of interacting genes that perform given functions in the cell and whose alteration is required to develop the disease. Several methods have been designed to identify cancer alterations using a-priori defined pathways or interaction information in the form of large interaction networks.

One of the main remaining challenges is the interpretation of such alterations, in particular identifying alterations with functional impact or with relevance to therapy.

Recent projects have characterized drug sensitivity in hundreds of cancer cell lines for a large number of drugs. This data, together with information about the genetic alterations in these cell lines, can be used to understand how genomic alterations impact drug sensitivity.

While the success of network based methods in other cancer domains suggests that such approaches should be also useful in the studies of drug response, most of previous approaches focused on discrete phenotypic traits – e.g., cancer vs. healthy, good or bad prognosis, or cancer subtypes – and therefore, cannot be directly applied to the analysis of continuous features such as drug sensitivity.

To address these challenges, this dissertation presents a collection of mathematical models and algorithms at the intersection of optimization and machine learning to identify the genetic causes of different drug responses in cancer patients.

The remainder of this dissertation is organized as follows. In chapter 2 we study the problem of finding groups of mutually exclusive alterations associated with a quantitative target. *Mutual exclusivity* has been employed by several recent methods that have shown its effectivenes in characterizing gene sets associated to cancer. Furthermore the availability of quantitative target profiles, for instance from drug sensitivity experiments, provides additional information that can be leveraged to improve the identification of cancer related gene sets by discovering groups with complementary functional associations with such targets. We propose a combinatorial formulation for the problem, and prove that the associated computation problem is computationally hard. We design two optimization algorithms to solve the problem and implement them in our tool UNCOVER. We provide analytic evidence of the effectiveness of UNCOVER in finding high-quality solutions and show experimentally that UNCOVER finds sets of alterations significantly associated with functional targets in a variety of scenarios. In particular, we show that our algorithms find sets which are better than the ones obtained by the state-of-the-art method. In addition, our algorithms are much faster than the state-of-the-art, allowing the analysis of large datasets of thousands of target profiles from cancer cell lines. While we formulate this as a more general computational problem, we use UNCOVER to analyze drug response data, identifying sets of mutations associated with drug sensitivity.

In Chapter 3 we discuss a computational method, named NETPHIX (NETwork-to-PHenotype assocIation with eXclusivity), which aims to identify subnetworks of genes whose genetic alterations are associated with a continuous cancer phenotype. Leveraging the properties of cancer mutations and the interactions among genes, we formulate the problem as an integer linear program and solve it optimally to obtain a set of associated genes. Note that this algorithm solves a related but different mathematical problem than the one considered in Chapter 2 since it also takes into account functional relationship among genes, which can be captured as an input network. Additionally NETPHIX, unlike UNCOVER, allows to pick up mixed sensitivity modules. Applied to a large-scale drug screening dataset, NETPHIX uncovered gene modules significantly associated with drug responses, and many of the modules are also validated in another independent dataset. Utilizing interaction information, NETPHIX modules are functionally coherent, and can thus provide important insights into drug action.

Chapter 4 presents a case study that provides novel biological insight obtained from NETPHIX and expression correlation analysis to investigate the genetic mutations associated with mutational signatures. Studies of cancer mutations have typically focused on identifying

cancer driving mutations that confer growth advantage to cancer cells. However, cancer genomes accumulate a large number of passenger somatic mutations resulting from various causes, including normal DNA damage and repair processes as well as mutations triggered by carcinogenic exposures. Different mutagenic processes often produce characteristic mutational patterns called mutational signatures. Identifying mutagenic processes underlying mutational signatures shaping a cancer genome is an important step towards understanding tumorigenesis. Specifically, our analysis aims to answer the following two complementary questions: (i) what are functional pathways whose gene expression activities correlate with the strengths of mutational signatures, and (ii) are there pathways whose genetic alterations might have led to specific mutational signatures. Analyzing a breast cancer dataset, we identified pathways associated with mutational signatures on both expression and mutation levels, elucidating differences between related signatures . This work investigated, for the first time, a network level association of mutational signatures and dysregulated pathways. The identified pathways and subnetworks provide novel insights into mutagenic processes that the cancer genomes might have undergone and important clues for developing personalized drug therapies.

While the methods outlined in Chapter 2 and 3 could be used as basis for prediction models and can lead to useful insight towards personalized drug treatment, in Chapter 5 we are concerned with the more practical problem of predicting drug response based on all the information oncologists have available about the patient, this can include genetic information but is more often based on the demographic, histology, baseline labs and medical history recorded in the patient's Electronic Health Record. We provide a framework to simultaneously predict multiple outcomes for each treatment ie we are not only concerned with the expected survival time of the patient, other relevant factors such as quality of life, side effects and tumor shrinkage are also considered as important quantifiable outcomes. Such outcomes are captured by a mix of different data types, for instance overall survival is usually measured with censored data, the presence of a particular side effect can be recorded as a binary or categorical outcome and tumor shrinkage is a continuous measure. If these outputs were unrelated then the obvious solution would be to predict each individual outcome separately with a suitable prediction method but in the clinical settings these outcomes are heavily correlated and one can leverage this property to improve prediction performance. While substantial work has been done on multiple output prediction for the regression and classification case very little work has been done on predicting multiple outcomes where the output data type is censored data or a mix of continuous and categorical variables. Furthermore to the authors best knowledge there is no current published work or package that is able to handle a mix of survival, continuous and categorical outcomes. The goal of this chapter is closing this gap by providing a unified framework for prediction of heterogeneous outcomes in a clinical setting, leveraging an ensemble learning method known as Random Forest. This method operates by constructing many decision trees and using the average prediction across all trees, we propose an updated node splitting rule that captures the heterogeneity of our outcomes.

Chapter 2

Efficient algorithms to discover alterations with complementary functional association in cancer

2.1 Introduction

Recent advances in sequencing technologies now allow to collect genome-wide measurements in large cohorts of cancer patients (e.g., [Brennan et al., 2013; Cancer Genome Atlas Network, 2015; Cancer Genome Atlas Research Network, 2013, 2014; Network et al., 2017a,b]). In particular, they allow the measurement of the entire complement of somatic (i.e., appearing during the lifetime of an individual) alterations in all samples from large tumour cohorts. The study of such alterations has lead to an unprecedented improvement in our understanding of how tumours arise and progress [Garraway and Lander, 2013]. One of the main remaining challenges is the interpretation of such alterations, in particular identifying alterations with functional impact or with relevance to therapy [McGranahan and Swanton, 2017].

Several computational and statistical methods have been recently designed to identify *driver* alterations, associated to the disease, and to distinguish them from random, *passenger* alterations not related with the disease [Raphael et al., 2014]. The identification of genes associated with cancer is complicated by the extensive *intertumour heterogeneity* [Vandin, 2017], with large (100-1000's) and different collections of alterations being present in tumours from different patients and no two tumours having the same collection of alterations [Vandin, 2017; Vogelstein et al., 2013]. Two main reasons for such heterogeneity are that i) most mutations are passenger, *random* mutations, and, more importantly, ii) driver alterations target cancer *pathways*, groups of interacting genes that perform given functions in the cell and whose alteration is required to develop the disease. Several methods have been designed to identify cancer genes using *a-priori* defined pathways [Vaske et al., 2010] or interaction information in the form of large interaction networks [Creixell et al., 2015; Leiserson et al., 2015b].

Recently several methods (see Section Related work) for the *de novo* discovery of mutated cancer pathways have leveraged the *mutual exclusivity* of alterations in cancer pathways. Mutual exclusivity of alterations, with sets of genes displaying at most one alteration for each patient, has been observed in various cancer types [Garraway and Lander, 2013; Hanahan and Weinberg, 2011; Kandoth et al., 2013; Vogelstein et al., 2013]. The mutual exclusivity property is due to the complementarity of genes in the same pathway, with alterations in different members of a pathway resulting in a similar impact at the functional level, while mutations in different members of the same pathway may not provide further selective advantage or may even lead to a disadvantage for the cell (e.g., in synthetic lethality). Even if mutual exclusivity of alterations is neither a sufficient nor a necessary property of cancer pathways, it has been successfully used to identify cancer pathways in large cancer cohorts [Ciriello et al., 2012; Kandoth et al., 2013; Leiserson et al., 2015a].

An additional source of information that can be used to identify genes with complementary functions are quantitative measures for each samples such as: functional profiles, obtained for example by genomic or chemical perturbations [Aguirre et al., 2016; Cowley et al., 2014; Tsherniak et al., 2017]; clinical data describing, obtained for example by (quantitative) indicators of response to therapy; activation measurements for genes or sets of genes, as obtained for example by single sample scores of Gene Set Enrichment Analysis [Mootha et al., 2003; Subramanian et al., 2005]. The employment of such quantitative measurements is crucial to identify meaningful complementary alterations since one can expect mutual exclusivity to reflect in functional properties (of altered samples) that are specific to the altered samples. For example, consider a scenario (Fig. 2.1) in which there are two altered molecular mechanisms: one that is altered in almost all samples and one that is altered in much fewer samples, but is related to the response to a given therapy (for example by interacting with a drug target). Methods that ignore therapy response information will report the first mechanism as significantly altered, while the second mechanisms, altered in a smaller fraction of all samples, is identified only by considering the therapy response information.

Related work

Several recent methods have used mutual exclusivity signals to identify sets of genes important for cancer [Yeang et al., 2008]. RME [Miller et al., 2011] identifies mutually exclusive sets using a score derived from information theory. Dendrix [Vandin et al., 2012b] defines a combinatorial gene set score and uses a Markov Chain Monte Carlo (MCMC) approach for identifying mutually exclusive gene sets altered in a large fraction of the patients. Multi-Dendrix [Leiserson et al., 2013] extends the score of Dendrix to multiple sets and uses an integer linear program (ILP) based algorithm to simultaneously find multiple sets with mutually exclusive alterations. CoMET [Leiserson et al., 2015a] uses a generalization of Fisher exact test to higher dimensional contingency tables to define a score to characterize mutually exclusive gene sets altered in relatively low fractions of the samples. WExT [Leiserson et al., 2015a] generalizes the test from CoMET to incorporate individual gene weights (probabilities) for each alteration in each sample. WeSME [Kim et al., 2016c] introduces a test that

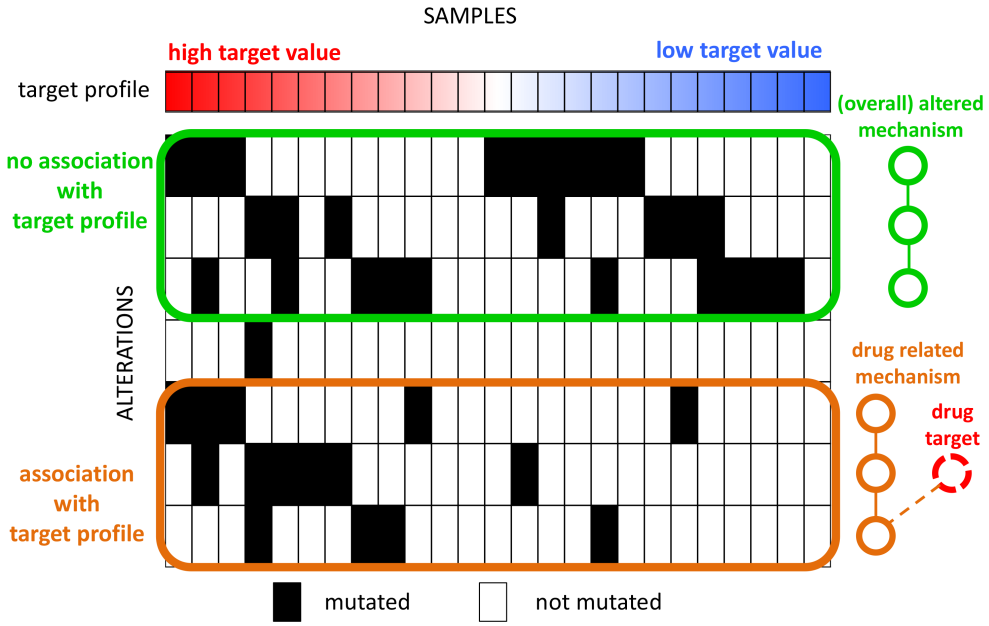


Figure 2.1. Alterations in the green set have high mutual exclusivity but no association with the target profile (e.g., a molecular mechanism commonly altered in cancer). Alterations in the orange set have lower mutual exclusivity but strong association with the target profile (e.g., genes in the same pathway of the drug target). Methods that find mutually exclusive sets of alterations without considering the target profile will identify the green set as the most important gene set.

incorporates the alteration rates of patients and genes and uses a fast permutation approach to assess the statistical significance of the sets. TiMEx [Constantinescu et al., 2015] assumes a generative model for alterations and defines a test to assess the null hypothesis that mutual exclusivity of a gene set is due to the interplay between waiting times to alterations and the time at which the tumor is sequenced. MEMo [Ciriello et al., 2012] and the method from [Babur et al., 2015] employ mutual exclusivity to find gene sets, but use an interaction network to limit the candidate gene sets. The method by [Raphael and Vandin, 2015] and PathTIMEx [Cristea et al., 2016] introduce an additional dimension to the characterization of inter-tumor heterogeneity, by reconstructing the order in which mutually exclusive gene sets are mutated. None of these methods take quantitative targets into account in the discovery of significant gene sets and sets showing high mutual exclusivity may not be associated with target profiles (Fig. 2.1).

[Kim, 2016] recently developed the repeated evaluation of variables conditional entropy and redundancy (REVEALER) method, to identify mutually exclusive sets of alterations associated with functional phenotypes. REVEALER uses as objective function (to score a set of alterations) a re-scaled mutual information metric called *information coefficient* (IC).

REVEALER employs a greedy strategy, computing at each iteration the conditional mutual information (CIC) of the target profile and each feature, conditioned on the current solution. REVEALER can be used to find sets of mutually exclusive alterations starting either from a user-defined seed for the solution or from scratch, and [Kim, 2016] shows that REVEALER finds sets of meaningful cancer-related alterations.

Contribution

In this chapter we study the problem of finding sets of alterations with complementary functional associations using alteration data and a quantitative (functional) target measure from a collection of cancer samples. Our contributions in this regard are fivefold. First, we provide a rigorous combinatorial formulation for the problem of finding groups of mutually exclusive alterations associated with a quantitative target and prove that the associated computational problem is NP-hard. Second, we develop two efficient algorithms, a greedy algorithm and an ILP-based algorithm to identify the set of k genes with the highest association with a target; our algorithms are implemented in our method fUNctional Complementarity of alteratiOns discoVERy (UNCOVER). Third, we show that our algorithms identify highly significant sets of genes in various scenarios; in particular, we compare UNCOVER with REVEALER on the same datasets used in [Kim, 2016], showing that UNCOVER identifies solutions of higher quality than REVEALER while being on average two order of magnitudes faster than REVEALER. Interestingly, the solutions obtained by UNCOVER are better than the ones obtained by REVEALER even when evaluated using the objective function (IC score) optimized by REVEALER. Fourth, we show that the efficiency of UNCOVER enables the analysis of large datasets, and we analyze a large dataset from Project Achilles, with thousands of genetic dependencies measurements and tens of thousands of alterations, and a large dataset from the Genomics of Drug Sensitivity in Cancer (GDSC) project, with hundreds of drug sensitivity measurements and tens of thousands of alterations. On such datasets UNCOVER identifies several statistically significant associations between target values and mutually exclusive alterations in genes sets, with an enrichment in well-known cancer genes and in known cancer pathways.

2.2 Materials and methods

This section describes the problem we study and the algorithms we designed to solve it, that are implemented in our tool UNCOVER. We also describe the data and computational environment for our experimental evaluation.

UNCOVER: Functional complementarity of alterations discovery

The workflow of our algorithm UNCOVER is presented in Fig. 2.2. UNCOVER takes in input information regarding 1. the alterations measured in a number of samples (e.g.,

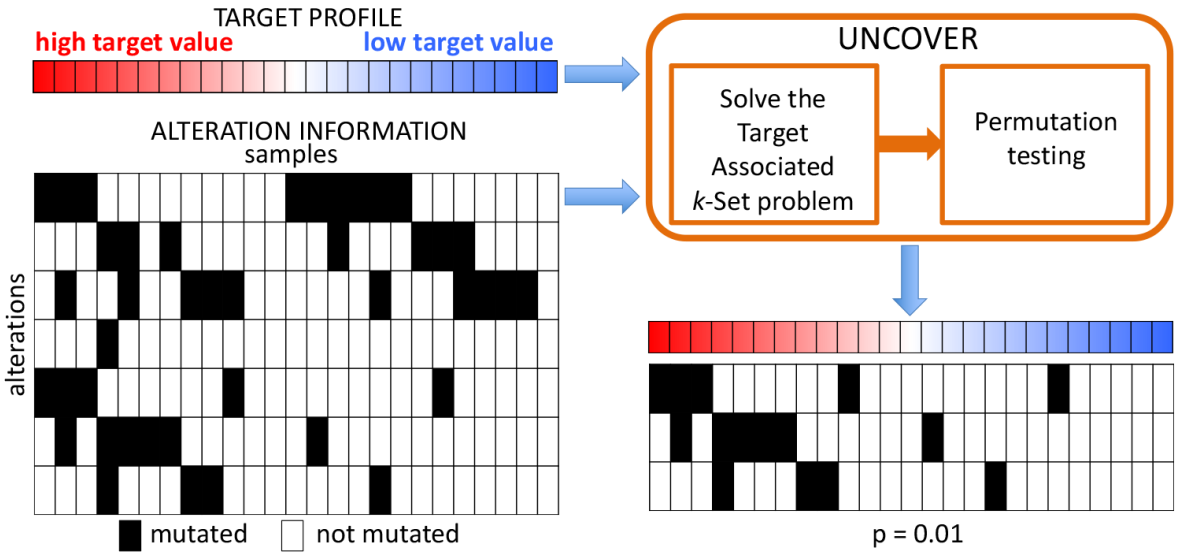


Figure 2.2.
UNCOVER takes in input the alterations information and a target profile for a set of samples, and identifies the set of complementary alterations with the highest association to the target by solving the Target Associated k-Set problem and performing a permutation test.

patients or cell lines), and 2. the value of the *target* measure for each patient. UNCOVER then identifies the set of mutually exclusive alterations with the highest association to the target, and employs a permutation test to assess the significance of the association. Details regarding the computational problem and the algorithms used by UNCOVER are described in the following sections. The implementation of UNCOVER is available at `https://github.com/VandinLab/UNCOVER`.

Computational problem

Let $J = \{j_1, \ldots, j_m\}$ be the set of samples and let $G = \{g_1, \ldots, g_n\}$ be the set of genes for which we have measured alterations in J. We are also given a *target profile*, that is for each sample $j \in J$ we have a target value $w_j \in \mathbb{R}$ which quantitatively measures a functional phenotype (e.g., pathway activation, drug response, etc.). For each sample $j \in J$ we also have information on whether each $g \in G$ is altered or not in j. Let A_g be the set of patients in which gene $g \in G$ is mutated. We say that a patient $j \in J$ is covered by gene $g \in G$ if $j \in A_g$ i.e. if gene g is mutated in sample j. Given a set of genes $S \subset G$, we say that sample $j \in J$ is covered by S if $j \in \cup_{g \in S} A_g$.

The goal is to identify a set S of at most k genes, corresponding to k subsets $S_1, S_2, \ldots S_k$ where for each subset S_i we have that $S_i \subseteq J$, such that the sum of the weights of the

elements covered by S is maximized. We also penalize overlaps between sets when an element is covered more than once by S by assigning a penalty p_j for each of the additional times j is covered by S. As penalty we use the positive average of the normalized target values if the original weight of the element was positive. If the original weight of the element was negative we assign a penalty equal to its weight.

Let $c_S(j)$ be the number of sets in $S_1, \ldots, S_k$ that cover element $j \in J$. Therefore for a set S of genes, we define its weight $W(S)$ as:

$$W(S) = \sum_{j \in \cup_{s \in S} A_s} w_j - \sum_{j \in \cup_{s \in S} A_s} (c_S(j) - 1) p_j \tag{2.1}$$

The Target Associated k-Set problem: *G*iven a set J of samples, sets $A_{g_1}, \ldots, A_{g_n}$ describing alterations of genes $G = \{g_1, \ldots, g_n\}$ in the set J, weights w_j and penalties $p_j > 0$ for each sample $j \in J$ find the S of $\leq k$ elements maximizing $W(S)$.

The following results defines the computational hardness of the problem above.

Theorem 1. *The Target Associated k-Set problem is NP-hard.*

Proof. The proof is by reduction from the Maximum Weight Submatrix Problem (MWSP) defined and proved to be NP-hard in [Vandin et al., 2012c]. The MSWP takes as input an $m \times n$ binary matrix A and an integer $k > 0$ and requires to find the $m \times k$ column sub-matrix $\hat{M}$ of A that maximizes the objective function $|\Gamma(M)| - \omega(M)$, where $\Gamma(M)$ is the set of rows with at least one 1 in columns of M and $\omega(M) = \sum_{g \in M} |\Gamma(\{g\})| - |\Gamma(M)|$.

Given an instance of Maximum Weight Submatrix Problem, we define an instance of the Target Associated k-Set as follow: the set of samples J corresponds to the rows of A, the set of genes G corresponds to the columns of A, and the set S_g of samples covered by gene $g \in G$ is the subset of the rows in which g has a 1. By setting $w_j = 1$ and $p_j = 1$ for all $j \in J$, we have that the objective function of MWSP corresponds to the weight $W(S)$ for the Target Associated k-Set therefore the optimal solution of the Target Associated k-Set corresponds to the optimal solution of MWSP. □

ILP formulation

In this subsection we provide an ILP formulation for the Target Associated k-Set problem. Let x_i be a binary variable equal to 1 if set $i \in G$ is selected and $x_i = 0$ otherwise. Let z_j be a binary variable equal to 1 if element j is covered and $z_j = 0$ otherwise. Let y_j denote the number of sets in the solution covering element j. Finally, let w_j be the weight of element j and p_j be the penalty for element j

In our ILP formulation, the following constraints need to be satisfied by a valid solution:

- the total number of sets in the solution is at most k: $\sum_i x_i \leq k$
- for each element $j \in J$ we have: $y_j = \sum_{i:j \in S_i} x_i$

- for each element $j \in J$, if j is covered by the current solution then the number of times j is covered in the solution is at least 1: $y_j \geq z_j$
- for each element $j \in J$, if j is covered by at least one element in the current solution then j is covered: $z_j \geq y_j/k$.

With the variables defined above, the score for a given solution is

$$z(q) = \max \sum_{j=1}^{m} (w_j + p_j) z_j - \sum_{j=1}^{m} p_j y_j. \tag{2.2}$$

$z(q)$ constitutes the objective function of our ILP formulation.

Greedy algorithm

Since solving ILPs can be impractical for very large datasets, we also design a k-stage greedy algorithm to solve the Target Associated k-Set problem. During each stage the algorithm picks 1 set A_i to be part of the solution; this is done by first computing the total weight of each subset which is defined as the sum of the weights of its elements $W(A_i) = \sum_{j \in A_i} w_j$. Then the algorithm finds the subset A_i of maximum positive weight and adds it to the solution. It may be that at some stage ℓ no additional set of positive weight can be selected, in this case, the solution obtained after stage $\ell - 1$ will be our output. At the end of the iteration the weight of every element j that belonged to the chosen set A_i is set to the negative of penalty p_j, in order to penalize future selections of the same elements. The greedy algorithm is described in Algorithm 1.

Input: A set of elements J (samples), a class I of subsets of J (genetic alterations) and an integer k (maximum number of alterations in the solution). Each element $j \in J$ has an associated weight w_j (target profile) and a penalty p_j.

Output: k subsets, $S_1, S_2, \ldots S_k$, where each subset selected is a member of I, such that the sum of the weights of the elements in the selected sets is maximized and the overlap between selected sets is minimized.

for $\ell \leftarrow 1$ *to* k **do**
  for $i \leftarrow 1$ *to* n **do** $W(A_i) \leftarrow \sum_{j \in A_i} w_j$;
  $S_\ell \leftarrow \arg\max_{A_i > 0}\{W(A_i)\}$;
  for $j \in S_\ell$ **do** $w_j \leftarrow -p_j$;
end
return $S_1 \ldots S_k$;

Algorithm 1: GREEDY Coverage

We note that our greedy algorithm is analogous to the greedy algorithm for the Maximum k-Coverage problem [Hochbaum and Pathria, 1998] with the difference that rather than eliminating the elements already selected we change their weight to a penalty. Also, assuming

it is acceptable to return less than k sets, we only pick a set if it has a positive weight. The running time of the algorithm is $O(kmn)$ where m = number of samples and n = number of alterations.

While the greedy algorithm may not return the optimal solution, we prove that it provides guarantees on the weight of the solution it provides.

Proposition 1. *Let S^* the optimal solution of the Target Associated k-Set and $\hat{S}$ be the solution returned by the greedy algorithm. Then $W(\hat{S}) \geq W(S^*)/k$.*

Proof. Note that the weight of subsets in the optimal solution $W(S^*)$ can only be lower compared to the original weight of the subsets, since the only weight update operation performed is to substitute positive weights of elements selected with a negative penalty.

The first subset $\hat{S}_1$ selected by our algorithm is the set of maximum weight out of all subsets and therefore $W(\hat{S}_1) \geq W(S^*_\ell)$ for $\ell = 1..k$. By the pigeonhole principle, one of these subsets in the optimal solution must cover at least $W(S^*)/k$ worth of elements. Thus $W(\hat{S}_1) \geq W(S^*)/k$. Therefore the first subset selected by the algorithm already gives a $1/k$ approximation of the optimal solution. In subsequent iterations of the algorithm we only pick additional sets if they have a positive weight so our approximation ratio can only improve. □

We also prove that the bound above is tight

Proposition 2. *There are instances of the Target Associated k-Set such that $W(\hat{S}) = W(S^*)/k$.*

The proof is in Appendix A.

While the proposition above is based on an extreme example, our experimental analysis shows that in practice the greedy algorithm works well and often identifies the optimal solution. We therefore analyze the greedy algorithm under a generative model in which there is a set H of k genes with mutually exclusive alterations associated with the target, while each genes $g \in G \setminus H$ is mutated in sample j with probability p_g independently of all other events. We also assume that the weights w_j are such that $\sum_{j \in J} w_j = 0$ and for each $j : |w_j \leq 1|$. (In practice this is achieved by normalizing the target values before running the algorithm, by subtracting to each w_j the average value $\sum_{j \in J} w_j/m$ and then dividing the result by the maximum of the absolute values of the transformed w_j's.) Note that this last condition implies that $|p_j| \leq 1$ for all j. We also assume that for genes in H the following assumptions hold:

- the set H has an association with the target, i.e.: $\mathbf{E}[W(H)] \geq \frac{m}{c'}$ for a constant $c' \geq 1$.
- each gene of H contributes to the weight of H, i.e. for each $S \subset H$ and each $g \in H \setminus S$ we have $\mathbf{E}[W(S \cup \{g\})] - \mathbf{E}[W(S)] \geq \frac{W(H)}{kc''}$ for a constant $c'' \geq 1$.

The following shows that, if enough samples from the generative model are considered, the greedy algorithm finds the set H associated with the target with high probability.

Proposition 3. *If $m \in \Omega\left(k^2 \ln(n/\delta)\right)$ samples from the generative model above are provided to the greedy algorithm, then the solution of the greedy algorithm is H with probability $\geq \delta$.*

The proof is in Appendix A.

Statistical significance

To assess the significance of the solution reported by our algorithms we use a permutation test in which the dependencies among alterations in various genes are maintained, while the association of alterations and the target is removed. In particular, a permuted dataset under the null distribution is obtained as follows: the sets $A_g, g \in \mathcal{G}$ are the same as observed in the data; the values of the target are randomly permuted across the samples.

To estimate the p-value for the solutions obtained by our methods we used the following standard procedure: 1) we run an algorithm (ILP or greedy) on the real data $\mathcal{D}$, obtaining a solution with objective function $o_{\mathcal{D}}$; 2) we generate N permuted datasets as described above; 3) we run the same algorithm on each permuted dataset; 4) the p-value is then given by $(e+1)/(N+1)$, where e is the number of permuted datasets in which our algorithm found a solution with objective function $\geq o_{\mathcal{D}}$.

Data and computational environment

Alteration Data. We downloaded data for the Cancer Cell Line Encyclopedia on 25^{th} September, 2017 from `http://www.broadinstitute.org/ccle`. In particular we used the mutation (single nucleotide variants) and copy number aberrations (CNAs) which are derived from the original Cancer Cell Line Encyclopedia (CCLE) mutations and CNA datasets. The file we used is `CCLE_MUT_CNA_AMP_DEL_0.70_2fold.MC.gct`. It consists of a binary (0/1) matrix across 1,030 samples and 48,270 gene alterations in the form of mutations, amplifications and deletions, with a 1 meaning that the alteration is present in a sample, and a 0 otherwise. For the GDSC experiments [Barretina et al., 2012a; Stransky et al., 2015], we used the alteration provided at `https://depmap.org/portal/download/all/`. We downloaded the data on July 6^{th} 2018. In particular we used mutation data from `portal-mutation-2018-06-21.csv` that includes binary entries for 18652 mutations. Additionally we considered 22746 amplifications and 22746 deletions computed from the gene copy number data in `portal-copy_number_relative-2018-06-21.csv`, with an amplification defined by a copy number above 2 and a deletion defined by a copy number below -1.

Target Data. In terms of target values we use the same datasets used by [Kim, 2016] to compare the performance of UNCOVER with REVEALER. In particular we used the following files available through the Supplementary Material of [Kim, 2016]: `CTNBB1_transcriptional_reporter.gct`, which consists of measurements of a β-catenin reporter in 81 cell lines; `NFE2L2_activation_profile.gct`, which includes NFE2L2 enrichment profiles for 182 lung

cell lines; `MEK_inhibitor_profile.gct`, which contains MEK-inhibitor PD-0325901 sensitivity profile in 493 cancer cell lines from the Broad Novartis CCLE14l; and `KRAS_essentiality_profile.gct`, which corresponds to the feature KRAS from a subset of 100 cell lines from the Achilles project dataset. In all these cases we considered the same direction of association (positive or negative) between alterations and the target as in [Kim, 2016]. Since our algorithm is very efficient we then decided to run it on a large dataset on genetic dependencies from Project Achilles (`https://portals.broadinstitute.org/achilles`), that uses genome-scale RNAi and CRISPR-Cas9 perturbations to silence or knockout individual genes. In particular, we use the whole 2.4.2 Achilles dataset (`Achilles_QC_v2.4.3.rnai.Gs.gct`) available from the project website. This dataset provides phenotype values for 5711 targets, corresponding to genes silenced by shRNA. The phenotype values correspond to ATARiS [Shao et al., 2013] gene (target) level scores, quantifying the cell viability when the target gene is silenced by shRNA. These scores are provided for 216 cell lines [Cowley et al., 2014], with 205 of them appearing in CCLE. We also used UNCOVER to analyze a large datasets from the Genomics of Drug Sensitivity in Cancer (GDSC) project (`https://www.cancerrxgene.org/`) which provides drug sensitivity data generated from high-throughput screening using fluorescence-based cell viability assays following 72 hours of drug treatment. In particular, we considered the area under the curve for each experiment as target. These scores are provided in the file `portal-GDSC_AUC-2018-06-21.txt`, available trough the DepMap portal (`https://depmap.org`) [Yang et al., 2013b] for 265 compounds and 743 cell lines, with 736 having alteration data in DepMap.

Data Preprocessing. To be consistent with REVEALER we discarded features with high or low frequency, in particular features present in less than 3 samples or more than 50 samples were excluded from our analyses. The only exception was the MEK-inhibitor example, where the high frequency threshold was changed to be 100 since the number of original samples was substantially higher (i.e., 493) for this case. From the Achilles dataset we excluded targets that have at least one missing value, in particular in this case we exclude 21 of the 5711 sets of target scores. From the GDSC dataset, since many samples have at least one target with a missing value, for every target we excluded samples with missing value for that target, that results in a different number of samples for each target. The number of samples varied between 240 and 705. We filtered alterations to only have alterations with frequencies between 0.1 and 0.25, removing in this way genes that have high alteration frequency due to genomic features not important for to the disease (e.g., gene length) [Raphael et al., 2014]. In all our experiments we normalized the target values before running the algorithm, by subtracting to each weight w_j the average value $\sum_{j \in J} w_j/m$ and dividing the result by the standard deviation of the (original) w_j's, in order to have both positive and negative target values.

Simulated Data. We investigated how effective UNCOVER is at finding selected alterations in a controlled setting, where the ground truth is known. We generated target values

according to a normal distribution with mean 0 and standard deviation 1. We tested dataset with 200, 600, 1000 and 10000 samples. For each dataset we considered the 38002 gene alterations present in CCLE and for each of them we placed alterations in the samples independently of all other events with the same frequencies as they appear in CCLE. To be consistent with the preprocessing done on real data we filtered alterations to only have alterations with frequencies between 0.1 and 0.25. We also generated a set T of 5 features to have an association with the target values. This association was varied throughout the experiments to cover different percentages of positive and negative targets. In particular we generated the selected features to cover 100%, 80%, 60%, 40% of the positive target values and 5%, 10%, 15%, 20% of the negative target values respectively, choosing random subsets of samples with positive or negative target values. We will refer to the parameter indicating the percentage of samples with positive target values selected as P and to the parameter for the percentage of samples with negative target values selected as N. We divided the number of targets covered by each of the 5 mutations equally.

Computing Environment and Solver Configuration. To describe and solve an ILP we used AMPL 20150516 and CPLEX 12.6.3. All parameters in CPLEX were left at their default values. We implemented our greedy algorithm in Python 3.6.1. We run our experiments on the same datasets considered by REVEALER [Kim, 2016] and on the Achilles project dataset on a MacBook Air with 1.7 GHz Intel Core i7 processor, 8 GB RAM and 500 GB of local storage. Experiments on simulated data were conducted on local nodes of a computing cluster. Each node had the following configuration: four 2.27 GHz CPUs, 5.71 GB RAM and 241 GB of storage. Experiments on the GDSC dataset for UNCOVER and REVEALER were conducted on an iMac with 3.4 GHz Intel Core i5 processor and 16 GB RAM. For the time comparison between UNCOVER and REVEALER we run the R code provided in [Kim, 2016] on the same machine used for UNCOVER, using R 3.5.1. All the parameters were left at their given values except for the number of permutations used to calculate their p-value, which we changed in order to compare the running time of the methods excluding the time needed to compute p-values.

2.3 Results and discussion

We tested UNCOVER on a number of cancer datasets in order to compare its results to the ones obtained without using the target, to state-of-the-art algorithms, and to test whether UNCOVER allows the analysis of large datasets. In particular, we first assessed the impact of the target values on the results of UNCOVER. We then compared UNCOVER with REVEALER using four datasets described in [Kim, 2016] as well as the GDSC project dataset described above. We then used simulated data to asses the performance of UNCOVERin finding groups of alterations associated with a target. We then performed a scalability test using a large dataset from the Achilles project and alterations from the Cancer Cell Line

Encyclopedia (CCLE). Finally, we used UNCOVER to analyze a drug sensitivity dataset from the GDSC project.

Impact of Target

We ran UNCOVER on the GDSC dataset for $k = 3$ and compared the results obtained when the target values are not considered in the analysis, obtained running UNCOVER ILP with $k = 3$ while setting the target values to 1 for all the samples considered in the analysis of a target. The latter analysis corresponds to the extraction of sets with high mutual exclusivity (e.g., by [Vandin et al., 2012c]). As expected, the solutions obtained in the two cases are very different: the solution obtained without considering the target values has one alteration in common with the solution obtained by UNCOVER using either positive or negative values of the target for only 11 targets of the 265 in the GDSC dataset, and for no target the solutions share more than 1 alteration. An example of the solutions obtained target using UNCOVER and without considering the target values are shown in Fig. 2.3. We observe that while the solutions obtained considering the target values display an association with the target profile (positive or negative), the solution obtained when the target values are not considered, while covering a large set of samples, does not display any positive or negative association with the target profile.

To asses the association between target values and alterations more consistently we calculated the point biserial coefficient for all 265 solutions. The coefficient varies between -1 and $+1$ with 0 implying no correlation. The average value obtained when ignoring the target is -0.02 with standard deviation 0.05, while the the average value obtained by UNCOVER is 0.20 with standard deviation 0.05. These results show that a mutual exclusivity analysis that disregards the values of the target does not identify sets of mutually exclusive alterations associated with target values.

In addition, the genes in solution identified by considering the drug target have a much more significant enrichment in known cancer genes, as reported in [Vogelstein et al., 2013], than the genes in solution identified disregarding the values of the target ($p = 3 \times 10^{-12}$ vs $p = 10^{-2}$).

Comparison with REVEALER

We run the greedy algorithm and the ILP from UNCOVER on the same four datasets considered by the REVEALER publication [Kim, 2016]. We used the same values of k used in [Kim, 2016], that is $k = 3$ for all the datasets, except from the KRAS dataset where $k = 4$ was used. For each dataset we recorded the solution reported by the greedy algorithm, the solution reported by the ILP, the value of the objective functions for such solutions and the running time to obtain such solutions. For ILP solutions, we also performed the permutation test (see Materials and methods) to compute a p-value using 1000 permutations. The results are reported in Table 2.1, in which we also show the results from REVEALER (without

initial seeds). Fig. 2.4 shows alteration matrices and the association with the target for the solutions identified by UNCOVER.

We can see that the greedy algorithm identifies the same solution of the ILP based algorithm in three out of four cases, and that the runtime of the ILP and the runtime of greedy algorithm are comparable and very low (< 40 seconds) in all cases. In contrast, the running time of REVEALER is much higher (> 1000 seconds in most cases). (We included all preprocessing in the reported UNCOVER runtimes in Table 2.1 to ensure a fair comparison with REVEALER; not including preprocessing our running times are all under 10 seconds.) Comparing the alteration matrices of the solutions by UNCOVER and the ones of solutions by REVEALER (Fig. S1) we note that alterations in solutions by UNCOVER tend to have higher mutual exclusivity and to be more concentrated in high weight samples than alterations in solutions by REVEALER. As expected, the value of the objective function we use is much lower for solutions from REVEALER than for solutions from our algorithm.

Table 2.1. Comparison of UNCOVER with REVEALER on REVEALER's datasets.

	NFE2L2 activation	**MEK-inhibitor**	**KRAS essentiality**	**β-catenin activation**
UNCOVER(ILP) solution	KEAP1.MC MUT ATP11B AMP SPINT4 DEL	BRAF.V600E MUT KRAS.G12 13 MUT NRAS MUT	KRAS.G12 13 MUT ZNF385B AMP ATP8A2 AMP C8orf22 AMP	APC.MC MUT CTNNB1.MC MUT SLITRK1 AMP
Objective value	46.17	108.32	28.00	22.97
IC score	0.58	0.49	0.63	0.67
p-value	0.000999	0.000999	0.025974	0.1068931
Running time (s)	14	39	9	9
UNCOVER(Greedy) solution	KEAP1.MC MUT ATP11B AMP SPINT4 DEL	BRAF MUT KRAS.G12 13 MUT NRAS MUT	KRAS.G12 13 MUT ZNF385B AMP ATP8A2 AMP C8orf22 AMP	APC.MC MUT CTNNB1.MC MUT SLITRK1 AMP
Objective value	46.17	104.29	28.00	22.97
IC score	0.58	0.5	0.63	0.67
Running time (s)	15	35	9	8
REVEALER solution	KEAP1.MC MUT LRP1B DEL OR4F13P AMP	BRAF MUT KRAS.G12 13 MUT NRAS MUT	KRAS.G12 13 MUT ZNF385B AMP LINC00340 DEL NUP153 MUT	APC.MC MUT CTNNB1.MC MUT ITGBL1 AMP
Objective value	30.35	104.29	21.86	22.12
IC score	0.54	0.5	0.6	0.7
Running time (s)	1615	4965	1243	787

For each of the four targets (NFE2L2 activation, MEK-inhibitor, KRAS essentiality, β-catenin activation) considered in [Kim, 2016], the set of alterations of cardinality k reported by our ILP algorithm, by our greedy algorithm, and by REVEALER (without seeds) is reported. k is chosen as in [Kim, 2016]. For each pair (algorithm, target) we also report the (objective) value of our objective function for the solution, the value of the IC score (that is, the objective function used in [Kim, 2016]), and the running time of the algorithm for the target. For solutions found by our ILP we also report the p-value computed by permutation test using 1000 permutations.

We then compared the solutions obtained by our algorithms with the solutions from REVEALER in terms of the *information coefficient* (IC), that is the target association score used in [Kim, 2016] as a quality of the solution. Surprisingly, in two out of four datasets UNCOVER, which does not consider the IC score, identifies solutions with IC score *higher* (by at least 5%) than the solutions reported by REVEALER. For the other two cases, in one dataset the IC score is very similar (0.50 vs 0.49) while in the other case the IC score by REVEALER is higher (0.7 vs 0.67) but the solution reported by REVEALER differs from the solution reported by UNCOVER by 1 gene only. Interestingly, the latter is the only case where the solution from the ILP has a p-value > 0.1 ($p < 0.03$ in all other cases), and therefore the solutions (by our methods and by REVEALER) for such dataset may be, at least in part, due to random fluctuations of the data.

In terms of biological significance, in most cases the solutions by UNCOVER and by REVEALER are very similar, with cancer relevant genes identified by both methods. For NFE2L2 activation, both methods identify KEAP1, a repressor of NFE2L2 activation [Solis et al., 2010]. For MEK-inhibitor, both methods find BRAF, KRAS, and NRAS, three well known oncogenic activators of the MAPK signaling pathway, which contains MEK as well. For KRAS essentiality, both methods report mutations in KRAS in the solution. For β-catenin activation, both methods identify CTNNB1 mutations and APC mutations, that is known to be associated to β-catenin activation [Minde et al., 2011]. These results show that UNCOVER identifies relevant biological solutions that are better than the ones identified by REVEALER when evaluated using our objective function *and* also when evaluated according to the objective function of REVEALER with a running time that is on average two orders of magnitude smaller than required by REVEALER. Since UNCOVER and REVEALER consider two different objective functions, it is unclear whether the improvement in running time comes from differences in implementation choices or from a inherently different computational complexity. However, since UNCOVER's objective function is easier to compute than REVEALER's objective function, we believe that the use of our objective function plays an important role in the efficiency of UNCOVER.

Table 2.2. Comparison of UNCOVER with REVEALER on GDSC dataset.

	Number of genes	Avg. effect size	Cancer genes enrichment p-value (fold enrich.)	Enriched KEGG pathways
REVEALER	570	0.11	2×10^{-4} (3)	11
UNCOVER	491	0.20	3×10^{-12} (7)	22

For each algorithm we report the distinct number of genes in its solutions, the average effect size of the algorithm's solutions, the p-value and fold enrichment for known cancer genes, and the number of KEGG pathways enriched for genes in the solutions by the algorithm.

We also compared the solutions obtained by UNCOVER and by REVEALER on the GDSC dataset. For both algorithms we obtained the solutions for $k = 3$. For UNCOVER, we considered the solution returned by the ILP. For REVEALER, we could only obtain

solutions for 246 targets, since for the other targets REVEALER terminated with an error message. Due to the high running time of REVEALER, we only obtained sets of alterations associated with positive values of the target (Table 2.2). For 33 targets the solution by UNCOVER and the solution by REVEALER share 1 alteration, while for 33 targets the solution by UNCOVER and the solution by REVEALER share 2 alterations; for no target UNCOVER and REVEALER report the same solution. This shows that the two methods identify completely different solution in most ($> 73\%$) of the cases. We compared the solutions obtained by UNCOVER and by REVEALER using the IC score considered by REVEALER but not from UNCOVER: surprisingly, in more than 50% of the cases (113 out of 208) the IC score of the solution from UNCOVER is higher than the IC of the solution from REVEALER. On the other hand, for all targets the solution by REVEALER is worst than the solution by UNCOVER when the UNCOVER objective function is considered. We also compared UNCOVER and REVEALER evaluating the association between target values and alterations in the solutions using a measure of association that is not considered by the two algorithms. In particular, we considered the point biserial correlation coefficient [**?**]. In more than 95% of the cases (199 out of 208) the point biserial correlation coefficient between the solution from UNCOVER and the target is higher than the point biserial correlation coefficient between the solution from REVEALER and the target, that is, the solution from UNCOVER has an higher association with the target than the solution from REVEALER. On average, the solution from UNCOVER has a point biserial correlation coefficient that is 37% higher than the point biserial correlation coefficient of the solution from REVEALER. Moreover, the average effect size of solutions from UNCOVER is more than 80% higher than the average effct size of solutions from REVEALER (Table 2.2). In addition, the genes in solutions from UNCOVER have a much higher enrichment ($p = 3 \times 10^{-13}$; 7-fold enrichment) for known cancer genes than solutions from REVEALER ($p = 2 \times 10^{-4}$; 3-fold enrichment). Analogously, more KEGG pathways display a significant enrichment in genes from UNCOVER solutions than from REVEALER solutions (22 vs 11). We also compared the running time of the two methods: UNCOVER required 3 hours to complete the analysis, while REVEALER required 9 days. Overall, these results show that UNCOVER obtains better results than REVEALER not only in terms of the UNCOVER objective function but also in terms of the score from REVEALER as well as in terms of a independent measure of association, while being 70 times faster than REVEALER.

Results on simulated data

For each combination we generated 10 simulated datasets as described in Materials and methods. Each dataset contains a *planted* set of 5 alterations associated with the target. We used both the greedy algorithm and the ILP from UNCOVER with $k = 5$ to attempt to find the 5 correct alteration, and evaluated our algorithms both in terms of fraction of the correct (i.e., planted) solution reported and running time.

As shown in Fig. 2.5, the greedy algorithm is faster than the ILP for all datasets, and the difference in running time increases as the number m of samples increases, with the runtime of the greedy algorithm being almost two orders of magnitude smaller than the runtime of the ILP for $m = 1000$ samples. In addition, for a fixed number of samples and alterations, the running time of the greedy algorithm is constant, that is it does not depend on the properties of the planted solution, while the running time of the ILP varies greatly depending on these parameters. For $m = 10,000$ samples the running time of the ILP becomes extremely high, so we restricted to consider only two sets of parameters ($p - n = 0.95$ and $p - n = 0.2$). In this case the ILP took between 44 minutes and 7 hours to complete, while the greedy algorithm terminates in 5 minutes.

In terms of the quality of the solutions found, as expected the ILP outperforms the greedy (Fig. 2.6) but the difference among the two tends to disappear when the number of samples is higher. In addition, since the ILP finds the optimal solution, we can see that for a limited number of samples we may not reliably identify the planted solution with 200 samples unless the planted solution appears almost only in positive targets and in almost all of them ($p - n = 0.95$), while for m=1000 we can reliably identify the planted solution using both the ILP and the greedy algorithm even when the association with the target is weaker ($p - n = 0.6$). When $m = 10,000$, both the ILP and the greedy algorithm perform well in terms of the quality of the solution: the ILP finds the correct alterations on every experiment and the greedy identifies the whole planted solution in all experiments but one for $p - n = 0.2$, for which it still reports a solution containing 4 genes in the planted solution.

These results show that for a large number of samples the greedy algorithm reliably identifies sets of alterations associated with the target, as predicted by our theoretical analysis, and is much faster than the ILP. For smaller sample size the ILP identifies better solutions than the greedy and has a reasonable running time.

Analysis of Achilles project data

The efficiency of UNCOVER renders the analysis of a large number of targets, such as the ones available through the Achilles project, possible. After preprocessing the dataset included 5690 functional phenotypes as targets, and for each of these the CCLE provides alteration information for 205 samples and 31137 alterations. In total we have therefore run 11380 instances (i.e., 5690 targets screened for positive and for negative associations) looking for both positive and negative association with target values. Since the number of samples (205) is relatively small, we have run only the ILP from UNCOVER on the whole Achilles dataset and looked for solutions with $k = 3$ genes. The runtime of UNCOVER to find both positive and negative associations, including preprocessing, is 24 hours. Based on the runtime required on the instances reported in [Kim, 2016] (see the Section Comparison with REVEALER), running REVEALER on this dataset would have required about 5 months of compute time.

To identify statistically significant associations with targets in the Achilles project dataset we used a nested permutation test. We first run the permutation test with 10 permutations on all instances (i.e., on all targets for both positive association and negative association). We

then considered all the instances with the lowest p-value (1/11) and performed a permutation test with 100 permutations only for such instances. We the iterated such procedure once more, selecting all the instances with lowest p-value (1/101) and performing a permutation test with 1000 permutations only for such instances. For positive association we found 60 solutions with p-value < 0.001, and for negative association we found 102 solutions with p-value < 0.001.

The genes in the solutions by UNCOVER with p-value 1/1001 are enriched ($p = 2 \times 10^{-12}$ by Fisher exact test; 8 fold enrichment) for well-known cancer genes. We also tested whether genes in solutions by UNCOVER (with p-value 1/1001) are enriched for interactions, by comparing the number of interactions in `iRefIndex` [Razick et al., 2008] among genes in such solution with the number of interactions in random sets of genes of the same cardinality. Genes in solutions by UNCOVER are significantly enriched in interactions ($p = 7 \times 10^{-3}$ by permutation test; 2 fold enrichment). In addition, the genes in solutions by UNCOVER are also enriched in genes in well-known pathways: 12 KEGG pathways [Kanehisa et al., 2017] have a significant (corrected $p \leq 0.05$) overlap with genes in solutions by UNCOVER and four of these (endometrial cancer, glioma, hepatocellular carcinoma, EGFR tyrosine kinase inhibitor resistance) are cancer related pathways. In addition, the *targets* (i.e., genes) with solutions of p-value 1/1001 are enriched ($p = 5 \times 10^{-4}$ by permutation test; 6 fold enrichment) for interactions in `iRefIndex` and for well-known cancer genes ($p = 2 \times 10^{-12}$ by Fisher exact test; 8 fold enrichment) as reported in [Vogelstein et al., 2013]. These results show that UNCOVER enables the identification of groups of well known cancer genes with significant associations to important targets in large datasets of functional target profiles. For example, for target (i.e., silenced gene) TSG101, related to cell growth, UNCOVER identifies the gene set shown in Fig. 2.7 as associated to reduced cell viability. ERBB2 is a well known cancer gene and CDH4 is frequently mutated in several cancer types, and both are associated to cell growth.

Analysis of GDSC project data

We use UNCOVER to analyze the GDSC project data, identifying sets of alterations associated with drug sensitivity. After preprocessing, the dataset included 64144 alterations and 265 targets, and for each of these the number of cell lines with available data varied between 240 and 705. In total we have therefore run 530 instances (i.e., 265 targets screened for positive and for negative associations) looking for both positive and negative association with target values.

We used the UNCOVER ILP for all instances to obtain solutions with $k = 3$ genes. For each solution, we use 100 permutations to compute its p-value. For positive association we found 51 solutions with p-value < 0.01, and for negative association we found 41 solutions with p-value < 0.01. We used the following procedure to focus on the most significant solutions: we run UNCOVER with $k = 4$ and computed the p-values for the solutions using 100 permutations; we then identified targets whose solution for $k = 3$ have p-value < 0.01 and are contained in the solution for the same target with $k = 4$ and have p-value $p < 0.01$ for

$k = 4$. In total, this procedure identifies 23 solutions for positive association and 22 solutions for negative associations.

The genes in the solutions identified as above are enriched ($p = 9 \times 10^{-10}$ by Fisher exact test; 20 fold enrichment) for well-known cancer genes, as reported in [Vogelstein et al., 2013]. We also tested whether these genes in solutions are enriched for interactions, by comparing the number of interactions in `iRefIndex` [Razick et al., 2008] among genes in such solution with the number of interactions in random sets of genes of the same cardinality. Genes in solutions by UNCOVER are significantly enriched in interactions ($p = 2 \times 10^{-2}$ by permutation test; 6 fold enrichment). In addition, these genes are also enriched in genes in well-known pathways: 21 KEGG pathways [Kanehisa et al., 2017] have a significant (corrected $p \leq 0.05$) overlap with genes in solutions by UNCOVER and 19 of these are cancer related pathways (e.g., ErbB signaling pathway) or related to drug resistance (e.g., EGFR tyrosine kinase inhibitor resistance).

For Palbociclib, UNCOVER identifies RB1 mutations, GRB7 amplifications, and RB1 deletions with significant association with reduced sensitivity to drug. RB1 is a well known cancer gene. The alterations are shown in Fig. 2.3a.While RB1 mutations and RB1 deletions are significantly associated when considered in isolation (the association of single alterations with drug sensitivity and the drug targets have been obtained from `https://www.cancerrxgene.org/`), GRB7 amplification is not associated with the target values when considered in isolation. GRB7 encodes a growth factor receptor-binding protein that interacts with epidermal growth factor receptor (EGFR). Both RB1 and EGFR are related to the cell cycle pathway, that is the pathway target of the compound, and the drug targets (CDK4, CDK6) as well EGFR are members of the PI3K-AKT pathway. For Sunitinib, UNCOVER identifies mutations in SETD2, ARHGAP19, and RB1, with significant association with reduced sensitivity to drug. The alterations are shown in Fig. 2.8a. RB1 is a well known cancer gene and SETD2 has tumor suppressor functionality. None of these alterations have significant association with drug sensitivity when considered in isolations. RB1 and SETD2 are involved in protein localization to chromatin, and ARHGAP19 is part of Rho mediated remodeling. For PLX-4720-2, UNCOVER identifies mutations in BRAF, CD244, and ARSB with significant association to increased sensitivity to drug. The alterations are shown in Fig. 2.8b. BRAF is a well-known cancer gene; it is the target of the compound and BRAF mutations have significant association to increased sensitivity to the compound, while the other two alterations do not. BRAF and CD244 are part of natural killer cell mediated cytotoxicity pathway, while ARSB is involved in the regulation of cell adhesion, cell migration and invasion in colonic epithelium [**?**], and is also part of metabolism related pathways. For VX-11e, UNCOVER identifies mutations in BRAF, KRAS, and NRAS, with significant association to increased sensitivity to drug. The alterations are shown in Fig. 2.8c. Only BRAF mutations have significant association with the target when considered in isolation. The pathway target for the compound is the ERK MAPK signaling pathway, to which all three alterations are related. All three genes have well identified roles in cancer. These results show that UNCOVER enables the identification of groups of relevant genes, many related to cancer, with significant associations to important targets in large datasets of drug sensitivity profiles.

Conclusion

In this work we study the problem of identifying sets of mutually exclusive alterations associated with a quantitative target profile.

We provide a combinatorial formulation for the problem, proving that the corresponding computational problem is NP-hard. We design two efficient algorithms, a greedy algorithm and an ILP-based algorithm, for the identification of sets of mutually exclusive alterations associated with a target profile. We provide a formal analysis for our greedy algorithm, proving that it returns solutions with rigorous guarantees in the worst-case as well under a reasonable generative model for the data. We implemented our algorithms in our method UNCOVER, and showed that it finds sets of alterations with a significant association with target profiles in a variety of scenarios. By comparing the results of UNCOVER with the results of REVEALER on four target profiles used in the REVEALER publication [Kim, 2016] and on a large dataset from the GDSC project, we show that UNCOVER identifies better solutions than REVEALER, even when evaluated using REVEALER objective function. Moreover, UNCOVER is much faster than REVEALER, allowing the analysis of large datasets such as the dataset from project Achilles and from the GDSC project, in which UNCOVER identifies a number of associations between functional target profiles and gene set alterations.

Our tool UNCOVER (as well as REVEALER) relies on the assumption that the mutual exclusivity among alterations is due to functional complementarity. Another explanation for mutual exclusivity is the fact that each cancer may comprise different subtypes, with different subtypes being characterized by different alterations [Leiserson et al., 2013]. UNCOVER can be used to identify sets of mutually exclusive alterations associated with a specific subtype whenever the subtype information is available, by assigning high weight to samples of the subtype of interest and low weight to samples of the other subtypes. In addition, while we consider a penalty based on mutual exclusivity, other types of penalties may be used to identify sets of alterations associated with a target profile. The study of the theoretical properties of the problem and the analysis of the results with different penalties are interesting directions of future research.

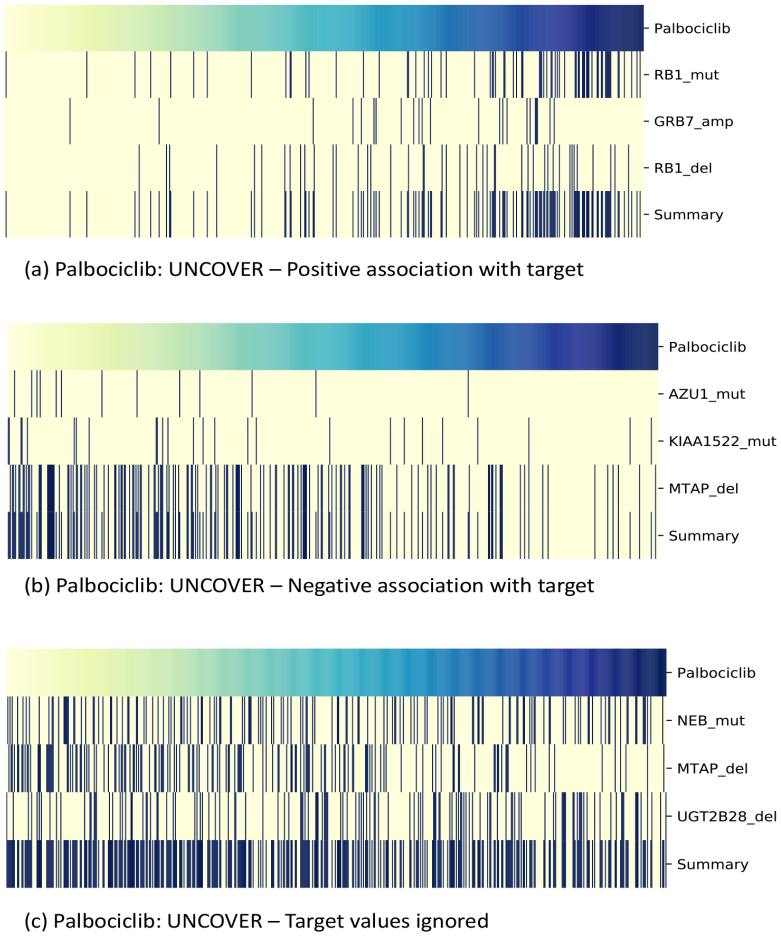


Figure 2.3.
UNCOVER results for target Palbociclib considering or ignoring target values. (a) Solution found by UNCOVER looking for an association with samples with high target values. (b) Solution found by UNCOVER looking for association with low target values. (c) Solution found by UNCOVER when the target values are ignored. Each panel shows the value of the target (top row) for various samples (columns), with yellow being negative and blue being positive values. For each gene in the solution, alterations in each sample are shown in dark blue, while samples not altered are in yellow. The last row shows the alteration profile of the entire solution.

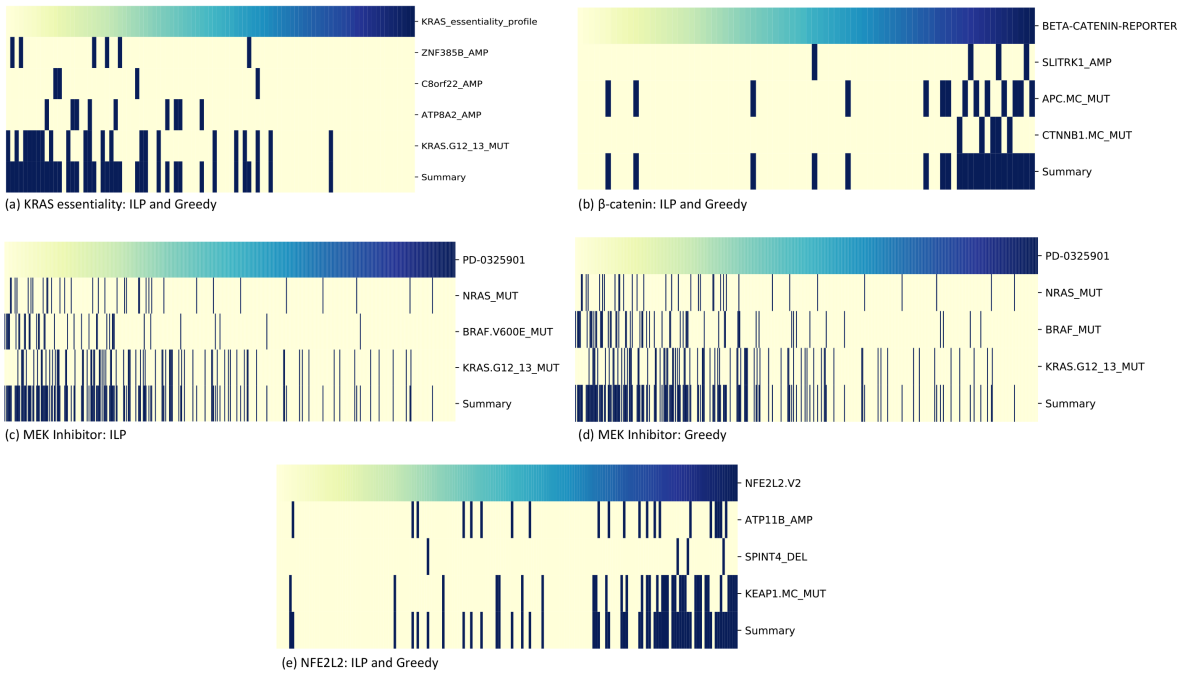


Figure 2.4.

(a) Solution found by ILP and greedy for KRAS essentiality target. (b) Solution found by ILP and greedy for β-catenin activation target. (c) Solution found by ILP for MEK inhibitor target. (d) Solution found by greedy for MEK inhibitor target. (e) Solution found by ILP and greedy for NFE2L2 activation target. Each panel shows the value of the target (top row) for various samples (columns), with yellow being negative and blue being positive values. For each gene in the solution, alterations in each sample are shown in dark blue, while samples not altered are in yellow. The last row shows the alteration profile of the entire solution.

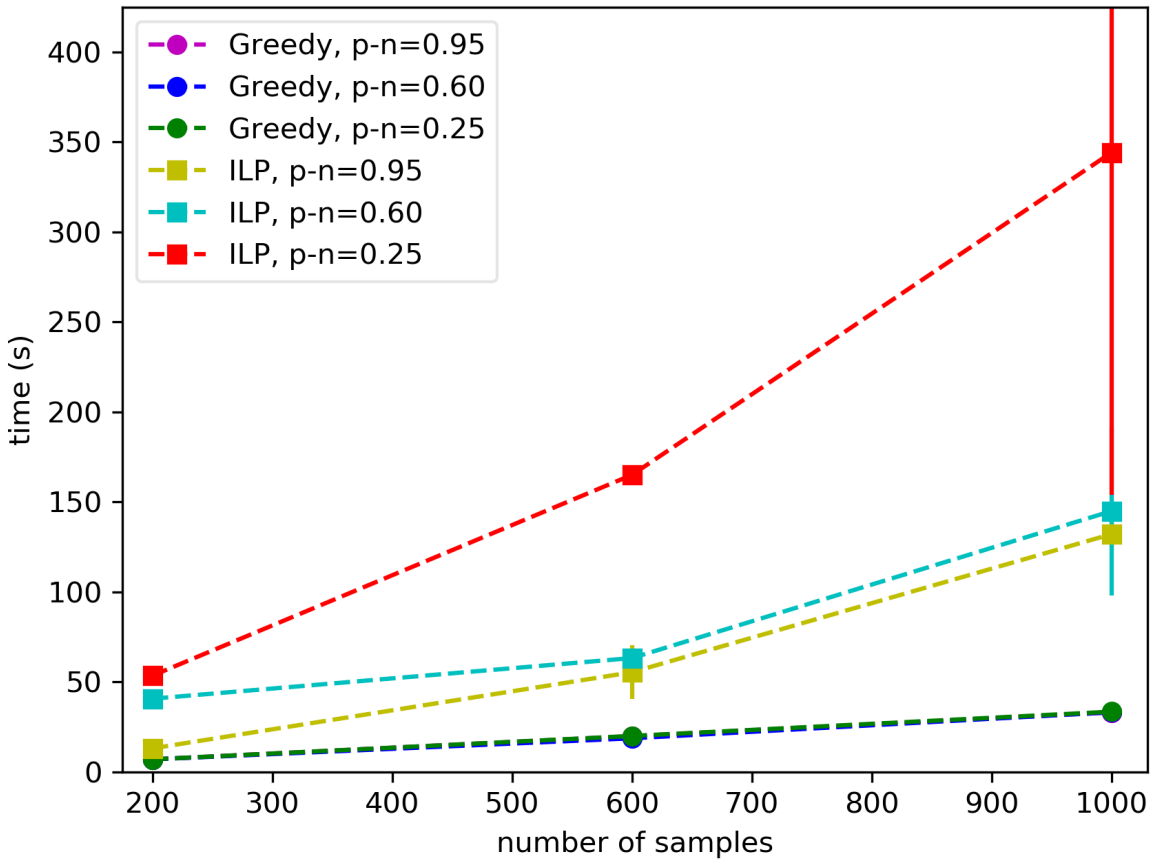


Figure 2.5. The running time (expectation and standard deviation) of the greedy algorithm and of the ILP approach are shown for different number of samples and the difference $p - n$ between the fraction p of samples with positive target and the fraction n of samples with negative target covered by the the correct solution.

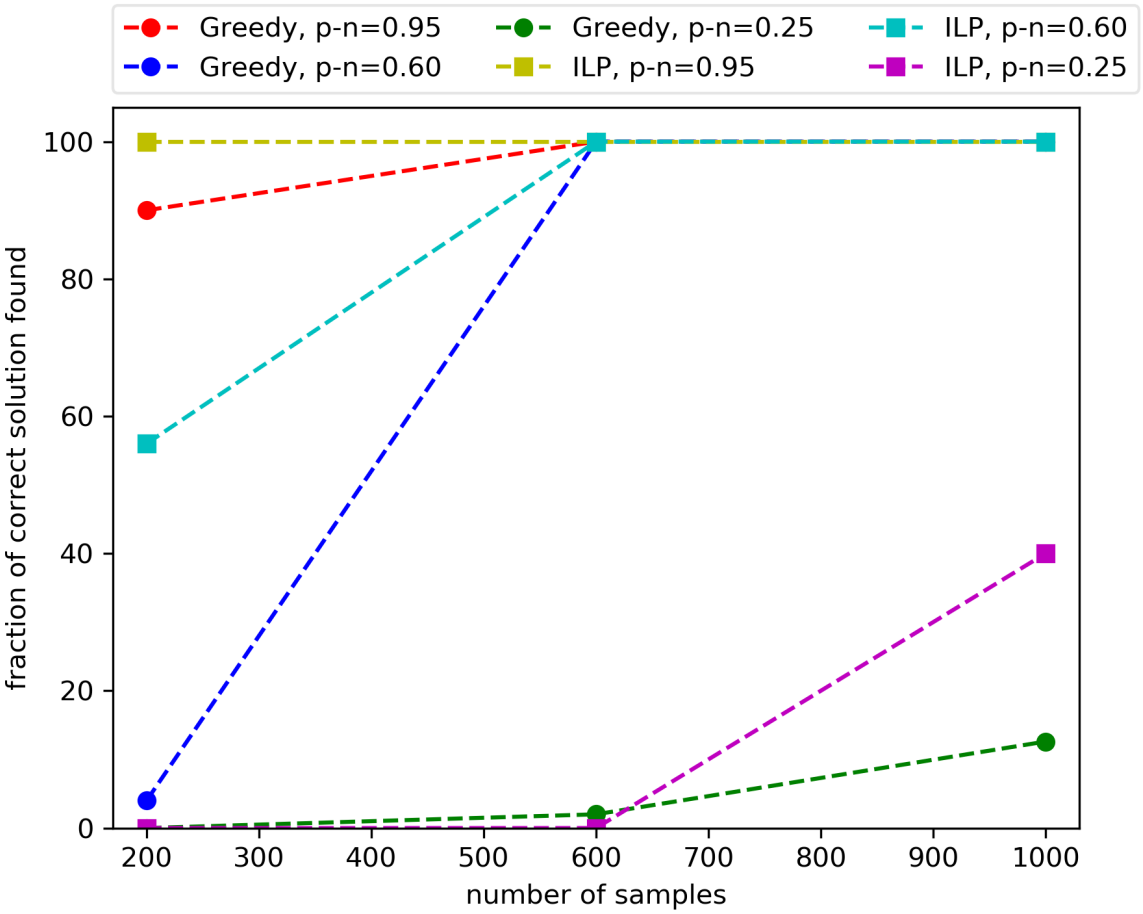


Figure 2.6. The fraction of genes in the planted (i.e., correct) solution found by the greedy algorithm and by the ILP approach are shown for different number of samples and the difference $p - n$ between the fraction p of samples with positive target and the fraction n of samples with negative target covered by the the correct solution.

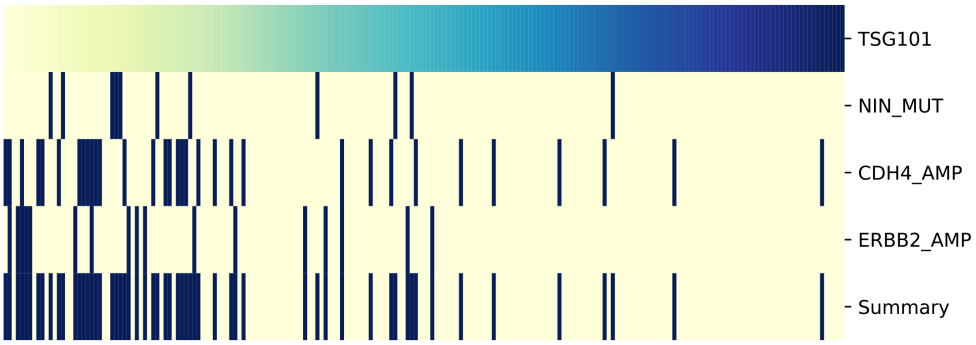
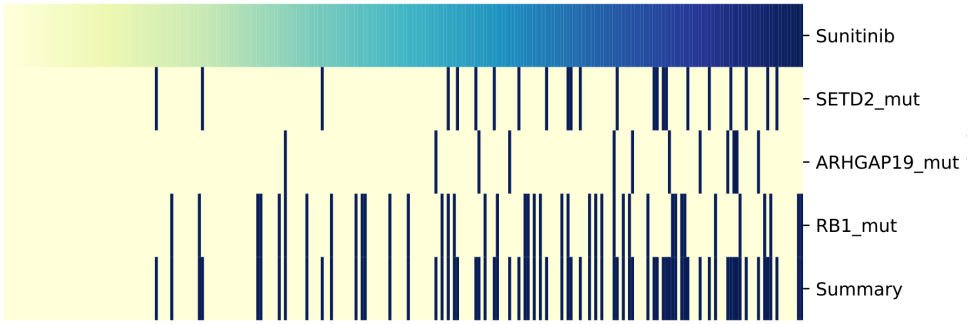
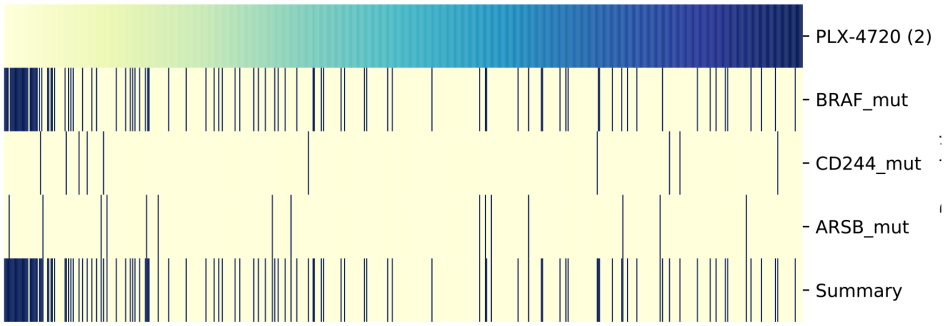


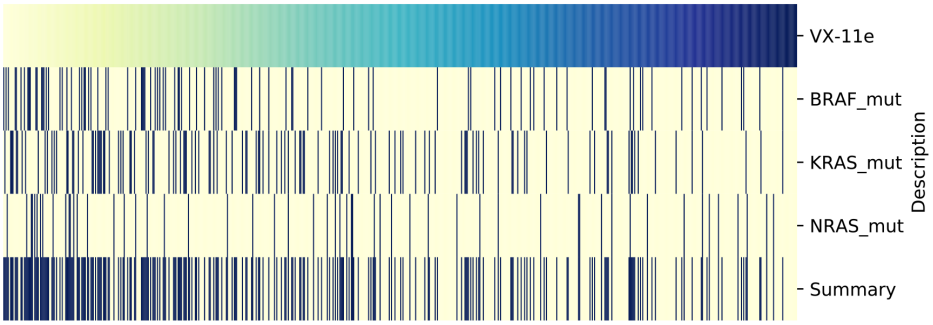
Figure 2.7. The alteration matrix of genes in the solution identified by UNCOVER as associated to reduced cell viability is reported. The value of the target (top row) for various samples (columns) is shown, with yellow being negative and blue being positive values. For each gene in the solution, alterations in each sample are shown in dark blue, while samples not altered are in yellow. The last row shows the alteration profile of the entire solution.



(a) Sunitib: UNCOVER – association with reduced sensitivity



(b) PLX-4720(2): UNCOVER – association with increased sensitivity



(c) VX-11e: UNCOVER – association with increased sensitivity

Figure 2.8. The alteration matrix of genes in some solutions identified by UNCOVER as associated to drug sensitivity for different targets. (a) Solution for reduced sensitivity to Sunitinib. (b) Solution for increased sensitivity to PLX-4720-2. (c) Solution for increased sensitivity to VX-11e. Each panel shows the value of the target (top row) for various samples (columns), with yellow being negative and blue being positive values. For each gene in the solution, alterations in each sample are shown in dark blue, while samples not altered are in yellow. The last row shows the alteration profile of the entire solution.

Chapter 3

Identifying Drug Sensitivity Subnetworks with NETPHIX

3.1 Introduction

Genetic alterations in cancer are associated with diverse phenotypic properties such as drug response or patient survival. However, the identification of mutations causing specific phenotypes and the interpretation of the phenotype-genotype relationships remain challenging due to a large number of passenger mutations and cancer heterogeneity. Indeed, the relationships between genotype and phenotype in most tumors are complex and different mutations in functionally related genes can lead to the same phenotype. The pathway-centric view of cancer [Garraway and Lander, 2013; Hanahan and Weinberg, 2011; Vogelstein et al., 2013] suggests that cancer phenotype should be considered from the context of dysregulated pathways rather than from the perspective of mutations in individual genes. Such a pathway-centric view significantly advanced the understanding of the mechanisms of tumorigenesis. Many computational methods to identify cancer driving mutations have been developed based on pathway-centric approaches [Chuang et al., 2007; Dao et al., 2017b; Hofree et al., 2013; Kim et al., 2013, 2016b; Vandin et al., 2012a]. Network based approaches have been further applied to find subnetworks associated with various disease phenotypes [Carter et al., 2013; Gilman et al., 2012; Hofree et al., 2013; Kim et al., 2016b; Zhang et al., 2018]. Those methods have been developed aiming to find genes whose mutations are associated specifically with given phenotypes rather than finding general cancer drivers.

Recent projects have characterized drug sensitivity in hundreds of cancer cell lines for a large number of drugs [Barretina et al., 2012b; Yang et al., 2013a]. This data, together with information about the genetic alterations in these cell lines, can be used to understand how genomic alterations impact drug sensitivity. While the success of network based methods in other cancer domains suggests that such approaches should be also useful in the studies of drug response, most of previous approaches focused on discrete phenotypic traits – e.g., cancer vs. healthy, good or bad prognosis, or cancer subtypes – and therefore, cannot be

directly applied to the analysis of continuous features such as drug sensitivity.

To address these challenges, we introduce a computational tool named NETPHIX (NETwork-to-PHenotype assocIation with eXclusivity). With the goal of identifying mutated subnetworks that are associated with a continuous phenotype, our algorithm builds on combinatorial optimization techniques involving *connected set cover*. The objective function of NETPHIX allows to find subnetworks with a mix of genes associated with increased or decreased sensitivity simultaneously, considering interactions between resistance and sensitivity alterations. In addition, we designed the objective function to allow for a preferential selection of mutually exclusive genes in the solution. Based on the observation that cancer related mutations tend to be mutually exclusive [Ciriello et al., 2013; Constantinescu et al., 2015; Kim et al., 2015, 2016c; Leiserson et al., 2015a; Vandin et al., 2012a], we hypothesized that mutual exclusivity may also be useful for the identification of gene modules associated with drug response. This approach together with selecting significantly associated modules allows to leave out passenger mutations from the sensitivity networks.

Several algorithms have been previously developed for the identification of mutations associated with drug response [Kim, 2016; Knijnenburg et al., 2016; Sarto Basso et al., 2019] but without considering functional relationships among genes. For example, REVEALER used a re-scaled mutual information metric to iteratively identify a set of genes associated with the phenotype [Kim, 2016]. UNCOVER employs an integer linear programming formulation based on the set cover problem, by designing the objective function to maximize the association with the phenotype and preferentially select mutually exclusive gene sets [Sarto Basso et al., 2019]. While UNCOVER uses a similar objective function as NETPHIX, it does not allow to pick up mixed sensitivity modules nor utilize network information. LOBICO [Iorio et al., 2016; Knijnenburg et al., 2016] is designed to identify a set of genes whose alterations are associated with differences in drug response. The algorithm is formulated as an integer linear program, based on logic models of binary input features that explain a continuous phenotype variable. However, none of the algorithms mentioned above utilize network interaction information. Since perturbations in functionally related genes are likely to lead to similar phenotypes, functional interaction information can be helpful for the identification of phenotype associated genes.

There have been related studies combining GWAS analysis with network constraints [Azencott et al., 2013; Jia et al., 2011; Li and Li, 2008; Liu et al., 2017]. While these methods generally perform well at broadly pointing to disease related genes, they do not consider complex properties of cancer mutations such as the aforementioned mutual exclusivity of cancer drivers, and are not designed to zoom in on subnetworks that are specific enough to help understand drug action. As discussed later, the genomic landscape related to drug response can be complex and mutations in different genes in the same pathway can affect the response differently. Pharmaceutical drugs are often developed to target specific genes, and the response depends on the function and the mutation status of the gene as well as other genes in the same pathway.

We evaluated NETPHIX and other related methods using simulations and showed that NETPHIX outperforms all the competing methods. Applying NETPHIX to a large scale drug

response data (Genomics of Drug Sensitivity in Cancer(GDSC)), we identified sensitivity-associated subnetworks for many of the drugs. We were also able to validate many of the identified modules with an independent drug screening datase (The Cancer Therapeutics Response Portal (CTRP)). These subnetworks provided important insights into drug action. Effective computational methods to discover these associations will improve our understanding of the molecular mechanism of drug sensitivity, help to identify potential drug combinations, and have a profound impact on genome-driven, personalized drug therapy. NETPHIX is available at `https://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/index.cgi#netphix`

3.2 Methods and Results

NETPHIX overview

Given gene alteration information and drug sensitivity profiles (or any cancer-related, continuous phenotype) for the same set of cancer samples (or cell lines), NETPHIX aims to identify genetic alterations underlying the phenotype of interest (Fig. 3.1a). Based on the assumption that genes whose mutations lead to the same phenotype must be functionally related, NETPHIX also utilizes functional interaction information among genes as an input, and enforces the identified genes to be highly *connected* in the network. The problem is formulated as an integer linear program (ILP) based on a connected set cover approach. Below we briefly describe the connected set cover based algorithm. For the formal definition of the problem and the detailed ILP formulation, see Section B.1 and B.2, respectively. By running ILP instances with different parameters and obtaining optimal solutions using CPLEX [cpl], we generate candidate modules whose aggregate alterations may be associated with a given drug response. Statistical significance of candidate modules is then assessed with a permutation test, and a set of maximal modules are selected as final sensitivity modules.

Connected set cover. For the first step, to obtain candidate subneworks, we design our algorithm based on connected set cover to maximize the total association with drug response (Fig. 3.1b). Connected set cover approaches have been used successfully for the identification of cancer driving mutations, to overcome the challenges posed by the heterogeneity of cancer mutations and to help uncover relevant genes with rare or medium mutation frequencies [Chowdhury and Koyuturk, 2010; Hristov and Singh, 2017; Kim et al., 2011, 2013, 2015; Sarto Basso et al., 2019; Ulitsky et al., 2010]. Since we are interested in finding alterations associated with drug response, we seek a connected set of genes that maximizes the total weight where the weights are assigned based on drug sensitivity profile.

It has been observed that different patients can harbor mutations in different but functionally related genes. This heterogeneity may arise when mutations in two different genes lead to dysregulation of the same cancer pathway and the role of the two genes for cancer progression is redundant. Building on this observation, NETPHIX identifies a connected set of genes S that maximizes the sum of phenotypic weights of the patients who have alterations

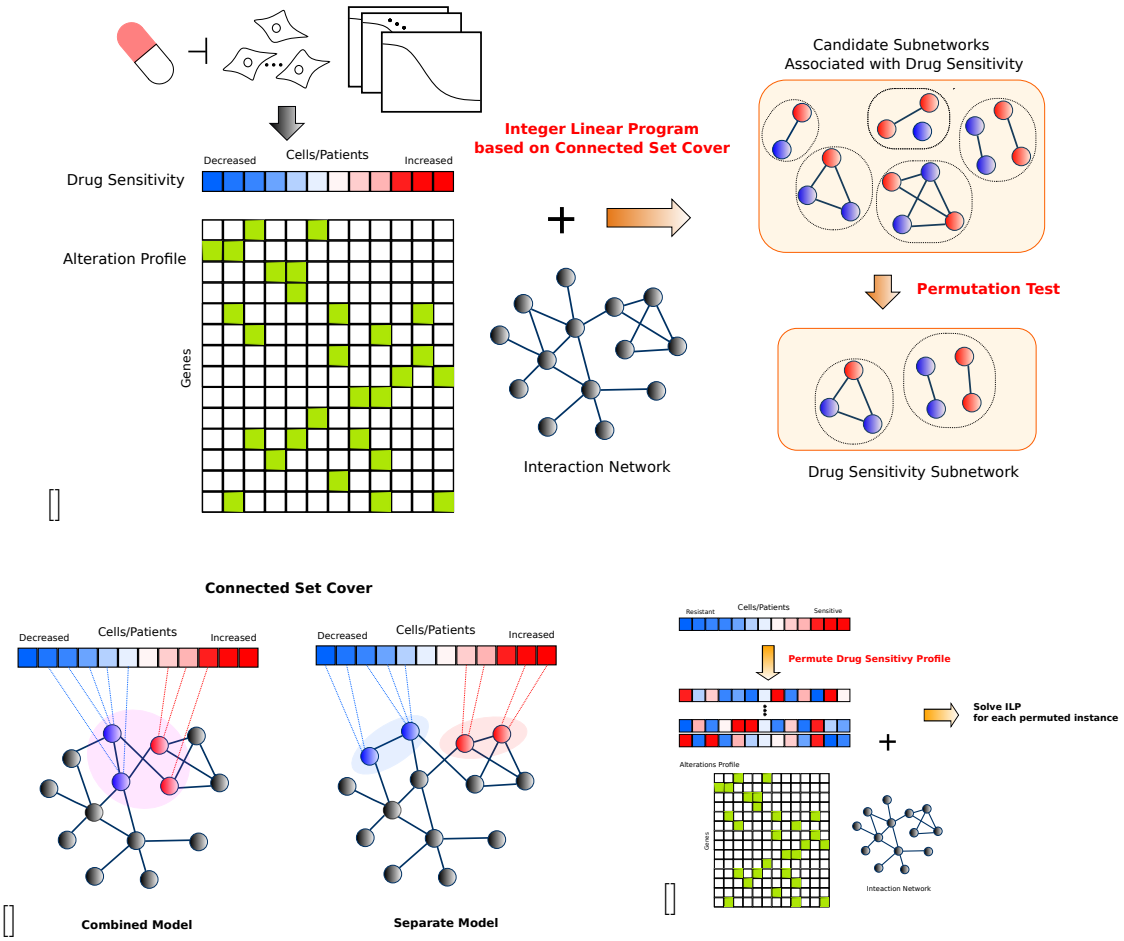


Figure 3.1. (a) Overview: NETPHIX takes a drug sensitivity profile (continuous phenotype values) and alterations status for the same set of samples as input. In addition, the algorithm utilizes interaction information among genes. Using a connected set cover based ILP algorithm, we first generate a set of candidate modules. The final set of modules are selected based on a permutation test and include only maximal modules among statistically significant optimal solutions. Genes associated with decreased and increased sensitivity are marked as blue and red respectively. (b) NETPHIX finds a connected set of genes for which corresponding mutations are associated with phenotype values (red colors in the drug response profile indicate increased sensitivity values and blue colors are for decreased sensitivity values). We considered the *combined* model in which all the selected genes are connected regardless of the directions of association and the *separate* model in which two subnetworks are identified for increased and decreased sensitivity module separately (c) The significance of identified modules are assessed using a permutation test in which drug sensitivity profiles are permuted.

in any genes in S. In addition, NETPHIX allows to add penalties in the objective function for overlapping mutations to enforce mutual exclusivity in the selected modules. While set cover approach can capture the heterogeneity of mutations, the penalties can be used to further reinforce the property of mutual exclusivity.

There may be genes associated with either directions of drug response - genes whose alterations correlate with increased sensitivity to the drug (decreased cell survival) and genes whose alterations correlate with decreased sensitivity to the drug (increased cell survival). Our algorithm is designed to find a module that includes both types of genes simultaneously. When modeling the connectivity constraints, we considered two different models – the combined model and the separate model (Fig. 3.1b). In the "combined" model, we identify one connected subnetwork which include all genes associated with either direction, assuming that alterations in different genes belonging to the same functional module can lead to different directions of drug response. In the "separate" model, we seek the solutions with two connected subnetworks, one for increased sensitivity and one for decreased sensitivity separately. This model can capture the case when two different functional modules affect drug response in different ways. As shown later in the results, the combined model finds more associated modules in general than the separate model, although there are a few drugs whose responses are associated more significantly with two separate subnetworks.

Selecting final modules. We run multiple ILP instances for different module sizes and connectivity options to obtain candidate modules. Once we obtain the optimal gene module for each parameter combination, we assess the significance of the identified module by performing a permutation test (Fig. 3.1c). Note that our algorithm is designed to identify the modules associated specifically with a given phenotype (e.g., drug sensitivity to each drug) rather than finding general cancer drivers, and therefore, a permutation test was performed by permuting the drug sensitivity profile so that the significance of the association is assessed in comparison with randomly generated phenotype. Among all significantly associated subnetworks, we obtain the final drug sensitivity modules by selecting maximal modules to remove redundancy. See Section B.3 for the details of the permutation test and maximal module selection.

Method evaluation on simulated data

We generated a set of simulated instances where we planted phenotype associated modules with varying parameters onto the background of real cancer cell mutation data (Section B.4). We then compared the performance of NETPHIX and three related methods – LOBICO, UNCOVER and SigMOD. LOBICO is a logic model based algorithm, developed to identify a set of genes whose alterations are related to drug response [Knijnenburg et al., 2016]. UNCOVER [Sarto Basso et al., 2019] was proposed as a method to identify a set of phenotype-associated genes by taking a set cover approach similar to ours. While both LOBICO and UNCOVER find an optimal solution using an integer linear program but neither algorithms utilizes interaction network information. SigMOD is a recently proposed module identification algorithm combining GWAS and network based approach, and was found to outperform other

related methods [Liu et al., 2017]. SigMOD requires individual association scores of genes to a phenotype as an input, for which we used the p-value of association of each gene to a given phenotype by performing t-tests on the coefficients of univariate linear regression.

For the evaluation purposes, we considered simple cases where alterations are associated with only one direction (either increased or decreased). Even though NETPHIX can identify subnetworks with mixed associations simultaneously, UNCOVER considers each direction separately. In addition, the logic models of LOBICO become more complicated and difficult to solve when both sides of associations are present. We planted modules of size 3, 4, and 5 and evaluated the accuracy of the three methods in identifying the planted modules (Fig. 3.2ab). For all algorithms except SigMOD, we ran the algorithm for different k's (k is a parameter for the size of a module searched by the algorithms), while SigMOD automatically adjusted its parameters to find the best module. Also for all ILP based algorithm, we limit the running time up to 24 hours, meaning the algorithms will stop and output the current solution (which may be suboptimal) when the time limit reaches.

As shown in Fig. 3.2a, only NETPHIX shows very low rate of false positives, i.e., falsely identified genes for all module sizes including when bigger modules than Afatinib and selumetinib planted are searched. NETPHIX usually does not extend the best module with spurious genes even if we searched for modules bigger than planted while UNCOVER and LOBICO tends to add more genes when increasing k. SigMOD identified a large number of false positives along with the planted modules (approx. 100-180 genes) that are not associated with phenotypes. In general, the algorithms uncovered the planted modules in most of instances (Fig. 3.2b) as long as the size of serached modules are at least as big as the planted module sizes. However, LOBICO solutions miss true positives more often possibly due to the fact that the algorithm returns suboptimal solutions after the time limit reaches. Note that the LOBICO results previously reported have been obtained with pre-selected genes/pathways (consisting of 1,000 elements) while in this simulation we used genome-wide alteration profiles without prefiltering. Both UNCOVER and NETPHIX found optimal solutions well within the time limit (See Fig. B.2).

Analysis of drug screening dataset

Using NETPHIX, we analyzed a large scale drug screening dataset (GDSC) for which genomic alteration profiles for hundreds of cell lines and drug sensitivity data for 265 drugs are available (Section B.4). Application of NETPHIX to the dataset resulted in identifying a total of 476 modules for 194 drugs (for the remaining drugs no modules with significant association were identified. Since there can be multiple functional modules affecting drug efficacy, our method allows to identify multiple associated modules for a specific drug. Out of 476 identified modules, 258 modules are one connected modules based on the combined model (for 163 drugs) and 218 modules consist of two connected components based on the separate model (for 136 drugs). See Appendix B.3 for detailed description on how the final modules are selected.

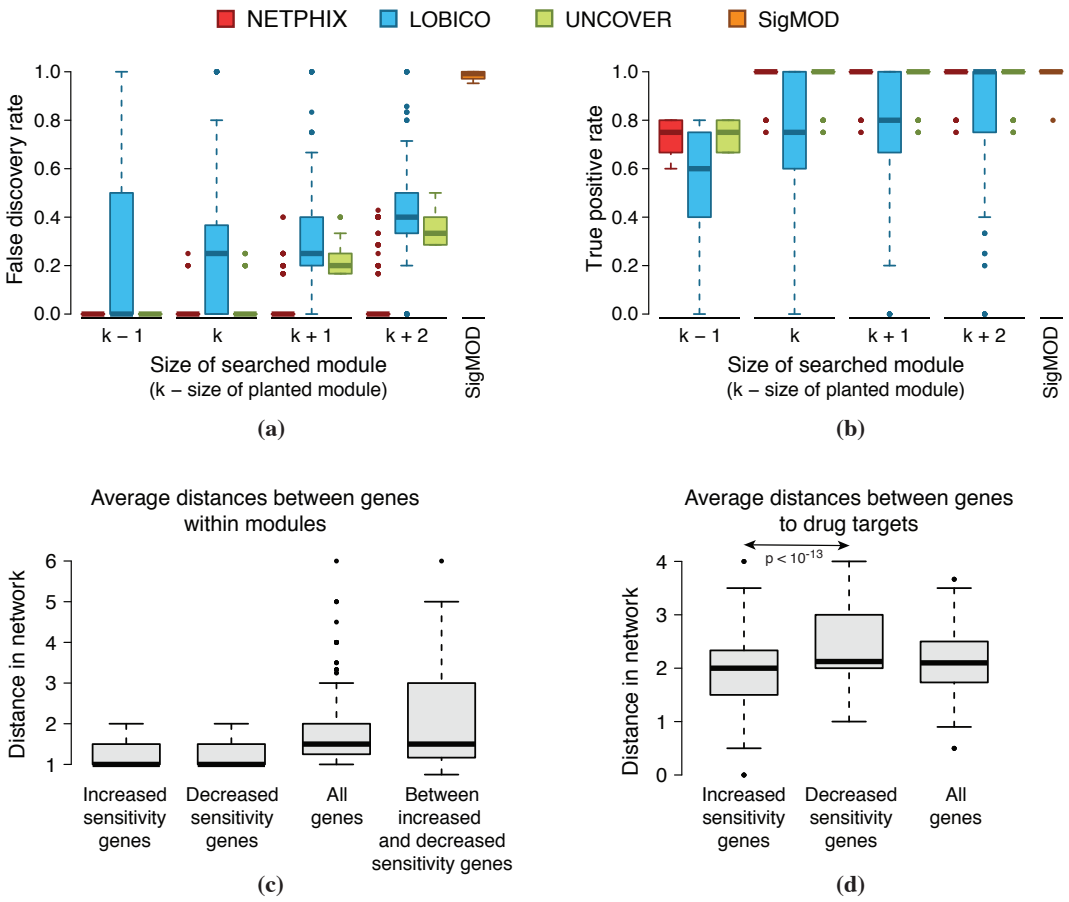


Figure 3.2. (a-b) Method comparison on simulated data. (a) False discovery rate and (b) True positive rate for the modules identified by NETPHIX (red), LOBICO (blue), UNCOVER (green), and SigMOD (orange). **(c-d) Properties of NETPHIX modules.** (c) Average distances between genes in the selected modules. Distances for genes associated with increased sensitivity (Inc), decreased sensitivity (Dec), all genes in the selected modules (All), and distances between increased sensitivity genes and decreased sensitivity (Between) are shown (d) Average distances between selected genes and the corresponding drug targets. Distances for genes associated with increased sensitivity (Inc), decreased sensitivity (Dec), all genes in the selected modules (All) are shown

NETPHIX modules are functionally related and close to the drug target. NETPHIX is designed to choose modules in which genes are highly connected. As we will discuss in Section 3.2, we observed that utilizing network information and finding connected gene sets indeed helps identify drug sensitivity modules (Fig. 3.3(f)). Since NETPHIX modules are connected, the genes in the selected modules are naturally close in terms of distance in the network (the mean of avg. distances = 1.78, Fig. 3.2c). The proximity in the network means that the genes are more likely to be functionally related. In addition, the genes with the same direction of association (decreased or increased) tend to be closer to each other than the genes associated in opposite direction although the two groups of genes are still close in the network (the mean of avg. distances between the two groups = 1.96).

To examine the relationship between the sensitivity modules and the targets for the corresponding drugs, we also computed the distances between the drug targets and the selected genes for each drug (Fig. 3.2d). We found that the genes in drug sensitivity modules are located near drug targets in the network (the mean of avg. distances = 2.12). Interestingly, we observed that the genes associated with increased sensitivity are closer to drug targets than the genes associated with decreased sensitivity (the mean of avg. distance 1.97 vs. 2.37, $p < 10^{-13}$, t-test), indicating that having perturbations in genes closer to drug targets could potentially improve the efficacy of the drugs.

NETPHIX identified biomarkers for drugs. Many of the modules identified by NETPHIX provide interesting insights related to drug action. In particular, we analyzed the response to drugs targeting the RAS/MAPK pathway (Fig. 3.3e). This pathway regulates growth, proliferation and apoptosis and is often dysregulated in various cancers. Among the most common mutations of this pathway are mutations of BRAF/KRAS/NRAS. Interestingly, NETPHIX identified modules including those genes (mostly BRAF, KRAS and sometimes NRAS) as associated with increased sensitivity to all MEK inhibitors (Selumetinib, Trametinib, CI-1040, PD0325901, and Refametinib) and an ERK inhibitor (VX-11e). All these six drugs act by blocking MEK1/MEK2 or ERK genes that are immediately downstream of BRAF/KRAS/NRAS and the increased sensitivity attributed to the alterations in this subnetwork is consistent with the action of these drugs. Modules associated with decreased sensitivity to the drugs are more diverse but NETPHIX frequently selected the module of genes ERBB2 (amplification), MYC and RB1 (mutations) or the module with TP53 mutations. All the genes in the modules are related to the MAPK/ERK signaling pathway. The mutation status of BRAF and KRAS, the core members of the pathway, were previously identified as predictors of MEK inhibitors although KRAS mutations can affect drug responses differently depending on the mutation types [Li et al., 2018; Nakayama et al., 2008; Sun et al., 2014]. ERBB2 is a receptor protein that signals through this pathway, while MYC, RB1 and TP53 are downstream of the MAPK/ERK signaling pathway. RB1 was found to be associated to the resistance to MEK inhibitors [Gong et al., 2019] and MYC degradation by inhibition of MEK leads to an increase in both ERBB2 and ERBB3 mRNA expression, causing intrinsic drug resistance [Sun et al., 2014]. TP53 mutations are associated

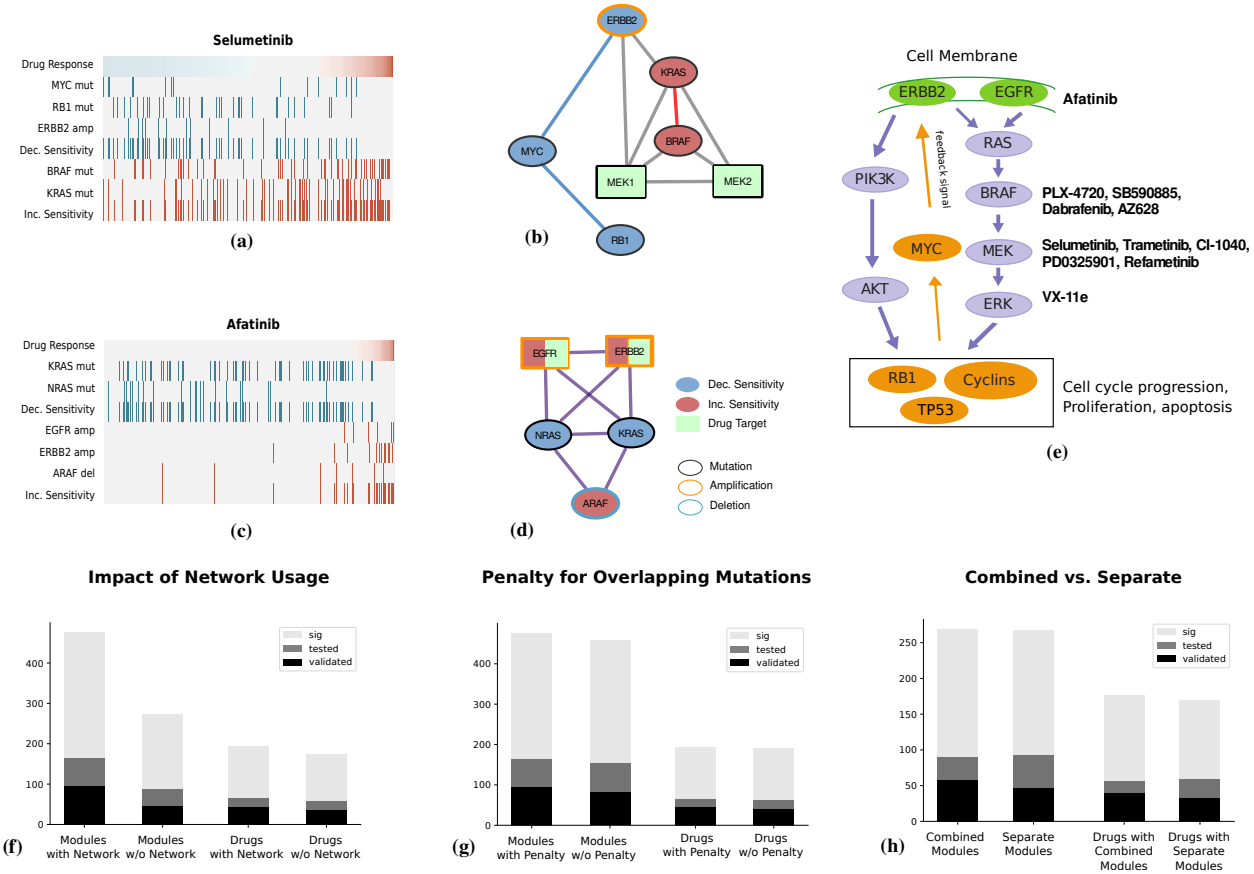


Figure 3.3. (a-d) Sensitivity networks identified by NETPHIX. Alternation profile and connectivity between the drug target and the genes for Selumetinib (a-b) and Afatinib (c-d). In the alteration profile, the panel shows the values of the phenotype (i.e., drug response, top row) for all samples (columns), with blue being decreased sensitivity values and red being increased sensitivity values. For each gene, alteration status in each sample are shown in red/blue (with the summary covers for decreased and increased sensitivity separately), while samples not altered are shown in grey. The module for Selumetinib is identified based on the separate connectivity model and the module for Afatinib is selected based on the combined connectivity model. **(e) Schematic diagram of MAPK/ERK and AKT signaling pathways** with drugs and their drug targets annotated. **(f-h) Comparison between different options.** (f) Comparison between runs with and without network information. The number of significant modules/drugs (Sig), the number of tested modules/drugs with CTRP (Tested) and the number of confirmed modules/drugs with CTRP (Validated) are shown. Drugs are counted when there is at least one associated modules that are significant, tested or validated, respectively. (f) Comparison between runs with and without penalty.(h) Comparison between the combined and separate connectivity models.

with multiple drug resistance [Keshelava et al., 2001; Najem et al., 2017]. These findings indicate that the alterations in different components of the same pathway can contribute to drug sensitivity in different ways.

In contrast to the response to MEK1/2 and ERK2 inhibitors, the drugs directly targeting BRAF are associated with more heterogeneous modules. While all BRAF inhibitors (except HG6-64-1) commonly exhibit increased sensitivity in BRAF mutant cell lines, KRAS mutations help the action of Type II BRAF inhibitors (AZ628) but develop resistance to Type I inhibitors such as Dabrafenib, PLX-4720, and SB590885, which is consistent with the previous findings [Sanchez-Laorden et al., 2014]. This suggests that patient specific mutational profiles can provide important clues in predicting drug response.

NETPHIX modules suggest candidates for drug combination therapy. We hypothesize that pairs of drugs can potentially be candidates for combination therapy if they are associated with similar modules but the genes in the modules are associated with drug responses in opposite directions. By analyzing the identified modules for pairs of drugs with such property, we identified 169 drug pairs. Although the systematic validation could not be performed due to the lack of validation dataset, we found the evidences in literature for the efficacy of several drug combinations. For example, Afatinib, a pan-ErbB inhibitor, have associated modules of KRAS, NRAS (mutations) for decreased sensitivity and EGFR, ERBB2 (amplification) and ARAF (deletion) for increased sensitivity. This suggests that it might be beneficial to use Afatinib in combination with MEK 1/2 and ERK2 targeting drugs discussed above (Fig. 3.3a-d). Indeed, studies showed that Afatinib and Selumetinib work synergistically [Sun et al., 2014] and clinical trials for combination therapy are currently underway [afa]. In addition, the efficacy of PD0325901 and Afatinib combination is also reported [Lin et al., 2019]. Both Selumetinib and PD0325901 have associated modules similar to Afatinib but in opposite direction.

Another example is the combination of Lapatinib and Vorinostat (Fig. B.1cd). Lapatinib is a drug that inhibits EGFR/ERBB2 and Vorinostat is a histone deacetylase (HDAC) inhibitor. Vorinostat has been shown to improve how well Lapatinib kills cancer cells in clinical trials [lap] and Lapatinib in general enhances the antitumor activity of the histone deacetylase inhibitor synergistically [LaBonte et al., 2011].

Validation of identified sensitivity modules with independent datasets

To examine if the alterations in the modules identified by NETPHIX indeed lead to different responses for the corresponding drugs, we performed the validation of the identified subnetworks with CTRP (Cancer Therapeutics Response Portal), an independent drug screening data. Among 194 drugs for which NETPHIX identified at least one module, 65 drugs have the drug response profiles reported in CTRP dataset, and many of the drugs have multiple associated modules, resulting in 164 modules available for validation. We divided the cell

lines in the dataset depending on the alteration status in the identified genes for each drug and tested if the cell survival rates differ between different groups. We found that 44 out of 65 drugs (68%) have at least one module having a statistically significant difference ($p < 0.05$ (ANOVA), FDR $< 9\%$ (BH)), and 94 modules in total (out of 164, 57 %) were confirmed in the validation.

Impact of NETPHIX design choices on the results

While the benefits of using gene interactions for the identification of cancer drivers are generally accepted, it was not investigated before how much gain the network usage provides in the context of drug responses. Here we use the same validation set to compare the performance of our algorithm with and without network information. Similarly, we also investigated the impact of using penalty that reinforces mutual exclusivity and different connectivity models by measuring the difference in performance in terms of the number of validated instances in the independent dataset.

Network information helps identify drug sensitivity modules. NETPHIX find a set of connected genes that are associated with drug response. To investigate the effects of using network information on the performance, we ran the algorithm without connectivity constraints and compared the solutions with NETPHIX modules.

As shown in Fig. 3.3f, NETPHIX finds more significant modules when network information is used (476 modules compared to 274 modules without network). The number of drugs with at least one associated modules is also larger (194 drugs with network vs. 175 drugs without network). We further compared the effects of penalties in terms of the number of drugs/modules confirmed in CTRP dataset. Without network, only 46 out of 88 tested modules (52%) were confirmed whereas 94 out of 164 tested modules (57%) were confirmed when connectivity was imposed. In addition, only 36 drugs had at least one confirmed modules (62% of tested drugs) without network compared with 44 drugs with network (68% of tested drugs). Overall the results show that network information significantly helps find more modules that are statistically significant, and the identified modules has a higher percentage of true positives as demonstrated in the validation using the independent dataset.

Imposing penalty on overlapping mutations may improve drug sensitivity module identification. Our objective function in ILP has an option to include penalties to further penalize overlapping mutations and enforce mutually exclusivity between mutations. We investigated the effects of using penalties on the performance by running the algorithm *without penalties* in the objective function and compared with our results obtained when penalties are used (Fig. 3.3g). We observe that NETPHIX finds a lesser number of significant modules when penalties are not included (459 modules compared to 476 modules in the original solution) although the numbers of drugs with at least one associated modules are similar (191 drugs without penalty vs. 194 drugs with penalty). In terms of the number

of drugs/modules confirmed in CTRP dataset, only 81 modules out of 152 tested modules (54%) were confirmed without penalty whereas 94 modules out of 164 tested modules (56%) were confirmed with penalty. In addition, 39 drugs had at least one confirmed modules (63% of tested drugs) without penalty compared with 44 drugs with penalty (68% of tested drugs).

Combined vs. separate connectivity model. Next, we compared the performance of two connectivity models – the combined model where all selected genes are connected and the separate model in which two modules are identified for increased and decreased sensitivity separately (Fig. 3.3h). It was not immediately obvious which approach would be more successful; On one hand, one can hypothesize that genes responsible for either type of response are functionally related. However, it was also possible that mutations associated with drug resistance may occur in a separate module such as genes related to drug metabolism.

We found that a similar number of modules were identified with the combined model and the separate model (269 vs. 268 modules) for 177 vs. 170 drugs. However, the combined connectivity model has a higher percentage of confirmed modules/drugs. Among 90 and 93 modules tested with CTRP dataset for the combined and separate model respectively, 58 (64%) and 47 modules (51%) were confirmed. In terms of the number of drugs with at least one confirmed modules, the combined model has 39 out of 57 drugs (68%) confirmed whereas the separate model has 32 drugs confirmed out of 59 (54%).

Among the drugs with only the combined model being confirmed in CTRP dataset is IGF-1R inhibitors, BMS-754807. The module has KRAS mutations associated with increased sensitivity and PTEN (mutations and deletions) with decreased sensitivity (Fig. B.1a). Both genes have been previously shown as biomarkers for IGF-1R inhibitors [Dillon and Miller, 2014; King et al., 2014; Molina-Arcas et al., 2013; Patel et al., 2014] although there are conflicting reports depending on cancer types and molecular status of other genes [Huang et al., 2015; King et al., 2014]. The module that NETPHIX identified for BMS-754807 was confirmed in the CTRP dataset (Fig. B.1a), showing a significant different in the cell survival rates of the two groups ($p < 0.00035$, ANOVA).

We hypothesize that there may be cases where activation and repression genes are less tightly related and the separate connectivity model may capture the two modules simultaneously. Several modules identified with the separate connectivity model include ADAM22 amplification. ADAM22 has been previously found as a prognostic and therapeutic drug target in endocrine-resistant breast cancer [Bolger and Young, 2013; McCartan et al., 2012]. On the other hand, for some drugs such as Cytarabine, both the combined and the separate model identified statistically significant modules, which are confirmed in CTRP dataset. In particular, the module associated with Cytarabine in the separate model (Fig. B.1b) includes UGT2B17 amplification and CYP2E1 deletion associated with decreased sensitivity. Both enzymes are hypothesized to be important players in the metabolism of common drugs [Garcia-Suastegui et al., 2017; Guillemette et al., 2014].

In summary, while the combined connectivity model works better in terms of the number of significant and validated modules, there are several drugs for which the separate model

provides different insights on drug action.

3.3 Discussion

We developed a new computational method, NETPHIX (NETwork-to-PHenotpe assocIation with eXclusivity), for the identification of mutated subnetworks that are associated with a continuous phenotype. Using simulations and analyzing a large scale drug screening dataset, we showed that NETPHIX can uncover the subnetworks associated with response to cancer drugs with high precision. We found many statistically significant and biologically relevant modules associated with drug response, including MAPK/ERK signaling related modules associated with opposite response to drugs targeting RAF, MEK and ERK genes. The genetic alteration status in many of identified modules indeed make differences in cell survival rates, as validated with an independent dataset. Overall, the modules identified by NETPHIX are in good correspondence with the action of the respective drugs, suggesting that NETPHIX can correctly identify relevant modules and the modules can thus be used to predict potential patient-specific drug combinations and to provide guidance to personalized treatment.

We demonstrate that the preferential selection of mutually exclusive genes was important for a better performance of the method. Interestingly, although one might assume that genes affecting drug resistance are not necessarily functionally related to the genes increasing drug sensitivity, we found that the combined connectivity model outperforms the separate connectivity model, indicating that the two groups of genes in fact might be related.

The applicability of NETPHIX can go far beyond the drug response discussed in this paper, to any continuous cancer phenotypes. We expect that NETPHIX will find broad applications in many types of network-to-phenotype association studies.

Chapter 4

Network-based approaches to elucidate differences between mutational signatures in Breast Cancer

4.1 Introduction

Cancer genomes accumulate a high number of mutations, only a small portion of which are cancer driving mutations. Most of such mutations are passenger somatic mutations, not directly contributing to cancer development. Analyses of large scale cancer genome data revealed that these passenger mutations often exhibit characteristic mutational patterns called "mutational signatures" [Alexandrov et al., 2013a]. Importantly, these characteristic mutational signatures are often linked to specific mutagenic processes, making it possible to infer which mutagenic processes have been active in the given patient. This information often provides important clues about the nature of the diseases. For example, the presence of specific signatures associated with homologous recombination repair deficiency (HRD) can help identify patients who can benefit from PARP inhibitor treatment [Davies et al., 2017]. With the increased interest in the information on mutagenic processes acting on cancer genomes, several computational approaches have been developed to define mutational signatures in cancer [Alexandrov and Stratton, 2014; Alexandrov et al., 2013a,b; Fischer et al., 2013; Goncearenco et al., 2017; Helleday et al., 2014], to identify patients whose genome contains given signatures [Fischer et al., 2013; Goncearenco et al., 2017; Huang et al., 2018b], to map patient mutations to these signatures [Huang et al., 2018a] and to identify superposition of several mutagenic processes [Wojtowicz et al., 2020].

Despite the importance of understanding cancer mutational signatures, the etiology of many signatures is still not fully understood. It is believed that mutational signatures may arise not only as a result from exogenous carcinogenic exposures (e.g., smoking, UV exposures)

but also due to endogenous causes (e.g., HRD signature mentioned above). That is, human genomes are protected by multiple DNA maintenance and repair mechanisms in the presence of various types of DNA damage but aberrations or other malfunctions in such mechanisms can leave errors not repaired, generating specific patterns of mutations [Knijnenburg et al., 2018].

From the perspective of individual patients, it is important to determine mutational signatures imprinted on each patient's genome and the strength of the (sometimes unknown) mutagenic processes underlining the signatures. Signature strength can be measured by the number of mutations that are attributed to the given signature and thus can be considered as a continuous phenotype. With this view in mind, we investigate the relation of this phenotype with other biological properties of cancer patients. In this study, we focus on the relation of mutational signature strength with gene expression in biological processes and gene alteration in subnetworks.

The hypothesis that mutational signatures can be related to aberrant gene expression or alterations in DNA repair genes is well supported. For example, the deactivation of MUTYH gene in cancer patients is associated with a specific mutational signature [Chae et al., 2016; Knijnenburg et al., 2018; Ma et al., 2018]. Previous studies identified correlations between several mutational signatures and some cancer drivers and acknowledged that the cause-effect relation between signatures and cancer drivers can be in either direction [Poulos et al., 2018]. On the other hand, like many other cancer phenotypes, the causes of mutational signatures can be heterogeneous and the same signature can arise due to different causes. For example, the above mentioned signature caused by the inactivation of the MUTYH gene was also found in cancers that do not harbor this aberration [Viel et al., 2017]. With the observation that different mutations in functionally related genes can lead to the same cancer phenotype [Garraway and Lander, 2013; Hanahan and Weinberg, 2011; Vogelstein et al., 2013], cancer phenotypes are increasingly considered in the context of genetically dysregulated pathways rather than in the context of individual genes [Chuang et al., 2007; Dao et al., 2017b; Hofree et al., 2013; Kim et al., 2013, 2016b; Vandin et al., 2012a]. Hence, we postulated that identifying mutated subnetworks and differentially expressed gene groups that are associated with mutational signatures can provide new insights on the etiology of mutational signatures.

In this study, we focused on mutational signatures in breast cancer, for which a large data set is available, including whole genome mutation profiles as well as expression data [Nik-Zainal et al., 2016]. The mutagenic landscape of this cancer type is complex and is yet to be fully understood. For example, previously defined *COSMIC* signatures present in breast cancer [Nik-Zainal et al., 2016] include two signatures (Signatures 1 and 5) as age related (clock-like) and two signatures associated with the activities of APOBEC enzyme (Signatures 2 and 13). The mechanisms underlying the differences between two distinct signatures with similar etiology are not fully understood.

The clock-like signatures (COSMIC Signatures 1 and 5) have been found correlated with the age of patients but the strengths of correlation differ between the two signatures and vary across different cancer types [Alexandrov et al., 2015]. Signature 1 is considered to arise from an endogenous mutational process initiated by spontaneous deamination of 5-methylcytosine

while the etiology of Signature 5 is less understood. Therefore, it is important to understand what processes, other than patient's age, contribute to each of these signatures.

APOBEC signatures have been the subject of particular attention [Buisson et al., 2017; Burns et al., 2013; Cescon and Haibe-Kains, 2016; Green et al., 2016; Leonard et al., 2013; Nik-Zainal et al., 2014; Seplyarskiy et al., 2016; Shimizu et al., 2018; Wang et al., 2018]. The proteins encoded by APOBEC gene family (known to be involved in immune response) deaminate cytosines in single-stranded DNA (ssDNA). Such deamination, if not properly repaired, can lead to C>T (Signature 2) or C>G (signature 13) mutations depending on how the resulting lesion is repaired or bypassed during the replication [Morganella et al., 2016]. Thus the final imprint of APOBEC related mutations on the genome depends on several factors: expression level of APOBEC genes, the amount of accessible ssDNA, and the lesion bypass mechanism. In particular, clustered APOBEC-induced mutations (*kataegis*) in breast cancer are assumed to be a result of the mutation opportunity offered by single-stranded DNA during repair of double-stranded breaks (DSBs). However, ssDNA regions can also emerge for other reasons such as topological stress. Thus, although several aspects contributing to the APOBEC signatures have been known for some time, we are yet to uncover the full complexity of the APOBEC derived signatures.

To address these challenges, we took two complementary pathway based approaches: one focused on gene modules whose expression correlates with signature strength and the second based on the identification of subnetworks of genes whose alterations are associated with mutational signatures.

Our study provides several new insights on the mutagenic processes in breast cancer including: (i) association of the NER pathway and oxidation processes with the strength of clock-like Signature 5 (ii) differences between the two clock-like signatures with respect to their associations with cell cycle (iii) differences in mutated subnetworks associated with different signatures including APOBEC related signatures. We demonstrate that our findings are consistent with the results from recent studies and provide additional insights that are important for understanding mutagenic processes in cancer and developing anti-cancer drugs.

4.2 Methods

Overview

In this study we consider mutational signatures in cancer patients and attempt to identify genes and pathways whose expression and/or genetic alterations are potentially causative of differences in mutational signature strength. We utilized the somatic mutations in the cohort of 560 breast cancer (BRCA) whole-genomes [Nik-Zainal et al., 2016]. We used 12 COSMIC signatures indited as active in BRCA in previous studies (Signatures 1, 2, 3, 5, 6, 8, 13, 17, 18, 20, 26, and 30). Since recent studies revealed that mutations occurring in close proximity to each other, referred to here as cloud mutations, have distinct properties from dispersed mutations [Huang et al., 2018a; Supek and Lehner, 2017], we additionally

subdivided all mutations (and subsequently their attributed signatures) into two groups – close-by **C**loud mutations and **D**ispersed mutations (see Methods section: Data)

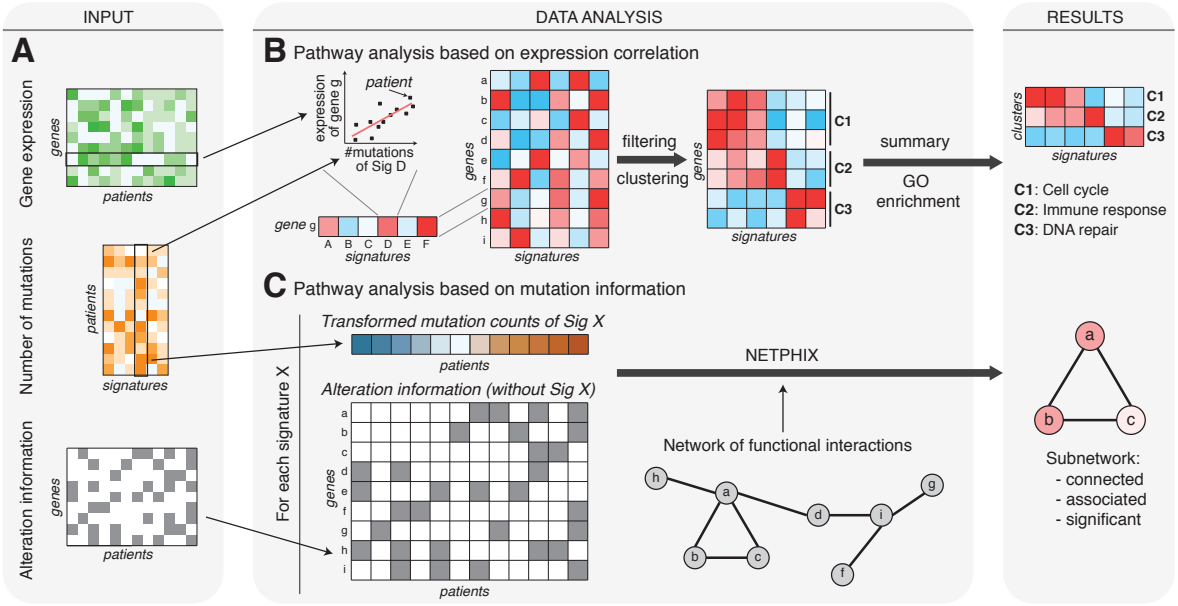


Figure 4.1. **(A)** The input data for this study consist of gene expression, mutational signature counts, and gene alteration across a number of cancer patients. **(B)** The functional pathways whose gene expression levels are associated with mutational signatures were found by computing correlations between expression levels of all genes and signature mutation counts, filtering out weak correlations, clustering expression correlation profiles, and performing GO enrichment analysis of the identified clusters. **(C)** The pathways whose gene alterations are associated with mutational signatures were found by applying NETPHIX to the transformed signature mutation counts (z-score of log-transformed counts), gene-patient alteration matrix and a known functional interactions network.

In the first part of the analysis, we looked for the genes whose expression levels are significantly correlated with mutational signature strength (Fig. 4.1A,B). Specifically, we first selected genes exhibiting significant correlation with at least one mutational signature by computing the correlation coefficient of the expression profile and mutation counts for each pair of genes and signatures. The selected genes were clustered based on their expression correlation patterns across mutational signatures (see Methods section: Expression correlation analysis).

The second part of the analysis involves uncovering subnetworks of genes whose alterations are associated with mutational signature strength (Fig. 4.1A,C). We hypothesize that a certain mutational signature can arise when a related pathway (e.g. DNA damage repair mechanism) is dysregulated. Due to the complex nature of cancer driving mutations, we adapted the NETPHIX method – a recently developed network based method to identify mutated subnetworks associated with continuous phenotypes [Kim et al., 2019]– to identify

such pathways. In this analysis, we consider the mutation count of a mutational signature in a whole cancer genome to be a cancer phenotype and aim to identify a subnetwork of genes whose alterations are associated with the phenotype. Importantly, when assessing association between gene level alterations and a mutational signature, the mutations attributed to the given mutational signature were not incorporated into the alteration information (Fig. 4.1C; Methods section: Mutation analysis, and Additional file 1: Supplemental Methods) in order to increase the likelihood of uncovered subnetworks being drivers of the signatures rather than their effect.

Data

We analyzed the somatic mutations in the cohort of 560 breast cancer (BRCA) whole-genomes published by Nik-Zainal *et al.* [Nik-Zainal et al., 2016]. The mutation data (single base substitutions and small indels) were downloaded from the ICGC data portal (release 22) [icg]. The most likely assignments of 3,479,652 individual point mutations to mutational signatures were generated with SigMa [Huang et al., 2018a] using 12 predefined COSMIC signatures (version 2; `https://cancer.sanger.ac.uk/cosmic/signatures_v2`) known to be active in BRCA (Signatures 1, 2, 3, 5, 6, 8, 13, 17, 18, 20, 26, and 30) [Nik-Zainal et al., 2016]. SigMa is a probabilistic model of sequential dependency for mutation signatures that allows for an accurate assignment of mutations to predefined signatures (it does not infer new signatures). To ensure SigMa's robustness with respect to random initialization used in its learning process, we computed the majority assignments over 31 random initialization runs. SigMa relies on the observation that adjacent mutations in a given cancer genome are more likely to be the result of the same mutation signature and that mutations that are assigned to the same signature can have distinct properties when being isolated versus being localized in clusters [Morganella et al., 2016; Nik-Zainal et al., 2016; Supek and Lehner, 2017]. Thus, it divides all mutations into two groups – close-by (clustered) **C**loud mutations and **D**ispersed (sky) mutations. The sequential dependencies between close-by mutations are modeled by a Hidden Markov model, while for dispersed mutations we use a multinomial mixture model. Here, we treat cloud and dispersed mutations, and their associated signatures, separately. For each patient, we computed signature profiles based on the patient mutation counts assigned to each specific signature, separating cloud and dispersed mutations. The mutational signature profiles were used as phenotype profiles in the expression correlation and mutated pathway analyses (Fig. 4.1A). For further analysis, we used only sufficiently abundant mutational signatures for cloud or dispersed mutations whose overall exposure levels are above 10% within both groups of mutations. This created 10 different phenotype profiles for Signatures 1D, 2C/D, 3C/D, 5D, 8C/D, and 13C/D, where the numbering refers to the COSMIC signature index and C/D denotes signatures attributed to close-by cloud and dispersed mutations.

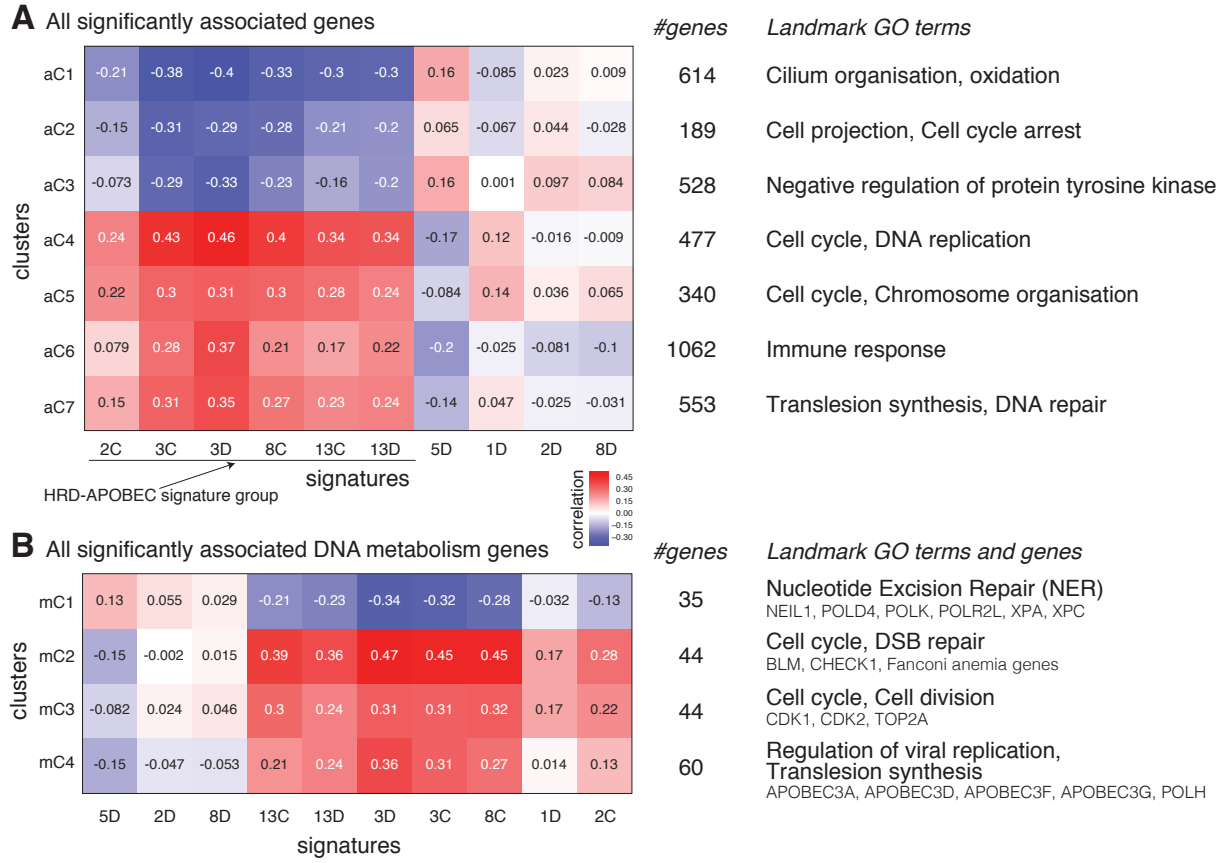


Figure 4.2. **(A)** All genes significantly correlated with at least one signature ($|corr| \geq 0.3$ and adjusted $pv \leq 0.005$). **(B)** DNA metabolic process genes, based on Gene Ontology (GO), significantly correlated with at least one signature. For both (A and B), we show a heatmap of mean expression correlation for each cluster and signature (left), number of genes in each cluster (middle), and representative GO terms enriched in cluster genes (right). For the DNA metabolic process, we also show representative genes for each cluster.

Expression correlation analysis

To identify expression based pathways that are associated with signatures, we downloaded the normalized gene expression data for 266 BRCA patients from Table S7 in [Nik-Zainal et al., 2016] and used correlation analysis followed by clustering of correlation patterns. Specifically, we first computed the Spearman correlation coefficient of the expression level and mutation count for each pair of genes and mutational signatures. We then selected the genes exhibiting significant correlation with at least one of 10 mutational signatures; the expression of a gene is considered significantly correlated with a signature if $|corr| \geq 0.3$ and adjusted $pv \leq 0.005$ ($corr$ is Spearman correlation coefficient, BH corrected p-value). The procedure selected 3,763 genes. We then clustered the genes based on their correlation pattern using a consensus K-means algorithm; running K-means clustering 100 times with random start and varying k

from 5 to 50 and subsequently running hierarchical clustering with consensus matrix from 100 runs of K-means. GO enrichment analysis was performed using hypergeometric test and significant terms were selected with nominal p-value < 0.05. The final 7 clusters and enrichment analysis results are summarized in Fig. 4.2A.

To take a closer look at DNA repair genes, we performed similar analysis with genes in GO DNA metabolic process. 184 genes are selected with the same significance cut-offs. The hierarchical clustering of the consensus clustering for 100 K-means ($k = 2$ to 20) generated 4 clusters shown in Fig. 4.2B. The enrichment analysis was performed using hypergeometric test with only the genes in GO DNA metabolic process as the background, and only for the GO terms with significant overlaps with GO DNA metabolic process (at least 2 genes in common and p-value of the intersection < 0.05).

Mutation analysis

To find alteration based pathways for signatures, we adapted the method proposed in Section 2, NETPHIX, which identifies mutated subnetworks associated with a continuous phenotype [Kim et al., 2019]. Given gene alteration information of cancer samples and continuous phenotype values for the same samples, NETPHIX aims to identify a connected subnetwork whose aggregated alterations are associated with the phenotype of interest (mutation counts for cancer mutational signatures in this study). NETPHIX utilizes functional interaction information among genes and enforces the identified genes to be *connected* in the network while, at the same time, making sure that the aggregated alterations of these genes are significantly *associated* with the given phenotype. In addition, in its integer linear program formulation, NETPHIX recognizes that cancer driving mutations tend to be mutually exclusive [Ciriello et al., 2013; Constantinescu et al., 2015; Kim et al., 2015, 2016c; Leiserson et al., 2015a; Vandin et al., 2012a] and incorporates this property in its objective function [Kim et al., 2019]. The detailed description of NETPHIX is given in Chapter 3.

For the gene level alteration information (the bottom matrix in Fig. 4.1A), we utilized all somatic point mutations and small indels for the same 560 patients data. In processing the somatic mutation data, we defined a gene to be altered if it has at least one non-silent mutation in its genomic region. In addition to somatic mutations, DNA repair genes can undergo alternative mechanisms of inactivation including pathogenic germline variants and promoter hypermethylation. A recent paper highlighted the importance of these mechanisms in inactivating the homologous recombination pathway [Davies et al., 2017]. To account for these additional sources of inactivation, we also defined a gene to be altered in a patient if the gene is annotated as being biallelic inactivated for the patient in Supplementary Table 4a and 4b from [Davies et al., 2017]. The gene alteration information is used to find mutated subnetworks associated with each signature (Fig. 4.1C). When computing association with a specific signature, we further refined the information to increase the likelihood that the association is causative (i.e., gene alteration causes mutational signatures, not vice versa). Specifically, the gene alteration information for the association analysis with a specific mutational signature were constructed after excluding the mutations attributed to the given

mutational signature (see Additional file 1: Supplemental Methods for details). Similarly, we removed all indels when we considered the associations with Signature 3 and 8 as these signatures are believed to lead to a high burden of indels. The assignment of mutations to signatures was performed using SigMa (see above).

For each mutational signature, we normalized the mutation counts by taking log and subsequently computing z-scores, and used the profiles as phenotype inputs to NETPHIX. For functional interactions among genes, we used the data downloaded from STRING database version 10.0 [str], only including the edges with high confidence scores (≥ 900 out of 1,000). The alteration tables were constructed as described above and genes altered in less than 1% of patients were removed from further consideration. We ran NETPHIX for each mutational signature with density constraint of 0.5 and for a fixed size modules k from 1 to 7. The appropriate k was selected by examining the increase of the objective function values and the significance of the solution using permutation tests. Specifically, the best k was selected to be maximal index for which the optimal objective function increased more than 5% with respect to previous index and the permutation p-value did not increase, with this property holding for all smaller indices ($k' < k$). The permutation test is computed by permuting the phenotype (the mutation counts for each signature in this case) and comparing the objective function value to the ones obtained with the permuted phenotypes. We define the identified module to be significant if the FDR adjusted p-value is less than 0.1.

For the analyses with BRCA subtypes, we utilized AIMS subtypes provided in Supplementary Table 18 [Nik-Zainal et al., 2016]. The association analyses with gene alteration information were performed with 78, 111, and 64 samples categorized as Luminal A, B and Basal subtypes, respectively (There are only 10 samples in HER2 subtype hence the results are not reported).

4.3 Results

Expression analysis to identify biological processes associated with mutational signatures

In order to identify biological processes associated with individual signatures, we clustered gene expression-signature correlation profiles as described in the Methods section. To obtain a bird's eye view, we first used all genes whose expression is correlated with at least one signature (Fig. 4.2A and Additional File 1: Fig. S1; see Methods section). Next, to obtain a finer scale expression modules related to DNA repair, we zoomed in on genes involved in Gene Ontology DNA metabolic process (Fig. 4.2B).

The first striking observation is the similarity of gene expression patterns among both variants of Signatures 3 and 13 and all other cloud signatures (2C and 8C). Since Signature 3 and 13 are considered to be associated with homologous recombination deficiency and APOBEC activity respectively, in what follows we refer to this group of signatures as HRD-APOBEC signature group. Note that Signature 2 is also known as an APOBEC related

signature but the group includes only Signature 2C but not 2D. Below, we will discuss insights obtained for the age-related signatures and the APOBEC signatures, and also provide independent supporting evidence from literature. Given expression correlation similarity within the members of the HRD-APOBEC group (all positively correlated with cell cycle, DNA repair, and immune response), we defer the analysis of this group to the next section where we look at this group through the lenses of mutated subnetworks.

The expression correlation analysis reveals important differences between the APOBEC signatures. Surprisingly, among 4 APOBEC related signatures (Signature 2C/D and 13C/D) Signature 2D has strikingly different correlation patterns compared to the remaining three APOBEC signatures. APOBEC activities are considered to be related to immune response. While the expression correlation patterns of all other APOBEC signatures are consistent with such understanding, Signature 2D exposure level has slightly negative correlation with immune response (4.2A, aC6). This is consistent with our previous observation that there is no positive correlation between Signature 2D and APOBEC expression [Huang et al., 2018a].

In addition, Signature 2 exposure level is either not correlated (2D) or has a weak correlation (2C) with the cluster enriched with translesion synthesis (4.2, aC7 and mC4) whereas both Signature 13C and 13D show positive correlation. This last observation supports the previous claim that the difference between Signatures 2 and 13 is related to differences in the repair mechanism [Morganella et al., 2016]. Specifically, it has been suggested that mutations in Signatures 13 emerge when lesions created by APOBEC activity are repaired by DNA translesion polymerase, which inserts 'C' opposite to the damaged base while Signatures 2 occurs when the damaged base is simply paired with 'A'.

Clock-like signatures 1D and 5D have different expression associations suggesting differences in their etiology. Although weaker than the correlation with the HRD-APOBEC Signature group, two clusters enriched in cell cycle function are positively correlated with Signature 1D (Fig. 4.2A, aC4 and aC5), which is consistent with the previous observation that Signature 1 is associated with aging [Alexandrov et al., 2015], and thus postulated to be correlated with the number of cell divisions. Consistent with this interpretation, many cancer types with high level of Signature 1 are derived from normal epithelia with high turnover such as stomach and colorectum [Alexandrov et al., 2015].

On the other hand, Signature 5D is not positively correlated with the expression of cell cycle genes despite the fact that Signature 5 is also considered to be a clock-like signature. This suggests that accumulation of mutations attributed to Signature 5 is related to the exposure to naturally occurring environmental/external processes. Interestingly, Signature 5D has a positive correlation with the cluster enriched in oxidative processes (Fig. 4.2A, aC1) and the cluster enriched in nucleotide excision repair (NER) pathway (Fig. 4.2B, mC1). The accumulation of oxidation base lesions is also assumed to be age-related [Hamilton et al., 2001], suggesting that Signature 5 might be related to oxidative damage. NER pathway

is involved in neutralizing oxidative DNA damage [Melis et al., 2013] and Signature 5 has been also associated with smoking [Alexandrov et al., 2016], which itself is associated with oxidative damage. Indeed Signature 5 was linked to the NER pathway in a recent study [Kim et al., 2016a]. Finally, comparative analysis of Signature 5 mutation rates in various types of kidney cancers supports the hypothesis that continuous exposure to ubiquitous metabolic mutagens may underlie Signature 5 mutations [Alexandrov et al., 2015].

The positive correlation of Signature 1 with the expression of cell cycle genes and lack of such correlation for Signature 5 may explain the stronger association of Signature 5 with the age of patients than Signature 1 in breast cancer [Alexandrov et al., 2015; Huang et al., 2018a] because cancer related cell division might obscure the association of Signature 1 with a patient's age.

Identifying mutated subnetworks associated with mutational signatures

The analysis of expression correlation clusters revealed different biological processes associated with some signatures but the signatures in HR-APOBEC group have largely similar expression patterns and require further investigation. Complementary to the expression analysis, we next searched for possible associations with subnetworks of mutated genes. Some mutational signatures can arise due to endogenous causes; aberrations in genes responsible for different DNA repair mechanisms can lead to the malfunctioning of the corresponding repair process, leaving errors not repaired and in turn generating specific patterns of mutations. We applied NETPHIX, a method to identify phenotype associated subnetworks, which can help to uncover a subnetwork of genes whose alterations are potentially causative of specific mutational signatures directly or indirectly. Note that not all mutational signatures have such association with mutated pathways. Mutational signatures arising from environmental exposure, age, or other external factors are not necessarily expected to have casual associations with mutated subnetworks.

Figure 4.3 shows all statistically significant subnetworks (phenotype permutation test; see Methods section) identified by NETPHIX and their alteration profiles. See Methods section (Mutation analysis) for how the module for each signature was selected. The extended subnetworks obtained with less stringent cutoffs are shown in Additional file 1: Fig. S2.

As expected, no modules are found to be significantly associated with the age related signatures 1D and 5D. This is consistent with the current understanding that these signatures can accumulate due to naturally occurring processes. In addition, consistently with the previous studies that linked the genes underlying the HRD to Signature 3 in breast cancer [Polak et al., 2017], the subnetworks identified for Signature 3 C/D contain BRCA1 and BRCA2 genes, two important genes in HR-mediated double-strand break (DSB) repair.

The agreement of the modules identified by NETPHIX with the current knowledge confirms its ability to correctly infer mutated subnetworks associated with signatures.

Encouraged by the results, we examined the remaining subnetworks identified by NET-

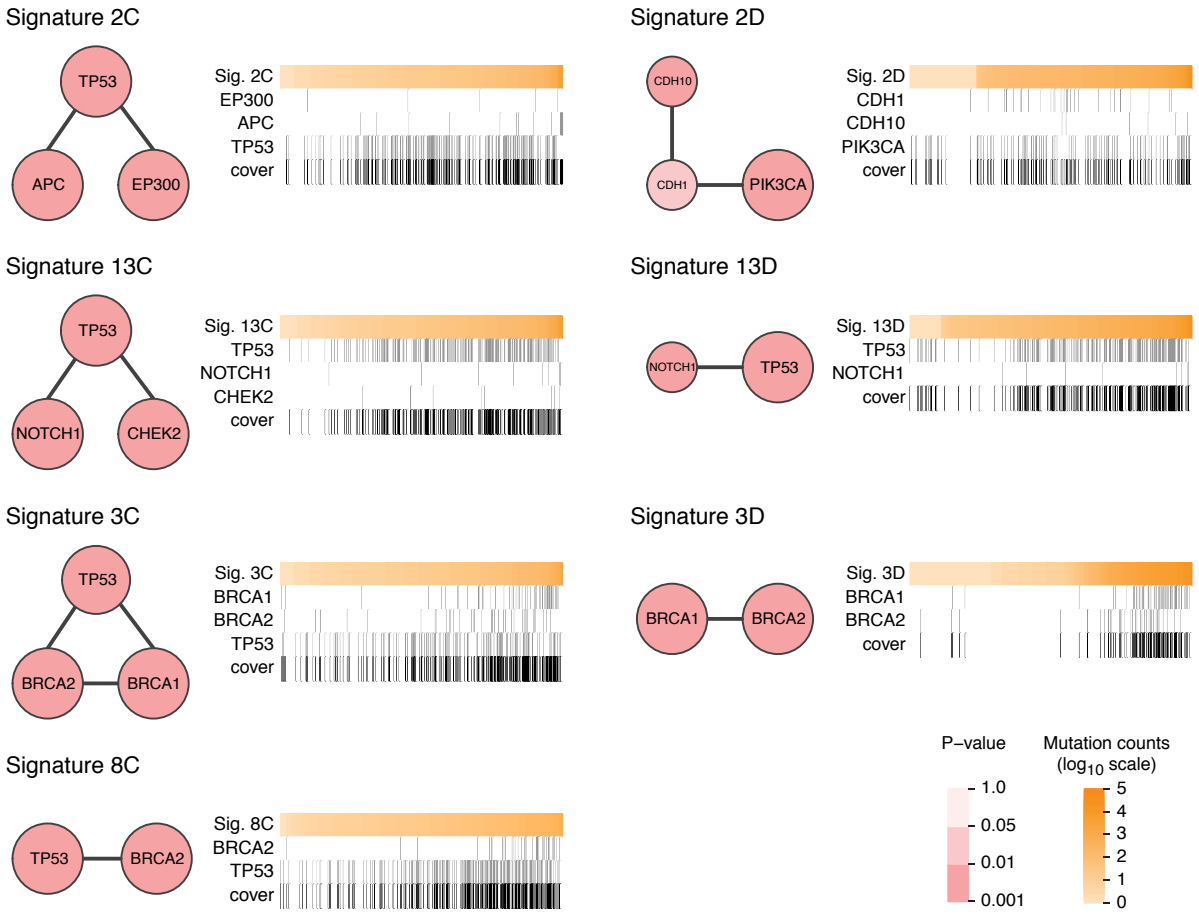


Figure 4.3. Panel for each signature consists of a network view of a module (left) and a heatmap showing an association of module gene alterations with signature strength across patients (right). The network node size indicates the gene robustness (regarding NETPHIX results for different random initialization runs of SIGMA) while the darkness of red color represents its individual association score (empirical p-value based on phenotype permutation test). Each heatmap shows the number of mutations attributed to a given signature for all patients (orange; top row; $\log_{10}$ scale) sorted from low to high (columns). For each gene in the module, gene alteration information observed in each patient are shown in gray, while patients not altered are in white. The last row shows the alteration profile of the entire subnetwork in black. Only subnetworks significant in phenotype associations for mutational Signatures 2C, 2D, 13C, 13D, 3C, 3D, and 8C are shown; results for Signatures 1D and 5D were not significant.

PHIX. Among statistically significant modules, TP53 was included in all modules associated with cloud signatures. TP53 is known to play a crucial role in DNA damage responses, including DSB repair. We note that its dysfunction could contribute to increased mutation burden and in turn to the emergence of cloud mutations independently of mutagenic processes

underlying individual signatures. However, whether or not TP53 mutations are causal or are a result of yet another mutagenic process cannot be concluded from this study. Complicating this picture, a recent study demonstrated that p53 controls expression of the DNA deaminase APOBEC3B suggesting a possible mechanism by which mutations in p53 can promote APOBEC expression [Periyasamy et al., 2017] and thus APOBEC related mutations. Hence the reason for the strong association of TP53 with cloud mutational signatures requires further investigation.

Compared to the modules obtained from expression analysis, the analysis with genetic alterations offers a better differentiation among the signatures in the HRD-APOBEC group. While most of the signatures in the group contain TP53, they also include different genes in the modules. In the subnetworks associated with Signatures 13 C/D, TP53 is accompanied by NOTCH1; NOTCH pathway regulates many aspects of metazoan development, including the control of proliferation and differentiation. CHEK2 is selected in addition to TP53 and NOTCH1 for Signature 13C. CHEK2 is a tumor suppressor regulating a cell cycle checkpoint and mutations in the gene confer an increased risk for breast cancer [Desrichard et al., 2011; Meijers-Heijboer et al., 2002]. CHEK2 plays multiple roles in DNA damage response [Zannini et al., 2014], including DSB repair in the emergence of clustered APOBEC related mutations.

In the subnetwork associated with Signature 2C, TP53 is accompanied by APC (Adenomatous Polyposis Coli), which is a tumor-suppressor gene frequently mutated in Colorectal cancer (CRC) and involved in the Wnt signalling pathway. A recent study linked APC to several DNA repair mechanisms, including the base excision repair (BER) pathway [Jaiswal and Narayan, 2008], DSB repair [Kouzmenko et al., 2008] and genomic stability [Fodde et al., 2001; Meniel et al., 2015].

Finally, the subnetwork for Signature 2D (dispersed, APOBEC related signature) consists of PIK3CA, CDH1 and CDH10 genes and is completely different from the subnetworks corresponding to the cloud variant of Signature 2 and other HR-APOBEC related signatures. Previous studies have found that some recurring mutations in PIK3CA are consistent with Signature 2 and may result from APOBEC activities [Poulos et al., 2018; Temko et al., 2018]. However, our analysis associated PIK3CA mutations with Signature 2 even after removing point mutations attributed to Signature 2, suggesting a more complex relation between Signature 2 and PIK3CA mutations.

In addition to PIK3CA, the subnetwork associated with Signature 2D has two Cadherin genes: CDH1 and CDH10. Cadherins are important in maintenance of cell adhesion and polarity, and alterations of these functions can contribute to tumorigenesis. CDH1 germline mutations have been associated with hereditary lobular breast cancer [Masciari et al., 2007] and hereditary diffuse gastric cancer [Hansford et al., 2015; Kaurah et al., 2007], while a recent study linked mutations in CDH1 and PIK3CA to the immune-related invasive lobular carcinoma of the breast [An et al., 2018]. In breast cancer, mutations in CDH1-PIK3CA module are mutually exclusive with mutations in TP53 and are strongly enriched in Luminal A subtype [Dao et al., 2017a]. Indeed, our analyses of individual subtypes show that the association of a PIK3CA module with Signature 2D is significant only with Luminal A subtype. Interestingly, the module identified in Luminal A contains, in addition to PIK3CA,

PTEN gene which is known to be a negative regulator of the PIK3CA [Carracedo and Pandolfi, 2008]. This, combined with the differences in expression correlations noted in the previous section, suggest that the etiology of Signature 2D is different from the other APOBEC mutational signatures (Signature 2C and 13)

4.4 Discussion

In order to gain insights into the etiology of mutational processes in cancer, we propose two complementary computational approaches and apply them to gain insights into the etiology of mutational processes in breast cancer. Both approaches leverage the idea of network level association of mutation signatures with gene networks and pathways but differ in the type of utilized data and mathematical formulation. The first approach uses gene expression data the second approach is focused on the identification of subnetworks of genes whose alterations are associated with each signature.

The expression correlation based approach allowed us to uncover important differences between clock-like signatures. Clock-like signatures can occur from life long exposure to naturally occurring mutagenic processes, thus related to aging. The most prominent clock-like signatures are Signature 1 and 5. Signature 1, a relatively well characterized clock-like signature, is considered to be the result of an endogenous mutational process related to spontaneous deamination of 5-methylcytosine. Each cell division provides an opportunity for such mutations to occur. This explains why many cancer types with high mutation rates of Signature 1 are derived from normal epithelia with high turnover [Alexandrov et al., 2015]. The correlation of Signature 1 mutation counts with the expression level of cell cycle genes observed in this study provides further support for this explanation. The etiology of Signature 5 was less clear. Our expression based analysis revealed that, differently from Signature 1, Signature 5 is not positively correlated with expression of cell cycle genes. Instead, we found an association of Signature 5 with oxidation process. This observation is consistent with several previous findings. In particular, our findings support the hypothesis that cell proliferation rate may not be a major factor for Signature 5 [Alexandrov et al., 2015]. In addition, accumulation of oxidation base lesions is assumed to be related to aging [Hamilton et al., 2001] as well as smoking, while the association of Signature 5 with smoking was observed in a previous study [Alexandrov et al., 2016]. More supporting evidence is provided by the association of Signature 5 with the nucleotide excision repair (NER) pathway which was shown to be involved in neutralizing oxidative DNA damage [Melis et al., 2013]. These results support the view that the correlation of Signature 5 with age is related to a continuous exposure to an environmental/metabolic mutagen.

While expression based analysis was very valuable for understanding the differences between Signatures 1 and 5, many signatures especially in the HRD-APOBEC signature group exhibit similar expression correlation patterns. The mutated pathway analysis provided additional insights into the differences among these signatures. In particular, both cloud and dispersed Signature 3 are associated with BRCA 1/2 genes while the subnetwork associated

with Signature 3C additionally contains TP53. The results of mutated subnetwork analysis also revealed the association of mutations in tumor suppressor APC for two different cloud signatures (Signature 2C and Signature 8C with a lenient cutoff) and NOTCH1 mutations for both variants of Signature 13.

In order to increase the probability that inferred mutated subnetworks are causal, we removed the mutations attributed to the signature of interest. This eliminates the possibility that the mutations resulted directly from the mutagenic process underlying the signature although it still does not guarantee causality. In particular, the consistent presence of TP53 in the subnetworks associated with cloud signatures makes it tempting to speculate that mutations in TP53 generally increase the mutation rates leading to an increase in cloud mutations. However, other indirect reasons for this association cannot be ruled out. Our analysis also showed unique properties of Signature 2D relative to the remaining APOBEC signatures. This signature is the only signature associated with PIK3CA and not TP53. Previous studies have found that several recurring mutations in PIK3CA are consistent with Signature 2 [Poulos et al., 2018; Temko et al., 2018]. However, our analysis indicates that even after removing mutations attributed to Signature 2, the association between PIK3CA mutations and Signature 2D remains. Another known cancer gene present in this subnetwork is CDH1. CDH1 was previously linked to hereditary lobular breast cancer [Corso et al., 2016] and hereditary diffuse gastric cancer and in particular, about 40% of hereditary diffuse gastric cancer patients are found to have mutations in CDH1 [Hansford et al., 2015; Kaurah et al., 2007]. Invasive lobular carcinoma is characterized by a unique immune signature [Du et al., 2018] which might provide additional insights to the etiology of Signature 2. Our previous studies with breast cancer demonstrated that mutations in CDH1-PIK3CA module are mutually exclusive with mutations in TP53 and are enriched in Luminal A subtype [Dao et al., 2017a]. Consistent with the observation, the subtype specific analysis using NETPHIX indicated that the association between signature 2D and subnetwork involving PIK3CA is particularly significant in the Luminal A subtype. Importantly, the module identified with samples in Luminal A subtype contains PTEN (in addition to PIK3CA), a known negative regulator of PIK3CA [Carracedo and Pandolfi, 2008]. These results suggest that the relation between Signature 2 mutations and the activation of PI3K pathway might be more complex than previously suggested.

Although our goal in this study was to investigate the genomic causes of mutational signatures regardless of cancer subtypes, we also performed the analysis for each subtype separately to examine the potential differences between subtypes. While generally consistent with the results using all samples, the results based on individual subtypes suggest that some associations are subtype specific and, as exemplified by the discussion of the PI3K-PTEN pathway above, can provide additional insights to the relation between mutagenic processes and mutated pathways.

Conclusions

Patterns of somatic mutations in a cancer genome can shed light on mutagenic processes acting on the genome. However, uncovering specific mutagenic processes underlying a given pattern of mutations is challenging. Previous studies demonstrated that network-centric approaches can be helpful for finding genotypic causes of diseases, classifying disease subtypes, and identifying drug targets [Kim et al., 2016b]. In addition, a recent study demonstrated that, within the same cancer type, different gene modules can be enriched in diffident mutational signatures [Dao et al., 2017b]. However, a broader utility of network based approaches for understating of mutagenic processes in caner was yet to be demonstrated. To fill this gap, we developed two complementing computational approaches and performed the first network level association analysis of mutation signatures with dysregulated pathways. Based on gene expression data, we identified gene modules whose expression correlates with mutation counts attributed to mutational signatures. Further analysis of these modules provided important insights into the mutagenic processes underlying specific signatures. Complementing expression analysis, we developed an ILP based method to identify subnetworks of genes whose alterations are associated with each signature. This analysis provided information about potential differences in the etiology of the signatures that could not be gained from the expression analysis alone.

Taken together, our study demonstrates the utility of these two complementary approaches for studying mutational signatures in cancer and provided several new insights into the etiology of mutational signatures.

Chapter 5

Discussion and Future Work

5.1 Discussion

The oncology pharmaceutics and genome sequencing industries are booming, and a significant challenge in modern medicine is analyzing the influx of data from these sources. Thousands of cancer drugs are currently in development, and patient genomic data is becoming readily accessible to physicians. There is little room for trial and error when treating cancer, and it is imperative that physicians make informed decisions when prescribing drugs. By combining drug response data and genomic information, physicians can make informed treatment decisions based on the unique biology of each patient. In this dissertation, multiple methods and algorithms to identify mutations that affect drug therapy response were presented.

Once we identify mutations that affect drug response, machine learning methods can be applied to train classifiers used to predict the sensitivity of drugs in cancerous cell lines. Methods outlined in Chapter 2 and 3 could be used as basis for prediction models and can lead to useful insight towards personalized drug treatment.

In the remainder of this chapter we present the outline for an approach to solve the practical problem of predicting drug response based on all the information oncologists have available about the patient, this can include genetic information but is also based on the demographic, histology, baseline labs and medical history recorded in the patient's Electronic Health Record. This is of critical importance in a world where more and more anti-cancer drugs are being developed, and doctors face the ever-increasing challenge of prescribing the right drug that will work best for an individual patient.

5.2 Predicting multiple treatment outcomes using random forests

Introduction

Due to disease heterogeneity, the effectiveness of any specific cancer therapy, such as chemotherapy or radiation, widely differs between individual patients. This inherent variability of cancer lends itself to the growing field of precision and personalized medicine. Personalized cancer therapy is an emerging treatment strategy based on the ability to predict which patients are more likely to respond well to specific treatments. It involves the systematic use of the information available about an individual patient to optimally select a course of treatment.

We present a framework for the problem of predicting multiple heterogenous clinical outcomes based on all the information oncologists have available about the patient, this can include genetic information but is more often based on the demographic, histology, baseline labs and medical history recorded in the patient's Electronic Health Record.

When oncologists make decisions about different treatment options they take into account various factors, besides the obvious expected survival time under different regimens. For instance one might also be interested in predicting early treatment discontinuation due to side effects, or the expected shrinkage of the tumor. Ideally during the decision making process the physician would have access to a complete picture of the outcomes for each treatment option (survival time, quality of life, side effects etc)

The aforementioned different treatment outcomes are recorded as heterogeneous data types. For instance overall survival is usually measured as censored data, the presence of a particular side effect can be recorded as a binary or categorical outcome and tumor shrinkage is a continuous measure.

If these outputs were unrelated then the obvious solution would be to predict each individual outcome separately with ad hoc methods. But in the clinical settings these outcomes are heavily correlated, for instance the expected survival time of a patient is often correlated with a potential tumor progression or the insurgence of serious side effects. Predicting multiple correlated outcomes simultaneously has shown in many cases to improve the prediction performance compared to predicting each outcome separately.

While substantial work has been done on multiple output prediction for the regression and classification case (for the classification case the problem is called multi-label), very little work has been done on predicting multiple outcomes where the data type is censored data [Su and Fan, 2004] or a mix of continuous and categorical variables [Ishwaran and Kogalur, 2020], [Moreno-Muñoz et al., 2018] and [Au et al., 2019]. Furthermore to the authors best knowledge there is no current published work or package that is able to a handle a mix of survival, continuous and categorical outcomes. The goal of this work is to close this gap by providing a unified framework for prediction of heterogeneous outcomes in a clinical setting, leveraging an ensemble learning method known as random forests. This method operates by constructing many decision trees and using the average prediction across all trees, see the

next section for a detailed description.

Materials and Method

Background

Most frequently, supervised learning problems involve the prediction of a single outcome. However for some applications we are required to predict multiple outcomes simultaneously, i.e. multiple outputs can be assigned to each observation and we need predictions for all of them. If these outputs were unrelated then the obvious solution would be to predict each individual outcome separately. But in many cases these outcomes present a certain degree of correlation and this property can be exploited to improve prediction accuracy. The problems of multi-output regressions and multi-output classification have been object of numerous studies, see [Borchani et al., 2015] for a review of the former.

There are two main approaches for solving multi-outcome problems: problem transformation methods that transform the multi-output problem into independent single output problems and algorithm adaptation methods that adapt specific single-output methods to handle multi-output data. In the second category we find several extensions of standard machine learning algorithms including statistical methods, support vector machines, kernel methods and decision trees.

Our method belongs to this last category. Decision trees are constructed by performing a series of splits on the values of predictor variables with the goal of creating groups of observations at each node that have similar outcome variable value. To construct the tree we need to iteratively identify the predictor and optimal splitting point. For continuous outcome variables this is often achieved by minimizing the sums of squared errors of each daughter node.

One way to improve the prediction performance of a single tree is to use an ensemble of decision trees and aggregate their predictions. There are two primary methods to create such ensembles, bagging and boosting. In bagging a set of trees is fit to bootstrap samples and the prediction is calculated as the average across all trees. This improves the the prediction error by decreasing the variance of the model and the influence of extreme observations. In boosting we build trees sequentially, each trying to focus on observations that were poorly predicted by the previous tree. In this chapter we focus on an extension of the bagging algorithm, an approach known as random forests [Breiman, 2001] where we randomly select the set of predictors evaluated for each tree split decision.

Method Description

For simplicity we use the same notation as [Ishwaran and Kogalur, 2020], since our work is an extension of their R package. Denote the response for patent i by $Y_i \in \mathbb{R}$. Suppose the

proposed split for the root node is of the form $x_i <= c$ and $x_i > c$ for a continuous x-variable x_i, and a split value of c. The split rule is to minimize

$$D(x,c) = \frac{1}{n_l}\sum_{x_i \le c}(Y_i - \bar{Y}_l)^2 + \frac{1}{n_r}\sum_{x_i > c}(Y_i - \bar{Y}_r)^2. \tag{5.1}$$

where the subscripts l and r indicate left and right daughter node membership respectively.

A similar approach can be used for classification problems. In this case, an averaged standardized Gini splitting rule can be used. In this model, the y-outcome associated with an individual i is a single categorical value. Let $p = (p_1, \ldots, p_J)$ be the class proportions for the classes 1 through J respectively, for the y-outcome in the node. The impurity of the node is defined as

$$\phi(\mathbf{p}) = \sum_{j=1}^{J} p_j(1-p_j) = 1 - \sum_{j=1}^{J} p_j^2. \tag{5.2}$$

The Gini index for a split c on x is

$$G(x,c) = \frac{n_l}{n}\phi(\mathbf{p}_l) + \frac{n_r}{n}\phi(\mathbf{p}_r), \tag{5.3}$$

where n_l and n_r are the number of cases in the daughters such that $(n = n_l + n_r)$.

The first approach to be proposed for extending decision trees to multiple outcomes prediction is [De'ath, 2002]. They present an extension of the classification and regression tree algorithm (CART) to the multi-output regression problem. They follow the same steps as CART: start with all instances at the root node, then iteratively finding the optimal split and partitioning the leaves accordingly until a pre-defined stopping criterion is reached. The only difference from CART is the redefinition of the node splitting rule as the sum of squared error over the multi-variate response.

Suppose we have M outcomes with associated response variable: $Y_{im} \in \mathbb{R}$. The split rule is to minimize

$$\frac{1}{n_l}\sum_{i:x_i>c}\sum_{m:x_i>c}(Y_{im} - \bar{Y}_{lm})^2 + \frac{1}{n_r}\sum_{i:x_i>c}\sum_{m:x_i>c}(Y_{im} - \bar{Y}_{rm})^2. \tag{5.4}$$

A similar rule can be defined for the the multi-output classification case by generalizing (5.3).

In [Ishwaran, 2015] they show that for mixed continuous and categorical outcomes we can form a combined splitting rule. The regression and classification settings present similar theoretical splitting properties and are combined in [Ishwaran and Kogalur, 2020] to form

a composite splitting rule. The new rule is a composite standardized split rule of the mean-squared error (5.4) and the Gini index splitting (5.3):

$$\theta(x, c) = D^*(x, c) + G^*(x, c) \tag{5.5}$$

where $D^*(x, c)$ and $G^*(x, c)$ are maximization multi-output equivalents of (1) and (3) respectively. See [Tang and Ishwaran, 2017] for details.

Note that while this approach can handle the prediction of continuous and binary clinical outcomes, it does not currently permit the inclusion of censored outputs. The prediction of expected survival time for a specific therapy is a key metric to consider in the decision-making process.

In survival analysis, the response associated with individual i is a pair of values specifying a non-negative survival time and censoring information. Denote the response for i by $Y_i = (T_i, \delta_i)$. An individual is said to be right censored at time T_i if $\delta_i = 0$ and is said to have died at time T_i if $\delta_i = 1$. An individual i who is right censored at T_i simply means that the individual is known to have been alive at T_i, but the exact time of death is unknown. The Cox proportional hazard regression model and its extensions are very often used to study survival outcomes with censoring. Survival trees and forests are popular alternatives to such models. They offer great flexibility and can automatically detect certain types of interactions without the need to specify them beforehand and a single tree can naturally group subjects according to their survival behavior based on their covariates. Moreover, survival trees are ideal candidates for combination by means of an ensemble method such as the aforementioned survival forests. There are multiple split rules that can be used to grow survival trees. In [Ishwaran and Kogalur, 2020] they use Log-rank splitting. In this case the best split is the one that maximizes survival differences between the two daughter nodes.

Let $t_1 < t_2 < ... < t_m$ be the distinct times of death in the parent node h, and let $d_{k,l}$ and $Y_{k,l}$ equal the number of deaths and individuals at risk respectively at time t_k in the left daughter node for $k \in 1, \ldots, m$. Similarly let $d_{k,r}$ and $Y_{k,r}$ refer to the right daughter node. Note that $Y_{k,s}$ is the number of individuals in daughter $s \in l, r$ who are alive at time t_k, or who have an event (death) at time t_k. More precisely,

$$Y_{k,l} = \#\{i : T_i \geq t_k,\ x_i \leq c\}, \quad Y_{k,r} = \#\{i : T_i \geq t_k,\ x_i > c\},$$

where x_i is the value of x for individual $i = 1, \ldots, n$. Define $Y_k = Y_{k,l} + Y_{k,r}$ and $d_k = d_{k,l} + d_{k,r}$. Let n_s be the total number of observations in daughter s, Thus,$n = n_l + n_r$, where $n_l = \#\{i : x_i \leq c\}$ and $n_r = \#\{i : x_i > c\}$. The log-rank test for a split at the value c for an variable x is

$$L(x, c) = \frac{\displaystyle\sum_{k=1}^{m} \left(d_{k,l} - Y_{k,l} \frac{d_k}{Y_k} \right)}{\sqrt{\displaystyle\sum_{k=1}^{m} \frac{Y_{k,l}}{Y_k} \left(1 - \frac{Y_{k,l}}{Y_k} \right) \left(\frac{Y_k - d_k}{Y_k - 1} \right) d_k}}. \tag{5.6}$$

The value $|L(x, c)|$ is a measure of node separation. Larger values of $|L(x, c)|$ imply a greater the difference between the two groups, and the better the split.

Going back to our original problem to predict continuous, categorical and censored outputs simultaneously, one could integrate the log-rank splitting rule in (6) in the composite splitting rule (5) introduced by [Ishwaran and Kogalur, 2020]. The proposed approach is to implement a new composite splitting rule

$$\theta(x, c) = w_1 D^*(x, c) + w_2 G^*(x, c) + w_3 |L(x, c)| \tag{5.7}$$

where w_1, w_2, w_3 can be input parameters either to be picked by the user to control the overall contribution of each different outcome (i.e. one might be more interested in the overall survival then accurately predicting the presence of nausea) or these can be parameters to be tuned to improve overall prediction performance.

Future work includes the testing of this method on real and simulated data and the comparison of with the standard approach of predicting each outcome separately.

5.3 Conclusion

In this work we presented a new set of data-driven models and algorithms with practical applications to problems in personalized cancer therapy. We briefly summarize the key contributions of each chapter here.

In Chapter 2 we study the problem of identifying sets of mutually exclusive alterations associated with a quantitative target profile. We provide a combinatorial formulation for the problem, proving that the corresponding computational problem is NP-hard. We design two efficient algorithms, a greedy algorithm and an ILP-based algorithm, for the identification of sets of mutually exclusive alterations associated with a target profile. We provide a formal analysis for our greedy algorithm, proving that it returns solutions with rigorous guarantees in the worst-case as well under a reasonable generative model for the data. We implemented our algorithms in our method UNCOVER, and showed that it finds sets of alterations with a significant association with target profiles in a variety of scenarios. By comparing the results of UNCOVER with the results of REVEALER on four target profiles used in the REVEALER publication and on a large dataset from the GDSC project, we show that UNCOVER identifies better solutions than REVEALER, even when evaluated using REVEALER objective function. Moreover, UNCOVER is much faster than REVEALER, allowing the analysis of large datasets such as the dataset from project Achilles and from the GDSC project, in which UNCOVER identifies a number of associations between functional target profiles and gene set alterations.

In Chapter 3 we developed a new computational method, NETPHIX (NETwork-to-PHenotpe assocIation with eXclusivity), for the identification of mutated subnetworks that are associated with a continuous phenotype. Using simulations and analyzing a large scale drug screening dataset, we showed that NETPHIX can uncover the subnetworks associated with response to cancer drugs with high precision. We found many statistically significant and

biologically relevant modules associated with drug response, including MAPK/ERK signaling related modules associated with opposite response to drugs targeting RAF, MEK and ERK genes. The genetic alteration status in many of identified modules indeed make differences in cell survival rates, as validated with an independent dataset. Overall, the modules identified by NETPHIX are in good correspondence with the action of the respective drugs, suggesting that NETPHIX can correctly identify relevant modules and the modules can thus be used to predict potential patient-specific drug combinations and to provide guidance to personalized treatment. We demonstrate that the preferential selection of mutually exclusive genes was important for a better performance of the method. Interestingly, although one might assume that genes affecting drug resistance are not necessarily functionally related to the genes increasing drug sensitivity, we found that the combined connectivity model outperforms the separate connectivity model, indicating that the two groups of genes in fact might be related.

In Chapter 4 we developed two complementing computational approaches and performed the first network level association analysis of mutation signatures with dysregulated pathways. Based on gene expression data, we identified gene modules whose expression correlates with mutation counts attributed to mutational signatures. Further analysis of these modules provided important insights into the mutagenic processes underlying specific signatures. Complementing expression analysis, we deploy the ILP based method outlined in Chapter 3 to identify subnetworks of genes whose alterations are associated with each signature. This analysis provided information about potential differences in the etiology of the signatures that could not be gained from the expression analysis alone. Taken together, our study demonstrates the utility of these two complementary approaches for studying mutational signatures in cancer and provided several new insights into the etiology of mutational signatures.

Appendix A

A.1 Proofs from Chapter 2

Proposition 4. *There are instances of the Target Associated k-Set such that* $W(\hat{S}) = W(S^*)/k$.

Proof. To see that the bound is tight just consider the following example. We want to pick k sets out of n sets $A_1...A_n$. Sets $A_1...A_k$ include 2 elements of respective weight $a \geq 0$ and $b = a/(k-1)$. Subset A_{k+1} includes all the elements of weight b from the previous k sets and one element with a small weight ϵ. Each of the remaining sets $A_{k+2}...A_n$ include an arbitrary number of elements with overall weight ≤ 0. We choose a penalty of value a. Note that one can choose the weights of elements in sets $A_{k+2}...A_n$ in such a way that the average of all positive normalized weights is equal to a. Clearly the optimal solution to the Target Associated k-Set problem consists of sets $A_1...A_k$ with an objective value of $k(a+b)$. The greedy algorithm will pick set A_{k+1} at the first iteration and then assign a new weight to its elements equal to $-a$. The updated weight of sets $A_1...A_k$ is now 0 and the algorithm will stop and output A_{k+1} as the solution, giving an approximation ratio of

$$\frac{kb+\epsilon}{k(a+b)} = \frac{1}{k} + \frac{\epsilon}{kb}$$

□

Proposition 5. *If* $m \in \Omega\left(k^2 \ln(n/\delta)\right)$ *samples from the generative model above are provided to the greedy algorithm, then the solution of the greedy algorithm is H with probability* $\geq \delta$.

Proof. We prove that in iteration i of the greedy algorithm, conditioning on the current solution being a set S with $S \subset H$, then the greedy algorithm adds a gene in $H \setminus S$ to the solution with probability $\geq delta/k$, and that the first gene added by the greedy algorithm is $g \in H$. The result then follows by union bound on the k iterations of the greedy algorithm.

Consider the first iteration of the greedy algorithm and consider a gene $g \in G$. Note that if $g \notin H$ then $\mathbf{E}[W(\{g\})] \leq 0$, since $\mathbf{E}[\sum_{j \in A_g} w_j] = 0$ because the samples in which g is mutated are taken uniformly at random while $\sum_{j \in A_g}(c_S(j) - 1) \geq 0$. If $g \in H$ by the assumptions of

the model we have $\mathbf{E}[W(\{g\})] \geq \frac{m}{kc'''}$ for a constant $c''' \geq 1$. Note that $W(\{g\})$ can be written as the sum $\sum_{i=1}^{m} X_i$ of random variables (r.v.'s) X_i where X_i is the contribution of sample i to $W(\{g\})$ with $X_i \in [-1, 1]$. By the Azuma-Hoeffding inequality [Mitzenmacher and Upfal, 2017] and union bound (on the n genes) the first gene chosen by the greedy algorithm is not gene $g \in H$ with probability $\leq e^{-\frac{2m^2}{4mk^2(c''')^2}}$ which is $\leq \delta/k$ when $m \in \Omega\left(k^2 \ln(nk/\delta)\right))$.

Now assume that in iteration i, for the current solution $S \subset H$. Consider a gene $g \in G \setminus H$, then $\mathbf{E}[W(S \cup \{g\}) - W(S)] \leq 0$, since $\mathbf{E}[\sum_{j \in \cup_{s \in S \cup g} A_s} w_j - \sum_{j \in \cup_{s \in S} A_s} w_j] \leq 0$ (by the assumptions of the model $W(S) > 0$ and the fact that alterations in $\{g\}$ are placed uniformly at random among samples) and $\mathbf{E}[\sum_{j \in \cup_{s \in S \cup g}} (c_S(j) - 1) - \sum_{j \in \cup_{s \in S}} (c_S(j) - 1)] \geq 0$ (because for each sample i, the number of alterations of $S \cup \{g\}$ in i is a superset of the number of alterations of S in i). Consider now a gene $g \in H \setminus S$: by the assumptions of the model $\mathbf{E}[W(S \cup \{g\}) - W(S)] \leq \frac{m}{kc'''}$ for a constant $c''' > 1$. Note that $\mathbf{E}[W(S \cup \{g\}) - W(S)$ can be written as the sum of $\sum_{i=1}^{m} X_i$ of random variables (r.v.'s) X_i where X_i is the contribution of sample i in the increase in weight from $W(S)$ to $W(S \cup \{g\})$, where $X_i \in [-1, 1]$. By the Azuma-Hoeffding inequality and union bound (on the $< n$ genes considered for addition by the greedy algorithm) the gene g added to S by the greedy algorithm in iteration i is not in $H \setminus S$ with probability $\leq e^{-\frac{2m^2}{4mk^2(c''')^2}}$ which is $\leq \delta/k$ when $m \in \Omega\left(k^2 \ln(nk/\delta)\right))$. □

Appendix B

B.1 Formal definition of the computational problem

We are given a graph $G = (V, E)$, with vertices $V = \{1, \ldots, n\}$ representing genes and edges E representing interactions among genes. Let P denote the set of m patients (or cell lines). For each sample $j \in P$, we are also given a phenotype profile value $w_j \in \mathbb{R}$ which quantitatively measures a phenotype (e.g., drug response in our study). Let $P_i \subseteq P$ be the set of patients in which gene $i \in V$ is altered. We say that a patient $j \in P$ is *covered* by gene $i \in V$ if $j \in P_i$ i.e. if gene i is altered in sample j. We say that a sample $j \in P$ is *covered* by a subset of genes (or vertices) $S \subseteq V$, if there exists at least one vertex v in S such that $j \in P_v$.

For simplicity of description, we start with the formulation in the case where the association is in one direction, for example, with increased drug sensitivity. Later we will show how to extend the problem to accommodate the case where mixed associations are allowed in the same module. Our goal is to identify a connected subgraph S of G of at most k vertices such that the sum of the weights of the samples covered by S is maximized. The weights are computed based on drug sensitivity. Since we are interested in functionally complementary mutations, we also penalize coverage overlap when a sample is covered more than once by S by assigning a penalty p_j for each of the additional times sample j is covered by S. Let $c_S(j)$ be the number of times element $j \in P$ is covered by S. For a set S of genes, we define its weight $W(S)$ as:

$$W(S) = \sum_{j \in \cup_{s \in S} P_s} w_j - \sum_{j \in \cup_{s \in S} P_s} (c_S(j) - 1) p_j \tag{B.1}$$

Thus, we define the optimization problem for one-side association as follows: Given a graph G defined on a set of n vertices V, a set P, a family of subsets $P = \{P_1, \ldots, P_n\}$ where for each i, $P_i \subseteq P$ is associated with $i \in V$, weights w_j and penalties $p_j \geq 0$ for each sample $j \in P$, find the subset $S \subseteq V$ of $\leq k$ connected vertices maximizing $W(S)$.

Since genetic alterations may affect the increase or decrease of drug sensitivity, we extend the problem to identify genes with associations in both directions in one module. Considering genes with increased and decreased sensitivity simultaneously can pick up stronger signals of

associations and allow to take into account the interactions between alterations affecting drug responses in different ways. Let I include the genes associated with increased sensitivity overall (i.e., genes i with positive total weights, $\sum_{j \in P_i} w_j \geq 0$) and D is the set of genes associated with decreased sensitivity overall (i.e., genes i with negative total weights, $\sum_{j \in P_i} w_j < 0$). Our objective function is then defined as follows:

$$W(S) = \sum_{j \in \cup_{s \in S \bigcap I} P_s} w_j^I - \sum_{j \in \cup_{s \in S \bigcap I} P_s} (c_{S \bigcap I}(j) - 1) p_j^I + (\sum_{j \in \cup_{s \in S \bigcap D} P_s} w_j^D - \sum_{j \in \cup_{s \in S \bigcap D} P_s} (c_{S \bigcap D}(j) - 1) p_j^D) \tag{B.2}$$

where we define $w_j^I = w_j$ and $w_j^D = -w_j$. We considered two versions of connectivity constraints among the associated genes as illustrated in Fig. 3.1b. In the first model, we insisted that all selected genes should be connected whether they are associated with increased or decreased sensitivity. In the second model, we ensured the connectivity of genes with the same direction of association, resulting in two connected components in a solution (one for increased and the other for decreased sensitivity).

Although the problem is NP-hard (by a reduction to set cover) even for the simple one-sided case without network constraints, we formulated it as an integer linear program as described in the next subsection, and solved it to optimality using CPLEX, which can be run in a reasonable amount of time (See Fig. B.2 for running times for the simulation instances with different k's). For the instances requiring a large amount of resources solving ILP, we set the time limit of 24h and the memory space limit of 10 GB.

B.2 ILP formulation

Let x_i be a binary variable (denoted with $x_i \in \mathbb{B}$) equal to 1 if gene $i \in V$ is selected and $x_i = 0$ otherwise. Let z_j^I(resp., z_j^D) be a binary variable equal to 1 if sample j is covered by a gene $i \in I$ (resp., $i \in D$) and 0 otherwise. Let $y_j^I(resp., y_j^D)$ denote the number of genes in I (resp., D) cover sample j in the solution. Finally, let w_j be the weight of sample j and p_j be the penalty for sample j. When sample j is covered by a gene in I, the weight and penalty remain the same $w_j^I = w_j$. When j is covered by a gene in D, $w_j^D = -w_j$. Our ILP formulation for the combined model is defined as follows:

$$z(q) = \max \quad \sum_j (w_j^I + p_j^I) z_j^I - \sum_j p_j^I y_j^I + \sum_j (w_j^D + p_j^D) z_j^D - \sum_j p_j^D y_j^D \tag{B.3}$$

$$\text{s.t.} \sum_i x_i \leq k, \tag{B.4}$$

$$y_j^I = \sum_{i:j \in P_i, i \in I} x_i, \qquad \forall j \tag{B.5}$$

$$y_j^D = \sum_{i:j \in P_i, i \in D} x_i, \qquad \forall j \tag{B.6}$$

$$y_j^I \geq z_j^I, \qquad \forall j \tag{B.7}$$

$$y_j^D \geq z_j^D, \qquad \forall j \tag{B.8}$$

$$z_j^I \geq y_j^I / k, \qquad \forall j \tag{B.9}$$

$$z_j^D \geq y_j^D / k, \qquad \forall j \tag{B.10}$$

$$x_i, z_j \in \mathbb{B}, y_j \in \mathbb{D} \qquad \forall i, j \tag{B.11}$$

$$\sum_{l:il \in E} x_l \geq C(k-1)(x_i - 1) + C\left(\sum_{l \in V} x_l - 1\right) \qquad \forall i \in V \tag{B.12}$$

Constraint (B.4) impose that the total number of sets (i.e., selected genes) in the solution is at most k. Constraints (B.5) and (B.6) define how many times each sample has been covered by genes in I and D, respectively. Constraints (B.7) (resp., Constraints (B.8)) ensure that for each sample $j \in P$, if j is covered by increased (resp., decreased) sensitivity genes in the current solution then the number of times j is covered by I (resp., D) in the solution is at least 1. Constraints (B.9) (resp., Constraints (B.10)) impose that for each element (sample) $j \in P$, if j is covered by at least one increased (resp., decreased) sensitivity gene in the current solution then j is covered by I (resp., D).

Constraints (B.12) were used to ensure the high connectivity of a selected module (the combined connectivity model). Specifically, the constraints enforce that each selected gene is connected with at least C fraction of genes in the selected module (other than the gene itself). Note that if $C \geq 0.5$, the module is a connected subgraph since for any two non-adjacent vertices, they must have a common neighbor ($C = 0.5$ is used in our analysis). In our study, we used a functional interaction network (from STRING database), which is relatively dense. For sparse networks where highly connected components are rare, we may use an alternative approach based on a branch-and-cut algorithm to ensure the connectivity [Bomersbach et al., 2016; Fischetti et al., 2017; Wang et al., 2017].

Note that Constraints (B.12) forces the connectivity among all selected genes regardless of the directions of association. For the separate connectivity model, we identify candidate modules so that the connectivity is only enforced among the genes in I and D, separately. In this case, we replace the connectivity constraints given in (B.12) with the following constraints.

$$\sum_{l:il\in E, l\in I} x_l \geq C(k-1)(x_i-1) + C\left(\sum_{l\in I} x_l - 1\right) \quad \forall i \in I \tag{B.13}$$

$$\sum_{l:il\in E, l\in D} x_l \geq C(k-1)(x_i-1) + C\left(\sum_{l\in D} x_l - 1\right) \quad \forall i \in D \tag{B.14}$$

B.3 Selecting final modules.

By computing the optimal solutions of ILP instances with different sizes k ($k = 1$ to 5) and two connectivity options (the combined and separate model), we first obtain a pool of candidate modules. For each candidate module, we run a permutation test to assess the statistical significance of association and select maximal modules among significantly associated ones. Note that we allow to choose multiple modules associated with a drug in the final solution because it is possible that multiple functional components are associated with drug response.

Permutation test: For each candidate module, we assess the statistical significance of the association between their alteration profile and drug response by a phenotype permutation test. In the phenotype permutation, the dependencies among alterations in genes are maintained, while the association between alterations and the phenotype is removed. Specifically, a permuted dataset under the null distribution is obtained as follows: the graph $G = (V, E)$ and the sets $P_i, i \in V$ are the same as observed in the data; the values of the phenotype are randomly permuted across the samples (Fig. 3.1c). Once we find the optimal solution for the original instance, we can run ILP as a feasibility test simply checking if a permuted instance has a solution with objective value that is greater than or equal to the optimal.

To estimate the p-value for the solutions obtained by ILP, we used the following standard procedure: 1) we run an algorithm on the real data $\mathcal{D}$, obtaining a solution with objective function $o_{\mathcal{D}}$; 2) we generate N permuted datasets as described above; 3) we run the same algorithm on each permuted dataset; 4) the p-value is then given by $(e+1)/(N+1)$, where e is the number of permuted datasets in which our algorithm found a solution with objective function $\geq o_{\mathcal{D}}$. We used $N = 100$ permutations in our analysis and considered the modules with $p < 0.05$ (FDR $< 10\%$, BH) as significantly associated modules.

Selecting maximal modules: Among all significantly associated modules obtained based on the permutation test, we remove redundant modules by selecting only maximal modules. In other words, let $M_1, M_2, ..., M_t$ be the set of significantly associated modules for a drug. For any two modules M_i and M_j such that $M_i \subset M_j$ then we only include M_j in the final solution for the drug.

B.4 Datasets and Method Details

Drug sensitivity dataset: The Genomics of Drug Sensitivity in Cancer Project (`https://www.cancerrxgene.org/`) consists of drug sensitivity data generated from high-throughput screening using fluorescence-based cell viability assays following 72 hours of drug treatment. In particular, we considered the area under the curve for each experiment as a phenotype. These scores are provided in the file `portal-GDSC_AUC-201806-21.txt` available through the DepMap data portal (`https://depmap.org`) for 265 compounds and 743 cell lines, with 736 having alteration data available through the DepMap portal. For the DepMap experiments [Barretina et al., 2012a; Stransky et al., 2015], we used the alteration provided at `https://depmap.org/portal/download/all/`. We downloaded the data on July 6^{th} 2018. In particular we used mutation data from the file `portal-mutation-201806-21.csv` that includes binary entries for 18,652 gene-level mutations. Additionally, we considered 22,746 amplifications and 22,746 deletions computed from the gene copy number data in `portal-copy_number_relative-2018-06-21.csv`, with an amplification defined by a copy number above 2 and a deletion defined by a copy number below -1.

Preprocessing drug sensitivity data: For every drug response profile, we excluded samples with missing values for that phenotype, which results in a different number of samples for each phenotype. The number of samples varied between 240 and 705. To generate drug sensitivity values for the patients, we took the negatives of cell viability (i.e., increased cell survival indicates decreased sensitivity to the drug and vice versa) and then normalized the phenotype values before running the algorithm, by using standard z-scores (subtracting the average value $\sum_{j\in J} w_j/m$ from each weight w_j and dividing the result by the standard deviation of the (original) w_j's), in order to have both positive and negative phenotype values. We excluded genes with low (present in less than 1% samples from our analyses. As penalty for increased sensitivity p_j^I, we use the average of the positive phenotype values if the original value of the element was positive ($w_j > 0$) and assign a penalty equal to its absolute value otherwise. The penalty for decreased sensitivity p_j^D is computed in the opposite way. The negative of the average of the negative phenotype values is used if the original value of the element was negative ($w_j < 0$) and assign a penalty equal to its absolute value otherwise

Interaction network and computing distances in the network: For functional interactions among genes, we used the data downloaded from STRING database version 10.0 [str]. We only included the edges with high confidence scores (≥ 900 out of 1000) as an input to NETPHIX. The resulting interaction network includes 9,215 nodes and 160,249 edges.

For the average distances within modules, we computed the pairwise shortest distances within modules and take the average distances. For each drug, we only used the drug targets present in the functional network that are reachable from the selected modules and computed the average distance for all pairs of genes.

Running simulated experiments: For the background of simulation data, we use the same gene alteration table and interactions from drug sensitivity dataset described previously in this section. The phenotype values for individual samples are randomly drawn from normal distribution $N(0, 1)$. We then planted randomly generated phenotypes and associated modules to the background as follows.

Phenotypes: α fraction of patients $P(\alpha)$ ($\alpha = 0.1, 0.2$, and 0.3) were randomly selected and assigned phenotype values drawn randomly from $N(z, 0.5)$ where z is a z-score corresponding to a cumulative p-value p ($p = 0.005$, 0.1, 0.99, and 0.995).

Associated gene modules: we randomly selected a gene set $S(k)$ of size k ($k = 3, 4$, and 5) and added random alterations in $S(k)$ for patients $P(\alpha)$ so that each patient in $P(\alpha)$ has an alteration in exactly one gene in $S(k)$. Therefore, the added alterations among the patients $P(\alpha)$ are mutually exclusive although there may be overlapping mutations due to the background alterations. We also added random edges among the genes $S(k)$ so that they satisfy the density constraints ($C = 0.5$)

We generated 10 random instances for each combination of parameters (k, α, z) and ran the module identification algorithms.

For LOBICO [Knijnenburg et al., 2016], we used its R implementation [rlo] with the default parameter settings, except the logic function parameters (K and M) and the maximum running time. The OR logic model with $K = k$ and $M = 1$ was used for increased sensitivity modules and the AND logic module with $K = 1$ and $M = k$ for decreased sensitivity modules, where k is the size of the searched module. We limited the running time of LOBICO to be 24h and reported the best current solution (which may be suboptimal) when the program stops.

Validation of identified modules with CTRP dataset: For validation of NETPHIX modules, we utilized an independent drug response dataset from the Cancer Therapeutics Response Portal (CTRP) [Seashore-Ludlow et al., 2015]. The drug screening results were downloaded from https://portals.broadinstitute.org/ctrp/ (Version 2). The area under the curve (AUC) values were used for drug response phenotypes. For the alteration profiles for the cell lines, we used `CCLE_MUT_CNA_AMP_DEL_binary_Revealer.gct` downloaded from `https://portals.broadinstitute.org/ccle/data` (08/21/2017).

We found the drug response profiles for 76 drugs in both CTRP and GDSC datasets, among which 69 drugs have at least one drug sensitivity module identified by NETPHIX. 821 cell lines having both drug response and gene alteration profiles were used for validation. To test if the alteration status of selected genes are associated with different drug response, we divided the cell lines into three groups; The cell lines (C_I) with alterations in increased sensitivity genes but no alterations in decreased sensitivity genes, the cell lines (C_D) with alterations in decreased sensitivity genes but no alterations in increased sensitivity genes, and the cell lines (C_N) with no mutations in the identified genes. We then performed ANOVA test for the cell survival rates for the three groups (C_I, C_D, and C_N).

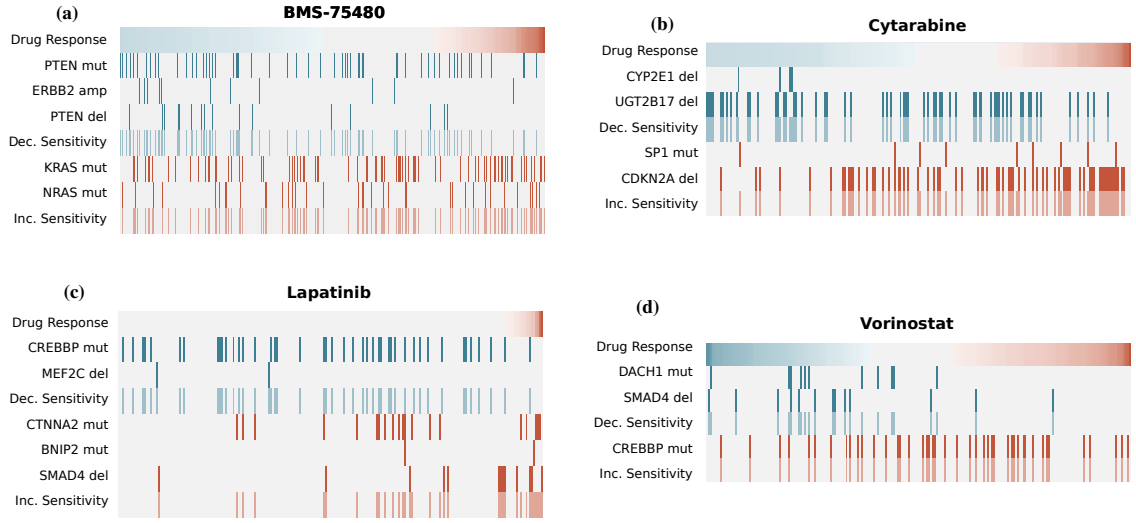


Figure B.1. (a) Sensitivity module for Cytarabine identified based on the combined connectivity model. (b) Sensitivity module for JQ1 identified based on the separate connectivity model. (c-d) Sensitivity module for Lapatinib (c) and Vorinostat (d). The two modules associated with the drugs are similar but they are associated with opposite directions. The efficacy of combination therapy with Lapatinib and Vorinostat is confirmed in clinical trials.

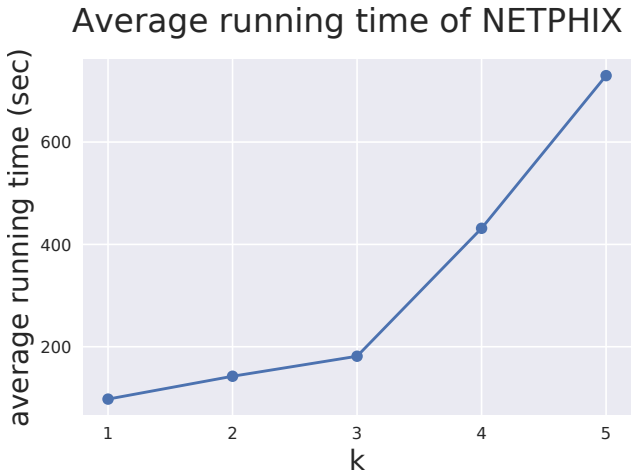


Figure B.2. The average running times of NETPHIX over different k's

Appendix C

C.1 Supplemental Methods for Chapter 4

Construction of Gene Alteration Table

The gene level alteration information for the input to NETPHIX is constructed by utilizing all somatic point mutations and small indels for the same 560 patients data. In general, we defined a gene g to be altered for a patient p if it has at least one "valid" mutation in the genomic region of g for p. The definition of "valid" mutations can be different for each signature as we further refined the information by removing mutations attributed to the signature. For example, the input alteration table used for the association with Signature 2 is constructed after removing all somatic mutations assigned to Signature 2. Formally, for the alteration table ALT_i used for association with Signature i, a gene g in ALT_i is defined to be altered only if it has at least one non-silent mutation in the genomic region of g that is not attributed to Signature i. For ALT_3 and ALT_8, we additionally removed all indels as these signatures are believed to lead to a high burden of indels. Finally, we augmented the alteration table if the gene is annotated as being biallelic inactivated (Supplementary Table 4a and 4b from Davies et al. [2017]).

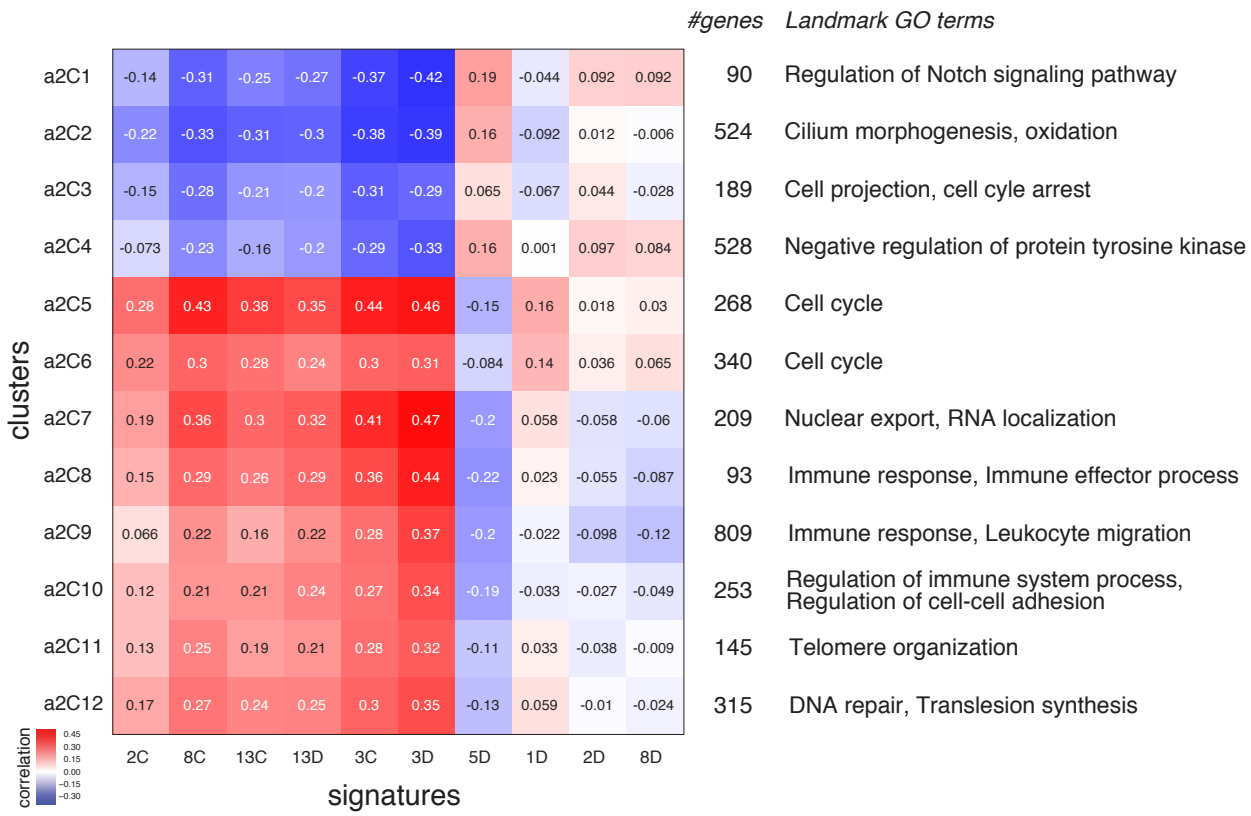


Figure C.1. Gene expression correlation modules. Clustering of all genes significantly correlated with at least one of the signatures. This shows a more fine-grained clustering (12 clusters) than in Fig. 4.2. A heatmap of mean expression correlation for each cluster and signature (left), number of genes in each cluster (middle), and representative GO terms enriched in each cluster of genes (right) are shown.

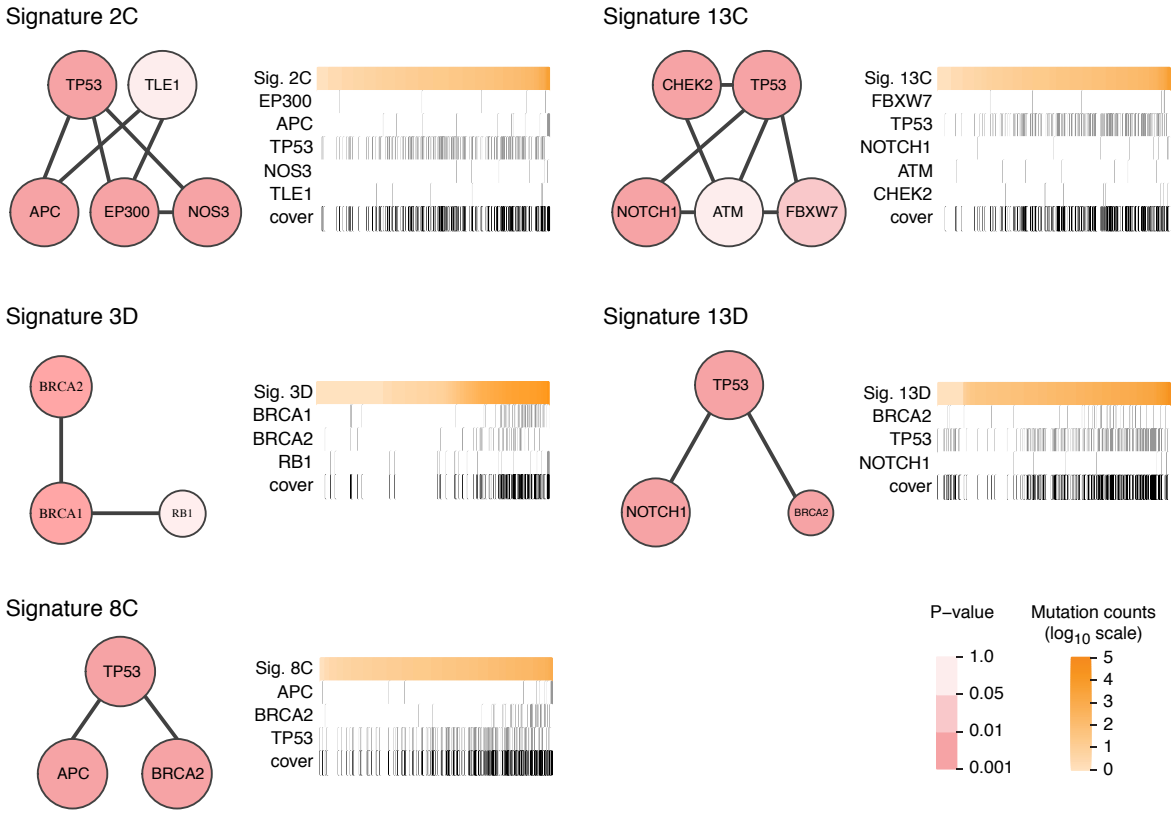


Figure C.2. Subnetworks identified by NETPHIX using less stringent cut-off (refers to Fig. 4.3). The best m (the module size) using less stringent cut-offs was selected as maximal index for which the optimal objective function increased more than 1% with respect to previous index and the phenotype p-value did not increase. Panel for each signature consists of a network view of a module (left) and a heatmap showing the association of selected gene alterations with signature strength across patients (right). The network node size indicates the gene robustness (regarding NETPHIX results for different random initialization runs of SigMa) while the darkness of red color represents its individual association score (p-value). Each heatmap shows the number of mutations attributed to a given signature for all samples (orange; top row; $\log_{10}$ scale) sorted from low to high (columns). For each gene in the module, gene mutations observed in each sample caused by other signatures are shown in gray, while samples not altered are in white. The last row shows the mutation profile of the entire subnetwork in black. Only subnetworks that changed with respect to the normal cut-offs (see Fig. 4.3 and Materials and Methods) are shown. Results for Signatures 2D and 3C did not change with respect to the normal cut-offs and results for Signatures 1D and 5D stayed insignificant (the FDR adjusted p-value above 0.1).

Table C.1. Subnetwork associated with mutational signatures for each subtype

Signature	**Subtype**	**Subnetwork**	***P*-value**
2C	Lum B	APC, TP53, SMAD4, PTEN	0.004
2D	Lum A	PIK3CA, PTEN	0.0049
3C	Lum B	BRCA2, TP53, MAP9	0.001
3C	Basal	BRCA1, BRCA2	0.023
3D	LumA	BRCA1, ARID1A, BRCA2, TP53, NF1	0.002
3D	LumB	BRCA2, TP53, MAP9	0.001
3D	Basal	BRCA1, BARD1, BRCA2, FANCA	0.002
8C	LumB	BRCA2, TP53, KRT19	0.002
13C	LumA	CASP8, TP53, AR, SIN3A, HDAC2	0.023
13C	LumB	CREBBP, BRCA2, TP53	0.003
13D	LumA	HIF1A, BRCA2, TP53, ATM, HDAC2	0.021
13D	LumB	CREBBP, BRCA2, TP53	0.001

Bibliography

Afatinib and selumetinib in advanced kras mutant and pik3ca wildtype non-small cell lung cancer (m14afs). https://clinicaltrials.gov/ct2/show/NCT02450656.

IBM ILOG CPLEX Optimization studio. https://www.ibm.com/analytics/cplex-optimizer.

Icgc data portal. https://dcc.icgc.org.

Vorinostat and lapatinib in advanced solid tumors and advanced breast cancer to evaluate response and biomarkers. https://clinicaltrials.gov/ct2/show/NCT01118975.

rlobico : An r/c++ package for building logical models. https://github.com/clareli9/rlobico, release 2018/7/27.

STRING : protein-protein association networks. https://string-db.org.

A. J. Aguirre, R. M. Meyers, B. A. Weir, F. Vazquez, C. Z. Zhang, U. Ben-David, A. Cook, G. Ha, W. F. Harrington, M. B. Doshi, M. Kost-Alimova, S. Gill, H. Xu, L. D. Ali, G. Jiang, S. Pantel, Y. Lee, A. Goodale, A. D. Cherniack, C. Oh, G. Kryukov, G. S. Cowley, L. A. Garraway, K. Stegmaier, C. W. Roberts, T. R. Golub, M. Meyerson, D. E. Root, A. Tsherniak, and W. C. Hahn. Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting. *Cancer Discov*, 6(8):914–929, Aug 2016.

L. B. Alexandrov and M. R. Stratton. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr. Opin. Genet. Dev.*, 24:52–60, Feb 2014.

L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. Aparicio, S. Behjati, et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, 2013a. ISSN 0028-0836. doi: 10.1038/nature12477. URL http://dx.doi.org/10.1038/nature12477.

L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, and M. R. Stratton. Deciphering signatures of mutational processes operative in human cancer. *Cell Reports*, 3(1):246–259, 2013b. ISSN 2211-1247. doi: 10.1016/j.celrep.2012.12.008. URL http://dx.doi.org/10.1016/j.celrep.2012.12.008.

L. B. Alexandrov, P. H. Jones, D. C. Wedge, J. E. Sale, P. J. Campbell, et al. Clock-like mutational processes in human somatic cells. *Nature Genetics*, 47(12):1402–1407, 2015. ISSN 1061-4036. doi: 10.1038/ng.3441. URL `http://dx.doi.org/10.1038/ng.3441`.

L. B. Alexandrov, Y. S. Ju, K. Haase, P. Loo, I. Martincorena, et al. Mutational signatures associated with tobacco smoking in human cancer. *Science (New York, N.Y.)*, 354(6312): 618–622, 2016. ISSN 0036-8075. doi: 10.1126/science.aag0299. URL `http://dx.doi.org/10.1126/science.aag0299`.

Y. An, J. R. Adams, D. P. Hollern, A. Zhao, S. G. Chang, M. S. Gams, P. E. D. Chung, X. He, R. Jangra, J. S. Shah, J. Yang, L. A. Beck, N. Raghuram, K. J. Kozma, A. J. Loch, W. Wang, C. Fan, S. J. Done, E. Zacksenhaus, C. J. Guidos, C. M. Perou, and S. E. Egan. Cdh1 and Pik3ca Mutations Cooperate to Induce Immune-Related Invasive Lobular Carcinoma of the Breast. *Cell Rep*, 25(3):702–714, Oct 2018.

Q. Au, D. Schalk, G. Casalicchio, R. Schödel, C. Stachl, and B. Bischl. Component-wise boosting of targets for multi-output prediction. *CoRR*, abs/1904.03943, 2019. URL `http://arxiv.org/abs/1904.03943`.

C. A. Azencott, D. Grimm, M. Sugiyama, Y. Kawahara, and K. M. Borgwardt. Efficient network-guided multi-locus association mapping with graph cuts. *Bioinformatics*, 29(13): i171–179, Jul 2013.

Ö. Babur, M. Gönen, B. A. Aksoy, N. Schultz, G. Ciriello, C. Sander, and E. Demir. Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome biology*, 16(1):45, 2015.

J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehar, G. V. Kryukov, D. Sonkin, A. Reddy, M. Liu, L. Murray, M. F. Berger, J. E. Monahan, P. Morais, J. Meltzer, A. Korejwa, J. Jane-Valbuena, F. A. Mapa, J. Thibault, E. Bric-Furlong, P. Raman, A. Shipway, I. H. Engels, J. Cheng, G. K. Yu, J. Yu, P. Aspesi, M. de Silva, K. Jagtap, M. D. Jones, L. Wang, C. Hatton, E. Palescandolo, S. Gupta, S. Mahan, C. Sougnez, R. C. Onofrio, T. Liefeld, L. MacConaill, W. Winckler, M. Reich, N. Li, J. P. Mesirov, S. B. Gabriel, G. Getz, K. Ardlie, V. Chan, V. E. Myer, B. L. Weber, J. Porter, M. Warmuth, P. Finan, J. L. Harris, M. Meyerson, T. R. Golub, M. P. Morrissey, W. R. Sellers, R. Schlegel, and L. A. Garraway. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, Mar 2012a.

J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehar, G. V. Kryukov, D. Sonkin, A. Reddy, M. Liu, L. Murray, M. F. Berger, J. E. Monahan, P. Morais, J. Meltzer, A. Korejwa, J. Jane-Valbuena, F. A. Mapa, J. Thibault, E. Bric-Furlong, P. Raman, A. Shipway, I. H. Engels, J. Cheng, G. K. Yu, J. Yu, P. Aspesi, M. de Silva, K. Jagtap, M. D. Jones, L. Wang, C. Hatton, E. Palescandolo, S. Gupta, S. Mahan, C. Sougnez, R. C. Onofrio, T. Liefeld, L. MacConaill, W. Winckler, M. Reich,

N. Li, J. P. Mesirov, S. B. Gabriel, G. Getz, K. Ardlie, V. Chan, V. E. Myer, B. L. Weber, J. Porter, M. Warmuth, P. Finan, J. L. Harris, M. Meyerson, T. R. Golub, M. P. Morrissey, W. R. Sellers, R. Schlegel, and L. A. Garraway. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, Mar 2012b.

J. C. Bolger and L. S. Young. ADAM22 as a prognostic and therapeutic drug target in the treatment of endocrine-resistant breast cancer. *Vitam. Horm.*, 93:307–321, 2013.

A. Bomersbach, M. Chiarandini, and F. Vandin. An efficient branch and cut algorithm to find frequently mutated subnetworks in cancer. In *Algorithms in Bioinformatics - 16th International Workshop, WABI 2016, Aarhus, Denmark, August 22-24, 2016. Proceedings*, pages 27–39, 2016. doi: 10.1007/978-3-319-43681-4_3. URL `https://doi.org/10.1007/978-3-319-43681-4_3`.

H. Borchani, G. Varando, C. Bielza, and P. Larrañaga. A survey on multi-output regression. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 5(5):216–233, 2015. doi: 10.1002/widm.1157. URL `https://doi.org/10.1002/widm.1157`.

L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324. URL `https://doi.org/10.1023/A:1010933404324`.

C. W. Brennan, R. G. W. Verhaak, A. McKenna, B. Campos, H. Noushmehr, S. R. Salama, S. Zheng, D. Chakravarty, J. Z. Sanborn, S. H. Berman, R. Beroukhim, B. Bernard, C.-J. Wu, G. Genovese, I. Shmulevich, J. Barnholtz-Sloan, L. Zou, R. Vegesna, S. A. Shukla, G. Ciriello, W. K. Yung, W. Zhang, C. Sougnez, T. Mikkelsen, K. Aldape, D. D. Bigner, E. G. Van Meir, M. Prados, A. Sloan, K. L. Black, J. Eschbacher, G. Finocchiaro, W. Friedman, D. W. Andrews, A. Guha, M. Iacocca, B. P. O'Neill, G. Foltz, J. Myers, D. J. Weisenberger, R. Penny, R. Kucherlapati, C. M. Perou, D. N. Hayes, R. Gibbs, M. Marra, G. B. Mills, E. Lander, P. Spellman, R. Wilson, C. Sander, J. Weinstein, M. Meyerson, S. Gabriel, P. W. Laird, D. Haussler, G. Getz, L. Chin, and TCGA Research Network. The somatic genomic landscape of glioblastoma. *Cell*, 155(2):462–77, Oct 2013. doi: 10.1016/j.cell.2013.09.034.

R. Buisson, M. S. Lawrence, C. H. Benes, and L. Zou. APOBEC3A and APOBEC3B Activities Render Cancer Cells Susceptible to ATR Inhibition. *Cancer Res.*, 77(17):4567–4578, 09 2017.

M. B. Burns, N. A. Temiz, and R. S. Harris. Evidence for apobec3b mutagenesis in multiple human cancers. *Nature Genetics*, 45(9):977–983, 2013. ISSN 1061-4036. doi: 10.1038/ng.2701. URL `http://dx.doi.org/10.1038/ng.2701`.

Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, 517(7536):576–82, Jan 2015. doi: 10.1038/nature14129.

Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, 499(7456):43–9, Jul 2013. doi: 10.1038/nature12222.

Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. *Cell*, 159(3):676–90, Oct 2014. doi: 10.1016/j.cell.2014.09.050.

A. Carracedo and P. P. Pandolfi. The PTEN-PI3K pathway: of feedbacks and cross-talks. *Oncogene*, 27(41):5527–5541, Sep 2008.

H. Carter, M. Hofree, and T. Ideker. Genotype to phenotype via network analysis. *Curr. Opin. Genet. Dev.*, 23(6):611–621, Dec 2013.

D. W. Cescon and B. Haibe-Kains. DNA replication stress: a source of APOBEC3B expression in breast cancer. *Genome Biol.*, 17(1):202, 09 2016.

Y. K. Chae, J. F. Anker, B. A. Carneiro, S. Chandra, J. Kaplan, A. Kalyan, C. A. Santa-Maria, L. C. Platanias, and F. J. Giles. Genomic landscape of DNA repair genes in cancer. *Oncotarget*, 7(17):23312–23321, Apr 2016.

S. A. Chowdhury and M. Koyuturk. Identification of coordinately dysregulated subnetworks in complex phenotypes. *Pac Symp Biocomput*, pages 133–144, 2010.

H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee, and T. Ideker. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, 3:140, 2007.

G. Ciriello, E. Cerami, C. Sander, and N. Schultz. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res*, 22(2):398–406, Feb 2012. doi: 10.1101/gr.125567.111.

G. Ciriello, E. Cerami, B. A. Aksoy, C. Sander, and N. Schultz. Using MEMo to discover mutual exclusivity modules in cancer. *Curr Protoc Bioinformatics*, Chapter 8:Unit 8.17, Mar 2013.

S. Constantinescu, E. Szczurek, P. Mohammadi, J. Rahnenführer, and N. Beerenwinkel. Timex: a waiting time model for mutually exclusive cancer alterations. *Bioinformatics*, page btv400, 2015.

G. Corso, M. Intra, C. Trentin, P. Veronesi, and V. Galimberti. CDH1 germline mutations and hereditary lobular breast cancer. *Fam. Cancer*, 15(2):215–219, Apr 2016.

G. S. Cowley, B. A. Weir, F. Vazquez, P. Tamayo, J. A. Scott, S. Rusin, A. East-Seletsky, L. D. Ali, W. F. Gerath, S. E. Pantel, P. H. Lizotte, G. Jiang, J. Hsiao, A. Tsherniak, E. Dwinell, S. Aoyama, M. Okamoto, W. Harrington, E. Gelfand, T. M. Green, M. J. Tomko, S. Gopal, T. C. Wong, T. C. Wong, H. Li, S. Howell, N. Stransky, T. Liefeld, D. Jang, J. Bistline, B. Hill Meyers, S. A. Armstrong, K. C. Anderson, K. Stegmaier, M. Reich, D. Pellman, J. S. Boehm, J. P. Mesirov, T. R. Golub, D. E. Root, and W. C. Hahn. Parallel genome-scale

loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Sci Data*, 1:140035, 2014.

P. Creixell, J. Reimand, S. Haider, G. Wu, T. Shibata, M. Vazquez, V. Mustonen, A. Gonzalez-Perez, J. Pearson, C. Sander, et al. Pathway and network analysis of cancer genomes. *Nature methods*, 12(7):615, 2015.

S. Cristea, J. Kuipers, and N. Beerenwinkel. pathtimex: Joint inference of mutually exclusive cancer pathways and their progression dynamics. *Journal of Computational Biology*, 2016.

P. Dao, Y. A. Kim, D. Wojtowicz, S. Madan, R. Sharan, and T. M. Przytycka. BeWith: A Between-Within method to discover relationships between cancer modules via integrated analysis of mutual exclusivity, co-occurrence and functional interactions. *PLoS Comput. Biol.*, 13(10):e1005695, Oct 2017a.

P. Dao, Y. A. Kim, D. Wojtowicz, S. Madan, R. Sharan, and T. M. Przytycka. BeWith: A Between-Within method to discover relationships between cancer modules via integrated analysis of mutual exclusivity, co-occurrence and functional interactions. *PLoS Comput. Biol.*, 13(10):e1005695, Oct 2017b.

H. Davies, D. Glodzik, S. Morganella, L. R. Yates, J. Staaf, et al. Hrdetect is a predictor of brca1 and brca2 deficiency based on mutational signatures. *Nature Medicine*, 23(4): 517–525, 2017. ISSN 1078-8956. doi: 10.1038/nm.4292. URL `http://dx.doi.org/10.1038/nm.4292`.

G. De'ath. Multivariate regression trees: A new technique for modeling species–environment relationships. *Ecology*, 83(4):1105–1117, 2002. doi: 10.1890/0012-9658(2002)083[1105:MRTANT]2.0.CO;2. URL `https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/0012-9658%282002%29083%5B1105%3AMRTANT%5D2.0.CO%3B2`.

A. Desrichard, Y. Bidet, N. Uhrhammer, and Y. J. Bignon. CHEK2 contribution to hereditary breast cancer in non-BRCA families. *Breast Cancer Res.*, 13(6):R119, 2011.

L. M. Dillon and T. W. Miller. Therapeutic targeting of cancers with loss of PTEN function. *Curr Drug Targets*, 15(1):65–79, Jan 2014.

T. Du, L. Zhu, K. M. Levine, N. Tasdemir, A. V. Lee, D. A. A. Vignali, B. V. Houten, G. C. Tseng, and S. Oesterreich. Invasive lobular and ductal breast carcinoma differ in immune response, protein translation efficiency and metabolism. *Sci Rep*, 8(1):7205, May 2018.

A. Fischer, C. J. Illingworth, P. J. Campbell, and V. Mustonen. Emu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biology*, 14 (4):1–10, 2013. ISSN 1474-760X. doi: 10.1186/gb-2013-14-4-r39. URL `http://dx.doi.org/10.1186/gb-2013-14-4-r39`.

M. Fischetti, M. Leitner, I. Ljubic, M. Luipersbeck, M. Monaci, M. Resch, D. Salvagnin, and M. Sinnl. Thinning out steiner trees: a node-based model for uniform edge costs. *Math. Program. Comput.*, 9(2):203–229, 2017. doi: 10.1007/s12532-016-0111-0. URL `https://doi.org/10.1007/s12532-016-0111-0`.

R. Fodde, J. Kuipers, C. Rosenberg, R. Smits, M. Kielman, C. Gaspar, J. H. van Es, C. Breukel, J. Wiegant, R. H. Giles, and H. Clevers. Mutations in the APC tumour suppressor gene cause chromosomal instability. *Nat. Cell Biol.*, 3(4):433–438, Apr 2001.

W. A. Garcia-Suastegui, L. A. Ramos-Chavez, M. Rubio-Osornio, M. Calvillo-Velasco, J. A. Atzin-Mendez, J. Guevara, and D. Silva-Adaya. The Role of CYP2E1 in the Drug Metabolism or Bioactivation in the Brain. *Oxid Med Cell Longev*, 2017:4680732, 2017.

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

S. R. Gilman, J. Chang, B. Xu, T. S. Bawa, J. A. Gogos, M. Karayiorgou, and D. Vitkup. Diverse types of genetic variation converge on functional gene networks involved in schizophrenia. *Nat. Neurosci.*, 15(12):1723–1728, Dec 2012.

A. Goncearenco, S. L. Rager, M. Li, Q. X. Sang, I. B. Rogozin, and A. R. Panchenko. Exploring background mutational processes to decipher cancer genetic heterogeneity. *Nucleic Acids Res.*, 45(W1):W514–W522, Jul 2017.

X. Gong, J. Du, S. H. Parsons, F. F. Merzoug, Y. Webster, P. W. Iversen, L. C. Chio, R. D. Van Horn, X. Lin, W. Blosser, B. Han, S. Jin, S. Yao, H. Bian, C. Ficklin, L. Fan, A. Kapoor, S. Antonysamy, A. M. Mc Nulty, K. Froning, D. Manglicmot, A. Pustilnik, K. Weichert, S. R. Wasserman, M. Dowless, C. Marugan, C. Baquero, M. J. Lallena, S. W. Eastman, Y. H. Hui, M. Z. Dieter, T. Doman, S. Chu, H. R. Qian, X. S. Ye, D. A. Barda, G. D. Plowman, C. Reinhard, R. M. Campbell, J. R. Henry, and S. G. Buchanan. Aurora A Kinase Inhibition Is Synthetic Lethal with Loss of the RB1 Tumor Suppressor Gene. *Cancer Discov*, 9(2):248–263, 02 2019.

A. M. Green, S. Landry, K. Budagyan, D. C. Avgousti, S. Shalhout, A. S. Bhagwat, and M. D. Weitzman. APOBEC3A damages the cellular genome during DNA replication. *Cell Cycle*, 15(7):998–1008, 2016.

C. Guillemette, E. Levesque, and M. Rouleau. Pharmacogenomics of human uridine diphosphoglucuronosyltransferases and clinical implications. *Clin. Pharmacol. Ther.*, 96(3):324–339, Sep 2014.

M. L. Hamilton, H. V. Remmen, J. A. Drake, H. Yang, Z. M. Guo, K. Kewitt, C. A. Walter, and A. Richardson. Does oxidative damage to dna increase with age? *Proceedings of the National Academy of Sciences*, 98(18):10469–10474, 1 2001. ISSN 0027-8424. doi: 10.1073/pnas.171202698. URL `http://dx.doi.org/10.1073/pnas.171202698`.

D. Hanahan and R. A. Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5): 646–674, 2011.

S. Hansford, P. Kaurah, H. Li-Chang, M. Woo, J. Senz, H. Pinheiro, K. A. Schrader, D. F. Schaeffer, K. Shumansky, G. Zogopoulos, T. A. Santos, I. Claro, J. Carvalho, C. Nielsen, S. Padilla, A. Lum, A. Talhouk, K. Baker-Lange, S. Richardson, I. Lewis, N. M. Lindor, E. Pennell, A. MacMillan, B. Fernandez, G. Keller, H. Lynch, S. P. Shah, P. Guilford, S. Gallinger, G. Corso, F. Roviello, C. Caldas, C. Oliveira, P. D. Pharoah, and D. G. Huntsman. Hereditary Diffuse Gastric Cancer Syndrome: CDH1 Mutations and Beyond. *JAMA Oncol*, 1(1):23–32, Apr 2015.

T. Helleday, S. Eshtad, and S. Nik-Zainal. Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.*, 15(9):585–598, Sep 2014.

D. Hochbaum and A. Pathria. Analysis of the Greedy Approach in Problems of Maximum k-Coverage. *Naval Research Logistics*, 45(6):615–627, September 1998.

M. Hofree, J. P. Shen, H. Carter, A. Gross, and T. Ideker. Network-based stratification of tumor mutations. *Nat. Methods*, 10(11):1108–1115, Nov 2013.

B. H. Hristov and M. Singh. Network-Based Coverage of Mutational Profiles Reveals Cancer Genes. *Cell Syst*, 5(3):221–229, 09 2017.

F. Huang, H. Chang, A. Greer, S. Hillerman, K. A. Reeves, W. Hurlburt, J. Cogswell, D. Patel, Z. Qi, C. Fairchild, R. P. Ryseck, T. W. Wong, F. G. Finckenstein, J. Jackson, and J. M. Carboni. IRS2 copy number gain, KRAS and BRAF mutation status as predictive biomarkers for response to the IGF-1R/IR inhibitor BMS-754807 in colorectal cancer cell lines. *Mol. Cancer Ther.*, 14(2):620–630, Feb 2015.

X. Huang, I. Sason, D. Wojtowicz, Y.-A. Kim, M. Leiserson, T. M. Przytycka, and R. Sharan. Hidden markov models lead to higher resolution maps of mutation signature activity in cancer. *bioRxiv*, 2018a. doi: 10.1101/392639. URL `https://www.biorxiv.org/content/early/2018/08/16/392639`.

X. Huang, D. Wojtowicz, and T. M. Przytycka. Detecting presence of mutational signatures in cancer with confidence. *Bioinformatics*, 34(2):330–337, 2018b. doi: 10.1093/bioinformatics/btx604. URL `http://dx.doi.org/10.1093/bioinformatics/btx604`.

F. Iorio, T. A. Knijnenburg, D. J. Vis, G. R. Bignell, M. P. Menden, M. Schubert, N. Aben, E. Goncalves, S. Barthorpe, H. Lightfoot, T. Cokelaer, P. Greninger, E. van Dyk, H. Chang, H. de Silva, H. Heyn, X. Deng, R. K. Egan, Q. Liu, T. Mironenko, X. Mitropoulos, L. Richardson, J. Wang, T. Zhang, S. Moran, S. Sayols, M. Soleimani, D. Tamborero, N. Lopez-Bigas, P. Ross-Macdonald, M. Esteller, N. S. Gray, D. A. Haber, M. R. Stratton, C. H. Benes, L. F. A. Wessels, J. Saez-Rodriguez, U. McDermott, and M. J. Garnett. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*, 166(3):740–754, Jul 2016.

H. Ishwaran. The effect of splitting on random forests. *Mach. Learn.*, 99(1):75–118, 2015. doi: 10.1007/s10994-014-5451-2. URL `https://doi.org/10.1007/s10994-014-5451-2`.

H. Ishwaran and U. Kogalur. *Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)*, 2020. URL `https://cran.r-project.org/package=randomForestSRC`. R package version 2.9.3.

A. S. Jaiswal and S. Narayan. A novel function of adenomatous polyposis coli (APC) in regulating DNA repair. *Cancer Lett.*, 271(2):272–280, Nov 2008.

P. Jia, S. Zheng, J. Long, W. Zheng, and Z. Zhao. dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. *Bioinformatics*, 27(1):95–102, Jan 2011.

C. Kandoth, M. D. McLellan, F. Vandin, K. Ye, B. Niu, C. Lu, M. Xie, Q. Zhang, J. F. McMichael, M. A. Wyczalkowski, et al. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471):333–339, 2013.

M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, 45(D1):D353–D361, Jan 2017.

P. Kaurah, A. MacMillan, N. Boyd, J. Senz, A. De Luca, N. Chun, G. Suriano, S. Zaor, L. Van Manen, C. Gilpin, S. Nikkel, M. Connolly-Wilson, S. Weissman, W. S. Rubinstein, C. Sebold, R. Greenstein, J. Stroop, D. Yim, B. Panzini, W. McKinnon, M. Greenblatt, D. Wirtzfeld, D. Fontaine, D. Coit, S. Yoon, D. Chung, G. Lauwers, A. Pizzuti, C. Vaccaro, M. A. Redal, C. Oliveira, M. Tischkowitz, S. Olschwang, S. Gallinger, H. Lynch, J. Green, J. Ford, P. Pharoah, B. Fernandez, and D. Huntsman. Founder and recurrent CDH1 mutations in families with hereditary diffuse gastric cancer. *JAMA*, 297(21):2360–2372, Jun 2007.

N. Keshelava, J. J. Zuo, P. Chen, S. N. Waidyaratne, M. C. Luna, C. J. Gomer, T. J. Triche, and C. P. Reynolds. Loss of p53 function confers high-level multidrug resistance in neuroblastoma cell lines. *Cancer Research*, 61(16):6185–6193, Aug. 2001. ISSN 0008-5472.

e. a. Kim, Jong. Characterizing genomic alterations in cancer by complementary functional associations. *Nature Biotechnology*, 34(5):539–546, May 2016.

J. Kim, K. W. Mouw, P. Polak, L. Z. Braunstein, A. Kamburov, D. J. Kwiatkowski, J. E. Rosenberg, E. M. Van Allen, A. D'Andrea, and G. Getz. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.*, 48(6): 600–606, 06 2016a.

Y. A. Kim, S. Wuchty, and T. M. Przytycka. Identifying causal genes and dysregulated pathways in complex diseases. *PLoS Comput. Biol.*, 7(3):e1001095, Mar 2011.

Y. A. Kim, R. Salari, S. Wuchty, and T. M. Przytycka. Module cover - a new approach to genotype-phenotype studies. *Pac Symp Biocomput*, pages 135–146, 2013.

Y. A. Kim, D. Y. Cho, P. Dao, and T. M. Przytycka. MEMCover: integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types. *Bioinformatics*, 31(12):i284–292, Jun 2015.

Y. A. Kim, D. Y. Cho, and T. M. Przytycka. Understanding Genotype-Phenotype Effects in Cancer via Network Approaches. *PLoS Comput. Biol.*, 12(3):e1004747, Mar 2016b.

Y.-A. Kim, S. Madan, and T. M. Przytycka. Wesme: uncovering mutual exclusivity of cancer drivers and beyond. *Bioinformatics*, page btw242, 2016c.

Y.-A. Kim, R. Sarto Basso, D. Wojtowicz, D. S. Hochbaum, F. Vandin, and T. M. Przytycka. Identifying drug sensitivity subnetworks with netphix. *bioRxiv*, 2019. doi: 10.1101/543876. URL `https://www.biorxiv.org/content/early/2019/02/08/543876`.

H. King, T. Aleksic, P. Haluska, and V. M. Macaulay. Can we unlock the potential of IGF-1R inhibition in cancer therapy? *Cancer Treat. Rev.*, 40(9):1096–1105, Oct 2014.

T. A. Knijnenburg, G. W. Klau, F. Iorio, M. J. Garnett, U. McDermott, I. Shmulevich, and L. F. Wessels. Logic models to predict continuous outputs based on binary inputs with an application to personalized cancer therapy. *Sci Rep*, 6:36812, 11 2016.

T. A. Knijnenburg, L. Wang, M. T. Zimmermann, N. Chambwe, and et al. Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Rep*, 23(1):239–254, Apr 2018.

A. P. Kouzmenko, K. Takeyama, Y. Kawasaki, T. Akiyama, and S. Kato. Truncation mutations abolish chromatin-associated activities of adenomatous polyposis coli. *Oncogene*, 27(36):4888–4899, Aug 2008.

M. J. LaBonte, P. M. Wilson, W. Fazzone, J. Russell, S. G. Louie, A. El-Khoueiry, H. J. Lenz, and R. D. Ladner. The dual EGFR/HER2 inhibitor lapatinib synergistically enhances the antitumor activity of the histone deacetylase inhibitor panobinostat in colorectal cancer models. *Cancer Res.*, 71(10):3635–3648, May 2011.

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.

M. D. Leiserson, H.-T. Wu, F. Vandin, and B. J. Raphael. Comet: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome biology*, 16(1): 160, 2015a.

M. D. M. Leiserson, F. Vandin, H.-T. Wu, J. R. Dobson, J. V. Eldridge, J. L. Thomas, A. Papoutsaki, Y. Kim, B. Niu, M. McLellan, M. S. Lawrence, A. Gonzalez-Perez, D. Tamborero, Y. Cheng, G. A. Ryslik, N. Lopez-Bigas, G. Getz, L. Ding, and B. J. Raphael. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet*, 47(2):106–14, Feb 2015b. doi: 10.1038/ng.3168.

B. Leonard, S. N. Hart, M. B. Burns, M. A. Carpenter, N. A. Temiz, A. Rathore, R. I. Vogel, J. B. Nikas, E. K. Law, W. L. Brown, Y. Li, Y. Zhang, M. J. Maurer, A. L. Oberg, J. M. Cunningham, V. Shridhar, D. A. Bell, C. April, D. Bentley, M. Bibikova, R. K. Cheetham, J. B. Fan, R. Grocock, S. Humphray, Z. Kingsbury, J. Peden, J. Chien, E. M. Swisher, L. C. Hartmann, K. R. Kalli, E. L. Goode, H. Sicotte, S. H. Kaufmann, and R. S. Harris. APOBEC3B upregulation and genomic mutation patterns in serous ovarian carcinoma. *Cancer Res.*, 73(24):7222–7231, Dec 2013.

C. Li and H. Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, May 2008.

S. Li, S. Liu, J. Deng, E. A. Akbay, J. Hai, C. Ambrogio, L. Zhang, F. Zhou, R. W. Jenkins, D. O. Adeegbe, P. Gao, X. Wang, C. P. Paweletz, G. S. Herter-Sprie, T. Chen, L. Gutiérrez-Quiceno, Y. Zhang, A. A. Merlino, M. M. Quinn, Y. Zeng, X. Yu, Y. Liu, L. Fan, A. J. Aguirre, D. A. Barbie, X. Yi, and K.-K. Wong. Assessing Therapeutic Efficacy of MEK Inhibition in a KRASG12c-Driven Mouse Model of Lung Cancer. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 24(19):4854–4864, 2018. ISSN 1078-0432. doi: 10.1158/1078-0432.CCR-17-3438.

X. Lin, J. Liao, Z. Yang, X. Fan, K. J. Cullen, L. Chen, and H. Dan. Inhibition of cisplatin-resistant head and neck squamous cell carcinoma by combination of Afatinib with PD0325901, a MEK inhibitor. *Am J Cancer Res*, 9(6):1282–1292, 2019.

Y. Liu, M. Brossard, D. Roqueiro, P. Margaritte-Jeannin, C. Sarnowski, E. Bouzigon, and F. Demenais. SigMod: an exact and efficient method to identify a strongly interconnected disease-associated module in a gene network. *Bioinformatics*, 33(10):1536–1544, May 2017.

J. Ma, J. Setton, N. Y. Lee, N. Riaz, and S. N. Powell. The therapeutic significance of mutational signatures from DNA repair deficiency in cancer. *Nat Commun*, 9(1):3292, 08 2018.

S. Masciari, N. Larsson, J. Senz, N. Boyd, P. Kaurah, M. J. Kandel, L. N. Harris, H. C. Pinheiro, A. Troussard, P. Miron, N. Tung, C. Oliveira, L. Collins, S. Schnitt, J. E. Garber, and D. Huntsman. Germline E-cadherin mutations in familial lobular breast cancer. *J. Med. Genet.*, 44(11):726–731, Nov 2007.

D. McCartan, J. C. Bolger, A. Fagan, C. Byrne, Y. Hao, L. Qin, M. McIlroy, J. Xu, A. D. Hill, P. O. Gaora, and L. S. Young. Global characterization of the SRC-1 transcriptome

identifies ADAM22 as an ER-independent mediator of endocrine-resistant breast cancer. *Cancer Res.*, 72(1):220–229, Jan 2012.

N. McGranahan and C. Swanton. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell*, 168(4):613–628, 2017.

H. Meijers-Heijboer, A. van den Ouweland, J. Klijn, M. Wasielewski, A. de Snoo, R. Oldenburg, A. Hollestelle, M. Houben, E. Crepin, M. van Veghel-Plandsoen, F. Elstrodt, C. van Duijn, C. Bartels, C. Meijers, M. Schutte, L. McGuffog, D. Thompson, D. Easton, N. Sodha, S. Seal, R. Barfoot, J. Mangion, J. Chang-Claude, D. Eccles, R. Eeles, D. G. Evans, R. Houlston, V. Murday, S. Narod, T. Peretz, J. Peto, C. Phelan, H. X. Zhang, C. Szabo, P. Devilee, D. Goldgar, P. A. Futreal, K. L. Nathanson, B. Weber, N. Rahman, and M. R. Stratton. Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat. Genet.*, 31(1):55–59, May 2002.

J. P. Melis, H. van Steeg, and M. Luijten. Oxidative DNA damage and nucleotide excision repair. *Antioxid. Redox Signal.*, 18(18):2409–2419, Jun 2013.

V. Meniel, M. Megges, M. A. Young, A. Cole, O. J. Sansom, and A. R. Clarke. Apc and p53 interaction in DNA damage and genomic instability in hepatocytes. *Oncogene*, 34(31): 4118–4129, Jul 2015.

C. A. Miller, S. H. Settle, E. P. Sulman, K. D. Aldape, and A. Milosavljevic. Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. *BMC Med Genomics*, 4:34, 2011. doi: 10.1186/1755-8794-4-34.

D. P. Minde, Z. Anvarian, S. G. Rudiger, and M. M. Maurice. Messing up disorder: how do missense mutations in the tumor suppressor protein APC lead to cancer? *Mol. Cancer*, 10: 101, Aug 2011.

M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge university press, 2017.

M. Molina-Arcas, D. C. Hancock, C. Sheridan, M. S. Kumar, and J. Downward. Coordinate direct input of both KRAS and IGF1 receptor to activation of PI3 kinase in KRAS-mutant lung cancer. *Cancer Discov*, 3(5):548–563, May 2013.

V. K. Mootha, C. M. Lindgren, K. F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, 34(3):267–273, Jul 2003.

P. Moreno-Muñoz, A. Artés-Rodríguez, and M. A. Álvarez. Heterogeneous multi-output gaussian process prediction. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 6712–6721, 2018. URL `http://papers.nips.cc/paper/7905-heterogeneous-multi-output-gaussian-process-prediction`.

S. Morganella, L. B. Alexandrov, D. Glodzik, X. Zou, H. Davies, et al. The topography of mutational processes in breast cancer genomes. *Nature Communications*, 7:11383, 2016. ISSN 2041-1733. doi: 10.1038/ncomms11383. URL `http://dx.doi.org/10.1038/ncomms11383`.

A. Najem, M. Krayem, F. Salès, N. Hussein, B. Badran, C. Robert, A. Awada, F. Journe, and G. E. Ghanem. P53 and MITF/Bcl-2 identified as key pathways in the acquired resistance of NRAS-mutant melanoma to MEK inhibition. *European Journal of Cancer (Oxford, England: 1990)*, 83:154–165, 2017. ISSN 1879-0852. doi: 10.1016/j.ejca.2017.06.033.

N. Nakayama, K. Nakayama, S. Yeasmin, M. Ishibashi, A. Katagiri, K. Iida, M. Fukumoto, and K. Miyazaki. KRAS or BRAF mutation status is a useful predictor of sensitivity to MEK inhibition in ovarian cancer. *British Journal of Cancer*, 99(12):2020–2028, Dec. 2008. ISSN 1532-1827. doi: 10.1038/sj.bjc.6604783.

C. G. A. R. Network et al. Integrated genomic characterization of oesophageal carcinoma. *Nature*, 541(7636):169–175, 2017a.

C. G. A. R. Network et al. Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer cell*, 32(2):185, 2017b.

S. Nik-Zainal, D. C. Wedge, L. B. Alexandrov, M. Petljak, A. P. Butler, N. Bolli, H. R. Davies, S. Knappskog, S. Martin, E. Papaemmanuil, M. Ramakrishna, A. Shlien, I. Simonic, Y. Xue, C. Tyler-Smith, P. J. Campbell, and M. R. Stratton. Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer. *Nat. Genet.*, 46(5):487–491, May 2014.

S. Nik-Zainal, H. Davies, J. Staaf, M. Ramakrishna, D. Glodzik, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54, 2016. ISSN 0028-0836. doi: 10.1038/nature17676. URL `http://dx.doi.org/10.1038/nature17676`.

M. Patel, N. C. Gomez, A. W. McFadden, B. M. Moats-Staats, S. Wu, A. Rojas, T. Sapp, J. M. Simon, S. V. Smith, K. Kaiser-Rogers, and I. J. Davis. PTEN deficiency mediates a reciprocal response to IGFI and mTOR inhibition. *Mol. Cancer Res.*, 12(11):1610–1620, Nov 2014.

M. Periyasamy, A. K. Singh, C. Gemma, C. Kranjec, R. Farzan, D. A. Leach, N. Navaratnam, H. L. Pálinkás, B. G. Vértessy, T. R. Fenton, J. Doorbar, F. Fuller-Pace, D. W. Meek, R. C. Coombes, L. Buluwela, and S. Ali. p53 controls expression of the dna deaminase apobec3b to limit its potential mutagenic activity in cancer cells. *Nucleic acids research*, 45(19):11056—11069, November 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx721. URL `http://europepmc.org/articles/PMC5737468`.

P. Polak, J. Kim, L. Z. Braunstein, R. Karlic, N. J. Haradhavala, G. Tiao, D. Rosebrock, D. Livitz, K. Kübler, K. W. Mouw, A. Kamburov, Y. E. Maruvka, I. Leshchiner, E. S. Lander, T. R. Golub, A. Zick, A. Orthwein, M. S. Lawrence, R. N. Batra, C. Caldas, D. A. Haber, P. W. Laird, H. Shen, L. W. Ellisen, A. D. D'Andrea, S. J. Chanock, W. D. Foulkes, and G. Getz. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nature Genetics*, 2017. ISSN 1061-4036. doi: 10.1038/ng.3934. URL `http://dx.doi.org/10.1038/ng.3934`.

R. C. Poulos, Y. T. Wong, R. Ryan, H. Pang, and J. W. H. Wong. Analysis of 7,815 cancer exomes reveals associations between mutational processes and somatic driver mutations. *PLoS Genet.*, 14(11):e1007779, 11 2018.

B. J. Raphael and F. Vandin. Simultaneous inference of cancer pathways and tumor progression from cross-sectional mutation data. *Journal of Computational Biology*, 22(6):510–527, 2015.

B. J. Raphael, J. R. Dobson, L. Oesper, and F. Vandin. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med*, 6(1):5, 2014.

S. Razick, G. Magklaras, and I. M. Donaldson. iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics*, 9:405, Sep 2008.

B. Sanchez-Laorden, A. Viros, M. R. Girotti, M. Pedersen, G. Saturno, A. Zambon, D. Niculescu-Duvaz, S. Turajlic, A. Hayes, M. Gore, J. Larkin, P. Lorigan, M. Cook, C. Springer, and R. Marais. BRAF inhibitors induce metastasis in RAS mutant or inhibitor-resistant melanoma cells by reactivating MEK and ERK signaling. *Sci Signal*, 7 (318):ra30, Mar 2014.

R. Sarto Basso, D. S. Hochbaum, and F. Vandin. Efficient algorithms to discover alterations with complementary functional association in cancer. *PLoS Comput. Biol.*, 15(5):e1006802, May 2019.

B. Seashore-Ludlow, M. G. Rees, J. H. Cheah, M. Cokol, E. V. Price, M. E. Coletti, V. Jones, N. E. Bodycombe, C. K. Soule, J. Gould, B. Alexander, A. Li, P. Montgomery, M. J. Wawer, N. Kuru, J. D. Kotz, C. S. Hon, B. Munoz, T. Liefeld, V. Dan?ik, J. A. Bittker, M. Palmer, J. E. Bradner, A. F. Shamji, P. A. Clemons, and S. L. Schreiber. Harnessing

Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discov*, 5(11): 1210–1223, Nov 2015.

V. B. Seplyarskiy, R. A. Soldatov, K. Y. Popadin, S. E. Antonarakis, G. A. Bazykin, and S. I. Nikolaev. APOBEC-induced mutations in human cancers are strongly enriched on the lagging DNA strand during replication. *Genome Res.*, 26(2):174–182, Feb 2016.

D. D. Shao, A. Tsherniak, S. Gopal, B. A. Weir, P. Tamayo, N. Stransky, S. E. Schumacher, T. I. Zack, R. Beroukhim, L. A. Garraway, A. A. Margolin, D. E. Root, W. C. Hahn, and J. P. Mesirov. ATARiS: computational quantification of gene suppression phenotypes from multisample RNAi screens. *Genome Res.*, 23(4):665–678, Apr 2013.

A. Shimizu, H. Fujimori, Y. Minakawa, Y. Matsuno, M. Hyodo, Y. Murakami, and K. I. Yoshioka. Onset of deaminase APOBEC3B induction in response to DNA double-strand breaks. *Biochem Biophys Rep*, 16:115–121, Dec 2018.

L. M. Solis, C. Behrens, W. Dong, M. Suraokar, N. C. Ozburn, C. A. Moran, A. H. Corvalan, S. Biswal, S. G. Swisher, B. N. Bekele, J. D. Minna, D. J. Stewart, and I. I. Wistuba. Nrf2 and Keap1 abnormalities in non-small cell lung carcinoma and association with clinicopathologic features. *Clin. Cancer Res.*, 16(14):3743–3753, Jul 2010.

N. Stransky, M. Ghandi, G. V. Kryukov, L. A. Garraway, J. Lehar, M. Liu, D. Sonkin, A. Kauffmann, K. Venkatesan, E. J. Edelman, M. Riester, J. Barretina, G. Caponigro, R. Schlegel, W. R. Sellers, F. Stegmeier, M. Morrissey, A. Amzallag, I. Pruteanu-Malinici, D. A. Haber, S. Ramaswamy, C. H. Benes, M. P. Menden, F. Iorio, M. R. Stratton, U. McDermott, M. J. Garnett, and J. Saez-Rodriguez. Pharmacogenomic agreement between two cancer cell line data sets. *Nature*, 528(7580):84–87, Dec 2015.

X. Su and J. Fan. Multivariate survival trees: a maximum likelihood approach based on frailty models. *Biometrics*, 60(1):93–99, Mar 2004.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, 102(43):15545–15550, Oct 2005.

C. Sun, S. Hobor, A. Bertotti, D. Zecchin, S. Huang, F. Galimi, F. Cottino, A. Prahallad, W. Grernrum, A. Tzani, A. Schlicker, L. F. Wessels, E. F. Smit, E. Thunnissen, P. Halonen, C. Lieftink, R. L. Beijersbergen, F. Di Nicolantonio, A. Bardelli, L. Trusolino, and R. Bernards. Intrinsic resistance to MEK inhibition in KRAS mutant lung and colon cancer through transcriptional induction of ERBB3. *Cell Rep*, 7(1):86–93, Apr 2014.

F. Supek and B. Lehner. Clustered mutation signatures reveal that error-prone dna repair targets mutations to active genes. *Cell*, 170(3):534–547.e23, 2017. ISSN 0092-8674. doi: 10.1016/j.cell.2017.07.003. URL `http://dx.doi.org/10.1016/j.cell.2017.07.003`.

F. Tang and H. Ishwaran. Random forest missing data algorithms. *Statistical Analysis and Data Mining*, 10(6):363–377, 2017. doi: 10.1002/sam.11348. URL `https://doi.org/10.1002/sam.11348`.

D. Temko, I. P. M. Tomlinson, S. Severini, B. Schuster-Bockler, and T. A. Graham. The effects of mutational processes and selection on driver mutations across cancer types. *Nat Commun*, 9(1):1857, 05 2018.

A. Tsherniak, F. Vazquez, P. G. Montgomery, B. A. Weir, G. Kryukov, G. S. Cowley, S. Gill, W. F. Harrington, S. Pantel, J. M. Krill-Burger, R. M. Meyers, L. Ali, A. Goodale, Y. Lee, G. Jiang, J. Hsiao, W. F. J. Gerath, S. Howell, E. Merkel, M. Ghandi, L. A. Garraway, D. E. Root, T. R. Golub, J. S. Boehm, and W. C. Hahn. Defining a Cancer Dependency Map. *Cell*, 170(3):564–576, Jul 2017.

I. Ulitsky, A. Krishnamurthy, R. M. Karp, and R. Shamir. DEGAS: de novo discovery of dysregulated pathways in human diseases. *PLoS ONE*, 5(10):e13367, Oct 2010.

F. Vandin. Computational methods for characterizing cancer mutational heterogeneity. *Frontiers in genetics*, 8:83, 2017.

F. Vandin, P. Clay, E. Upfal, and B. J. Raphael. Discovery of mutated subnetworks associated with clinical data in cancer. *Pac Symp Biocomput*, pages 55–66, 2012a.

F. Vandin, E. Upfal, and B. J. Raphael. De novo discovery of mutated driver pathways in cancer. *Genome Res*, 22(2):375–85, Feb 2012b. doi: 10.1101/gr.120477.111.

F. Vandin, E. Upfal, and B. J. Raphael. De novo discovery of mutated driver pathways in cancer. *Genome Res.*, 22(2):375–385, Feb 2012c.

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–i245, 2010.

A. Viel, A. Bruselles, E. Meccia, M. Fornasarig, M. Quaia, V. Canzonieri, E. Policicchio, E. D. Urso, M. Agostini, M. Genuardi, E. Lucci-Cordisco, T. Venesio, A. Martayan, M. G. Diodoro, L. Sanchez-Mete, V. Stigliano, F. Mazzei, F. Grasso, A. Giuliani, M. Baiocchi, R. Maestro, G. Giannini, M. Tartaglia, L. B. Alexandrov, and M. Bignami. A specific mutational signature associated with dna 8-oxoguanine persistence in mutyh-defective colorectal cancer. *EBioMedicine*, 20:39–49, 2017. ISSN 2352-3964. doi: 10.1016/j.ebiom.2017.04.022. URL `http://dx.doi.org/10.1016/j.ebiom.2017.04.022`.

B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler. Cancer genome landscapes. *science*, 339(6127):1546–1558, 2013.

S. Wang, M. Jia, Z. He, and X. S. Liu. APOBEC3B and APOBEC mutational signature as potential predictive markers for immunotherapy response in non-small cell lung cancer. *Oncogene*, 37(29):3924–3936, Jul 2018.

Y. Wang, A. Buchanan, and S. Butenko. On imposing connectivity constraints in integer programs. *Math. Program.*, 166(1-2):241–271, 2017. doi: 10.1007/s10107-017-1117-8. URL `https://doi.org/10.1007/s10107-017-1117-8`.

D. Wojtowicz, M. D. M. Leiserson, R. Sharan, and T. M. Przytycka. DNA Repair Footprint Uncovers Contribution of DNA Repair Mechanism to Mutational Signatures. *Pac Symp Biocomput*, 25:262–273, 2020.

W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, S. Ramaswamy, P. A. Futreal, D. A. Haber, M. R. Stratton, C. Benes, U. McDermott, and M. J. Garnett. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.*, 41(Database issue):D955–961, Jan 2013a.

W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. A. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, S. Ramaswamy, P. A. Futreal, D. A. Haber, M. R. Stratton, C. Benes, U. McDermott, and M. Garnett. Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, 41(Database-Issue):955–961, 2013b. doi: 10.1093/nar/gks1111. URL `https://doi.org/10.1093/nar/gks1111`.

C.-H. Yeang, F. McCormick, and A. Levine. Combinatorial patterns of somatic gene mutations in cancer. *The FASEB Journal*, 22(8):2605–2622, 2008.

L. Zannini, D. Delia, and G. Buscemi. CHK2 kinase in the DNA damage response and beyond. *J Mol Cell Biol*, 6(6):442–457, Dec 2014.

W. Zhang, J. Ma, and T. Ideker. Classifying tumors by supervised network propagation. *Bioinformatics*, 34(13):i484–i493, Jul 2018.