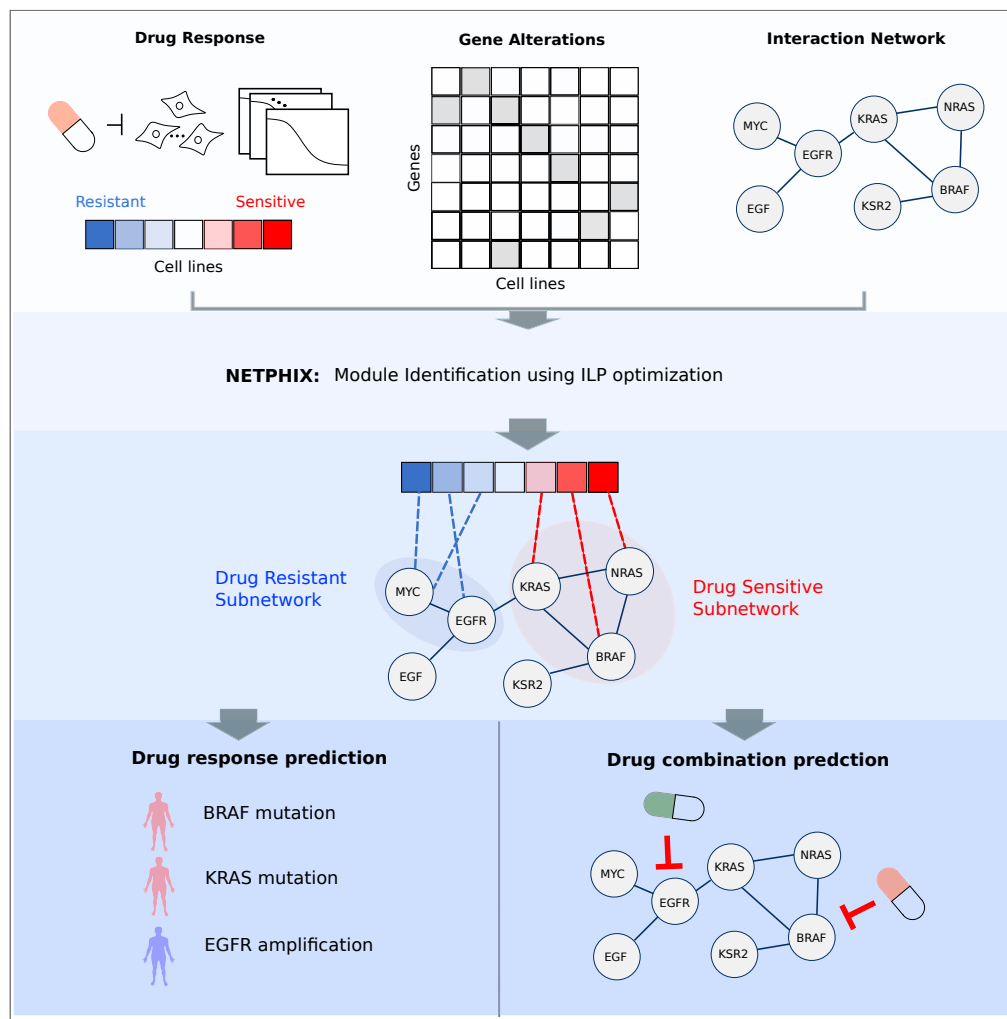**Article**

Identifying Drug Sensitivity Subnetworks with NETPHIX

Yoo-Ah Kim,
Rebecca Sarto
Basso, Damian
Wojtowicz,
Amanda S. Liu,
Dorit S.
Hochbaum, Fabio
Vandin, Teresa M.
Przytycka

kimy3@ncbi.nlm.nih.gov (Y.-A.K.)
przytyck@ncbi.nlm.nih.gov
(T.M.P.)

HIGHLIGHTS

NETPHIX identifies
mutated subnetworks
associated with a
continuous phenotype

It can aid the identification
of mutated modules
associated with drug
response

NETPHIX modules can be
used to predict drug
response and infer drug
combinations

NETPHIX uncovers
subnetworks using Integer
Linear Programming
optimization

Article

Identifying Drug Sensitivity Subnetworks with NETPHIX

Yoo-Ah Kim,[1,5,*] Rebecca Sarto Basso,[2,5] Damian Wojtowicz,[1] Amanda S. Liu,[3,7] Dorit S. Hochbaum,[2] Fabio Vandin,[4,6] and Teresa M. Przytycka[1,6,8,*]

SUMMARY

Phenotypic heterogeneity in cancer is often caused by different patterns of genetic alterations. Understanding such phenotype-genotype relationships is fundamental for the advance of personalized medicine. We develop a computational method, named NETPHIX (NETwork-to-PHenotype association with eXclusivity) to identify subnetworks of genes whose genetic alterations are associated with drug response or other continuous cancer phenotypes. Leveraging interaction information among genes and properties of cancer mutations such as mutual exclusivity, we formulate the problem as an integer linear program and solve it optimally to obtain a subnetwork of associated genes. Applied to a large-scale drug screening dataset, NETPHIX uncovered gene modules significantly associated with drug responses. Utilizing interaction information, NETPHIX modules are functionally coherent and can thus provide important insights into drug action. In addition, we show that modules identified by NETPHIX together with their association patterns can be leveraged to suggest drug combinations.

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 owing 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.

Predicting drug responses and identifying their genotypic causes is one of the most critical problems in cancer studies and crucial for developing personalized treatments. 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 or related pathways. Recently, several projects have characterized drug sensitivity in hundreds of cancer cell lines for a large number of drugs (Yang et al., 2013; Barretina et al., 2012). These data, together with information about the genetic alterations in these cell lines, provided unprecedented opportunities to understand how genetic alterations affect drug sensitivity.

The pathway-centric view of cancer (Hanahan and Weinberg 2011; Garraway and Lander 2013; 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; Vandin et al., 2012; Kim et al., 2013; Hofree et al., 2013; Kim et al., 2016a; Dao et al., 2017). Network-based approaches have been further extended to find genes whose mutations are associated specifically with given phenotypes rather than finding general cancer drivers (Gilman et al., 2012; Hofree et al., 2013; Carter et al., 2013; Kim et al., 2016a; Zhang et al., 2018). Although 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 the aforementioned network-based approaches focused on identifying mutations associated with discrete phenotypic traits, e.g., cancer versus healthy, good or bad prognosis, or cancer subtypes, and, therefore, cannot be directly applied to the analysis of continuous features such as drug sensitivity.

[1]National Center of Biotechnology Information, National Library of Medicine, NIH, Bethesda, MD 20894, USA

[2]Department of Industrial Engineering and Operations Research, University of California at Berkeley, Berkeley, CA 94709, USA

[3]Montgomery Blair High School, Silver Spring, MD 20901, USA

[4]Department of Information Engineering, University of Padova, Padova 35131, Italy

[5]These authors contributed equally

[6]Senior authors

[7]Present address: University of Maryland, College Park, MD 20742, USA

[8]Lead Contact

*Correspondence: kimy3@ncbi.nlm.nih.gov (Y.-A.K.), przytyck@ncbi.nlm.nih.gov (T.M.P.)

There have been several methods proposed for predicting drug response in cancer, including a few network-based approaches (Emad et al., 2017; Han et al., 2019; Wang et al., 2019; Ali and Aittokallio, 2019; Chen and Zhang, 2020). For example, ProGENI adopts a random walk approach on a gene interaction network to rank genes based on their association with drug response and their expression status, subsequently using highly ranked genes for drug response prediction Emad et al. (2017). However, ProGENI and most of the algorithms cited above focus on response prediction, primarily relying on transcriptomic information rather than identifying mutational biomarkers associated with drug responses. Computational methods that make a better use of CNV and point mutation data considering their unique properties are still in need of development (Ali and Aittokallio, 2019).

Another body of related studies is combining GWAS analysis with network constraints (Li and Li, 2008; Jia et al., 2011; Azencott et al., 2013; Liu et al., 2017). Although these methods generally perform well at broadly pointing to disease-related genes, they do not consider complex properties of cancer mutations such as frequently observed mutual exclusivity of cancer mutations, and are not designed to zoom in on subnetworks that are specific enough to help understand drug action.

Several algorithms have considered a problem closely related to our work—identifying mutations associated with drug response—but without including functional relationship information (Kim et al., 2016; Sarto Basso et al., 2019; Knijnenburg et al., 2016). For example, REVEALER used a re-scaled mutual information metric to iteratively identify a set of genes associated with a phenotype Kim et al. (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). Although UNCOVER uses a similar objective function as NETPHIX (NETwork-to-PHenotype association with eXclusivity), it does not utilize network information nor allows to pick up mixed sensitivity modules (i.e., simultaneously identify genes associated with increased and decreased drug sensitivity). LOBICO (Knijnenburg et al., 2016; Iorio et al., 2016) is another algorithm 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 utilizes 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.

Having the network-centric views in mind, we introduce a computational tool named for identifying mutated subnetworks that are associated with a continuous phenotype. Our algorithm builds on combinatorial optimization techniques involving *connected set cover* to find a connected set of genes associated with increased or decreased sensitivity. The objective function of NETPHIX allows different options to capture various properties of cancer mutations: First, both drug-resistant and -sensitive genes can be identified simultaneously, considering interactions between them. Second, by having an option to impose penalties for overlapping mutations, the algorithm can preferentially select mutually exclusive genes in the solution. Based on the observation that cancer-related mutations tend to be mutually exclusive (Ciriello et al., 2013; Kim et al., 2015, 2016b; Vandin et al., 2012; Leiserson et al., 2015; Constantinescu et al., 2015), we hypothesized that mutual exclusivity may also be useful for the identification of drug sensitivity modules.

We evaluated NETPHIX and other related methods using both simulations and real drug screening datasets and showed that NETPHIX outperforms 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, which provided important insights into drug action. We were also able to validate many of the identified modules with an independent drug screening dataset (The Cancer Therapeutics Response Portal [CTRP]). Finally, we show that properties of modules associated with drug can point to potential drug combinations. Effective computational methods to discover genetic alterations causal to drug sensitivity 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.

Besides drug sensitivity profiles, NETPHIX can be applied to other continuous phenotypes. For instance, a simplified version of our method was used to investigate the genetic aberrations associated with
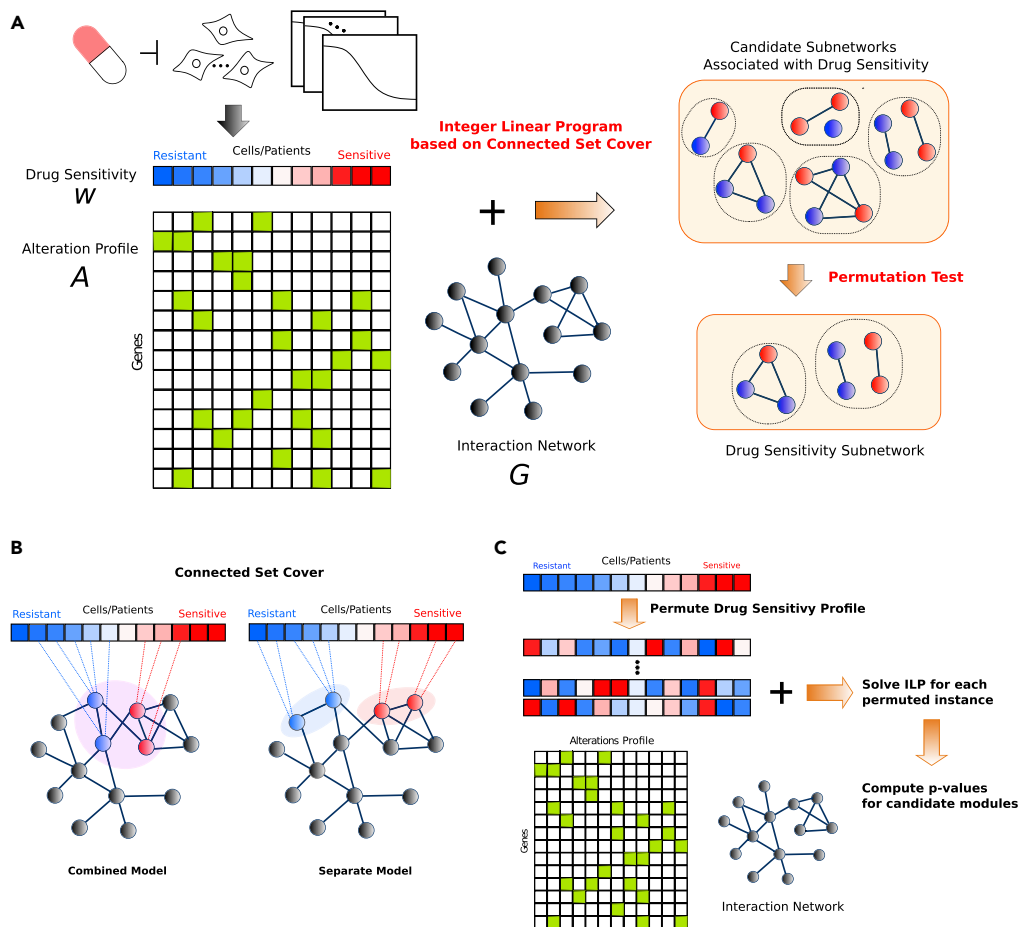
Figure 1. NETPHIX Method

(A) Overview: NETPHIX takes drug sensitivity profile, alterations status for the same set of samples, and interaction information among genes as inputs. Using a connected set cover-based ILP algorithm, we first generate a set of candidate modules. The final set of modules include only maximal modules among statistically significant solutions. Genes associated with decreased and increased sensitivity are marked as blue and red, respectively.

(B) NETPHIX finds a connected set of genes of which alterations 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 and the *separate* model in which two subnetworks are identified for increased and decreased sensitivity separately.

(C) The significance of identified modules is assessed using a permutation test by permuting drug sensitivity profiles.

mutational signatures, providing novel insights into mutagenic processes that the cancer genomes might have undergone Kim et al. (2020). Furthermore, the mutual exclusivity condition can be easily removed and therefore NETPHIX can also be used for the more general problem of linking continuous phenotypes (unrelated to cancer) to genetic alterations.

RESULTS

NETPHIX Method Overview

NETPHIX takes gene alteration information, drug response profiles, and interaction network as inputs and identifies genetic alterations underlying the phenotype of interest (Figure 1A). In the first phase, we generate candidate modules solving an optimization problem based on a connected set cover approach as described in section Connected Set Cover Based Algorithm for Selecting Candidate Modules. The problem is formulated as an integer linear program (ILP), and candidate modules are generated by obtaining the optimal solutions using CPLEX (https://www.ibm.com/analytics/cplex-optimizer) for ILP instances

with different parameters. In the second phase, the statistical significance of candidate modules is assessed with a permutation test, and final sensitivity modules are selected (see sections Selecting Final Modules and S1.1.4 for details).

Connected Set Cover-Based Algorithm for Selecting Candidate Modules

To obtain candidate subnetworks, we design our algorithm based on connected set cover to maximize the association with drug response (Figure 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 (Kim et al., 2013, 2015; Sarto Basso et al., 2019; Kim et al., 2011; Chowdhury and Koyuturk, 2010; Ulitsky et al., 2010; Hristov and Singh, 2017). Below we describe how the technique can be extended to identify modules associated with drug response. For the formal definition of the problem, see Section S1.1.1.

NETPHIX takes a gene alteration matrix A and drug response profile w for a set of patients P (or cell lines) as inputs. Alteration matrix A is a $|V| \times |P|$ binary matrix where $A(v, p) = 1$ if gene v is altered in patient p and 0 otherwise. w is a vector of length $|P|$ (Figure 1A). In addition, an interaction network $G = (V, E)$ is given where V represents genes. NETPHIX then aims to identify a set of connected genes $S \subseteq V$ so as to maximize

$$W(S) = \sum_{p \in P(S)} w(p) \tag{Equation 1}$$

where $P(S)$ is the set of patients who have alterations in any genes in S. In other words, the algorithm tries to *cover* drug-sensitive patients with genes having alterations in the patients while maximizing the total weight of covered patients and enforcing the identified genes to be *connected* in the network. The connected set cover approach can capture the heterogeneity of cancer mutations as different patients who are sensitive to a drug may have causal alterations in (or are covered by) different but functionally related genes.

In addition, the objective function of NETPHIX can include penalties for overlapping mutations in covered patients to reinforce the property of mutual exclusivity in the selected modules. Based on the observation that cancer-related mutations tend to be mutually exclusive, we hypothesized that the property may also be useful for the identification of drug sensitivity modules. To include penalties for overlapping mutations, we can extend the objective function as follows:

$$W(S) = \sum_{p \in P(S)} w(p) - \sum_{p \in P(S)} pt(p)(c(p, S) - 1) \tag{Equation 2}$$

where $c(p, S) = |\{g \in S | A(g, p) = 1\}|$ and $pt(p)$ is the penalty for overlapping mutations in patient p.

Although it is natural to assume that genes associated with a specific response are likely from a functional subnetwork, it was not clear what relation should be assumed between modules with association in opposite directions. Therefore, we considered two different connectivity models (Figure 1B) as there can be genes associated with either direction of drug response—genes whose alterations correlate with increased sensitivity to the drug (sensitive) and genes whose alterations correlate with decreased sensitivity to the drug (resistant). NETPHIX finds a module that includes both types of genes simultaneously using two different models—the combined model and the separate model. In the "combined" model, we identify one connected subnetwork that includes all genes associated with either direction (decreased or increased drug sensitivity). In the "separate" model, we seek to identify two subnetworks, one for increased sensitivity and one for decreased sensitivity separately. This model is designed to capture the case when two different functional modules affect drug response in different ways.

We create multiple ILP instances for different module sizes (i.e., the number of genes in the module) and connectivity options and obtain candidate modules by solving the ILP instances optimally with CPLEX. For the detailed ILP formulation and parameters, see Sections S1.1.2 and S1.1.3.

Selecting Final Modules

Once we obtain the optimal gene module for each parameter combination, the significance of the identified module is assessed by performing a permutation test (Figure 1C). Note that our algorithm is designed to identify the modules associated specifically with a given phenotype (e.g., drug sensitivity to each drug)
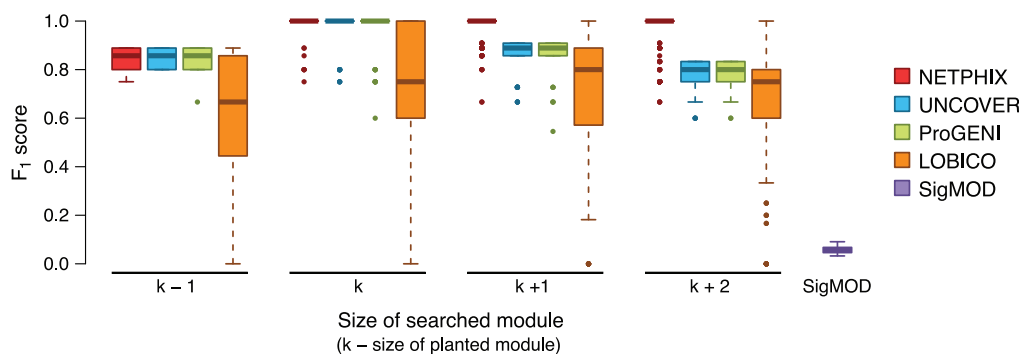
Figure 2. Method Comparison on Simulated Data
F_1 score for the modules identified by NETPHIX (red), UNCOVER (blue), ProGENI (green), LOBICO (orange), and SigMOD (purple).

rather than finding general cancer drivers. 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 phenotypes. Among all significantly associated subnetworks, we obtain the final drug sensitivity modules by selecting maximal modules to remove redundancy. In other words, for any two significant modules M_i and M_j such that $M_i \subset M_j$, only M_j is included in the final solution for the drug. See Section S1.1.4 for the details of the permutation test and maximal module selection.

Method Evaluation

Evaluation with Simulation Data

We first compared the performance of NETPHIX using simulation data with four other related methods—LOBICO, UNCOVER, SigMOD, and ProGENI. Of the four algorithms, SigMOD and ProGENI are network-based algorithms. ProGENI adopts a random walk approach in an interaction network and ranks genes based on their associations between drug response and gene aberrations Emad et al. (2017). SigMOD is a recently proposed module identification algorithm combining GWAS and a network-based approach 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. 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. Although both LOBICO and UNCOVER find an optimal solution using an integer linear program, neither algorithms utilize interaction network information.

The simulated instances are generated by randomly generating drug response profiles and planting a connected set of genes so that their alteration patterns are associated with the drug response. We generated the instances with varying parameters and planted on the background of real cancer cell line mutation data and the functional interaction network information downloaded from STRING database (https://string-db.org) (see Section S1.3.1 for the detailed description). Although 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. Therefore, in order not to disadvantage any of the methods, for the simulation-based evaluation we considered simple cases where alterations are associated with only one direction (either increased or decreased). We planted modules of sizes 3, 4, and 5 (the module size refers to the number of genes in the selected module) and evaluated the accuracy of the five methods in identifying the planted modules in terms of F1 scores (Figure 2; see Figures S1A and S1B for precision and recall). For all algorithms except SigMOD and ProGENI, we ran the algorithm for different k's (k' is a parameter for the size of a module searched by the algorithms). Since ProGENI ranks genes instead of selecting modules, we considered the top k' genes to be the selected modules. SigMOD automatically adjusts its parameters to find the best module. Also, for all ILP-based algorithm, we limit the running time up to 24 h, meaning the algorithms will stop and output the current solution (which may be suboptimal) when the time limit reaches.

As shown in Figure 2, NETPHIX outperforms other algorithms, especially when the algorithms search for bigger modules than planted. Note that, in real applications, the size of a module is unknown; therefore, the ability of correctly estimating the size of the module is an important component of a method. Since the genes should be connected, NETPHIX usually does not extend the best module with spurious genes even if we search for modules bigger than planted. On the other hand, other algorithms tend to add more genes when increasing k'. In general, in this basic scenario, the algorithms uncovered the planted modules in most instances (Figure S1B) as long as the size of searched modules are at least as big as the planted module sizes, but LOBICO solutions missed true positives more often compared with other algorithms. SigMOD identified a large number of false positives along with the planted modules (approximately 100–180 genes) that are not associated with phenotypes.

Evaluation with Real Drug Screening Dataset

We next evaluated the performance of algorithms with real drug screening dataset (GDSC) and further validated the results on an independent set (CTRP). We included ProGENI and UNCOVER for comparison since the performance of these methods on simulated data was fairly good.

In the GDSC dataset, the responses for 265 drugs are available for 240–705 cell lines depending on drugs. For the alteration table, 26,917 gene-level alteration profiles are collected, combining amplification, deletion, and somatic mutations (see Section S1.2 for a detailed description of the data, including criteria used to identify genomic alterations).

NETPHIX identified a total of 476 modules for 194 drugs (for the remaining drugs no modules with significant association were identified, Table S1). Since there can be multiple functional modules affecting drug efficacy, our method allows us to identify multiple associated modules for a specific drug. Of 476 identified modules, 258 modules consist of one connected module based on the combined model (for 163 drugs) and 218 modules consist of two connected components based on the separate model (for 136 drugs). UNCOVER identified modules for 127 drugs (p value < 0.05) and only those modules were tested; 114 drugs had both NETPHIX and UNCOVER modules identified. ProGENI ranks genes and does not provide significance scores, and therefore, top ranked genes for all 265 drugs were used for comparison.

We compared the performance in terms of several measures. First, the distances between the modules and the drug targets were computed to examine the relationship between them. Next, the predictive power of drug responses for the identified genes were measured using probabilistic concordance index (PCI). In addition, we computed p values with ANOVA test to examine if the alteration status of the selected genes is associated with drug response.

Distance to Drug Targets.

We first examined the relationship between the identified modules and drug targets. Specifically, given that we expect a mutated module to be related to the action of a drug, the module is likely to be close to the drug target(s) in the network. To test this, we computed the distances between the drug targets and the genes selected by each method for the corresponding drugs (Figure 3A). In case of NETPHIX the genes in drug sensitivity modules are located close to the corresponding drug targets in the network (the mean distances of 2.12, whereas the average distance to targets for randomly selected set of genes is greater than 2.97; $p<10^{-57}$, t test). For UNCOVER, the distances are larger than for NETPHIX (the mean distances of 2.5) but still smaller than random ($p<10^{-11}$). ProGENI genes have similar statistics as random genes.

As NETPHIX selects genes associated with both decreased and increased sensitivity, we also examined if there are differences between the two sets. Interestingly, we observed that the genes associated with increased sensitivity are closer to drug targets than the genes associated with decreased sensitivity (the mean distance 1.97 versus 2.37, $p<10^{-13}$, t test), indicating that the efficacy of a drug improves if it targets genes that are close to perturbed gene modules. We observed similar patterns in UNCOVER modules albeit less significant (the mean distance 2.23 versus 2.67, $p<10^{-3}$).

Predictive Power of Selected Modules Using Regression.

Next, we examined the ability of predicting drug responses for the modules identified by different algorithms. To this purpose, we performed random forest (RF) regression using the selected genes as features. As NETPHIX selects multiple modules for some drugs, we merged all modules significantly associated with the response for each drug and used all
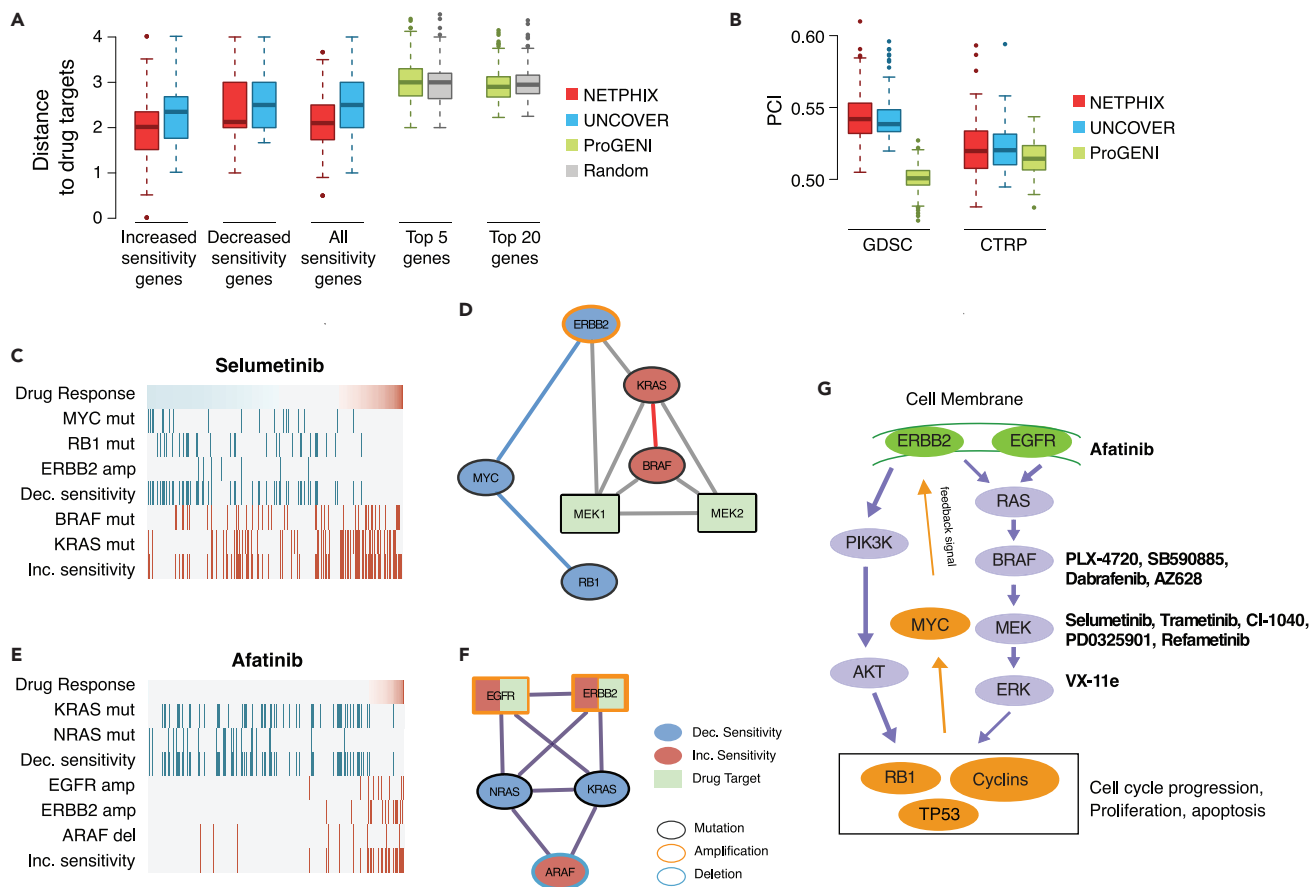
Figure 3. Evaluation Results and Modules Identified with Real Drug Screening Dataset

(A–B) Evaluation with drug screening dataset. (A) Average distances of selected genes to drug targets. Distances for genes associated with increased sensitivity, decreased sensitivity, all genes in the selected modules are shown for NETPHIX and UNCOVER. The distances with 5 and 20 top ranked genes for ProGENI and randomly selected genes are also shown. (B) Predictive power of the modules selected by NETPHIX, UNCOVER, and ProGENI. PCI scores with random forest regression tested within GDSC and with CTRP dataset are shown.

(C–F) Sensitivity networks identified by NETPHIX. Alternation profile and connectivity between the drug target and the genes for Selumetinib (C, D) and Afatinib (E, F). 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 is shown in red/blue (with the summary covers for decreased and increased sensitivity separately), whereas samples not altered are shown in gray. 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.

(G) Schematic diagram of MAPK/ERK and AKT signaling pathways with drugs and their drug targets annotated.

selected genes as features for RF regression. The Boolean indicators of mutation status were used as feature values. The number of features for each drug varies between 3 and 15 in total. UNCOVER finds at most one significant module for decreased and increased sensitivity, respectively, and we used genes in both directions as features (total of 3–6 genes). For ProGENI, we used 20 top ranked genes as features for regression. We first performed a nested cross-validation on GDSC dataset by learning the best model with training set and testing with the remaining set of cell lines within GDSC (see Section S1.3.2 for details). In the second set of evaluation, we learned hyperparameters from GDSC dataset and applied the best model to an independent dataset (CTRP) and examined the predictive power. As shown in Figure 3B, the predictive power of NETPHIX outperformed other algorithms in both measurements, although UNCOVER had comparable prediction results for CTRP.

Predictive Power of Selected Modules Using ANOVA Test. For another way to examine if the alterations in the identified modules indeed lead to specific responses for the corresponding drugs, we divided

cell lines into different groups depending on their alteration status and tested if the groups have statistically significant difference in drug responses using ANOVA.

Specifically, for each module identified using GDSC dataset for a given drug, we tested the predictive power of the module with CTRP dataset. We grouped the cell lines into three groups: (1) cell lines with alterations in *decreased sensitivity genes* only, (2) cell lines with *increased sensitivity genes* only, and (3) cell lines with *no alterations* in any of the selected genes. Subsequent ANOVA tests evaluated the statistical significance of differences in drug response among three groups of cell lines (see Section S1.3.2 for details). Since the test requires selected modules and their mutation status associated with either direction of sensitivity, we examined this measure only for NETPHIX and UNCOVER. ProGENI only ranks genes instead of identifying modules, and the directions of association (increased or decreased) are not specified.

For NETPHIX, 164 modules for 65 drugs were tested for the response with CTRP dataset (many drugs have multiple associated modules). We found that 102 modules have a statistically significant difference ($p < 0.05$ [ANOVA], FDR <9% [BH]) and 45 drugs have at least one significant module. On the other hand, only 29 of 44 UNCOVER modules had significant difference in the response (UNCOVER identifies at most one increased and decreased module, respectively, for a drug). In summary, the results confirm that NETPHIX can identify more gene modules that are predictive of drug responses.

Impact of NETPHIX Design Choices on the Results

Although the benefits of using protein interaction network for the identification of cancer drivers are generally accepted, it was not well investigated before how much gain the network usage provides in the context of drug responses. Here we use ANOVA tests 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 instances validated with ANOVA test as described in the previous section.

Network Information Helps Identify Drug Sensitivity Modules

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

NETPHIX finds more significant modules when network information is used (476 modules compared with 274 modules without network). The number of drugs with at least one associated module is also larger (194 drugs with network versus 175 drugs without network). As shown in Figure S2A, without network only 57 of 88 tested modules were confirmed, whereas 102 of 164 tested modules were confirmed when connectivity was imposed. In addition, only 40 drugs had at least one confirmed module without network compared with 45 drugs with network. Overall, the results show that network information helps find more modules that are predictive of drug sensitivity.

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 mutual 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 (Figure S2B). We observe that NETPHIX finds a lesser number of significant modules when penalties are not included (460 modules compared with 476 modules in the original solution), although the numbers of drugs with at least one associated modules are similar (192 drugs without penalty versus 194 drugs with penalty). In terms of the number of drugs/modules confirmed in CTRP dataset with ANOVA, 88 modules of 152 tested modules were confirmed without penalty, whereas 102 modules of 164 tested modules were confirmed with penalty. In addition, 41 drugs had at least one confirmed module without penalty compared with 45 drugs with penalty.

Combined versus 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 (Figure S2C). It was not immediately obvious which approach would be

more successful. On the 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. A drug may have multiple drug targets, which may lead to separate modules.

We found that a similar number of modules were identified with the combined model and the separate model (286 versus 278 modules) for 177 versus 170 drugs. However, the combined connectivity model has a slightly higher percentage of confirmed modules/drugs. Ninety-eight modules were tested with CTRP dataset for both combined and separate models, and 65 and 61 modules were confirmed, respectively. In terms of the number of drugs with at least one confirmed module, the combined model has 37 of 57 drugs confirmed, whereas the separate model has the same number of drugs confirmed out of 59.

Network Properties of NETPHIX Modules

NETPHIX is designed to choose modules in which genes are highly connected, which helps identify drug sensitivity modules as discussed above. The selected modules are relatively small (the average size of 3.86) and densely connected (the average edge density of 0.55). Since NETPHIX modules are densely connected, the genes in the selected modules are naturally close in terms of distance in the network (the mean of average distances = 1.78, Figure S2D). 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 average distances between the two groups = 1.96).

Case Studies: NETPHIX Identifies Mutated Subnetworks Associated with Drug Responses

Many of the modules identified by NETPHIX are putative drug biomarkers supported by previous studies and provide interesting further insights related to drug action. We analyzed the identified modules associated with a few drugs in more detail.

Drugs Targeting RAS/MAPK Pathways

RAS/MAPK pathway regulates growth, proliferation, and apoptosis and is often dysregulated in various cancers (Figure 3G). Among the most common mutations of this pathway are mutations of BRAF/KRAS/NRAS. Interestingly, all modules associated with increased sensitivity to MEK inhibitors (Selumetinib, Trametinib, CI-1040, PD0325901, Refametinib) and an ERK inhibitor (VX-11e) included BRAF and KRAS mutations. NRAS mutations were included in 12 of 20 such modules. 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 ERBB2 amplification, MYC and RB1 mutations (four times) or the module with TP53 mutations (eight times). 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, was previously identified as predictors of MEK inhibitors, although KRAS mutations can affect drug responses differently depending on the mutation types (Nakayama et al., 2008; Li et al., 2018; Sun et al., 2014). ERBB2 is a receptor protein that signals through this pathway, whereas 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 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. Although 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.

Drugs Targeting Histone Deacetylases

Histone deacetylase (HDAC) inhibitors are cytotoxic drugs, used to destroy cancer cells by inhibiting cell division and causing cell death. HDAC is frequently dysregulated in cancer, and although HDAC inhibitors, targeting HDAC proteins, have a wide range of effects including cell-cycle arrest and apoptosis, the mechanisms of drug action are still uncertain. NETPHIX identified several modules associated with HDAC inhibitor sensitivity. For example, Vorinostat is a pan-HDAC inhibitor and inhibits class I and class II HDAC enzymes. NETPHIX found that the drug response to Vorinostat is associated with CREBBP mutation for increased sensitivity and SMAD4 deletion and DACH1 mutation for decreased sensitivity (Figure S3A). Several studies found that cell lines with CREBBP mutations are sensitive to Vorinostat treatment (Mullighan et al., 2011; Andersen et al., 2012). SMAD4 inactivation is implicated in tumor malignancy and resistance to multiple drugs (Papageorgis et al., 2011; Chen et al., 2014).

On other hand, Belinostat and CUDC-101 are found to be associated with CDKN2D deletion (decreased sensitivity), whereas TP53 deletion and MAPK7, MYC mutations are associated with increased sensitivity of the drug. Consistent with our findings, studies found that HDACis were preferentially cytotoxic to cells with mutant TP53 (Blagosklonny et al., 2005) but CDKN2A deletion is linked to poor effectiveness in combination therapy of HDACi with another agent (Chen et al., 2010).

Drug resistance is often observed for HDAC inhibitors, and combined treatments of HDACi with other anti-cancer drugs have demonstrated promising effects in clinical studies (Suraweera et al., 2018). As discussed below, based on NETPHIX modules we identified multiple candidates of combination therapy with HDACi that are also supported by previous evidences.

NETPHIX Modules Suggest Candidates for Drug Combination Therapy

We hypothesize that pairs of drugs can potentially have synergistic effects if they are associated with similar modules but the genes in the modules are associated with opposite directions of drug responses. By analyzing the modules for pairs of drugs with such property, we identified 153 drug pairs (Table S3). Although the systematic validation could not be performed owing to the lack of validation dataset, we found evidences in the literature for the efficacy of many predicted drug combinations. In Figure 4, we show examples of the identified combination candidates of drugs for which we found supporting evidences from the literature. The drugs are categorized by the target pathways. For example, Afatinib, a receptor tyrosine kinase inhibitor, has 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 (Figures 3C–3F). Indeed, studies showed that Afatinib and Selumetinib work synergistically (Sun et al., 2014) and clinical trials for combination therapy are currently underway (https://clinicaltrials.gov/ct2/show/NCT02450656). In addition to Selumetinib, other MEK inhibitors, such as Refametinib and PD0325901, have associated modules similar to Afatinib but in opposite direction, and the efficacy of PD0325901 and Afatinib combination is also reported (Lin et al., 2019). Indeed, drugs targeting the same molecules are typically predicted to have similar drug combination partners (Figure 4).

HDAC inhibitors are found to improve drug efficacy when combined with other anti-cancer drugs (Figure 4D). An example is the combination of Vorinostat and Lapatinib (Figure S3A and S3B). Lapatinib is a drug that inhibits EGFR/ERBB2, and Vorinostat is a histone deacetylase (HDAC) inhibitor. Lapatinib in general enhances the antitumor activity of the histone deacetylase inhibitor synergistically (LaBonte et al., 2011). Vorinostat has been shown to improve how well Lapatinib kills cancer cells in clinical trials (https://clinicaltrials.gov/ct2/show/NCT01118975) and can also improve the therapeutic efficacy of other RTK inhibitors such as Erlotinib (Leone et al., 2015).

DISCUSSION

We developed a new computational method, NETPHIX (NETwork-to-PHenotype association with eXclusivity), for the identification of mutated subnetworks that are associated with a continuous phenotype. Using simulations and analyzing large-scale drug screening datasets, we showed that NETPHIX can uncover the subnetworks associated with response to cancer drugs. 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
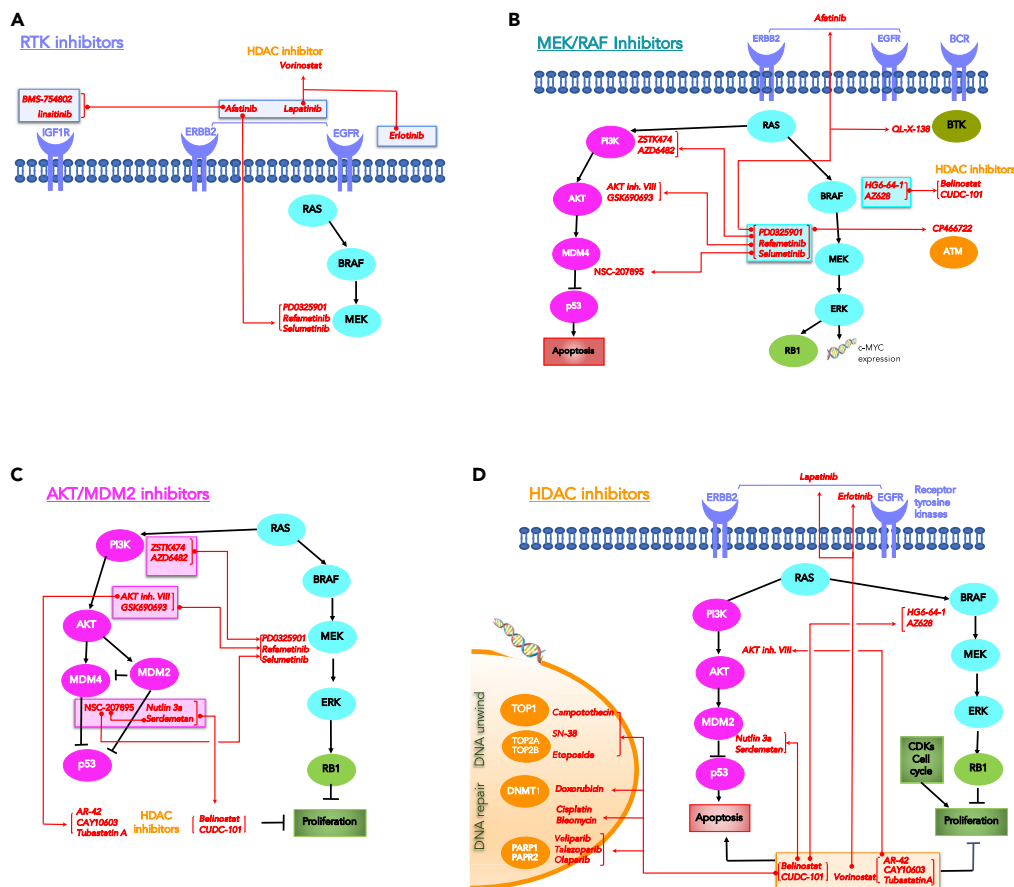
Figure 4. Examples of Predicted Drug Combinations Supported by Independent Evidence

The candidate drug pairs (connected by red lines) that are identified based on NETPHIX modules and supported by the literature survey are shown.

(A–D) The panels are organized by different drug target pathways: (A) RTK inhibitors (inside blue boxes); (B) MEK/RAF inhibitors (turquoise boxes); (C) AKT/MDM2 inhibitors (magenta boxes); (D) HDAC inhibitors (orange boxes). The name of each drug is displayed next to its target protein, and their predicted combination partners are connected by red lines. Drugs with the same interaction partners are connected by brackets. The genes/nodes are color coded as follows: orange (DNA processing), magenta (PIK/AKT/mTOR pathway), turquoise (MAPK pathway), green (cell cycle), blue half-moons (Receptors Tyrosine Kinases), dark green (BRT is a non-receptor Tyrosine Kinase).

status in many of the 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.

The mutual exclusivity property has been shown to be helpful for identifying cancer drivers, but the hypothesis that the property can also help define mutated subnetworks associated with drug response was not well investigated before. We demonstrate that the preferential selection of mutually exclusive genes can improve the performance of the method, although the improvement is moderate. We note that the set cover approach itself might favor the genes that are mutually exclusive to maximize the coverage with the smallest number of genes. We hypothesize that the improvement is related to the fact that mutual exclusivity increases the probability that selected genes are cancer related. However, further studies may be necessary regarding the relationship between mutually exclusive mutations and drug sensitivity. By removing the mutual exclusivity condition, NETPHIX can also be used for the more general problem of linking continuous phenotypes (not specific to cancer) with genetic alterations.

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 slightly outperforms the separate connectivity model, indicating that the two groups of genes in fact might be related. We also found that, even for the separate model, the sensitivity modules are relatively close to each other (Figure S2D), which suggests that the genes whose alterations are associated with the same drug response may belong to the same or related pathways.

Although the combined connectivity model works better in terms of the number of significant and validated modules, there are a few drugs for which the separate model provides different insights on drug action. For example, for Cytarabine, both the combined and the separate model identified statistically significant modules, which are also confirmed in CTRP dataset. In particular, the module associated with Cytarabine in the separate model (Figure S3C) includes UGT2B17 and CYP2E1 deletion associated with decreased sensitivity. Both enzymes are hypothesized to be important players in the metabolism of common drugs (Guillemette et al., 2014; Garcia-Suastegui et al., 2017). Since drug metabolism pathways are typically not part of cell growth-related cancer-driving pathways, the separate connectivity approach can provide insights that the combined model cannot provide in such cases.

The applicability of NETPHIX can go far beyond the drug response discussed in this paper, to any continuous cancer phenotypes. A simplified version of our method was used to investigate the genetic aberrations associated with mutational signatures, by which the mutagenic processes underlying the cancer genomes were elucidated (Kim et al., 2020). In that case, the phenotype of interest was the strength of a mutagenic process as measured by the mutation counts of the corresponding mutational signature. We expect that NETPHIX will find broad applications in many other types of network-to-phenotype association studies, especially in the context of cancer.

Limitations of the Study

NETPHIX is focused on the identification of mutated subnetworks associated with drug response. Such mutated modules might not always exist as, for example, some drugs are designed to reduce expression of overexpressed genes. In addition, there is no guarantee that the association identified by NETPHIX is causal, although the proximity of identified subnetworks to the drug targets suggests that causal relations are quite likely. Finally, in case of systematic co-occurrence of alterations (for example, when two genes are co-amplified) NETPHIX might report only one of these genes owing to the nature of set cover-based algorithms.

Resource Availability

Lead Contact

Further information and requests for resources and information should be directed to and will be fulfilled by the Lead Contact, Teresa M. Przytycka (przytyck@ncbi.nlm.nih.gov).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

The source code and the datasets used for and generated during this study are available at https://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/index.cgi#netphix.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.101619.

ACKNOWLEDGMENTS

REFERENCES

Ali, M., and Aittokallio, T. (2019). Machine learning and feature selection for drug response prediction in precision oncology applications. Biophys. Rev. 11, 31–39.

Andersen, C.L., Asmar, F., Klausen, T., Hasselbalch, H., and Grønbæk, K. (2012). Somatic mutations of the crebbp and ep300 genes affect response to histone deacetylase inhibition in malignant dlbcl clones. Leuk. Res. Rep. 2, 1–3.

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

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehar, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483, 603–607.

Blagosklonny, M.V., Trostel, S., Kayastha, G., Demidenko, Z.N., Vassilev, L.T., Romanova, L.Y., Bates, S., and Fojo, T. (2005). Depletion of mutant p53 and cytotoxicity of histone deacetylase inhibitors. Cancer Res. 65, 7386–7392.

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

Chen, J., and Zhang, L. (2020). A survey and systematic assessment of computational methods for drug response prediction. Brief. Bioinform. bbz164, https://doi.org/10.1093/bib/bbz164.

Chen, M., Voeller, D., Marquez, V.E., Kaye, F.J., Steeg, P.S., Giaccone, G., and Zajac-Kaye, M. (2010). Enhanced growth inhibition by combined dna methylation/hdac inhibitors in lung tumor cells with silenced cdkn2a. Int. J. Oncol. 37, 963–971.

Chen, Y.W., Hsiao, P.J., Weng, C.C., Kuo, K.K., Kuo, T.L., Wu, D.C., Hung, W.C., and Cheng, K.H.

(2014). Smad4 loss triggers the phenotypic changes of pancreatic ductal adenocarcinoma cells. BMC Cancer 14, 181.

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

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

Ciriello, G., Cerami, E., Aksoy, B.A., Sander, C., and Schultz, N. (2013). Using MEMo to discover mutual exclusivity modules in cancer. Curr. Protoc. Bioinform. 8, 8.17.

Constantinescu, S., Szczurek, E., Mohammadi, P., Rahnenführer, J., and Beerenwinkel, N. (2015). Timex: a waiting time model for mutually exclusive cancer alterations. Bioinformatics 32, 968–975.

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

Emad, A., Cairns, J., Kalari, K.R., Wang, L., and Sinha, S. (2017). Knowledge-guided gene prioritization reveals new insights into the mechanisms of chemoresistance. Genome Biol. 18, 153.

Garcia-Suastegui, W.A., Ramos-Chavez, L.A., Rubio-Osornio, M., Calvillo-Velasco, M., Atzin-Mendez, J.A., Guevara, J., and Silva-Adaya, D. (2017). The role of CYP2E1 in the drug metabolism or bioactivation in the brain. Oxid. Med. Cell. Longev. 2017, 4680732.

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

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

Gong, X., Du, J., Parsons, S.H., Merzoug, F.F., Webster, Y., Iversen, P.W., Chio, L.C., Van Horn, R.D., Lin, X., Blosser, W., et al. (2019). Aurora A kinase inhibition is synthetic lethal with loss of the RB1 tumor suppressor gene. Cancer Discov. 9, 248–263.

Guillemette, C., Levesque, E., and Rouleau, M. (2014). Pharmacogenomics of human uridine diphospho-glucuronosyltransferases and clinical implications. Clin. Pharmacol. Ther. 96, 324–339.

Han, Y., Wang, C., Dong, Q., Chen, T., Yang, F., Liu, Y., Chen, B., Zhao, Z., Qi, L., Zhao, W., et al. (2019). Genetic interaction-based biomarkers identification for drug resistance and sensitivity in cancer cells. Mol. Ther. Nucleic Acids 17, 688–700.

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

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

Hristov, B.H., and Singh, M. (2017). Network-based coverage of mutational profiles reveals cancer genes. Cell Syst. 5, 221–229.

Iorio, F., Knijnenburg, T.A., Vis, D.J., Bignell, G.R., Menden, M.P., Schubert, M., Aben, N., Goncalves, E., Barthorpe, S., Lightfoot, H., et al. (2016). A landscape of pharmacogenomic interactions in cancer. Cell 166, 740–754.

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

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

Kim, J., Botvinnik, O.B., Abudayyeh, O., Birger, C., Rosenbluh, J., Shrestha, Y., Abazeed, M.E., Hammerman, P.S., DiCara, D., Konieczkowski, D.J., et al. (2016). Characterizing genomic alterations in cancer by complementary functional associations. Nat. Biotechnol. *34*, 539–546.

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

Kim, Y.A., Cho, D.Y., and Przytycka, T.M. (2016a). Understanding genotype-phenotype effects in cancer via network approaches. PLoS Comput. Biol. *12*, e1004747.

Kim, Y.A., Madan, S., and Przytycka, T.M. (2016b). Wesme: uncovering mutual exclusivity of cancer drivers and beyond. Bioinformatics *33*, 814–821.

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

Kim, Y.A., Wojtowicz, D., Sarto Basso, R., Sason, I., Robinson, W., Hochbaum, D.S., Leiserson, M.D.M., Sharan, R., Vadin, F., and Przytycka, T.M. (2020). Network-based approaches elucidate differences within APOBEC and clock-like signatures in breast cancer. Genome Med. *12*, 52.

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

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

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

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

Leone, A., Roca, M.S., Ciardiello, C., Terranova-Barberio, M., Vitagliano, C., Ciliberto, G., Mancini, R., Di Gennaro, E., Bruzzese, F., and Budillon, A. (2015). Vorinostat synergizes with egfr inhibitors in nsclc cells by increasing ros via up-regulation of the major mitochondrial porin vdac1 and modulation of the c-myc-nrf2-keap1 pathway. Free Radic. Biol. Med. *89*, 287–299.

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

Li, S., Liu, S., Deng, J., Akbay, E.A., Hai, J., Ambrogio, C., Zhang, L., Zhou, F., Jenkins, R.W., Adeegbe, D.O., et al. (2018). Assessing therapeutic efficacy of MEK inhibition in a KRASG12c-driven mouse model of lung cancer. Clin. Cancer Res. Official J. Am. Assoc. Cancer Res. *24*, 4854–4864.

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

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

Mullighan, C.G., Zhang, J., Kasper, L.H., Lerach, S., Payne-Turner, D., Phillips, L.A., Heatley, S.L., Holmfeldt, L., Collins-Underwood, J.R., Ma, J., et al. (2011). Crebbp mutations in relapsed acute lymphoblastic leukaemia. Nature *471*, 235–239.

Najem, A., Krayem, M., Sals, F., Hussein, N., Badran, B., Robert, C., Awada, A., Journe, F., and Ghanem, G.E. (2017). P53 and MITF/Bcl-2 identified as key pathways in the acquired resistance of NRAS-mutant melanoma to MEK inhibition. Eur. J. Cancer *83*, 154–165.

Nakayama, N., Nakayama, K., Yeasmin, S., Ishibashi, M., Katagiri, A., Iida, K., Fukumoto, M., and Miyazaki, K. (2008). KRAS or BRAF mutation status is a useful predictor of sensitivity to MEK inhibition in ovarian cancer. Br. J. Cancer *99*, 2020–2028.

Papageorgis, P., Cheng, K., Ozturk, S., Gong, Y., Lambert, A.W., Abdolmaleky, H.M., Zhou, J.R., and Thiagalingam, S. (2011). Smad4 inactivation promotes malignancy and drug resistance of colon cancer. Cancer Res. *71*, 998–1008.

Sanchez-Laorden, B., Viros, A., Girotti, M.R., Pedersen, M., Saturno, G., Zambon, A., Niculescu-Duvaz, D., Turajlic, S., Hayes, A., Gore, M., et al. (2014). BRAF inhibitors induce metastasis in RAS mutant or inhibitor-resistant melanoma cells by reactivating MEK and ERK signaling. Sci. Signal. *7*, ra30.

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

Sun, C., Hobor, S., Bertotti, A., Zecchin, D., Huang, S., Galimi, F., Cottino, F., Prahallad, A., Grernrum, W., Tzani, A., et al. (2014). Intrinsic resistance to MEK inhibition in KRAS mutant lung and colon cancer through transcriptional induction of ERBB3. Cell Rep. *7*, 86–93.

Suraweera, A., O'Byrne, K.J., and Richard, D.J. (2018). Combination therapy with histone deacetylase inhibitors (hdaci) for the treatment of cancer: achieving the full therapeutic potential of hdaci. Front. Oncol. *8*, 92.

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

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

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

Wang, S., Huang, E., Cairns, J., Peng, J., Wang, L., and Sinha, S. (2019). Identification of pathways associated with chemosensitivity through network embedding. PLoS Comput. Biol. *15*, e1006864.

Yang, W., Soares, J., Greninger, P., Edelman, E.J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J.A., Thompson, I.R., et al. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. *41*, D955–D961.

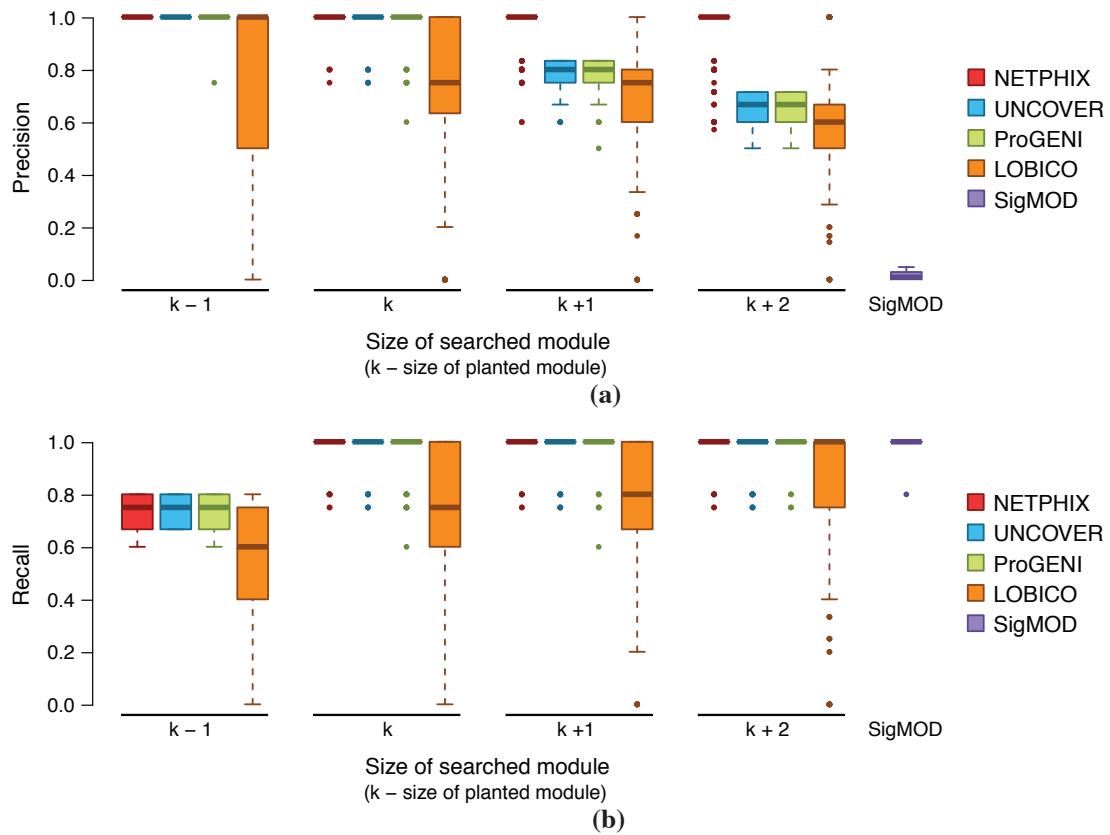Zhang, W., Ma, J., and Ideker, T. (2018). Classifying tumors by supervised network propagation. Bioinformatics *34*, i484–i493.

Supplemental Information

# Identifying Drug Sensitivity Subnetworks with NETPHIX

Yoo-Ah Kim, Rebecca Sarto Basso, Damian Wojtowicz, Amanda S. Liu, Dorit S. Hochbaum, Fabio Vandin, and Teresa M. Przytycka

Supplemental Information

Figure S1: Method comparison on simulated data (Related to Figure 2). (a) Precision and (b) Recall for the modules identified by NETPHIX (red), UNCOVER (blue), ProGENI (green), LOBICO (orange), and SigMOD (purple).
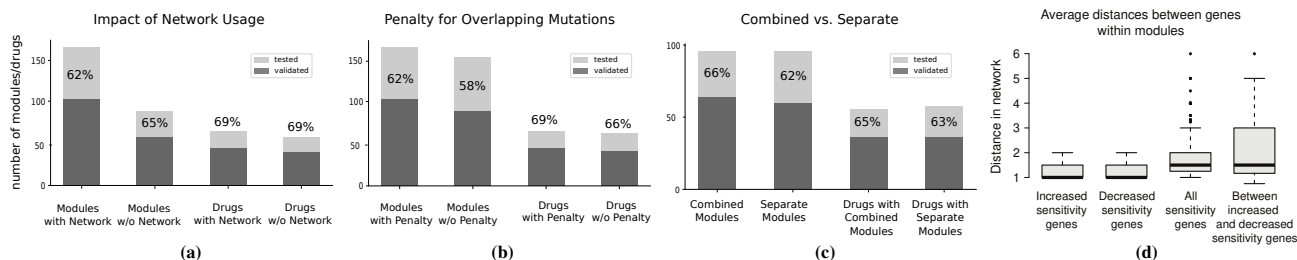
Figure S2: Impact of design choices on the performance of the algorithm (Related to Figure 3) (a) Comparison between runs with and without network information. The number of tested modules/drugs with CTRP (tested) and the number of confirmed modules (validated) with ANOVA test ($p < 0.05$) are shown. Drugs are counted when there is at least one associated modules that are validated (b) Comparison between runs with and without penalty promoting mutual exclusivity (c) Comparison between the combined and separate connectivity models. (d) Average distances between genes in the selected modules. Distances for genes associated with increased sensitivity, decreased sensitivity, all sensitivity, and between decreased and increased sensitivity genes are shown.
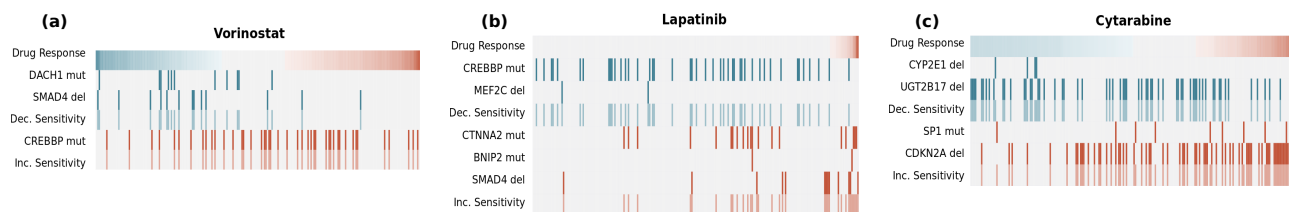
Figure S3: Modules identified by NETPHIX (Related to Figure 3). (a-b) Sensitivity module for Vorinostat (a) and Lapatinib (b). The two modules associated with the drugs are similar but they are associated with opposite directions. The efficacy of combination therapy with Lapatinib and Vorinostat is confirmed in clinical trials. (c) Sensitivity module for Cytarabine identified based on the separate connectivity model.
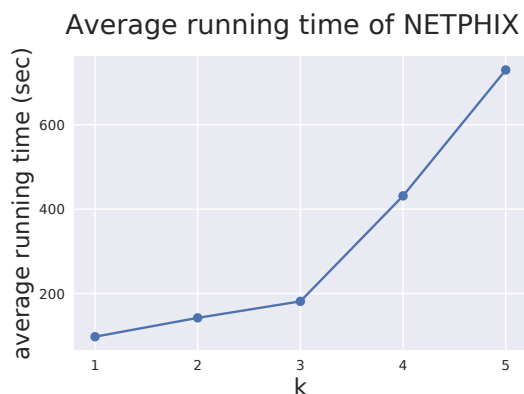
Figure S4: The average running times of NETPHIX over different k's (Related to Section S1.1.3)

S1   Transparent Methods

S1.1   NETPHIX method

S1.1.1   Formal definition of the computational problem for NETPHIX

We are given a graph $G = (V, E)$, with vertices $V = \{1, \dots, n\}$ 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. To identify functionally complementary mutations, we can 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:

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

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 $P = \{P_1, \dots, P_n\}$ 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:

$$W(S) = \sum_{j \in \cup_{s \in S \cap I} P_s} w_j^I - \sum_{j \in \cup_{s \in S \cap I} P_s} (c_{S \cap I}(j) - 1) p_j^I + \left(\sum_{j \in \cup_{s \in S \cap D} P_s} w_j^D - \sum_{j \in \cup_{s \in S \cap D} P_s} (c_{S \cap D}(j) - 1) p_j^D \right)$$

(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 Figure 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, which can be solved using a optimization software package such as CPLEX.

S1.1.2 ILP formulation of NETPHIX

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:

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

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

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

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

$$y_j^I \ge z_j^I, \qquad\qquad\qquad \forall j \tag{7}$$

$$y_j^D \ge z_j^D, \qquad\qquad\qquad \forall j \tag{8}$$

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

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

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

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

Constraint (4) impose that the total number of sets (i.e., selected genes) in the solution is at most k. Constraints (5) and (6) define how many times each sample has been covered by genes in I and D, respectively. Constraints (7) (resp., Constraints (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 (9) (resp., Constraints (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 (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 $C \ge 0.5$, 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 [Fischetti et al., 2017, Bomersbach et al., 2016, Wang et al., 2017].

Note that Constraints (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 (12) with the following constraints.

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

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

S1.1.3 Parameters

To obtain a pool of candidate modules for each drug, we generated ILP instances with different sizes k ($k = 1$ to 5) and two connectivity options (the combined and separate model). The objective function can include a penalty to reinforce mutual exclusivity. As for the 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. The penalties are set to be zero when no penalty is imposed.

We solved the ILP instances to optimality using CPLEX, which can be run in a reasonable amount of time (See Figure S4 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.

S1.1.4   Selecting final 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 (Figure 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 the 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 a feasibility test by simply checking if a permuted instance has a solution with the objective value greater than or equal to $o_\mathcal{D}$; 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 let p_{best} is the most significant p-value for the drug among different parameters. We only considered modules with p-value $= p_{best}$. If $p_{best} < 0.05$ (FDR $< 10\%$, BH), we considered those modules 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 $M_i \subset M_j$, we only include M_j in the final solution for the drug. Therefore, for two

overlapping modules, when one is not a proper subset of the other, both modules may be included.

S1.2 Datasets

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 [Stransky et al., 2015, Barretina et al., 2012], we used the alteration provided at `https://depmap.org/portal/download/all/`. We downloaded the data on July 6th 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. Removing genes not present in the interaction network (see below for the details of interaction network data), we collected 26,917 gene-level alteration profiles (combining amplification, deletion and mutation).

We also utilized an independent drug response dataset from the Cancer Therapeutics Response Portal (CTRP) for validation [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 in `v20.data.curves_post_qc.txt` were used for drug response phenotypes (from `CTRPv2.0_2015_ctd2_ExpandedDataset.zip` file downloaded on August 2nd, 2019).

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 mutation frequency (present in less than 1% samples) from our analyses.

Interaction network: For functional interactions among genes, we used the data downloaded from STRING database version 10.0 (`https://string-db.org`). The data integrates multiple types of interactions including physical interactions. We only included the edges with high confidence scores ($\geq$ 900 out of 1000) as an input to NETPHIX. The resulting interaction network includes 9,215 nodes and 160,249 edges.

S1.3 Evaluation Details

S1.3.1 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 Section S1.2. 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 (`https://github.com/clareli9/rlobico`,release 2018/7/27) 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 ILP instances to be 24h and reported the best current solution (which may be suboptimal) when the program stops.

For ProGENI, we downloaded the program from github (`https://github.com/KnowEnG/ProGENI`, release 2017/7/12). ProGENI originally utilized gene expression information for drug prediction but for comparison with NETPHIX, we used gene alteration profiles instead. The profiles for genes are computed by summing over mutation, deletion and amplification for each gene. Therefore, each entry in the matrix can have a value between 0 and 2 as deletions and amplifications cannot co-occur. We ran a robust version of ProGENI, in which gene prioritization is performed on randomly selected 80% of samples 50 times repeatedly and the resulting ranked lists are aggregated to produce the final ranking of genes.

S1.3.2 Method comparison with real drug screening dataset.

Identifying modules using different methods. UNCOVER modules are obtained by setting $k = 3$ (module size) as presented in [Sarto Basso et al., 2019] and for decreased and increased sensitivity separately. The significance of each module is assessed using permutation tests (using 100 permuted instances), and we consider the modules with $p < 0.05$ as significant modules and used for further analysis. The rankings of genes in ProGENI are computed by performing RobustProGENI with bootstrap sampling rate of 95% and 100 runs.

Computing distance information. To compute the distances from drug targets to the selected modules by the algorithms, we used the drug target information given in Table S1. 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. The identified modules are used for NETPHIX and UNCOVER while 5 and 20 top-ranked genes are used for ProGENI. We also included 5

and 20 randomly selected genes for control.

For NETPHIX modules, we also computed the average distances within modules. The distances are computed by computing the pairwise shortest distances within modules and take the average distances.

Response prediction using random forest regression. We ran RandomForestRegressor in scikit-learn package to learn and test the models. First, we ran 3-fold nested cross validation with GDSC dataset for each drug. For each of training sets, the best model is learned based on 3-fold cross validations inside the training set, with the best parameters estimated using GridsearchCV with combinations of parameters (n_estimators = (10, 100), max_depth = (None, 10, 100), and min_samples_split = (1, 2, 3)). To measure the performance, we used probabilistic concordance index (PCI) defined as in [Costello et al., 2014, Cokelaer et al., 2015, Emad et al., 2017]. The PCI metric compares the ranks of the predicted values and actual AUC scores to compute the prediction power.

For NETPHIX, we merged all modules significantly associated with the response for each drug and used them as features for regression. The number of features for each drug vares between 3 to 15 in total. UNCOVER finds at most one significant module for decreased and increased sensitivity respectively, and genes in both directions are used as features (total of 3 to 6 genes). For ProGENI, we used 20 top ranked genes as features for regression.

We found the drug response profiles for 76 drugs in both CTRP and GDSC datasets, among which 44 drugs with consistent drug response profiles were used (pearson correlation coefficient > 0.25) for validation. For the 44 drugs, we trained the model with GDSC datasets for the modules identified by the algorithm and tested with the responses in CTRP. The hyperparameters were learned using 4-fold cross validation in GDSC among the same parameter combinations given above and the best models were used to predict the drug response values in CTRP.

Validation with CTRP dataset using ANOVA test. To test if the alteration status of selected genes are associated with different drug responses, we also performed ANOVA tests by testing UNCOVER and NETPHIX modules identified from GDSC dataset in the drug responses in CTRP. For each module, 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 tests for the cell survival rates (AUC) in CTRP dataset for the three groups (C_I, C_D, and C_N), and the modules with $p < 0.05$ are considered as validated.

References

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

[Bomersbach et al., 2016] Bomersbach, A., Chiarandini, M., and Vandin, F. (2016). 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.

[Cokelaer et al., 2015] Cokelaer, T., Bansal, M., Bare, C., Bilal, E., Bot, B. M., Chaibub Neto, E., Eduati, F., de la Fuente, A., G?nen, M., Hill, S. M., Hoff, B., Karr, J. R., K?ffner, R., Menden, M. P., Meyer, P., Norel, R., Pratap, A., Prill, R. J., Weirauch, M. T., Costello, J. C., Stolovitzky, G., and Saez-Rodriguez, J. (2015). Dreamtools: a python package for scoring collaborative challenges. *F1000Research*, 4:1030.

[Costello et al., 2014] Costello, J. C., Heiser, L. M., Georgii, E., G?nen, M., Menden, M. P., Wang, N. J., Bansal, M., Ammad-ud din, M., Hintsanen, P., Khan, S. A., Mpindi, J.-P., Kallioniemi, O., Honkela, A., Aittokallio, T., Wennerberg, K., Community, N. D., Collins, J. J., Gallahan, D., Singer, D., Saez-Rodriguez, J., Kaski, S., Gray, J. W., and Stolovitzky, G. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. *Nature biotechnology*, 32(12):1202–12.

[Emad et al., 2017] Emad, A., Cairns, J., Kalari, K. R., Wang, L., and Sinha, S. (2017). Knowledge-guided gene prioritization reveals new insights into the mechanisms of chemoresistance. *Genome biology*, 18(1):153.

[Fischetti et al., 2017] Fischetti, M., Leitner, M., Ljubic, I., Luipersbeck, M., Monaci, M., Resch, M., Salvagnin, D., and Sinnl, M. (2017). Thinning out steiner trees: a node-based model for uniform edge costs. *Math. Program. Comput.*, 9(2):203–229.

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

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

[Seashore-Ludlow et al., 2015] Seashore-Ludlow, B., Rees, M. G., Cheah, J. H., Cokol, M., Price, E. V., Coletti, M. E., Jones, V., Bodycombe, N. E., Soule, C. K., Gould, J., Alexander, B., Li, A., Montgomery, P., Wawer, M. J., Kuru, N., Kotz, J. D., Hon, C. S., Munoz, B., Liefeld, T., Dan?ik, V., Bittker, J. A., Palmer, M., Bradner, J. E., Shamji, A. F., Clemons, P. A., and Schreiber, S. L. (2015). Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discov*, 5(11):1210–1223.

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

[Wang et al., 2017] Wang, Y., Buchanan, A., and Butenko, S. (2017). On imposing connectivity constraints in integer programs. *Math. Program.*, 166(1-2):241–271.