JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 29, Number 12, 2022

Pp. 1305–1323
DOI: 10.1089/cmb.2022.0336

NetMix2: A Principled Network Propagation Algorithm for Identifying Altered Subnetworks

UTHSAV CHITRA,[1,*] TAE YOON PARK,[1,2,*] and BENJAMIN J. RAPHAEL[1,2]

ABSTRACT

A standard paradigm in computational biology is to leverage interaction networks as prior knowledge in analyzing high-throughput biological data, where the data give a score for each vertex in the network. One classical approach is the identification of *altered subnetworks*, or subnetworks of the interaction network that have both outlier vertex scores and a defined network topology. One class of algorithms for identifying altered subnetworks search for high-scoring subnetworks in *subnetwork families* with simple topological constraints, such as connected subnetworks, and have sound statistical guarantees. A second class of algorithms employ *network propagation*—the smoothing of vertex scores over the network using a random walk or diffusion process—and utilize the global structure of the network. However, network propagation algorithms often rely on ad hoc heuristics that lack a rigorous statistical foundation. In this work, we unify the subnetwork family and network propagation approaches by deriving the *propagation family*, a subnetwork family that approximates the sets of vertices ranked highly by network propagation approaches. We introduce NetMix2, a principled algorithm for identifying altered subnetworks from a wide range of subnetwork families. When using the propagation family, NetMix2 combines the advantages of the subnetwork family and network propagation approaches. NetMix2 outperforms other methods, including network propagation on simulated data, pan-cancer somatic mutation data, and genome-wide association data from multiple human diseases.

Keywords: altered subnetworks, anomaly detection, cancer, network analysis, network propagation, GWAS, interaction network.

[1]Department of Computer Science, Princeton University, Princeton, New Jersey, USA.
[2]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey, USA.
*These authors contributed equally to this work.

An early version of this article was published as part of the 2022 Annual International Conference on Research in Computational Molecular Biology (RECOMB), and a preprint of the article was originally deposited to bioRxiv at https://doi.org/10.1101/2022.01.31.478575.

1. INTRODUCTION

A STANDARD PARADIGM IN COMPUTATIONAL BIOLOGY is to use an interaction network as prior knowledge for interpreting high-throughput, genome-scale data. Interaction networks such as protein-protein interaction networks or gene regulatory networks have informed the analysis of biological data in many different applications, including differential expression analysis (Cho et al, 2012; de la Fuente, 2010; Dittrich et al, 2008; Ideker et al, 2002; Ulitsky and Shamir, 2007; Vlaic et al, 2018; Xia et al, 2015), identification of driver mutations in cancer (Creixell et al, 2015; Leiserson et al, 2015; Hofree et al, 2013; Nibbe et al, 2010; Shrestha et al, 2017; Vandin et al, 2011), protein function prediction (Chua et al, 2006; Deng et al, 2003; Nabieva et al, 2005; Radivojac et al, 2013; Sharan et al, 2007), prioritization of germline variants (Califano et al, 2012; Hormozdiari et al, 2015; Huang et al, 2018; Lee et al, 2011; Leiserson et al, 2013; Robinson et al, 2017), and more (Berger et al, 2013; Cornish and Markowetz, 2014; Gligorijević and Pržulj, 2015; Halldórsson and Sharan, 2013; modENCODE Consortium et al, 2010; Wang et al, 2011; Chasman et al, 2016; Chiassian et al, 2015; Luo et al, 2017; Menche et al, 2015; Picart-Armada et al, 2019).

A classical approach for leveraging interaction networks in interpreting high-throughput omics data is the identification of *altered subnetworks*, also called *network modules* or *active subnetworks*. Given an interaction network and a score for each vertex (gene/protein) of the network (e.g., *p*-values from differential gene expression), the goal of the altered subnetwork identification problem is to identify subnetworks (modules) that contain high scoring vertices and conform to some topological condition—for example, connected subnetworks.[†]

Numerous methods for identifying altered subnetworks have been developed (see Berger et al, 2013; Creixell et al, 2015; Cowen et al, 2017; Dimitrakopoulos and Beerenwinkel, 2017; Jia and Zhao, 2014; Mitra et al, 2013) for reviews of these algorithms. Methods to identify altered subnetworks employ a diverse collection of techniques, but they can be grouped into two major classes. The first class of methods rely on the specification of a *subnetwork family*, or a family of subnetworks with a topological constraint; sometimes, the family is stated explicitly—for example, the early approaches such as jActiveModules (Ideker et al, 2001) or heinz (Dittrich et al, 2008) identify connected subnetworks—but in other methods, the subnetwork family is implicitly specified—for example, the optimization problems of Azencott et al (2013) and Liu et al (2017) penalize subnetworks with large cut-size and small edge-density, respectively.

The subnetwork family-based approach is closely related to the identification of *network anomalies* in the data-mining and machine-learning literature (Arias-Castro et al, 2011; Arias-Castro et al, 2008; Arias-Castro et al, 2006; Addario-Berry et al, 2010; Sharpnack et al, 2016; Sharpnack et al, 2013a,b). However, a major challenge with these approaches is to choose an appropriate subnetwork family. For example, connectivity is often too weak of an assumption for biological networks; for example, some methods that use connectivity identify large subnetworks (Nikolayeva et al, 2018) because of a statistical bias in a commonly used test statistic (Reyna et al, 2021; Chitra et al, 2021).

The second class of methods employ a mathematical framework known as *network propagation* (Cowen et al, 2017). Briefly, network propagation uses a random walk or diffusion process to ''smooth'' vertex scores across a network. By using these random walk/diffusion processes, network propagation methods simultaneously account for all possible paths between vertices, and thus fully utilize the *global* structure of the interaction network. Following Cowen et al (2017), we use the term network propagation to refer to the broad class of methods that smooth scores over a network using a random walk or diffusion process.

This includes not only popular processes such as the random walk with restart (Page et al, 1999), but also other processes including the heat kernel (Vandin et al, 2012b; Vandin et al, 2011) or diffusion state distance (Cao et al, 2013; Cowen et al, 2021). Network propagation was first applied in *network ranking* problems, for example, protein function prediction or disease-gene prioritization (Köhler et al, 2008; Weston et al, 2004), where one wants to rank vertices according to their similarity to a subset of vertices with a specific biological function.

These methods were inspired by the success of these random walk, diffusion, and graph kernel methods for ranking problems in statistics and machine learning, for example, the PageRank algorithm (Page et al,

[†]A related problem is the identification of altered subnetworks according to network topology alone. Many of the leading methods for this problem were benchmarked in a recent DREAM competition (Choobdar et al., 2019).

1999). Network propagation has since become the dominant approach for network ranking (Cowen et al, 2017), and it has even been shown to be asymptotically optimal for network ranking for some random graph models (Kloumann et al, 2017).

A major difference between network propagation and the subnetwork family approaches is that network propagation does not output altered subnetwork(s), but only a *ranking* of all genes. Thus, network propagation by itself does not estimate the vertices that are in altered subnetworks, nor even the size of the altered subnetwork(s). Several approaches have attempted to bridge the gap between subnetwork family-based approaches and network propagation, combining the modeling of global network topology from network propagation with heuristics to identify altered subnetwork(s) after performing network propagation.

For example, PRINCE (Vanunu et al, 2010) identifies altered subnetworks as edge-dense subnetworks whose vertices have large network propagated scores. The HotNet algorithms (Vandin et al, 2012a; Vandin et al, 2011; Leiserson et al, 2015; Reyna et al, 2018) identify altered subnetworks by finding clusters in a weighted and directed graph derived from network propagation. TieDIE (Paull et al, 2013) propagates two sets of vertex scores and aims at finding high-scoring subnetworks for both sets of propagated scores. More recently, the NetCore algorithm (Barel and Herwig, 2020) finds subnetworks whose vertices have large node ''coreness'' and large propagated scores.

However, current heuristics to combine network propagation and subnetwork family approaches lack explicit definitions of the subnetwork families, and consequently they do not have provable guarantees for altered subnetwork identification. In contrast, methods that explicitly rely on a well-defined subnetwork family often have statistical or theoretical guarantees, for example, jActiveModules (Ideker et al, 2001) computes a maximum likelihood estimator whereas our recent estimator NetMix is asymptotically unbiased (Chitra et al, 2021; Reyna et al, 2021). Thus, there remains a gap between network propagation and subnetwork-family approaches.

Another important practical issue is the evaluation of altered subnetwork methods. Most network algorithms demonstrate their performance by benchmarking their algorithm against existing network algorithms. Although these comparisons are useful, they may also hide biases shared between algorithms. For example, Lazareva et al (2021) observed that some well-known network algorithms have a bias toward high-degree vertices in the interaction network, whereas Levi et al (2021) observed a bias in GO term enrichment among well-known network algorithms.

To quantify the potential biases of altered subnetwork algorithms, these algorithms need to be compared against carefully selected baselines, including baselines that do not use the interaction network and baselines that do not use the vertex scores.

In this article, we introduce NetMix2, an algorithm which unifies the network propagation and subnetwork family approaches. NetMix2 generalizes NetMix (Reyna et al, 2021) to a wide range of subnetwork families and vertex score distributions. NetMix2 takes as input a wide variety of subnetwork families, including not only the family of connected subgraphs used by existing altered subnetwork methods (Dittrich et al, 2008; Ideker et al, 2002; Reyna et al, 2021) but also any subnetwork family defined by linear or quadratic constraints, such as subnetworks with high edge density or subnetworks with small cut-size.

We use this flexibility to investigate the topology of subnetworks identified by network propagation methods. We show empirically that network propagation does not correspond to standard topological constraints on altered subnetworks such as connectivity (Dittrich et al, 2008; Ideker et al, 2002; Reyna et al, 2021), cut-size (Azencott et al, 2013), or edge-density (Liu et al, 2017). Instead, we derive the *propagation family*, a subnetwork family that we show ''approximates'' the sets of vertices that are ranked highly by network propagation approaches and thereby unifies the two major network approaches in the literature: network propagation and subnetwork family approaches. NetMix2 also uses local false discovery rate (local FDR) methods (Efron, 2007a,b; Efron, 2004) to flexibly model vertex score distributions, in contrast to the strict parametric assumptions made by existing methods (Dittrich et al, 2008; Reyna et al, 2021).

On simulated data we show that NetMix2 outperforms network propagation for subnetworks from the propagation family and other common subnetwork families. Interestingly, NetMix2 outperforms network propagation by the largest margin for the propagation family. We then apply NetMix2 with the propagation family to cancer mutation data and genome-wide association studies (GWAS) data from several complex diseases. On cancer data, we show that NetMix2 outperforms existing network propagation and altered subnetwork methods in identifying cancer driver genes.

On GWAS data, we demonstrate that network propagation often has similar performance to simple baselines that only use the vertex scores or only use the network. However, in cases where network propagation outperforms these baselines, we show that NetMix2 outperforms network propagation.

2. METHODS

2.1. Altered subnetwork problem

We start by formalizing the problem of altered subnetwork identification. Let $G=(V, E)$ be an interaction network with a score X_v for each vertex v. We assume there is an *altered subnetwork* $A \subseteq V$ whose scores $\{X_v\}_{v\in A}$ are drawn independently from a different distribution than the scores $\{X_v\}_{v\notin A}$ of vertices not in the altered subnetwork A. The topology of the altered subnetwork is described by membership in a *subnetwork family* $\mathcal{S} \subseteq \mathcal{P}(V)$, where $\mathcal{P}(V)$ denotes the power set of all subsets of vertices V.

Following the exposition in Chitra et al (2021); Reyna et al (2021) we model the distribution of the scores $\mathbf{X}=\{X_v\}_{v\in V}$ as the altered subnetwork distribution (ASD).

2.1.1. Altered subnetwork distribution. Let $G=(V, E)$ be a graph, let $\mathcal{S} \subseteq \mathcal{P}(V)$ be a subnetwork family, and let $A \in \mathcal{S}$. We say $\mathbf{X}=(X_v)_{v\in V}$ is distributed according to the ASD $ASD_{\mathcal{S}}(A, \mathcal{D}_a, \mathcal{D}_b)$ provided the X_v are independently distributed as

$$X_v \sim \begin{cases} \mathcal{D}_{\mathrm{a}}, & \text{if } v \in A, \\ \mathcal{D}_{\mathrm{b}}, & \text{otherwise}, \end{cases} \tag{1}$$

where $\mathcal{D}_{\mathrm{a}}$ is the altered distribution and $\mathcal{D}_{\mathrm{b}}$ is the background distribution.

The distribution $\mathrm{ASD}_{\mathcal{S}}(A, \mathcal{D}_a, \mathcal{D}_b)$ is parameterized by four quantities: the altered subnetwork A, the subnetwork family $\mathcal{S}$, the altered distribution $\mathcal{D}_{\mathrm{a}}$, and the background distribution $\mathcal{D}_{\mathrm{b}}$.

Given the measurements $\mathbf{X} \sim \mathrm{ASD}_{\mathcal{S}}(A, \mathcal{D}_a, \mathcal{D}_b)$ and the subnetwork family $\mathcal{S} \subseteq \mathcal{P}(V)$, the goal of the *Altered Subnetwork Problem* is to identify the altered subnetwork A. We formalize this problem below.

2.1.2. Altered subnetwork problem. Given $\mathbf{X} \sim \mathrm{ASD}_{\mathcal{S}}(A, \mathcal{D}_a, \mathcal{D}_b)$ and subnetwork family $\mathcal{S}$, find A.

The altered subnetwork problem (ASP) describes a broad class of problems that are studied in many fields, including computational biology (Dittrich et al, 2008; Ideker et al, 2002; Reyna et al, 2021), statistics (Arias-Castro et al, 2011; Addario-Berry et al, 2010; Glaz et al, 2001; Kulldorff, 1997), and machine learning (Chitra et al, 2021; Cadena et al, 2019; Sharpnack et al, 2013a), with different problems making different choices for the distributions $\mathcal{D}_{\mathrm{a}}$, $\mathcal{D}_{\mathrm{b}}$ and the subnetwork family $\mathcal{S}$. Two prominent examples of distributions $\mathcal{D}_{\mathrm{a}}$, $\mathcal{D}_{\mathrm{b}}$ that have been previously studied in the biological literature are the following.

- **Normal distributions:** $\mathcal{D}_{\mathrm{a}}=N(\mu, 1)$ and $\mathcal{D}_{\mathrm{b}}=N(0, 1)$. Normal distributions are often used to model z-scores (Cai et al, 2007; Donoho and Jin, 2004; McLachlan et al, 2006; Pan et al, 2003; Reyna et al, 2021). We call the ASP and ASD with these distributions the *normally distributed ASP* and *normally distributed ASD*, respectively; for notational convenience, we use $\mathrm{NASD}_{\mathcal{S}}(A, \mu)$ to refer to the normally distributed ASD.
- **Beta-uniform distributions:** $\mathcal{D}_{\mathrm{a}}=\mathrm{Beta}(a, 1)$ and $\mathcal{D}_{\mathrm{b}}=\mathrm{Uni}(0, 1)$. Beta-uniform mixture distributions are another common model for p-value distributions (Dittrich et al, 2008; Pounds and Morris, 2003). We call the ASP, ASD with these distributions the *Beta-Uniform ASP* and *Beta-Uniform ASD*, respectively.

We also list several examples of subnetwork families $\mathcal{S}$, where each subnetwork family corresponds to a different topological assumption on the altered subnetwork A. Some of these families have been explicitly applied in biological settings, whereas other families formalize topological constraints that are implicitly made in the biological literature.

- $\mathcal{S}=\mathcal{C}_G$, the *connected family*, or the set of all connected subgraphs S of an interaction network G. Dittrich et al (2008), Ideker et al (2002), and Reyna et al (2021) identify altered subnetworks by solving the ASP for the connected family $\mathcal{C}_G$.

- $\mathcal{S} = \mathcal{E}_{G,p}$, the *edge-dense family*, or the set of all subgraphs S of G with edge-density $\frac{E(S)}{\binom{|S|}{2}} \geq p$, where $E(S) = |\{(u,v) \in E : u \in S, v \in S\}|$ is the number of edges between vertices in S. The edge-dense family $\mathcal{E}_{G,p}$ formalizes the topological constraints made by Guo et al (2007), Liu et al (2017), Vanunu et al (2010), which identify altered subnetworks that have large edge-density.
- $\mathcal{S} = \mathcal{T}_{G,\rho}$, the *cut family*, or the set of all subgraphs S of G with $\frac{\text{cut}(S)}{|S|} \leq \rho$, where $\text{cut}(S) = |\{(u,v) \in E : u \in S, v \notin S\}|$ is the number of edges with exactly one endpoint in S. The cut family $\mathcal{T}_{G,\rho}$ formalizes the topological constraints made by Azencott et al (2013), which identifies altered subnetworks that have small cut.
- $\mathcal{S} = \mathcal{Q}_{G,\rho}$, the *modularity family*, or the set of all subgraphs S of G with modularity $Q(S) \geq \rho$. The modularity family formalizes the topological constraints made by Ayati et al (2015), which identifies altered subnetworks that have high modularity.

We note that the ASP—with the subnetwork families $\mathcal{S}$ described above—describes the problem of identifying a single altered subnetwork in a network G. By creating a new subnetwork family consisting of the union of k disjoint subnetworks in family $\mathcal{S}$, the ASP also describes the problem of identifying *multiple* altered subnetworks.

Early methods for identifying altered subnetwork solved the ASP for the connected family $\mathcal{S} = \mathcal{C}_G$ and different choices of vertex score distributions $\mathcal{D}_a$, $\mathcal{D}_b$. For example, two seminal methods, jActiveModules (Ideker et al, 2002) and heinz (Dittrich et al, 2008), solve the normally distributed and Beta-Uniform ASP, respectively, with the connected family $\mathcal{S} = \mathcal{C}_G$. Recently, we showed that many existing methods, including jActiveModules and heinz, are *biased*, in the sense that they typically estimate subnetworks $\hat{A}$ that are much larger than the altered subnetwork A (Chitra et al, 2021; Reyna et al, 2021).

To this end, we derived the NetMix algorithm, which finds an asymptotically unbiased $\hat{A}_{\text{NetMix}}$ of the altered subnetwork A for the connected family $\mathcal{S} = \mathcal{C}_G$. However, as we demonstrate in a previous work (Reyna et al, 2021) and Section 3 next, many of these methods—including NetMix—have comparable performance to a naive ''scores-only'' baseline that does not use the network G.

2.2. Network propagation and the propagation family

Another strategy often used to incorporate interaction networks G with high-throughput biological data is *network propagation*. Network propagation involves the use of random walk or diffusion processes to ''smooth'' or ''propagate'' vertex scores X_v across a network (Cowen et al, 2017). Formally, given vertex scores X_v, the *network propagated scores* Y_v are computed as

$$Y_v = \sum_{w \in V} M_{v,w} X_w \tag{2}$$

where $M \in \mathbb{R}^{|V| \times |V|}$ is a similarity matrix on the vertices V of the network G typically derived from a random walk on G. One popular choice for the similarity matrix M is the random walk with restart (personalized PageRank) similarity matrix $M_{\text{PPR}} = r(I - (1-r)P)^{-1}$, where $r \in (0,1)$ is the restart probability, I is the identity matrix, and P is the transition matrix for a random walk with restart on G.

A few methods have attempted to use network propagation to identify the altered subnetwork A from propagated scores Y_v, for example, PRINCE (Vanunu et al, 2010) finds edge-dense subnetwork with large propagated scores Y_v. These methods implicitly assume that the propagated scores Y_v are larger for vertices $v \in A$ in the altered subnetwork A compared with vertices $v \notin A$ not in the altered subnetwork A.

However, we empirically find (Section 3.1) that this assumption generally does not hold for altered subnetworks $A \in \mathcal{S}$ from the connected family $\mathcal{S} = \mathcal{C}_G$, the edge-dense family $\mathcal{S} = \mathcal{E}_{G,p}$, and the cut family $\mathcal{S} = \mathcal{T}_{G,\rho}$, which suggests that network propagation methods do not solve the ASP with these subnetwork families $\mathcal{S}$.

Thus, we derive a subnetwork family $\mathcal{S}$ that approximates the sets of vertices that are ranked highly by network propagation methods. Informally, we first observe that network propagation methods identify altered subnetworks A whose vertices $v \in A$ have large propagated scores Y_v. By making the simplifying assumption that the vertex scores $X_v = 1_{\{v \in A\}}$ are binary, we observe that the propagated score $Y_v = \sum_{w \in A} M_{v,w}$ of a vertex v is large if the similarity $M_{v,w}$ is large for many $w \in A$. Intuitively, one natural

way to enforce that the similarities $M_{v,w}$ are large is to lower-bound them, that is, require that $M_{v,w} \geq \delta$ for many $w \in A$ and for some (large) constant $\delta > 0$.

This intuition motivates the formal definition of the *propagation family* $\mathcal{M}_{\delta,p}$, or the set of all subgraphs S such that at least a p fraction of tuples $(u, v) \in S$ have both $M_{u,v} \geq \delta$ and $M_{v,u} \geq \delta$ (We constrain both $M_{u,v}$ and $M_{v,u}$ since the similarity matrix M derived from network propagation is not necessarily symmetric.). We note that the propagation family $\mathcal{M}_{\delta,p}$ is equal to the edge-dense family $\mathcal{E}_{G_\delta,p}$ for the similarity threshold graph $G_\delta = (V, E_\delta)$, which has edge $(u, v) \in E_\delta$ if and only if $M_{u,v} \geq \delta$ and $M_{v,u} \geq \delta$.

We partially formalize our informal derivation of the propagation family $\mathcal{M}_{\delta,p}$ with the following result, which bounds the probability of the altered subnetwork A being the subset of vertices with the largest propagated scores, given data $\mathbf{X} \sim \text{NASD}_{\mathcal{M}_{\delta,p}}(A, \mu)$ from the normally distributed ASD with propagation family $\mathcal{M}_{\delta,p}$ and with density $p = 1$, that is, cliques in the similarity threshold graph G_δ.

Proposition 1. *Let $G = (V, E)$ be a graph and let $M \in [0, 1]^{|V| \times |V|}$ be a matrix indexed by vertices V. Define* $r = \min_{v \in V} M_{v,v}$, $c = \max_{v \notin A} \sum_{w \in A} M_{v,w}$, *and* $d = \max_{v,w \in V} \sum_{u \in V} (M_{w,u} - M_{v,u})^2$. *Let* $\mathbf{X} \sim \text{NASD}_{\mathcal{M}_{\delta,1}}(A, \mu)$ *where* $\mu \geq 1$ *and* $\delta > \frac{c - 2r + 4\sqrt{d \log n}}{|A| - 1}$. *Then with probability at least* $1 - \frac{1}{|V|}$, *the altered subnetwork A consists of the $|A|$ vertices with the largest propagated scores Y_v.*

2.3. NetMix2

We derive the NetMix2 algorithm, which solves the ASP for a wide range of subnetwork families $\mathcal{S}$ and distributions $\mathcal{D}_a$, $\mathcal{D}_b$ (Fig. 1). In particular, NetMix2 solves the ASP for the propagation family $\mathcal{M}_{\delta,p}$, and thus bridges the gap between the ASP and network propagation. NetMix2 consists of two steps.

Step One. The first step of NetMix2 is to estimate the number $|A|$ of vertices in the altered subnetwork A. Our previous method NetMix (Reyna et al, 2021) estimated $|A|$ by fitting the vertex scores $\{X_v\}_{v \in V}$ to a

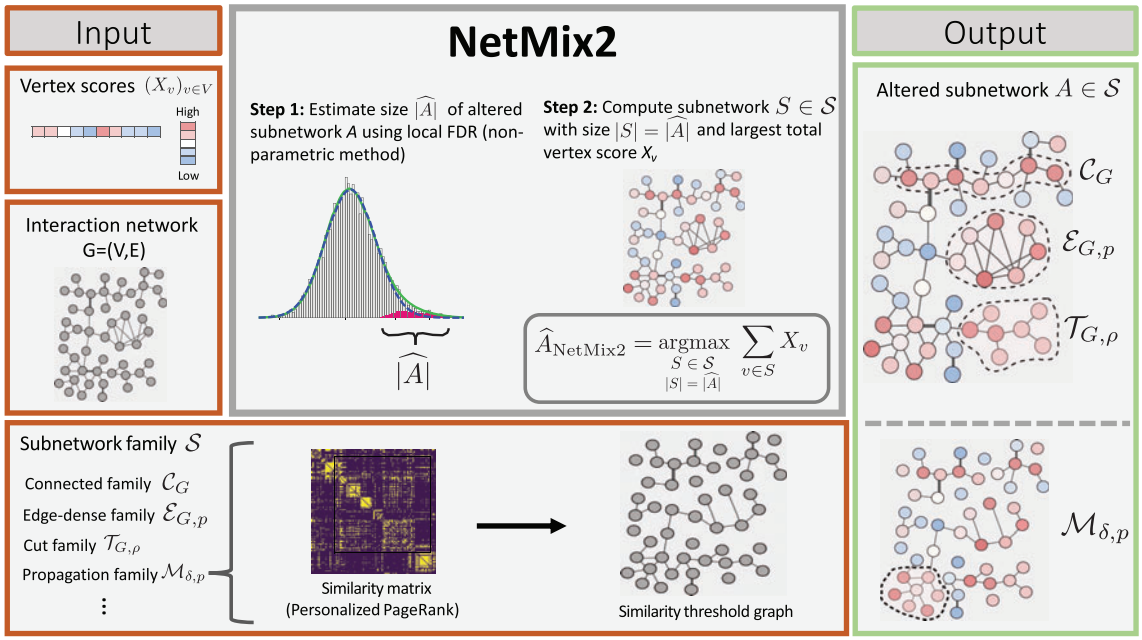

FIG. 1. Overview of the NetMix2 algorithm. The inputs to NetMix2 are a graph G, gene scores $\{X_v\}_{v \in V}$, and a subnetwork family $\mathcal{S}$. First, NetMix2 computes an estimate $\widehat{|A|}$ of the size $|A|$ of the altered subnetwork A using local false discovery rate (local FDR). Next, NetMix2 solves an optimization problem to identify the subnetwork $S \in \mathcal{S}$ size $|S| = \widehat{|A|}$ from the input subnetwork family $\mathcal{S}$ and with the largest total vertex score $\sum_{v \in S} X_v$. By default, NetMix2 uses the *propagation family* $\mathcal{S} = \mathcal{M}_{\delta,p}$. In this case, NetMix2 constructs an additional graph (the similarity threshold graph) based on vertex similarities quantified by Personalized PageRank from the input graph. The choice of subnetwork family $\mathcal{S}$ for NetMix2 is flexible and can be generalized to other families defined by linear or quadratic constraints, including the *connected family* $\mathcal{C}_G$, *edge-dense family* $\mathcal{E}_{G,p}$, and *cut family* $\mathcal{T}_{G,\rho}$.

Gaussian mixture model (GMM), under strict parametric assumptions on the altered distribution $\mathcal{D}_a = N(\mu, 1)$ and background distribution $\mathcal{D}_b = N(0, 1)$. However, not all vertex score distributions are well-fit by normal distributions of this form. Thus, in NetMix2, we extend NetMix by using local false discovery rate (local FDR) methods (Efron, 2007a,b; Efron, 2004) to estimate $|A|$, as local FDR methods make mild assumptions on the forms of the distributions $\mathcal{D}_a$, $\mathcal{D}_b$.

However, one practical issue is that the original local FDR methods (Efron, 2007a,b; Efron, 2004) assume that the altered distribution $\mathcal{D}_a$ is *two-sided*, whereas in our applications we assume that the altered distribution $\mathcal{D}_a$ is *one-sided*. Thus, we define the following heuristic to estimate the number of vertices in a one-sided altered distribution. First, we fit the vertex scores $\mathbf{X} = (X_v)_{v \in V}$ using local FDR methods (Efron et al, 2011; Efron, 2007a,b; Efron, 2004), which yields (1) an estimate $\hat{\mu}_b$ of the mean of the background distribution $\mathcal{D}_b$ and (2) estimates $\widehat{\text{fdr}}_v$ of $\text{fdr}_v = P(v \notin A | X_v)$ for each vertex v (the quantity fdr_v is known as the local FDR for vertex v). We then estimate the size $|A|$ of the altered subnetwork A as the number of vertices v with score $X_v > \hat{\mu}_b$ and estimated local FDR $\widehat{\text{fdr}}_v < 0.5$, that is,

$$\widehat{|A|} = |\{v \in V : X_v > \hat{\mu}_b \text{ and } \widehat{\text{fdr}}_v < 0.5\}|. \tag{3}$$

The first condition, $X_v > \hat{\mu}_b$, ensures that our estimate $\widehat{|A|}$ only counts vertices v with scores X_v that are larger than expected, consistent with our assumption that the altered distribution $\mathcal{D}_a$ is one-sided. The second condition, $\widehat{\text{fdr}}_v < 0.5$, is equivalent to $1 - \widehat{\text{fdr}}_v \geq 0.5$, whose left-hand side is approximately $1 - \widehat{\text{fdr}}_v \approx 1 - \text{fdr}_v = P(v \in A | X_v)$. Thus, the second condition ensures that our estimate of $|A|$ only counts vertices v with probability at least 0.5 of being in the altered subnetwork A.

Step Two. The second step of NetMix2 is to compute the subnetwork $S \in \mathcal{S}$ with size $|S| = \widehat{|A|}$ and the largest total vertex score X_v:

$$\widehat{A}_{\text{NetMix2}} = \underset{\substack{S \in \mathcal{S} \\ |S| \leq \widehat{|A|}}}{\text{argmax}} \sum_{v \in S} X_v. \tag{4}$$

This optimization problem can be formulated as an integer linear program or integer quadratic program for a number of subnetwork families, including the edge-dense family $\mathcal{E}_{G,p}$, the cut family $\mathcal{T}_{G,\rho}$, the connected family $\mathcal{C}_G$, and the propagation family $\mathcal{M}_{\delta,p}$. Note that Eq. (4) involves maximizing the sum $\sum_{v \in S} X_v$ of the vertex scores X_v, whereas the objective in the NetMix optimization problem (Reyna et al, 2021) is the sum $\sum_{v \in S} r_v$ of the vertex responsibilities $r_v = P(v \in A | X_v)$. In practice, we observe that maximizing the sum of the vertex scores X_v yields slightly better performance than maximizing the sum of the responsibilities r_v.

2.3.1. Parameter selection. The definition of the propagation family $\mathcal{M}_{\delta,p}$ depends on two parameters: the similarity threshold δ which defines the edges E_δ in the similarity threshold graph G_δ and the minimum edge density p of altered subnetworks. In our analyses below, we chose the values of these parameters to reflect the network properties of the input diseases. For all analyses, we set δ so that the number $|E_\delta|$ of edges in the similarity threshold graph is 40% of the original protein interaction network, rounded to the nearest 25,000.

We chose this parameter to have a large number $|E_\delta|$ of pairs of vertices in the similarity threshold graph G_δ with high pairwise similarity while also balancing the computational tractability of NetMix2, whose run-time increases with the number $|E_\delta|$ of edges. We set the minimum edge density parameter p according to the network properties of a set R of reference genes for each disease. See the Supplementary Notes and Eq. (16) in Supplementary Data S1 for more details, and see Table S2 for the parameter values we use in our analyses (Section 3).

2.3.2. Implementation, data, and code availability. We implement NetMix2 using Python 3. We use the Python implementation of locFDR R package from https://github.com/leekgroup/locfdr-python for the first step of NetMix2. We use the Gurobi optimizer (Gurobi Optimization, LLC, 2021) to solve the integer linear/quadratic program in Eq. (4). NetMix2 code as well as a tutorial (Jupyter notebook) for running NetMix2 are publicly available at https://github.com/raphael-group/netmix2.

2.4. Scores-only and network-only baselines

When evaluating any algorithm for the identification of altered subnetworks, we argue that it is essential to compare against two baselines: a ''scores-only'' baseline that only uses the vertex scores X_v and a ''network-only'' baseline that only uses the interaction network G. These two baselines quantify whether the altered subnetwork algorithm is outperforming simpler approaches that do not integrate vertex scores with a network.

These baselines should be evaluated on each dataset and match as closely as possible the inputs to the altered subnetwork problem. A scores-only baseline is straightforward: we rank the vertices v by their vertex scores X_v. Because this baseline outputs a ranked list of all vertices in the graph, we threshold the ranking when evaluating against other altered subnetwork algorithms by taking the k most highly ranked vertices for some integer k.

Defining a network-only baseline is a more subtle issue, and it was discussed in two recent articles (Levi et al, 2021; Lazareva et al, 2021). Levi et al (2021) benchmarks altered subnetwork algorithms on randomly permuted vertex scores $\widetilde{X}_v$ while keeping the network G fixed. The authors find that many existing methods output similar altered subnetworks (in terms of GO enrichment) on their permuted data, which suggests that these methods are utilizing the network G more than the vertex scores X_v.

Lazareva et al (2021) benchmarks altered subnetwork algorithms on randomly permuted networks with the same degree distribution as G while keeping the vertex scores X_v fixed. The authors find that many existing algorithms output similar altered subnetworks on permuted networks, indicating a degree bias in these methods. We propose a more direct network-only baseline: we rank vertices v by their network centrality score $N(v)$ for a network centrality measure N that is derived from the topological constraints used by the altered subnetwork algorithm.

For example, for an algorithm that relies on the connected subfamily, we propose that degree centrality $N(v)=d_v$ is an appropriate measure, as in Lazareva et al (2021). However, for network propagation algorithms that use random walk with restart, we claim that the PageRank centrality $N(v)=(M_{\text{PPR}}\cdot\mathbf{1})_v$, where $\mathbf{1}\in\mathbb{R}^n$ is an all-ones vector, is the more appropriate network-only baseline. This is because compared with degree centrality, PageRank centrality better captures how network propagation methods use the interaction network G.

3. RESULTS

We evaluated NetMix2 on simulated data and on real datasets, including somatic mutations in cancer and genome-wide association studies (GWAS) from several diseases. Unless indicated otherwise, we ran NetMix2 with the propagation family $\mathcal{M}_{\delta,p}$ using the personalized PageRank matrix M_{PPR} with restart probability $r=0.4$, using the parameters in Supplementary Table S1. We solved the integer program in Eq. (4) using the Gurobi optimizer (Gurobi Optimization, LLC, 2021).

We ran Gurobi for up to 24 hours, which typically results in a near-optimal solution for the protein-protein interaction networks G that we used. For all ranking methods (e.g., network propagation, scores-only, and network-only baselines), we estimated the altered subnetwork $\widehat{A}$ as the $|\widehat{A}_{\text{NetMix2}}|$ highest ranked vertices, where $\widehat{A}_{\text{NetMix2}}$ is the output of NetMix2.

3.1. Simulated data

We compare NetMix2, network propagation, a scores-only baseline, and a network-only baseline on simulated instances of the Altered Subnetwork Problem with various subnetwork families $\mathcal{S}$ derived from the HINT+HI protein interaction network $G=(V,E)$ (Leiserson et al, 2015; Reyna et al, 2021). This network contains 15,074 vertices and around 170,000 edges. The edges E are a union of the protein-protein and protein-complex interactions between proteins from the HINT network (Das and Yu, 2012) and the protein-protein interactions from the HI network (Rolland et al, 2014).

For each instance, we randomly selected a subnetwork $A\in\mathcal{S}$ of size $|A|=0.01n$ and drew a sample $\mathbf{X}\sim\text{ASD}_{\mathcal{S}}(A,\mathcal{D}_a,\mathcal{D}_b)$ with altered distribution $D_a=N(1.8,1.2)$ and background $D_b=N(0.1,0.9)$. Note that we use slightly different altered and background distributions compared with the normally distributed ASD, so as to reflect the systemic errors in measurement often found in real data (Efron, 2007b,a; Efron, 2004).

We implant altered subnetworks A from four different subnetwork families $\mathcal{S}$: the connected family $\mathcal{S}=\mathcal{C}_G$, the edge density family $\mathcal{S}=\mathcal{E}_{G,p}$ with density $p=0.15$, the cut family $\mathcal{S}=\mathcal{T}_{G,\rho}$ with cut size $\rho=6$, and the propagation family $\mathcal{S}=\mathcal{M}_{\delta,p}$ with δ chosen so that G_δ has 200,000 edges and with edge density $p=0.3$ in G_δ. The edge density (resp. cut-size) parameters are chosen to be at least three standard deviations above (resp. below) the average edge density (resp. cut-size) for a subnetwork A of size $|A|=0.01n$.

We ran each method on vertex scores $\mathbf{X}$, with NetMix2 also having the subnetwork family $\mathcal{S}$ as input, to obtain an estimate $\widehat{A}$ of the altered subnetwork A. We found that NetMix2 outperformed the other three methods—network propagation, scores-only, and network-only—for altered subnetworks $A \in \mathcal{S}$ from all four families $\mathcal{S}$ (Fig. 2). However, the advantage of NetMix2 over the other methods depended on the subnetwork family $\mathcal{S}$.

For example, NetMix2 only slightly outperformed the scores-only baseline for the connected family $\mathcal{C}_G$ and cut family $\mathcal{T}_{G,\rho}$, and it only moderately outperformed the scores-only baseline for the edge-dense family $\mathcal{E}_{G,p}$, suggesting that these families impose relatively weak constraints on the altered subnetwork A, as discussed for the connected family by Reyna et al (2021). On the other hand, NetMix2 substantially outperformed the scores-only baseline for the propagation family $\mathcal{M}_{\delta,p}$, demonstrating that the propagation family $\mathcal{M}_{\delta,p}$ provides strong topological constraints on the altered subnetwork.

We also observed (Fig. 2) that the performance of network propagation depends on the subnetwork family $\mathcal{S}$. In particular, network propagation had similar performance to the scores-only baseline for the connected family $\mathcal{C}_G$ and edge dense family $\mathcal{E}_{G,p}$ and worse performance for the cut family $\mathcal{T}_{G,\rho}$, suggesting that network propagation is not well suited to identifying altered subnetworks A from these subnetwork families.

On the other hand, network propagation had a substantial gain over the scores-only baseline for the propagation family $\mathcal{M}_{\delta,p}$. This demonstrates that the propagation family contains subnetworks whose vertices are likely to be highly ranked by network propagation. Interestingly, we observe that for altered subnetworks A from the propagation family $\mathcal{M}_{\delta,p}$, the network-only baseline also outperforms the scores-only baseline.

This is most likely due to the close relationship between the propagation family $\mathcal{M}_{\delta,p}$ and the network-only baseline: the network-only baseline ranks vertices by their PageRank centrality whereas the propagation family $\mathcal{M}_{\delta,p}$ is derived from the personalized PageRank matrix. Interestingly, ranking vertices by their degrees—as suggested by Lazareva et al (2021)—performed much worse than PageRank centrality (F-measure ≈ 0.05, not shown in Fig. 2) for altered subnetworks A from the propagation family $\mathcal{M}_{\delta,p}$, demonstrating the importance of using a network-only baseline that models the same topological properties as the network method.

3.2. *Somatic mutations in cancer*

Next, we compared the performance of NetMix2 on the task of identifying cancer driver genes against four other methods for identification of altered subnetworks: NetMix (Reyna et al, 2021), Heinz (Dittrich et al, 2008), NetSig (Horn et al, 2018), and Hierarchical HotNet (Reyna et al, 2018). We also compared

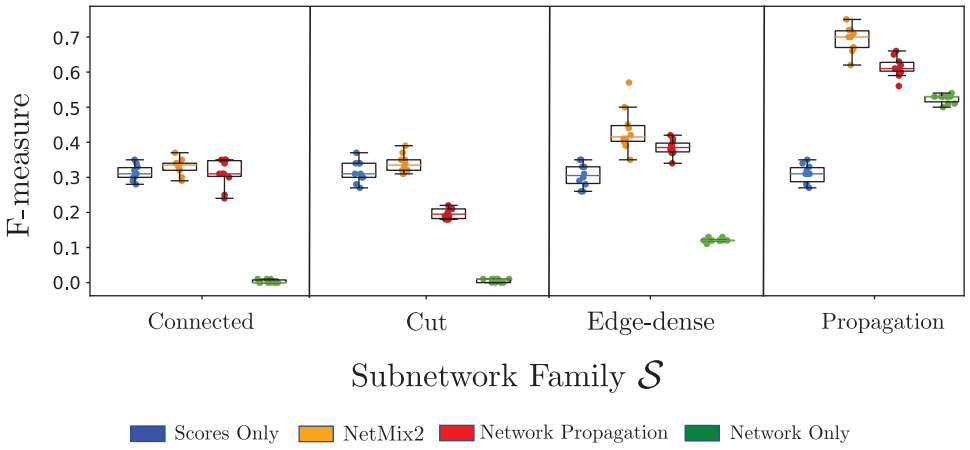

FIG. 2. Comparison between NetMix2, network propagation, a scores-only baseline, and the PageRank centrality network-only baseline in identifying altered subnetworks from different subnetwork families $\mathcal{S}$ implanted in the HINT+HI interaction network.

with four ranking methods: network propagation, the scores-only baseline, the network-only baseline, and the degrees-only baseline on MutSig2CV p-values from The Cancer Genome Atlas (TCGA) PanCanAtlas project using the STRING interaction network.

For each vertex (gene) v, the vertex score X_v is a z-score computed from p-values from MutSig2CV (Lawrence et al, 2014), a statistical method that predicts cancer driver genes based on the frequency that the gene is mutated in a cohort of cancer patients. We obtained these scores for 10,437 samples across 33 cancer types from the TCGA PanCanAtlas project (Bailey et al, 2018). We ran each method using the vertex scores X_v and the STRING protein-protein interaction network (Szklarczyk et al, 2015) and evaluated the performance by computing the overlap between genes in their reported subnetworks and reference lists of cancer driver genes from the COSMIC cancer gene census (CGC) (Tate et al, 2019), OncoKB (Chakravarty et al, 2017), and TCGA (Bailey et al, 2018). Further details on datasets and procedures for running each method are included in the Supplementary Data S1.

We found that NetMix2 using the propagation family outperformed other methods in F-measure for all three reference gene sets (Table 1). In addition, comparing NetMix2 using the propagation family and NetMix2 using the connected family (the second best method) shows that the altered subnetwork found using the propagation family contains several genes that are not found by using the connected family (Supplementary Fig. S1).

For example, NetMix2 using the propagation family identifies nine CGC driver genes that are not found by using the connected family including *PDGFRA*, an oncogene whose gain-of-function mutations promote cancer growth (Velghe et al, 2014) and *NCOR2*, a well-known tumor suppressor implicated in breast and prostate cancers (Battaglia et al, 2010); none of these genes are found by the baseline methods (Supplementary Fig. S2).

We also found that NetMix2 outperforms the altered subnetwork methods with different parameter settings, including Heinz with FDR = 0.005 as well as the network-only baseline with degree centrality (Supplementary Table S2). In particular, ranking vertices by degree had a substantially lower F-measure compared with ranking vertices by PageRank centrality—again demonstrating the importance of an appropriately chosen network-only baseline. We also observe that many network approaches, including NetSig and Hierarchical HotNet, had lower F-measure than the scores-only baseline. Although it is possible that we did not use the optimal parameters for these methods, our results suggest that these methods were over-fitting to the network compared with the vertex scores.

3.3. Genome-wide association studies

We next applied NetMix2 to data from genome-wide association studies (GWAS), another application where network approaches are often used to prioritize germline variants that are associated with biological

Table 1. Results of Altered Subnetwork Identification Methods Using MutSig2CV Cancer Driver Gene p-Values from TCGA Tumor Samples

		STRING network					
		CGC		*OncoKB*		*TCGA*	
Method	*Subnetwork size*	*Number*	F-*measure*	*Number*	F-*measure*	*Number*	F-*measure*
NetMix2	280	132	**0.3**	133	**0.313**	151	**0.546**
NetMix	313[a]	129	0.282	130	0.295	147	0.502
Heinz (FDR = 0.01)	335	139	0.297	138	0.306	156	0.513
NetSig	773	145	0.211	172	0.257	84	0.161
Hierarchical HotNet	246	73	0.172	70	0.172	74	0.285
Network propagation	280	86	0.195	89	0.210	98	0.354
Scores-only	280	126	0.286	127	0.3	145	0.524
Network-only	280	77	0.175	83	0.196	55	0.199

Subnetworks are evaluated using reference sets of cancer genes from CGC, OncoKB, and TCGA. The best scores are colored in bold red.

[a]GMM from NetMix overestimated the size of the subnetwork, thus we excluded genes with outlier scores as described by Reyna et al (2021).

CGC, cancer gene census; TCGA, The Cancer Genome Atlas.

traits (Greene et al, 2015; Jia and Zhao, 2014). We first analyzed GWAS data for eight different disease traits from Carlin et al (2019). The original study by Carlin et al (2019) introduced the NAGA method (also in companion article by Fong et al, 2019) that outputs a *ranked list* of genes by applying network propagation to gene scores X_v obtained by selecting the single nucleotide polymorphism (SNP) with minimum p-value in the neighborhood of each gene v. The genes are then ranked by their propagated scores Y_v, and these rankings are evaluated using reference lists of disease genes from the DisGeNET database (Piñero et al, 2020).

As a preliminary analysis, we first compared NAGA (network propagation) against our scores-only and network-only (PageRank centrality) baselines. The original study by Carlin et al (2019) claimed that network propagation outperforms the scores-only baseline (as well as existing network ranking methods for GWAS) in recovering reference genes for all eight diseases. They demonstrated their claim by comparing network propagation against other ranking methods (rather than as altered subnetwork methods) using the area under the receiver operating characteristic curve (AUROC) metric.

However, when the reference gene list is small (which is typically the case for GWAS data), the AUROC of a ranking algorithm is not necessarily representative of its performance; two network ranking algorithms may have similar AUROC even if they have noticeably different performance (Davis and Goadrich, 2006). Therefore, we re-evaluated the network ranking approaches from Carlin et al (2019) for prioritizing disease genes using area under the precision-recall curve (AUPRC), a metric that is reported to be a more appropriate metric for assessing ranking algorithms compared with AUROC when the reference sets are small (Davis and Goadrich, 2006).

We focused our comparison on network propagation (using the same parameters as NAGA), the scores-only baseline, and the network-only baseline using PageRank centrality. We used the PCNet interaction network G (Huang et al, 2018), the same network used in the original evaluation by Carlin et al (2019).

We found that—contrary to the claims of Carlin et al (2019)—network propagation does not always outperform the scores-only baseline (Fig. 3). Further, based on the additional comparison to the network-only baseline, we identified three distinct groups of GWAS datasets (Fig. 3). In the first group are three diseases (schizophrenia, hypertension, type 1 diabetes) where the network-only AUPRC was comparable

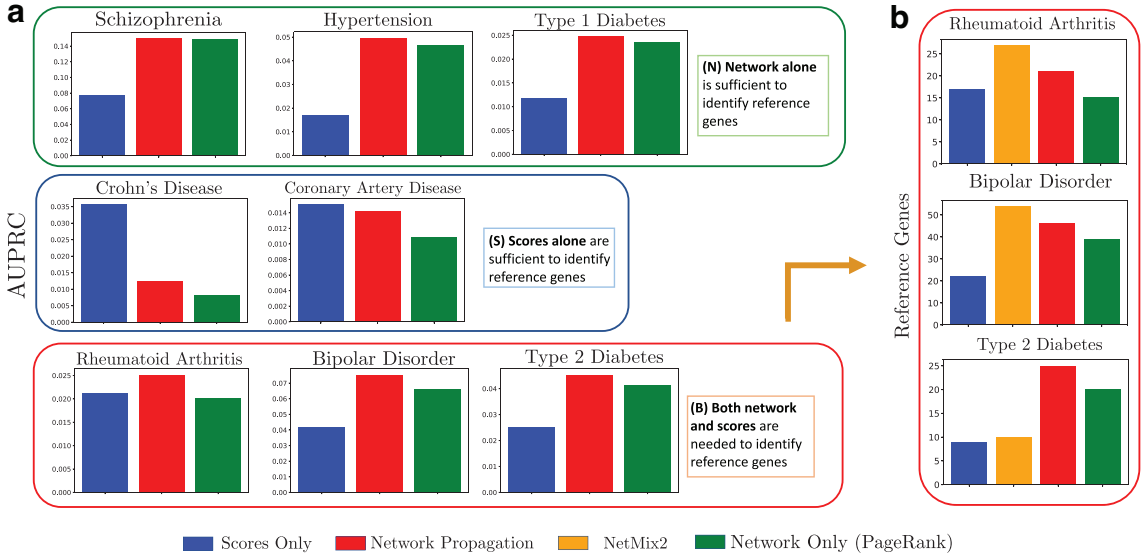

FIG. 3. **(a)** Comparison of three network ranking methods—network propagation, scores-only, and network-only (PageRank centrality)—on GWAS data from Carlin et al (2019). Network ranking methods are evaluated by their AUPRC using reference lists of disease genes (Piñero et al, 2020). **(b)** Comparison of four altered subnetwork identification methods—NetMix2, network propagation, scores-only, and network-only (PageRank centrality)—for diseases from **(a)** where network propagation has at least 10% larger AUPRC compared with the scores-only and network-only baselines. Methods are evaluated according to the number of reference genes in the estimated altered subnetwork. AUPRC, area under the precision-recall curve.

(<10% difference) to the network propagation AUPRC. In other words, the first group consists of datasets where the vertex scores X_v did not seem to add much value compared with using only the network G in identifying reference genes. Interestingly, this group includes schizophrenia, a disease that Barel and Herwig (2020), Carlin et al (2019) specifically highlight as a case where network propagation outperformed the scores-only baseline in AUROC.

The second group consists of two diseases (Crohn's disease, coronary artery disease) where the scores-only AUPRC was comparable to (or larger than) the network propagation AUPRC. This group consists of diseases where the network G seemingly did not add much value compared with using only the vertex scores X_v in identifying reference genes. The third and final group consists of three diseases (Rheumatoid arthritis, bipolar disorder, type 2 diabetes) where the network propagation AUPRC is noticeably larger (by at least 10%) than the AUPRC for both scores-only and network-only.

This group consists of diseases where using both the network G and scores X_v is better at identifying reference genes than using either alone, and it is the group of diseases where one would expect altered subnetwork approaches to perform well. We also note that for all eight diseases, ranking vertices by their degree (not shown in Fig. 3) had lower AUPRC than ranking vertices by their PageRank centrality, again illustrating the benefit of comparing with appropriate network-only baselines.

Next, we compared NetMix2 against network propagation and the two baselines for the three diseases in the latter group: rheumatoid arthritis, bipolar disorder, and type 2 diabetes (Fig. 3b). We found that for two out of three diseases (rheumatoid arthritis and bipolar disease) NetMix2 identified noticeably more reference genes than network propagation. These results are consistent with the simulations in Figure 2, and they demonstrate that NetMix2 using the propagation family outperforms network propagation for the diseases where *both* the scores and the network are important for identifying reference disease genes.

We also observed that for type 2 diabetes, our procedure for setting the minimum edge density parameter p appears to be sub-optimal, as NetMix2 using $p = 0.2$ identifies a subnetwork with 20 reference genes compared with the 10 reference genes identified with NetMix2 using $p = 0.05$ (Fig. 3b).

One empirical observation from comparing the subnetworks identified by different approaches is that the NetMix2 subnetwork was similar to the scores-only subnetwork, whereas the network propagation subnetwork was more similar to the network-only subnetwork (Supplementary Fig. S3). As a result, we expect that the quality of the gene scores X_v computed from the GWAS summary statistics of individual SNPs plays an important role for the performance of NetMix2 as well as the scores-only baseline.

Recent studies by Lamparter et al (2016), Nakka et al (2016) report that the minSNP method—which finds the most significant SNP in surrounding regions of each gene and was used by Carlin et al (2019) to compute the gene scores in this experiment—introduce bias toward longer genes and do not account for linkage disequilibrium, a well-known confounding factor in GWAS summary statistics (Uffelmann et al, 2021).

To address this issue, we examined another set of GWAS (Levi et al, 2021) where gene scores are computed using the *Pascal* method (Lamparter et al, 2016). Briefly, Pascal aggregates SNP p-values from GWAS into gene scores while correcting for confounders, including linkage disequilibrium and gene length. We applied NetMix2 to Pascal scores of six diseases from the GWAS data in Levi et al (2021).

We used the STRING interaction network (Szklarczyk et al, 2015), the same network used by Levi et al (2021), and compared against three ranking methods—network propagation, the scores-only baseline, and the network-only (PageRank centrality) baseline—as well as DOMINO, the altered subnetwork identification method presented in Levi et al (2021). Similar to the previous experiment, we evaluated the methods by computing the overlap between subnetworks identified by each method and DisGeNET reference disease gene sets (Piñero et al, 2020).

We found that the Pascal gene scores alone represent a strong signal for recovering disease reference genes as the scores-only baseline outperformed other methods in four of the six diseases, namely atrial fibrillation, coronary artery disease, age-related macular degeneration, and Crohn's disease (Fig. 4). Consistent with the results on the previous GWAS experiment, we also found that the NetMix2 subnetwork was more similar to the scores-only subnetwork, whereas the network propagation subnetwork was more similar to the network-only subnetwork (Supplementary Fig. S4).

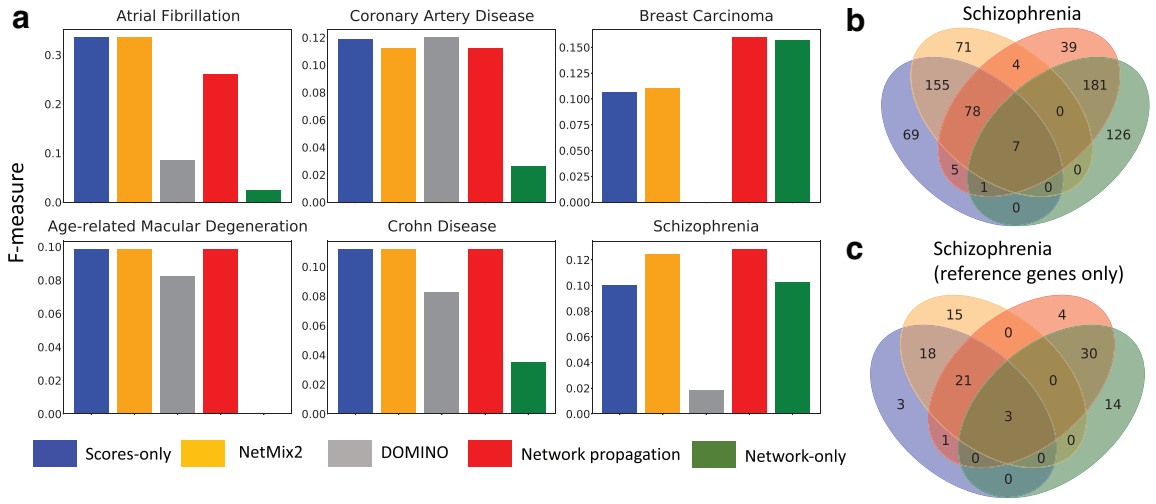

FIG. 4. **(a)** Comparison of NetMix2, DOMINO, network propagation, and the scores-only and network-only (PageRank centrality) baseline methods on GWAS data from Levi et al (2021) using the STRING interaction network (Szklarczyk et al, 2015). **(b)** Overlap in genes in altered subnetworks in schizophrenia identified by scores-only, NetMix2, network propagation, and network-only. **(c)** Schizophrenia reference genes in the altered subnetworks from **(b)**. GWAS, genome-wide association studies.

Thus, it is not surprising to see that NetMix2 performed nearly as well as the scores-only baseline in these four diseases. The strong performance of scores-only baseline using Pascal scores demonstrates that the quality of gene scores is an important component of the performance of altered subnetwork methods.

We next evaluated DOMINO against the scores-only and network-only baselines—comparisons that were missing from the DOMINO publication (Levi et al, 2021)—and against NetMix2. We found that DOMINO performed worse than the baseline methods as well as NetMix2 in five of the six diseases, with coronary artery disease being the lone exception (Fig. 4). On the other hand, NetMix2 was comparable to the best performing methods in five diseases (Fig. 4a). For schizophrenia, the altered subnetwork identified by NetMix2 contained 71 genes that were not found by the ranking methods (Fig. 4b).

These 71 genes are significantly enriched (p-value = 6.46e-06, fold enrichment = 3.34; hypergeometric test) for the reference genes for schizophrenia from the DisGeNet database (Fig. 4c). Further, the altered subnetworks identified by NetMix2 for schizophrenia as well as breast carcinoma are significantly enriched (schizophrenia: p-value = 2.04e-11, fold enrichment = 16.1; hypergeometric test, and breast carcinoma: p-value = 1.56e-22, fold enrichment = 6.38; hypergeometric test) for genes whose expression is significantly associated with the corresponding diseases according to recent studies of expression quantitative trait loci (eQTL) and GWAS data (Guo et al, 2018; Zhu et al, 2016).

We also find that the procedure for setting the minimum edge density parameter p is sub-optimal for schizophrenia, as NetMix2 using $p = 0.07$ outperforms all other methods, identifying a subnetwork with 61 reference genes compared with the 57 reference genes identified with NetMix2 using $p = 0.06$ (Supplementary Fig. S5). Lastly, we note that the results were qualitatively similar when we used the size of the subnetwork identified by DOMINO (instead of the subnetwork found by NetMix2) to threshold the ranking methods (Supplementary Fig. S6).

Taken together, these results demonstrate the importance of evaluating methods for identifying altered subnetwork against the scores-only or the network-only baseline methods, both to gauge the potential bias from either source of information and to comprehensively assess the performance of a network method.

4. DISCUSSION

We introduced NetMix2, an algorithm that unifies the network propagation and subnetwork family-based approaches for analyzing biological data using interaction networks. NetMix2 is inspired by network propagation, a standard approach for solving the *network ranking* problem, and attempts to bridge the gap

between two paradigms for using networks in the analysis of high-throughput genomic data—*network ranking* and the identification of *altered subnetworks*—in a principled way by explicitly deriving a new family of subnetworks called the *propagation family* that approximates the altered subnetworks found by network propagation methods. We showed that NetMix2 is effective in finding disease-associated genes using somatic mutation data in cancer and GWAS data from multiple diseases.

Our evaluation also revealed that simple baseline methods that use either only the vertex scores or only the interaction network sometimes perform surprisingly well, often outperforming more sophisticated network methods. Although publications describing new network methods typically benchmark against other network methods, these publications are wildly inconsistent in benchmarking against scores-only and network-only baselines, and only rarely does a publication contain benchmarks using both baselines.

Further, some publications use a network-only baseline that uses a different type of network information compared with the method under evaluation. For a fair comparison, it is essential that the network-only baseline and the proposed altered subnetwork method use the same network information; for example, PageRank centrality is a more appropriate benchmark for network propagation methods than vertex degree.

Although NetMix2 outperformed every other method we tested on cancer data, the performance of NetMix2—and other network methods—was generally underwhelming on the GWAS data; we observed only a modest improvement over baseline methods even in the diseases where NetMix2 worked well. There are several possible reasons for this discrepancy. First, the diseases in the GWAS data that we analyzed may have high genetic complexity. Indeed, multiple studies have suggested that the signal from GWAS data for complex traits are more widely dispersed throughout the genome (including non-coding genomic regions) than previously predicted, resulting in very small effect sizes of individual entities (SNPs/genes).

An extreme example is the omnigenic model (Boyle et al, 2017), which posits that nearly all genes have functional relevance to the GWAS trait. Such complexity coupled with various challenges in interpreting the GWAS data, for example, linkage disequilibrium, could result in a high number of false positives—genes with significant p-values that are not related to the trait—or a gene score distribution with an unusual shape that is not well fit by the semi-parametric models in local FDR methods. These complications could yield an inaccurate estimate of the size of the altered subnetwork, which, in turn, would be detrimental to the performance of NetMix2.

An inaccurate estimate of the altered subnetwork size would also affect our evaluation of ranking methods, which we evaluated using the same number of highly ranked genes as the size of the estimated altered subnetwork. Finally, it is possible that SNPs associated with some diseases affect multiple biological processes in distinct subnetworks, and that a single altered subnetwork is too restrictive to describe the genetic signal.

There are several directions for future work. The first direction is to extend NetMix2 to identify *multiple* altered subnetworks simultaneously. This can be done by running NetMix2 iteratively, or by modifying the integer program to output multiple solutions. However, solving the corresponding model selection procedure to choose the number and sizes of altered subnetworks without overfitting is a difficult problem. At the same time, allowing for multiple altered subnetworks may simplify the parameter selection for the propagation family, allowing for larger minimum edge densities in the similarity graph.

A second direction is to extend NetMix2 with an appropriate permutation test to evaluate the statistical significance of the altered subnetwork(s). Third, we observed that our parameter selection procedure for the minimum edge density parameter was sometimes sub-optimal, for example, in the GWAS evaluations, using a different minimum edge density would sometimes identify a larger number of reference genes. Developing a more biologically relevant parameter selection is an important future direction.

Finally, although we evaluated several network methods and simple baselines, there are numerous other network methods that could be included in these benchmarks. However, there are few gold standards to perform such a comprehensive evaluation as the reference disease gene sets remain relatively limited and potentially biased by their sources. Thus, a useful extension would be deriving a reliable evaluation scheme for network methods that accounts for various sources of bias including the ascertainment bias in current interaction networks and disease gene sets.

ACKNOWLEDGMENTS

The authors would like to thank Jasper C.H. Lee and Christopher Musco for helpful discussions, as well as Matthew Myers and Palash Sashittal for reviewing early versions of the article.

AUTHORS' CONTRIBUTIONS

Conceptualization: B.J.R.; methodology: U.C., B.J.R.; software: U.C., T.P.; validation: U.C., T.P.; formal analysis: U.C., T.P., and B.J.R.; investigation: U.C., T.P.; data curation: U.C., T.P.; writing—original draft: U.C., T.P., and B.J.R.; writing—review and editing: U.C., T.P., and B.J.R.; visualization: U.C., T.P.; supervision: B.J.R.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no conflicting financial interests.

FUNDING INFORMATION

U.C. is supported by NSF GRFP DGE 2039656. B.J.R. is supported by grant U24CA264027 from the National Cancer Institute (NCI).

SUPPLEMENTARY MATERIAL

Supplementary Data S1
Supplementary Figure S1
Supplementary Figure S2
Supplementary Figure S3
Supplementary Figure S4
Supplementary Figure S5
Supplementary Figure S6
Supplementary Figure S7
Supplementary Table S1
Supplementary Table S2

REFERENCES

Addario-Berry L, Broutin N, Devroye L, et al. On combinatorial testing problems. Ann Stat 2010;38(5):3063–3092.

Arias-Castro E, Candès EJ, Helgason H, et al. Searching for a trail of evidence in a maze. Ann Stat 2008;36(4):1726–1757.

Arias-Castro E, Candès EJ, Durand A. Detection of an anomalous cluster in a network. Ann Stat 2011;39(1):278–304.

Arias-Castro E, Donoho DL, Huo X. Adaptive multiscale detection of filamentary structures in a background of uniform random points. Ann Stat 2006;34(1):326–349.

Ayati M, Erten S, Chance MR, et al. MOBAS: Identification of disease-associated protein subnetworks using modularity-based scoring. EURASIP J Bioinform Syst Biol 2015;2015(1):7.

Azencott C-A, Grimm D, Sugiyama M, et al. Efficient network-guided multi-locus association mapping with graph cuts. Bioinformatics 2013;29(13):i171–i179.

Bailey MH, Tokheim C, Porta-Pardo E, et al. Comprehensive characterization of cancer driver genes and mutations. Cell 2018;173(2):371–385.

Barel G, Herwig R. NetCore: A network propagation approach using node coreness. Nucleic Acids Res 2020;48(17):e98–e98.

Battaglia S, Maguire O, Campbell MJ. Transcription factor co-repressors in cancer biology: Roles and targeting. Int J Cancer 2010;126(11):2511–2519.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Ser B (Methodol) 1995;57(1):289–300.

Berger B, Peng J, Singh M. Computational solutions for omics data. Nat Rev Genet 2013;14(5):333–346.

Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: From polygenic to omnigenic. Cell 2017;169(7):1177–1186.

Cadena J, Chen F, Vullikanti A. Near-optimal and practical algorithms for graph scan statistics with connectivity constraints. ACM Trans Knowl Discov Data 2019;13(2):20:1–20:33.

Cai TT, Jin J, Low MG. Estimation and confidence sets for sparse normal mixtures. Ann Stat 2007;35(6):2421–2449.

Califano A, Butte AJ, Friend S, et al. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. Nat Genet 2012;44(8):841–847.

Cao M, Zhang H, Park J, et al. Going the distance for protein function prediction: A new distance metric for protein interaction networks. PLoS One 2013;8(10).

Carlin DE, Fong SH, Qin Y, et al. A fast and flexible framework for network-assisted genomic association. iScience 2019;16:155–161.

Chakravarty D, Gao J, Phillips S, et al. OncoKB: A precision oncology knowledge base. JCO Precis Oncol 2017;1:1–16.

Chasman D, Siahpirani AF, Roy S. Network-based approaches for analysis of complex biological systems. Curr Opin Biotechnol 2016;39:157–166.

Chitra U, Ding K, Lee JCH, et al. Quantifying and reducing bias in maximum likelihood estimation of structured anomalies. In: Proceedings of the 38th International Conference on Machine Learning PMLR; 2021; pp. 1908–1919.

Cho D-Y, Kim Y-A, Przytycka TM. Chapter 5: Network biology approach to complex diseases. PLoS Comput Biol 2012;8(12):1–11.

Choobdar S, Ahsen ME, Crawford J, et al. Assessment of network module identification across complex diseases. Nat Methods 2019;16(9):843–852.

Chua HN, Sung W, Wong L. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. Bioinformatics 2006;22(13):1623–1630.

Cornish AJ, Markowetz F. SANTA: Quantifying the functional content of molecular networks. PLoS Comput Biol 2014;10(9):e1003808-.

Cowen L, Devkota K, Hu X, et al. Diffusion state distances: Multitemporal analysis, fast algorithms, and applications to biological networks. SIAM J Math Data Sci 2021;3(1):142–170.

Cowen L, Ideker T, Raphael BJ, et al. Network propagation: A universal amplifier of genetic associations. Nat Rev Genet 2017;18(9):551–562.

Creixell P, Reimand J, Haider S, et al. Pathway and network analysis of cancer genomes. Nat Methods 2015;12(7):615–621.

Das J, Yu H. HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Syst Biol 2012;6(1):92.

Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning. ICML '06 Association for Computing Machinery: New York, NY, USA; 2006; pp. 233–240.

de la Fuente A. From 'differential expression' to 'differential networking'—Identification of dysfunctional regulatory networks in diseases. Trends Genet 2010;26(7):326–333.

Deng M, Zhang K, Mehta S, et al. Prediction of protein function using protein–protein interaction data. J Comput Biol 2003;10(6):947–960.

Dimitrakopoulos CM, Beerenwinkel N. Computational approaches for the identification of cancer genes and pathways. WIREs Syst Biol Med 2017;9(1):e1364.

Dittrich MT, Klau G, Rosenwald A, et al. Identifying functional modules in protein–protein interaction networks: An integrated exact approach. Bioinformatics 2008;24(13):i223–i231.

Donoho D, Jin J. Higher criticism for detecting sparse heterogeneous mixtures. Ann Stat 2004;32(3):962–994.

Efron B. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. J Am Stat Assoc 2004;99(465):96–104.

Efron B. Correlation and large-scale simultaneous significance testing. J Am Stat Assoc 2007a;102(477):93–103.

Efron B. Size, power and false discovery rates. Ann Stat 2007b;35(4):1351–1377.

Efron B, Turnbull BB, Narasimhan B. Locfdr: Computes local false discovery rates. R package version 2011;1:1–7.

Fong SH, Carlin DE, Ozturk K, et al. Strategies for network GWAS evaluated using classroom crowd science. Cell Syst 2019;8(4):275–280.

Ghiassian SD, Menche J, Barabási A-L. A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. PLoS Comput Biol 2015;11(4):e1004120-.

Glaz J, Naus J, Wallenstein S. Scan Statistics. Springer-Verlag New York: New York, NY, USA; 2001.

Gligorijević V, Pržulj N. Methods for biological data integration: Perspectives and challenges. J R Soc Interface 2015;12(112):20150571.

Greene CS, Krishnan A, Wong AK, et al. Understanding multicellular function and disease with human tissue-specific networks. Nat Genet 2015;47(6):569–576.

Guo X, Lin W, Bao J, et al. A comprehensive Cis-EQTL analysis revealed target genes in breast cancer susceptibility loci identified in genome-wide association studies. Am J Human Genet 2018;102(5):890–903.
Guo Z, Li Y, Gong X, et al. Edge-based scoring and searching method for identifying condition-responsive protein–protein interaction sub-network. Bioinformatics 2007;23(16):2121–2128.
Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual. 2021.
Halldórsson BV, Sharan R. Network-based interpretation of genomic variation data. J Mol Biol 2013;425(21):3964–3969.
Hofree M, Shen JP, Carter H, et al. Network-based stratification of tumor mutations. Nat Methods 2013;10(11):1108–1115.
Hormozdiari F, Penn O, Borenstein E, et al. The discovery of integrated gene networks for autism and related disorders. Genome Res 2015;25(1):142–154.
Horn H, Lawrence MS, Chouinard CR, et al. NetSig: Network-based discovery from cancer genomes. Nat Methods 2018;15(1):61–66.
Huang JK, Carlin DE, Yu MK, et al. Systematic evaluation of molecular networks for discovery of disease genes. Cell Syst 2018;6(4):484–495.
Ideker T, Ozier O, Schwikowski B, et al. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 2002;18(Suppl. 1):S233–S240.
Ideker T, Thorsson V, Ranish JA, et al. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 2001;292(5518):929–934.
Jia P, Zhao Z. Network assisted analysis to prioritize GWAS results: Principles, methods and perspectives. Human Genet 2014;133(2):125–138.
Kloumann IM, Ugander J, Kleinberg J. Block models and personalized pagerank. Proc Natl Acad Sci 2017;114(1):33–38.
Kulldorff M. A spatial scan statistic. Commun Stat Theory Methods 1997;26(6):1481–1496.
Köhler S, Bauer S, Horn D, et al. Walking the interactome for prioritization of candidate disease genes. Am J Human Genet 2008;82(4):949–958.
Lamparter D, Marbach D, Rueedi R, et al. Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. PLoS Comput Biol 2016;12(1):e1004714.
Lawrence MS, Stojanov P, Mermel CH, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 2014;505(7484):495–501.
Lazareva O, Baumbach J, List M, et al. On the limits of active module identification. Brief Bioinform 2021;22(5).
Lee I, Blom UM, Wang PI, et al. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res 2011;21(7):1109–1121.
Leiserson MD, Eldridge JV, Ramachandran S, et al. Network analysis of GWAS data. Curr Opin Genet Dev 2013;23(6):602–610.
Leiserson MDM, Vandin F, Wu H-T, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet 2015;47(2):106–114.
Levi H, Elkon R, Shamir R. DOMINO: A network-based active module identification algorithm with reduced rate of false calls. Mol Syst Biol 2021;17(1):e9593.
Liu Y, Brossard M, Roqueiro D, et al. SigMod: An exact and efficient method to identify a strongly interconnected disease-associated module in a gene network. Bioinformatics 2017;33(10):1536–1544.
Luo Y, Zhao X, Zhou J, et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nat Commun 2017;8(1):573.
McLachlan GJ, Bean RW, Jones LB-T. A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. Bioinformatics 2006;22(13):1608–1615.
Menche J, Sharma A, Kitsak M, et al. Uncovering disease-disease relationships through the incomplete human interactome. Science 2015;347(6224):1257601–1257601.
Mitra K, Carvunis A-R, Ramesh SK, et al. Integrative approaches for finding modular structure in biological networks. Nat Rev Genet 2013;14(10):719–732.
modENCODE Consortium, Roy S, Ernst J, et al. Identification of functional elements and regulatory circuits by Drosophila ModENCODE. Science 2010;330(6012):1787–1797.
Nabieva E, Jim K, Agarwal A, et al. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics 2005;21:i302–i310.
Nakka P, Raphael BJ, Ramachandran S. Gene and network analysis of common variants reveals novel associations in multiple complex diseases. Genetics 2016;204(2):783–798.
Nibbe RK, Koyutürk M, Chance MR. An integrative-omics approach to identify functional sub-networks in human colorectal cancer. PLoS Comput Biol 2010;6(1):e1000639.

Nikolayeva I, Pla OG, Schwikowski B. Network module identification–a widespread theoretical bias and best practices. Methods 2018;132:19–25.

Page L, Brin S, Motwani R, et al. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab: Stanford, CA, USA; 1999.

Pan W, Lin J, Le CT. A mixture model approach to detecting differentially expressed genes with microarray data. Funct Integr Genom 2003;3(3):117–124.

Paull EO, Carlin DE, Niepel M, et al. Discovering causal pathways linking genomic events to transcriptional states using tied diffusion through interacting events (TieDIE). Bioinformatics 2013;29(21):2757–2764.

Picart-Armada S, Barrett SJ, Willé DR, et al. Benchmarking network propagation methods for disease gene identification. PLoS Comput Biol 2019;15(9):1–24.

Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res 2020;48(D1):D845–D855.

Pounds S, Morris SW. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of P-values. Bioinformatics 2003;19(10):1236–1242.

Radivojac P, Clark WT, Oron TR, et al. A large-scale evaluation of computational protein function prediction. Nat Methods 2013;10(3):221–227.

Reyna MA, Chitra U, Elyanow R, et al. NetMix: A network-structured mixture model for reduced-bias estimation of altered subnetworks. J Comput Biol 2021;28(5):469–484.

Reyna MA, Leiserson MD, Raphael BJ. Hierarchical HotNet: Identifying hierarchies of altered subnetworks. Bioinformatics 2018;34(17):i972–i980.

Robinson S, Nevalainen J, Pinna G, et al. Incorporating interaction networks into the determination of functionally related hit genes in genomic experiments with Markov random fields. Bioinformatics 2017;33(14):i170–i179.

Rolland T, Taşan M, Charloteaux B, et al. A proteome-scale map of the human interactome network. Cell 2014;159(5):1212–1226.

Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. Mol Syst Biol 2007;3:88–88.

Sharpnack J, Krishnamurthy A, Singh A. Near-optimal anomaly detection in graphs using Lovász extended scan statistic. In: Proceedings of the 26th International Conference on Neural Information Processing Systems, Volume 2. NIPS'13 2013a; pp. 1959–1967.

Sharpnack J, Rinaldo A, Singh A. Detecting anomalous activity on networks with the graph fourier scan statistic. IEEE Trans Signal Process 2016;64(2):364–379.

Sharpnack J, Singh A, Rinaldo A. Changepoint detection over graphs with the spectral scan statistic. In: Artificial Intelligence and Statistics 2013b; pp. 545–553.

Shrestha R, Hodzic E, Sauerwald T, et al. HIT'nDRIVE: Patient-specific multidriver gene prioritization for precision oncology. Genome Res 2017;27(9):1573–1588.

Szklarczyk D, Franceschini A, Wyder S, et al. STRING V10: Protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res 2015;43(D1):D447–D452.

Tate JG, Bamford S, Jubb HC, et al. COSMIC: The catalogue of somatic mutations in cancer. Nucleic Acids Res 2019;47(D1):D941–D947.

Uffelmann E, Huang QQ, Munung NS, et al. Genome-wide association studies. Nat Rev Methods Prim 2021;1(1):1–21.

Ulitsky I, Shamir R. Identification of functional modules using network topology and high-throughput data. BMC Syst Biol 2007;1(1):8.

Vandin F, Clay P, Upfal E, et al. Discovery of mutated subnetworks associated with clinical data in cancer. In: Pacific Symposium on Biocomputing 2012a; pp. 55–66.

Vandin F, Upfal E, Raphael BJ. Algorithms for detecting significantly mutated pathways in cancer. J Comput Biol 2011;18(3):507–522.

Vandin F, Upfal E, Raphael BJ. De novo discovery of mutated driver pathways in cancer. Genome Res 2012b;22(2):375–385.

Vanunu O, Magger O, Ruppin E, et al. Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol 2010;6(1):e1000641.

Velghe A, Van Cauwenberghe S, Polyansky A, et al. PDGFRA alterations in cancer: Characterization of a gain-of-function V536E transmembrane mutant as well as loss-of-function and passenger mutations. Oncogene 2014;33(20):2568–2576.

Vlaic S, Conrad T, Tokarski-Schnelle C, et al. ModuleDiscoverer: Identification of regulatory modules in protein-protein interaction networks. Sci Rep 2018;8(1):433.

Wang X, Terfve C, Rose JC, et al. HTSanalyzeR: An R/bioconductor package for integrated network analysis of high-throughput screens. Bioinformatics 2011;27(6):879–880.

Weston J, Elisseeff A, Zhou D, et al. Protein ranking: From local to global structure in the protein similarity network. Proc Natl Acad Sci 2004;101(17):6559–6563.

Xia J, Gill EE, Hancock REW. NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. Nat Protoc 2015;10(6):823–844.

Zhu Z, Zhang F, Hu H, et al. Integration of summary data from GWAS and EQTL studies predicts complex trait gene targets. Nat Genet 2016;48(5):481–487.

Address correspondence to:
Prof. Benjamin J. Raphael
Department of Computer Science
Princeton University
Princeton, NJ 08540
USA

E-mail: braphael@princeton.edu