JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 28, Number 5, 2021

Pp. 469–484
DOI: 10.1089/cmb.2020.0435

NetMix: A Network-Structured Mixture Model for Reduced-Bias Estimation of Altered Subnetworks

MATTHEW A. REYNA,[2,*] UTHSAV CHITRA,[1,*]
REBECCA ELYANOW,[1,3] and BENJAMIN J. RAPHAEL[1]

ABSTRACT

A classic problem in computational biology is the identification of *altered subnetworks*: subnetworks of an interaction network that contain genes/proteins that are differentially expressed, highly mutated, or otherwise aberrant compared with other genes/proteins. Numerous methods have been developed to solve this problem under various assumptions, but the statistical properties of these methods are often unknown. For example, some widely used methods are reported to output very large subnetworks that are difficult to interpret biologically. In this work, we formulate the identification of altered subnetworks as the problem of estimating the parameters of a class of probability distributions that we call the Altered Subset Distribution (ASD). We derive a connection between a popular method, jActiveModules, and the maximum likelihood estimator (MLE) of the ASD. We show that the MLE is *statistically biased*, explaining the large subnetworks output by jActiveModules. Based on these insights, we introduce NetMix, an algorithm that uses Gaussian mixture models to obtain less biased estimates of the parameters of the ASD. We demonstrate that NetMix outperforms existing methods in identifying altered subnetworks on both simulated and real data, including the identification of differentially expressed genes from both microarray and RNA-seq experiments and the identification of cancer driver genes in somatic mutation data.

Keywords: biological networks, altered subnetworks pathways, bias, differential gene expression, cancer.

1. INTRODUCTION

A STANDARD PARADIGM IN COMPUTATIONAL BIOLOGY is to use interaction networks as prior knowledge in the analysis of high-throughput omics data, with numerous applications, including: protein function prediction (Deng et al., 2003; Nabieva et al., 2005; Chua et al., 2006; Sharan et al., 2007; Radivojac et al., 2013), gene expression analysis (Ideker et al., 2002; Dittrich et al., 2008; de la Fuente, 2010; Cho et al., 2012; Xia et al., 2015), germline variants (Lee et al., 2011; Califano et al., 2012; Leiserson et al., 2013; Hormozdiari

[1]Department of Computer Science, Princeton University, Princeton, New Jersey, USA.
[2]Department of Biomedical Informatics, Emory University, Atlanta, Georgia, USA.
[3]Department of Computer Science, Brown University, Providence, Rhode Island, USA.
*These authors contributed equally.

et al., 2015; Huang et al., 2018), somatic variants in cancer (Nibbe et al., 2010; Vandin et al., 2011; Hofree et al., 2013; Creixell et al., 2015; Leiserson et al., 2015; Shrestha et al., 2017), and other analysis of other data (modENCODE Consortium *et al.*, 2010; Wang et al., 2011; Berger et al., 2013; Halldórsson and Sharan, 2013; Gligorijević and Pržulj, 2015; Chasman *et al.*, 2016; Cowen et al., 2017; Luo et al., 2017).

One classic approach is to identify *active*, or *altered*, subnetworks of an interaction network that contain unusually high or low measurements. The altered subnetwork problem takes as input: (1) an interaction network whose vertices are biological entities (e.g., genes or proteins) and whose edges represent biological interactions (e.g., physical or genetic interactions, co-expression, etc.); and (2) a measurement or score for each vertex. The goal is to find high-scoring subnetworks that correspond to groups of similarly altered vertices. This problem was introduced in Ideker *et al.* (2002) for gene expression analysis, where gene scores were derived from *p*-values of differential expression. Ideker *et al.* (2002) developed the jActiveModules algorithm to solve this problem and identify altered subnetworks of differentially expressed genes. Subsequently, Dittrich *et al.* (2008) introduced heinz as ''the first approach that really tackles and solves the original problem raised by Ideker *et al.* (2002) to optimality.'' jActiveModules and heinz have since become widely used tools with diverse applications; a few recent examples include mass-spectrometry proteomics (Kim and Hwang, 2016; Liu *et al.*, 2018), damaging *de novo* mutations in schizophrenia and other neurological disorders (Gulsuner *et al.*, 2013; Choi *et al.*, 2016), and single-cell RNA-seq (Soul *et al.*, 2015; Guo *et al.*, 2016; Klimm *et al.*, 2019).

In the past two decades, many algorithms have been developed to identify altered subnetworks in biological data (reviewed in Mitra et al., 2013; Creixell et al., 2015; Dimitrakopoulos and Beerenwinkel, 2017; Cowen et al., 2017). Each publication describing a new algorithm demonstrates the performance of their algorithm on specific biological datasets, and many of these publications also benchmark their algorithm against existing algorithms on real and/or simulated data. However, few of these publications prove theoretical guarantees for their algorithm's performance on a well-defined generative model of the data. Thus, the true performance of these algorithms is often unknown. Indeed, recent benchmarking studies (Batra et al., 2017; He et al., 2017) of several widely used network algorithms—including jActiveModules and heinz—show considerate disagreement between subnetworks identified by different methods on the same biological datasets. Moreover, these benchmarking studies (and many others) do not compare network algorithms against single-gene tests that do not use the network; thus, the tacit assumption that interaction networks necessarily improve gene prioritization is often not tested.

Separately, many publications in the statistics and machine-learning literature investigate the problem of *detecting* whether or not a network contains an anomalous subnetwork, or a *network anomaly* (Arias-Castro et al., 2006, 2008, 2018; Addario-Berry et al., 2010; Arias-Castro et al., 2011; Sharpnack and Singh, 2013; Sharpnack et al., 2013a, 2016). These papers describe specific generative models of network anomalies and use a rigorous hypothesis-testing framework to prove asymptotic results regarding the conditions under which it is possible to detect a network anomaly. Importantly, these papers also provide theoretical guarantees about conditions under which a network contributes to anomaly detection. However, the network anomaly literature does not address the specific altered subnetwork problem studied in computational biology, with three key differences. First, the *detection* problem of deciding whether or not an altered subnetwork exists is not the same as the *estimation* problem of identifying the vertices in an altered subnetwork. Second, biological networks have a finite size, and it is unclear what guarantees the asymptotic results provide for finite-size networks. Finally, the topological constraints on network anomalies are often different from those considered in computational biology.

In this article, we aim at bridging the gap between the theoretical guarantees in the network anomaly literature and the practical problem of identifying altered subnetworks in biological data. We provide a rigorous formulation of the *Altered Subnetwork Problem*, the problem that jActiveModules (Ideker et al., 2002), heinz (Dittrich et al., 2008), and other methods aim at solving. Our formulation of the Altered Subnetwork Problem is inspired by the generative model used in the network anomaly literature, but it requires that the altered subnetwork is a connected subnetwork, a constraint motivated by the topology of signaling pathways (Bhalla and Iyengar, 1999); Kelley et al., 2004) and by the seminal works of Ideker et al. (2002) and Dittrich et al. (2008).

We show that the Altered Subnetwork Problem is equivalent to estimating the parameters of a distribution, which we define as the *Altered Subset Distribution* (ASD). We prove that the jActiveModules problem (Ideker et al., 2002) is equivalent to finding a maximum likelihood estimator (MLE) of the parameters of the ASD for connected subgraphs. At the same time, we demonstrate that if (1) the size of the

altered subnetwork is moderately small and (2) the scores of vertices inside and outside of the altered subnetwork are not well separated, then the MLE is a *biased* estimator of the size of the altered subnetwork. This statistical bias provides a rigorous explanation for the large subnetworks produced by jActiveModules (Nikolayeva et al., 2018). We also show that the size of the altered subnetworks identified by heinz (Dittrich et al., 2008) is biased for most choices of its user-defined parameters.

We introduce a new algorithm, NetMix, that combines a Gaussian mixture model (GMM) and a combinatorial optimization algorithm to identify altered subnetworks. We show that NetMix is a reduced-bias estimator of the size of the altered subnetwork. We demonstrate that NetMix outperforms other methods for identifying altered subnetworks on simulated data, gene expression data, and somatic mutation data.

2. ALTERED SUBNETWORKS, ALTERED SUBSETS, AND MAXIMUM LIKELIHOOD ESTIMATION

2.1. Altered subnetwork problem

Let $G = (V, E)$ be a biological interaction network with a measurement, or score, X_v for each vertex $v \in V$. We assume that there is a connected subnetwork A in G, the *altered subnetwork*, whose scores are derived from a different distribution than the scores of the vertices not in A (Fig. 1). The goal of the Altered Subnetwork Problem is to find A. The problem is defined formally as follows.

Altered Subnetwork Problem (ASP). Let $G = (V, E)$ be a graph with vertex scores $\mathbf{X} = (X_v)_{v \in V}$, and let $A \subseteq V$ be a connected subgraph of G. Suppose that

$$X_v \overset{i.i.d.}{\sim} \begin{cases} D_A, & \text{if } v \in A, \\ D_B, & \text{if } v \in V \backslash A, \end{cases} \tag{1}$$

where D_A is the altered *distribution and* D_B is the background *distribution.* Given G and $\mathbf{X}$, find A.

Note that the Altered Subnetwork Problem (ASP) assumes that the interaction network G has a *single* altered subnetwork A. This is a reasonable assumption in some cases, for example, when the altered subnetwork A is large or the interaction network G has small diameter or high connectivity. More generally, when the network contains multiple altered subnetworks, one can recursively solve the ASP to identify more than one altered subnetwork.

The seminal algorithm for solving the ASP is jActiveModules (Ideker et al., 2002). jActiveModules takes as input a p-value p_v for each vertex v; for example, a p-value of differential gene expression. jActiveModules transforms the p-values into scores $X_v = \Phi^{-1}(1 - p_v)$, where Φ is the cumulative distribution function (CDF) of a standard normal distribution. Under the null hypothesis, the p-values p_v across genes are distributed according to the uniform distribution $U(0, 1)$, and thus by transforming p-values,

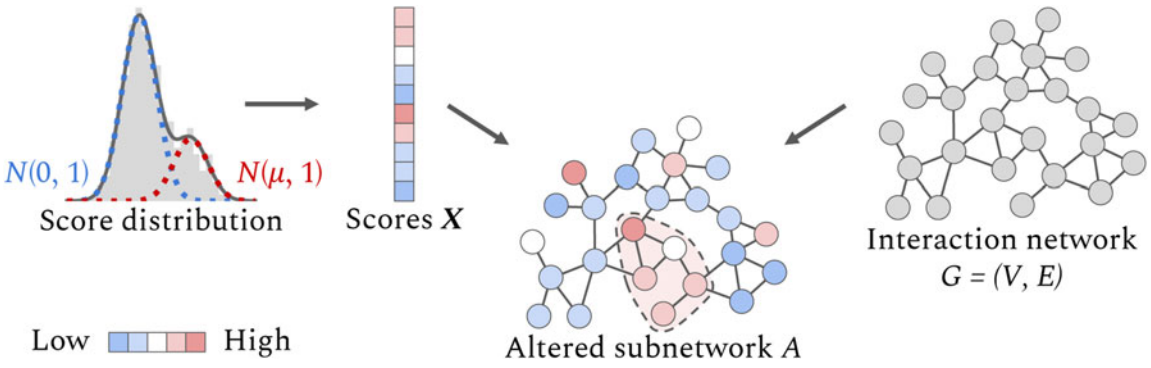

FIG. 1. Altered Subnetwork Problem. Measurements, or scores, $\mathbf{X}$ from a high-throughput experiment are drawn from one of two distributions: genes/proteins in an altered subnetwork A of an interaction network $G = (V, E)$ have scores drawn from an altered distribution $N(\mu, 1)$ with $\mu > 0$, whereas genes/proteins not in A have scores drawn from a background distribution $N(0, 1)$. The goal is to identify A from $\mathbf{X}$ and G; the difficulty of this problem depends on the separation μ between the distributions and the size $|A|$ of the altered subnetwork. ASD, Altered Subset Distribution.

jActiveModules solves the ASP with background distribution $D_B = N(0, 1)$. jActiveModules aims to find a connected subgraph $\hat{A}$ that maximizes* $\Gamma(S) = \frac{1}{\sqrt{|S|}} \sum_{v \in S} X_v$, that is,

$$\hat{A} = \underset{\text{connected } S \subseteq V}{\operatorname{argmax}} \Gamma(S) = \underset{\text{connected } S \subseteq V}{\operatorname{argmax}} \frac{1}{\sqrt{|S|}} \sum_{v \in S} X_v. \tag{2}$$

The presentation of jActiveModules in Ideker et al. (2002) does not specify the altered distribution D_A. However, in Section 2.2, we argue that the choice of the objective function in Equation (2) implicitly assumes that $D_A = N(\mu, 1)$ for some parameter $\mu > 0$. Thus, we define the normally distributed ASP as follows.

Normally Distributed Altered Subnetwork Problem. *Let $G = (V, E)$ be a graph with vertex scores $\mathbf{X} = (X_v)_{v \in V}$, and let $A \subseteq V$ be a connected subgraph of G. Suppose that for some $\mu > 0$,*

$$X_v \overset{i.i.d.}{\sim} \begin{cases} N(\mu, 1), & \text{if } v \in A, \\ N(0, 1), & \text{if } v \in V \backslash A. \end{cases} \tag{3}$$

Given G and $\mathbf{X}$, find A.

The Normally Distributed ASP has a sound statistical interpretation: if the p-values p_v of the genes are derived from an asymptotically normal test statistic, as is often the case, then the transformed p-values $X_v = \Phi^{-1}(1 - p_v)$ are distributed as $N(0, 1)$ for genes satisfying the null hypothesis and $N(\mu, 1)$ for genes satisfying the alternative hypothesis (Hung et al., 1997). Normal distributions have previously been used to model transformed p-values from differential gene expression experiments (Pan et al., 2003; McLachlan et al., 2006; Wang et al., 2009).

More generally, the Normally Distributed Altered Subnetwork Problem is related to a larger class of *network anomaly* problems, which have been studied extensively in the machine-learning and statistics literature (Arias-Castro et al., 2006, 2008, 2011, 2018; Addario-Berry et al., 2010; Sharpnack and Singh, 2013; Sharpnack et al., 2013a, b; Sharpnack et al., 2016). To better understand the relationships between these problems and the algorithms developed to solve them, we will describe a generalization of the Altered Subnetwork Problem. We start by defining the following distribution, which generalizes the connected subnetworks in the Normally Distributed Altered Subnetwork Problem to any family of altered subsets.

Normally Distributed ASD. *Let $n > 0$ be a positive integer, let $\mathcal{S}$ be a family of subsets of $\{1, \ldots, n\}$, and let $A \in \mathcal{S}$. $\mathbf{X} = (X_1, \ldots, X_n)$ be distributed according to the* Normally Distributed *$ASD_{\mathcal{S}}(A, \mu)$ provided*

$$X_i \overset{i.i.d.}{\sim} \begin{cases} N(\mu, 1), & \text{if } i \in A, \\ N(0, 1), & \text{if } i \notin A. \end{cases} \tag{4}$$

Here, $\mu > 0$ is the mean *of the ASD and A is the* altered *subset of the ASD.*

More generally, the ASD can be defined for any background distribution D_B and altered distribution D_A. We will restrict ourselves to normal distributions in accordance with the Normally Distributed Altered Subnetwork Problem, and we will subsequently assume normal distributions in both the Altered Subset Distribution (ASD) and the Altered Subnetwork Problem.

The distribution in the Altered Subnetwork Problem is the $\text{ASD}_{\mathcal{S}}(A, \mu)$, where the family $\mathcal{S}$ of subsets are connected subgraphs of the network G. In this terminology, the Altered Subnetwork Problem is the problem of estimating the parameters A and μ of the ASD given data $\mathbf{X} \sim \text{ASD}_{\mathcal{S}}(A, \mu)$ and knowledge of the parameter space $\mathcal{S}$ of altered subnetworks A. Thus, we generalize the Altered Subnetwork Problem to the Altered Subset Distribution Estimation Problem, defined as follows.

ASD Estimation Problem. *Let $\mathbf{X} = (X_1, \ldots, X_n) \sim \text{ASD}_{\mathcal{S}}(A, \mu)$. Given $\mathbf{X}$ and $\mathcal{S}$, find A and μ.*

*jActiveModules actually maximizes $\Gamma_{\text{norm}}(S) = (\Gamma(S) - \mu_{|S|})/\sigma_{|S|}$, a z-score normalized version of $\Gamma(S)$, where $\mu_{|S|}$ and $\sigma_{|S|}$ are the mean and standard deviation, respectively, of $\Gamma(T)$ over all subsets $T \subseteq V$ of size $|S|$. We prove in the supplement that maximizing $\Gamma_{\text{norm}}(S)$ is equivalent to maximizing the unnormalized $\Gamma(S)$ when the data are generated from normal distributions.

The ASD Estimation Problem is a general problem of estimating the parameters of a *structured* alternative distribution, sometimes known as a ''structured normal means'' problem in statistics (Sharpnack et al., 2013a). Different choices of $\mathcal{S}$ for the ASD Estimation Problem yield a number of interesting problems, some of which have been previously studied in the altered subnetwork literature.

- $\mathcal{S} = \mathcal{P}_n$, the power set of all subsets of $\{1, \ldots, n\}$. We call the distribution $\text{ASD}_{\mathcal{P}_n}(A, \mu)$ the *unstructured* ASD.
- $\mathcal{S} = \mathcal{C}_G$, the set of all connected subgraphs of a graph $G = (V, E)$. We call $ASD_{\mathcal{C}_G}(A, \mu)$ the *connected* ASD. The connected ASD Estimation Problem is equivalent to the Altered Subnetwork Problem described earlier.
- $\mathcal{S} = \mathcal{D}_G(\rho)$, the set of all subgraphs of a graph $G = (V, E)$ with edge density $\geq \rho$. Guo et al. (2007), Vanunu et al. (2010), and Ayati et al. (2015) identify altered subnetworks with high edge density, and Amgalan and Lee (2014) identifies altered subnetworks with edge density $\rho = 1$, that is, cliques.
- $\mathcal{S} = \mathcal{N}_G = \{\mathcal{N}(v) : v \in V\}$, the set of all first-order network neighborhoods of a graph $G = (V, E)$. Cho et al. (2016) and Horn et al. (2018) use first-order network neighborhoods to prioritize cancer genes.
- $\mathcal{S} \subseteq \mathcal{P}_n$, a family of subsets. Typically, $|\mathcal{S}| \ll |\mathcal{P}_n|$ and $\mathcal{S}$ is not defined in terms of a graph. A classic example is gene set analysis [see Hung et al. (2011) for a review].

2.2. Bias in maximum likelihood estimation of the ASD

One reasonable approach for solving the ASD Estimation Problem is to compute an MLE for the parameters of the ASD. We derive the MLE next and show that it has undesirable statistical properties. All proofs are in the Supplementary Data.

Theorem 1. *Let* $\mathbf{X} \sim \text{ASD}_{\mathcal{S}}(A, \mu)$. *The MLEs* $\hat{A}_{ASD}$ *and* $\hat{\mu}_{ASD}$ *of* A *and* μ, *respectively, are*

$$\hat{A}_{ASD} = \underset{S \in \mathcal{S}}{\text{argmax}}\, \Gamma(S) = \underset{S \in \mathcal{S}}{\text{argmax}} \frac{1}{\sqrt{|S|}} \sum_{v \in S} X_v \quad \text{and} \quad \hat{\mu}_{ASD} = \frac{1}{|\hat{A}_{ASD}|} \sum_{v \in \hat{A}_{ASD}} X_v. \tag{5}$$

The maximization of Γ over $\mathcal{S}$ in Equation (5) is a version of the *scan statistic*, a commonly used statistic to study point processes on lines and rectangles under various distributions (Kulldorff, 1997; Glaz et al., 2001). Comparing Equations (5) and (2), we see that jActiveModules (Ideker et al., 2002) computes the scan statistic over the family $\mathcal{S} = \mathcal{C}_G$ of connected subgraphs of the graph G. Thus, although jActiveModules (Ideker et al., 2002) neither specifies the anomalous distribution D_A nor provides a statistical justification for their subnetwork scoring function, Theorem 1 above shows that jActiveModules implicitly assumes that D_A is a normal distribution, and that jActiveModules aims to solve the Altered Subnetwork Problem by finding the MLE $\hat{A}_{\text{ASD}}$.

Despite this insight that jActiveModules computes the MLE, it has been observed that jActiveModules often identifies large subnetworks. Nikolayeva et al. (2018) note that the subnetworks identified by jActiveModules are large and ''hard to interpret biologically.'' They attribute the tendency of jActiveModules to identify large subnetworks to the fact that a graph typically has more large subnetworks than small ones. Although this observation about the relative numbers of subnetworks of different sizes is correct, we argue that this tendency of jActiveModules to identify large subnetworks is due to a more fundamental reason: The MLE $\hat{A}_{\text{ASD}}$ is a *biased* estimator of A.

First, we recall the definitions of bias and consistency for an estimator $\hat{\theta}$ of a parameter θ.

Definition 1. *Let* $\hat{\theta} = \hat{\theta}(\mathbf{X})$ *be an estimator of a parameter* θ *given observed data* $\mathbf{X} = (X_1, \ldots, X_n)$. *(a) The* bias *in the estimator* $\hat{\theta}$ *of* θ *is* $Bias_\theta(\hat{\theta}) = E[\hat{\theta}] - \theta$. *We say that* $\hat{\theta}$ *is a* biased *estimator of* θ *if* $Bias_\theta(\hat{\theta}) \neq 0$, *and it is an* unbiased *estimator of* θ *otherwise. (b) We say that* $\hat{\theta}$ *is a* consistent *estimator of* θ *if* $\hat{\theta} \xrightarrow{p} \theta$, *where* $\xrightarrow{p}$ *denotes convergence in probability as* $n \to \infty$, *and it is an* inconsistent *estimator of* θ *otherwise.*

When it is clear from context, we omit the subscript θ and write $\text{Bias}(\hat{\theta})$ for the bias of estimator $\hat{\theta}$.

Let $\mathbf{X} \sim \text{ASD}_{\mathcal{P}_n}(A, \mu)$ be distributed according to the unstructured ASD. We observe that the estimators $|\hat{A}_{\text{ASD}}|/n$ and $\hat{\mu}_{\text{ASD}}$ are both biased and inconsistent when both $|A|/n$ and μ are moderately small (Fig. 2). We summarize these observations in the following conjecture:

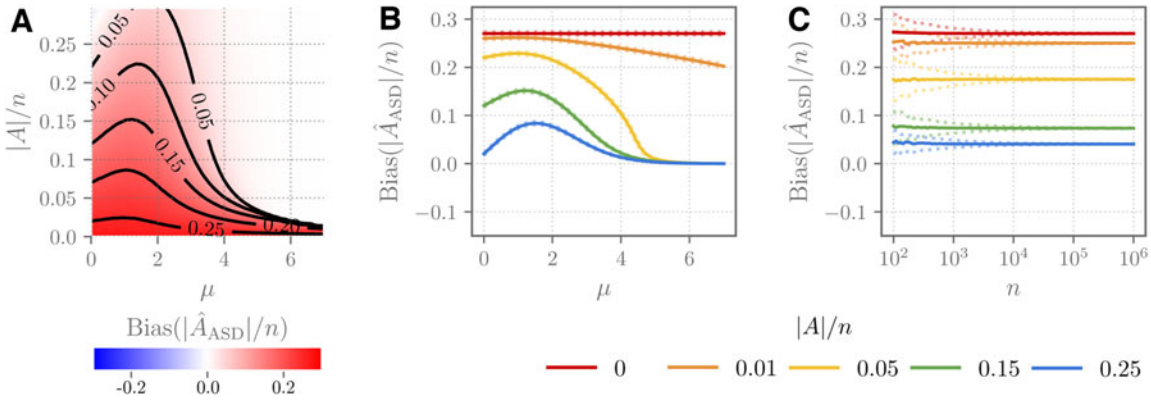

FIG. 2. Scores $\mathbf{X} \sim \mathrm{ASD}_{\mathcal{P}_n}(A, \mu)$ are distributed according to the unstructured ASD. **(A)** $\mathrm{Bias}(|\hat{A}_{\mathrm{ASD}}|/n)$ in the maximum likelihood estimate of $|A|/n$ as a function of the mean μ and altered subset size $|A|/n$ for $n = 10^4$. **(B)** $\mathrm{Bias}(|\hat{A}_{\mathrm{ASD}}|/n)$ for $n = 10^4$ and several values of $|A|/n$. Dotted lines indicate first and third quartiles in the estimate of the bias. **(C)** $\mathrm{Bias}(|\hat{A}_{\mathrm{ASD}}|/n)$ as a function of n for $\mu = 3$ and for several values of $|A|/n$.

Conjecture. *Let* $\mathbf{X} = (X_1, \ldots, X_n) \sim \mathrm{ASD}_{\mathcal{P}_n}(A, \mu)$. *Then, there exist* $\mu_0 > 0$ *and* $\beta > 0$ *such that, if* $\mu < \mu_0$ *and* $|A|/n < \beta$, *then* $|\hat{A}_{ASD}|/n$ *and* $\hat{\mu}_{ASD}$ *are biased and inconsistent estimators of* $|A|/n$ *and* μ, *respectively.*

Note that there are many examples in the literature of biased MLEs; for example, the MLE for the variance of a (univariate) normal distribution or the MLE for the inverse of the mean of a Poisson distribution (Firth, 1993). However, examples of inconsistent MLEs are somewhat rare (Ferguson, 1982).

Although we do not have a proof of the earlier conjecture, we prove the following results that partially explain the bias and inconsistency of the estimators $|A_{\mathrm{ASD}}|$ and μ_{ASD}. For the bias, we prove the following.

Theorem 2. *Let* $\mathbf{X} = (X_1, \ldots, X_n) \sim \mathrm{ASD}_{\mathcal{P}_n}(A, \mu)$ *with* $A = \emptyset$. *Then* $|\hat{A}_{ASD}| = cn$ *for sufficiently large* n *and with high probability, where* $0 < c < 0.35$ *is independent of* n.

Empirically, we observe $c \approx 0.27$, that is, $\hat{A}_{\mathrm{ASD}}$ contains more than a quarter of the scores (Fig. 2). This closely aligns with the observation in Nikolayeva et al. (2018) that jActiveModules reports subnetworks that contain $\sim 29\%$ of all vertices in the graph. Based on Theorem 2, one may suspect that $|\hat{A}_{\mathrm{ASD}}| \approx cn$ when μ or $|A|/n$ is sufficiently small, providing some intuition for why $|\hat{A}_{\mathrm{ASD}}|/n$ is biased. For inconsistency, we prove that the bias is independent of n.

Theorem 3. *Let* $\mathbf{X} = (X_1, \ldots, X_n) \sim \mathrm{ASD}_{\mathcal{P}_n}(A, \mu)$, *where* $|A| = \theta(n)$. *For sufficiently large* n, $Bias(|\hat{A}_{ASD}|/n)$ *and* $Bias(\hat{\mu}_{ASD})$ *are independent of* n.

3. THE NETMIX ALGORITHM

Following the observation that the MLEs of the distribution $\mathrm{ASD}_{\mathcal{P}_n}(A, \mu)$ are biased, we aim at finding a less biased estimator by explicitly modeling the distribution of the scores $\mathbf{X}$. In this section, we derive a new algorithm, NetMix, that solves the Altered Subnetwork Problem by fitting a GMM to $\mathbf{X}$.

3.1. Gaussian mixture model

We start by recalling the definition of a GMM.

Gaussian Mixture Model. *Let* $\mu > 0$ *and* $\alpha \in (0, 1)$. X *is distributed according to the* GMM (α, μ) *with parameters* α *and* μ *provided*

$$X \sim \alpha N(\mu, 1) + (1 - \alpha) N(0, 1). \tag{6}$$

An alternate interpretation of the GMM is to draw a latent variable $Z \sim \mathrm{Bernoulli}(\alpha)$ and select $X \sim N(\mu, 1)$ if $Z = 1$, and $X \sim N(0, 1)$ if $Z = 0$.

Given data $\mathbf{X} = (X_1, \ldots, X_n)$, we define $\hat{\mu}_{\mathrm{GMM}}$ and $\hat{\alpha}_{\mathrm{GMM}}$ to be the MLEs for μ and α, respectively, obtained by fitting a GMM to $\mathbf{X}$. In practice, $\hat{\mu}_{\mathrm{GMM}}$ and $\hat{\alpha}_{\mathrm{GMM}}$ are obtained by the expectation-maximization

(EM) algorithm (Dempster et al., 1977), which is known to converge to the MLEs as the number of samples goes to infinity (Xu et al., 2016; Daskalakis et al., 2017). Further, if $X_i \overset{\text{i.i.d.}}{\sim} \text{GMM}(\mu, \alpha)$ are distributed according to the GMM with $\alpha \neq 0$, then $\hat{\mu}_{\text{GMM}}$ and $\hat{\alpha}_{\text{GMM}}$ are consistent (and therefore asymptotically unbiased) estimators of μ and α, respectively (Chen, 2017).

Analogously, by fitting a GMM to data $\mathbf{X} \sim \text{ASD}_{\mathcal{P}_n}(A, \mu)$ from the unstructured ASD, we observe that $\hat{\alpha}_{\text{GMM}}$ is a less biased estimator of $|A|/n$ than $|\hat{A}_{\text{ASD}}|/n$ (Fig. 3A, B). We also observe that $\hat{\alpha}_{\text{GMM}}$ is a consistent estimator of $|A|/n$ (Fig. 3C). We summarize our findings in the following conjecture:

3.1.2. Conjecture. *Let* $\mathbf{X} \sim \text{ASD}_{\mathcal{P}_n}(A, \mu)$ *with* $|A| > 0$, *and let* $\hat{A}_{\text{ASD}}$ *be the MLE of* A *as defined in Equation* (5). *Let* $\hat{\alpha}_{\text{GMM}}$ *and* $\hat{\mu}_{\text{GMM}}$ *be the MLEs of* α *and* μ *obtained by fitting a GMM to* $\mathbf{X}$. *Then,* $\text{Bias}_{|A|/n}(\hat{\alpha}_{\text{GMM}}) < \text{Bias}_{|A|/n}(|\hat{A}_{\text{ASD}}|/n)$. *Moreover,* $\hat{\alpha}_{\text{GMM}}$ *and* $\hat{\mu}_{\text{GMM}}$ *are consistent estimators of* $|A|/n$ *and* μ, *respectively.*

Although we do not have a proof of the earlier conjecture, a partial justification is the following relationship between the unstructured ASD and the GMM distribution. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be drawn from a mixture of unstructured ASDs over all possible anomalous sets A of size k; in other words, $\mathbf{X} \sim B \cdot \sum_{|A|=k} \text{ASD}_{\mathcal{P}_n}(A, \mu)$, where $B = 1 \Big/ \binom{n}{k}$ is a normalizing constant. Let $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{GMM}(\alpha, \mu)$ be independent samples from the GMM for $\mu > 0$ and $\alpha = \frac{k}{n}$ with corresponding latent variables $Z_1, \ldots, Z_n$. Then, the joint distribution of the GMM samples $\mathbf{Y} = (Y_1, \ldots, Y_n)$ conditioned on $\sum_{i=1}^{n} Z_i = k$ is equal to the distribution of $\mathbf{X}$:

$$\mathbf{X} \overset{d}{=} \left(\mathbf{Y} \,\middle|\, \sum_{i=1}^{n} Z_i = k \right). \tag{7}$$

3.2. NetMix algorithm

We derive an algorithm, NetMix, that uses the MLEs $\hat{\mu}_{\text{GMM}}$ and $\hat{\alpha}_{\text{GMM}}$ from the GMM to solve the Altered Subnetwork Problem. Note that the GMM is not identical to the ASD, the distribution that generated the data. Despite this difference in distributions, the earlier conjecture provides justification that the GMM yields less biased estimators of A and μ than the MLEs of the ASD distribution.

Given a graph $G = (V, E)$ and scores $\mathbf{X} = (X_v)_{v \in V}$, NetMix first computes the *responsibility* $r_v = \Pr(v \in A | X_v)$, or the probability that $v \in A$, for each vertex $v \in V$. The responsibilities r_v are computed from the GMM MLEs $\hat{\mu}_{\text{GMM}}$ and $\hat{\alpha}_{\text{GMM}}$ (which are estimated by the EM algorithm) according to the formula

$$\hat{r}_v = \frac{\hat{\alpha}_{\text{GMM}} \cdot P(N(\hat{\mu}_{\text{GMM}}, 1) = X_v)}{\hat{\alpha}_{\text{GMM}} \cdot P(N(\hat{\mu}_{\text{GMM}}, 1) = X_v) + (1 - \hat{\alpha}_{\text{GMM}}) \cdot P(N(0, 1) = X_v)} \tag{8}$$

where $P(N(\mu, 1) = X_v)$ is the probability density function (PDF) of the normal distribution $N(\mu, 1)$ evaluated at X_v.

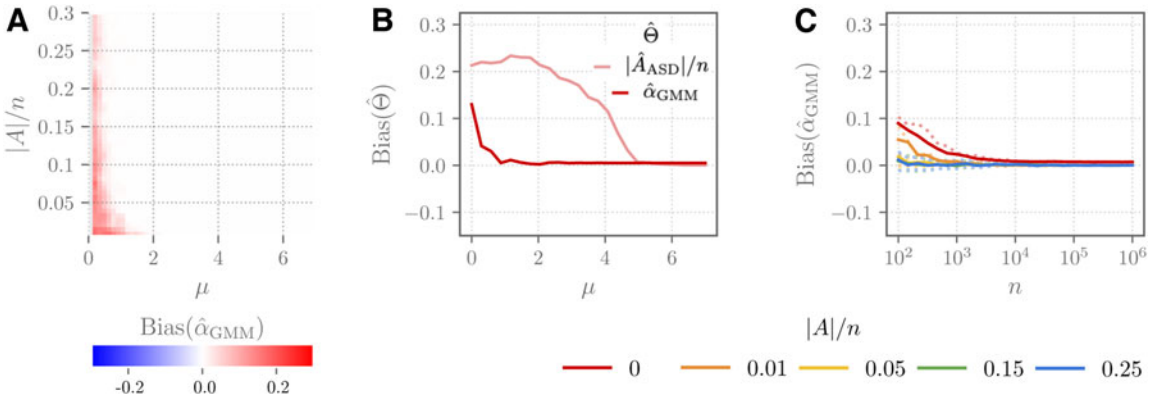

FIG. 3. Scores $\mathbf{X} \sim \text{ASD}_{\mathcal{P}_n}(A, \mu)$ are distributed according to the unstructured ASD, and parameters $\hat{\alpha}_{\text{GMM}}$ and $\hat{\mu}_{\text{GMM}}$ are obtained by the EM algorithm. **(A)** $\text{Bias}(\hat{\alpha}_{\text{GMM}})$ as a function of the mean μ and altered subnetwork size $|A|/n$ for $n = 10^4$. Compare with Figure 2A. **(B)** $\text{Bias}(\hat{\alpha}_{\text{GMM}})$ and $\text{Bias}(|\hat{A}_{\text{ASD}}|/n)$ as functions of the mean μ for $|A|/n = 0.05$ and $n = 10^4$. **(C)** $\text{Bias}(\hat{\alpha}_{\text{GMM}})$ as a function of n for mean $\mu = 3$ and several values of $|A|/n$. Compare with Figure 2C. EM, expectation maximization.

Next, NetMix aims at finding a connected subgraph C of size $|C| \approx n\alpha$ that maximizes $\sum_{v \in C} r_v$. To find such a subgraph, NetMix assigns a weight $w(v) = \hat{r}_v - \tau$ to each vertex v, where τ is chosen so that approximately $n\hat{\alpha}_{\text{GMM}}$ vertices have non-negative weights. NetMix then computes the maximum weight connected subgraph (MWCS) $\hat{A}_{\text{NetMix}}$ in G by adapting the integer linear program in Dittrich et al. (2008). The use of τ is motivated by the observation that, if $\hat{\alpha}_{\text{GMM}} \approx \alpha$, then we expect $|\hat{A}_{\text{NetMix}}| \approx n\hat{\alpha}_{\text{GMM}} \approx n\alpha = |A|$.

We formally describe the NetMix algorithm for solving the Altered Subnetwork Problem later.

NetMix algorithm. Given a network $G = (V, E)$ and vertex scores $\mathbf{X} = (X_v)_{v \in V}$,

1. Compute $\hat{\alpha}_{\text{GMM}}$ and $\hat{\mu}_{\text{GMM}}$, the MLEs of α and μ, by fitting a GMM to $\mathbf{X}$ using EM.
2. Compute the estimated responsibilities $\hat{r}_v$ for each vertex v by using Equation (8).
3. Compute τ such that $|\{v \in V : \hat{r}_v > \tau\}| = \lceil n\hat{\alpha}_{\text{GMM}} \rceil$, where $\lceil \cdot \rceil$ is the ceiling function.
4. Find the connected subgraph $\hat{A}_{\text{NetMix}}$ defined by

$$\hat{A}_{\text{NetMix}} = \underset{\text{connected} C \subseteq V}{\text{argmax}} \sum_{v \in C} (\hat{r}_v - \tau) \tag{9}$$

using integer linear programming.

NetMix bears some similarities to heinz (Dittrich et al., 2008), another algorithm to identify altered subnetworks. However, there are two important differences. First, heinz does not solve the Altered Subnetwork Problem defined in the previous section. Instead, heinz models the vertex scores (assumed to be p-values) with a beta-uniform mixture (BUM) distribution. The motivation for the BUM is based on an empirical goodness-of-fit in Pounds and Morris (2003); however, later work by the same author (Pounds and Cheng, 2004) observes that the BUM tends to underestimate the number of p-values drawn from the altered distribution.

Second, heinz requires that the user specify a false discovery rate (FDR) and shifts the p-values according to this FDR. We show later that nearly all choices of the FDR lead to a biased estimate of $|A|$. Moreover, the manually selected FDR allows users to selectively tune the value of this parameter to influence which genes are in the inferred altered subnetwork, analogous to ''p-hacking'' (Ioannidis, 2005; Head et al., 2015; Nuzzo, 2015). Indeed, recently published analyses using heinz use a wide range of FDR values, ranging anywhere from 10^{-25} to 0.05 (Liang et al., 2012; Choi et al., 2016; He et al., 2017; Klimm et al., 2019). See the Supplementary Data for more details on the differences between heinz and NetMix.

Despite these limitations, the ILP given in heinz to solve the MWCS problem is very useful for implementing NetMix and for computing the scan statistic [Eq. (2)] used in jActiveModules (Section 4).

4. RESULTS

We compared NetMix with jActiveModules (Ideker et al., 2002) and heinz (Dittrich et al., 2008) on simulated instances of the Altered Subnetwork Problem and on real datasets, including differential gene expression experiments from the Expression Atlas (Petryszak et al., 2015) and somatic mutations in cancer. jActiveModules is accessible only through Cytoscape (Shannon et al., 2003; Cline et al., 2007) and not a command-line interface, making it difficult to run on a large number of datasets. Thus, we implemented jActiveModules*, which computes the scan statistic [Eq. (5)] by adapting the integer linear program in heinz.[†] jActiveModules* output relies on the global optimum of the scan statistic, whereas jActiveModules relies on heuristics (simulated annealing and greedy search) to find a local optimum.

4.1. Simulated data

We compared NetMix, jActiveModules*, and heinz on simulated instances of the Altered Subnetwork Problem by using the HINT+HI interaction network (Leiserson et al., 2015), a combination of binary and

---

*The scan statistic [Eq. (2)] is the maximization of a non-linear objective function, but for a fixed subnetwork size $|S|$ the objective function is linear. We computed the scan statistic by modifying the ILP in heinz (Dittrich *et al.,* 2008) and running this ILP over all possible subnetwork sizes.

[†]Formally, μ is the smallest mean such that the hypotheses $H_0 : X \sim ASD_{\mathcal{C}_G}(\emptyset, 0)$ and $H_1 : X \sim ASD_{\mathcal{C}_G}(A, \mu)$ are asymptotically distinguishable. See Sharpnack *et al.* (2013a) for details.

co-complex interactions in High-quality INTeractones (HINT) (Das and Yu, 2012) with high-throughput derived interactions from the Human Interactone (HI) network (Rolland et al., 2014) as the graph G. For each instance, we randomly selected a connected subgraph $A \subseteq V$ with size $|A| = 0.05n$ by using the random walk method of Lu and Bressan (2012), and we drew a sample $\mathbf{X} \sim \mathrm{ASD}_{\mathcal{C}_G}(A, \mu)$. We ran each method on $\mathbf{X}$ and G to obtain an estimate $\hat{A}$ of the altered subnetwork A. We ran heinz with three different choices of the FDR parameter (FDR $= 0.001$, FDR $= 0.1$, and FDR $= 0.5$) to reflect the variety of FDRs used in practice.

We found that NetMix output subnetworks whose size $|\hat{A}_{\mathrm{NetMix}}|$ was very close to the true size of the altered subnetwork across all values of μ in the simulations (Fig. 4A). In contrast, jActiveModules* output subnetworks that were much larger than the altered subnetwork for $\mu < 5$. This behavior is consistent with our earlier conjectures about the large bias in the MLE $\hat{A}_{\mathrm{ASD}}$ for the unstructured ASD. Note that $\mu > 5$ corresponds to a large separation between the background and alternative distributions, and the network is not needed to separate these two distributions.

We also quantified the overlap between the true altered subnetwork A and the subnetwork $\hat{A}$ output by each method using the F-measure, finding that NetMix outperforms other methods across the full range of μ (Fig. 4B). heinz requires the user to select an FDR value, and we find that the size of the output subnetwork and the F-measure varies considerably for different FDR (Fig. 4A, B). When μ was small, a high FDR value (FDR $= 0.5$) yielded the best performance in terms of F-measure. However, when μ was large, a low FDR value (FDR $= 0.001$) gave better performance. Although there are FDR values where the performance of heinz is similar to NetMix, the user *does not know what FDR value to select* for any given input, as the values of μ and the size $|A|$ of the altered subnetwork are unknown.

The bias in $|\hat{A}|/n$ observed using jActiveModules* with the interaction network (Fig. 4A) is similar to the bias for the unstructured ASD (Fig. 2A). Thus, we also evaluated how much benefit the network provided for each method. For small μ, we found that NetMix had a small but noticeable gain in performance when using the network; in contrast, other methods had nearly the same performance with or without the network (Fig. 4C with further details in the Supplementary Data). These results emphasize the importance of evaluating network methods on simulated data *and* demonstrating that a network method outperforms a single-gene test; neither of these was done in the jActiveModules (Ideker et al., 2002) and heinz (Dittrich et al., 2008) papers, nor are they common in many other papers on biological network analysis.

4.2. *Differential gene expression subnetworks*

We compared NetMix, jActiveModules*, and heinz on gene expression data from the Expression Atlas (Petryszak et al., 2015). We analyzed 945 differential expression experiments, including 292 RNA-seq experiments and 653 microarray experiments. For 84% of these experiments, the GMM used by NetMix provided a better fit to the p-value distributions than the BUM (Pounds and Morris, 2003) used by heinz (see the Supplementary Data for more details). In addition, the GMM provided a better fit in 83/85

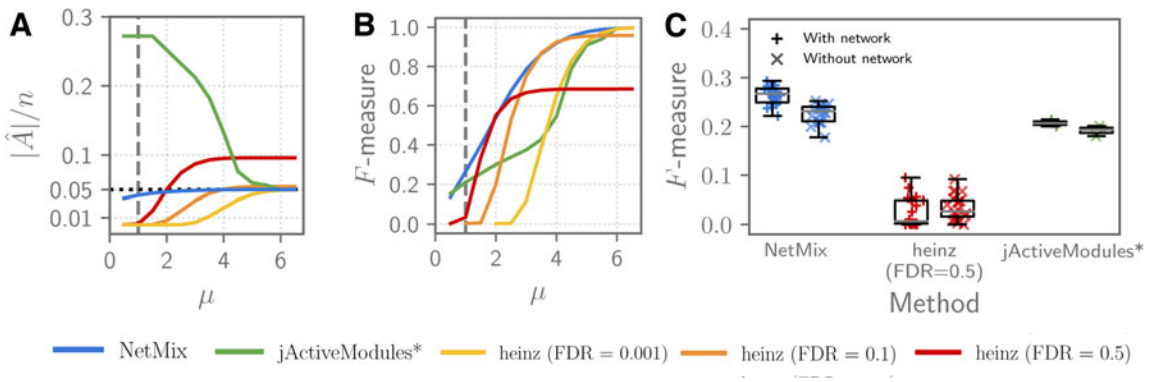

FIG. 4. Comparison of altered subnetwork identification methods on simulated instances of the Altered Subnetwork Problem using the HINT+HI interaction network with $n = 15,074$ vertices, and where the altered subnetwork A has size $|A| = 0.05n$. [Dashed vertical line ($\mu = 1$) represents the smallest μ such that one can detect whether G contains an altered subnetwork]. **(A)** Size $|\hat{A}|/n$ of identified altered subnetwork $\hat{A}$ as a function of mean μ. **(B)** F-measure for $\hat{A}$ as a function of μ. **(C)** F-measure for $\hat{A}$ at $\mu = 1$, comparing performance with the network (left series for each method) and without the network (right series for each method). HI, Human Interactone; HINT, High-quality INTeractone.

experiments, where the null proportion (fraction of genes not differentially expressed) estimated by the GMM and BUM differed by ≥ 0.25. In all 85 of these experiments, the BUM estimated a higher null proportion, consistent with the report in Pounds and Cheng (2004) that the BUM tends to overestimate the null proportion.

As many experiments had a small null proportion (i.e., most genes in the experiment were differentially expressed), we restricted our analysis to the 157 experiments from the Expression Atlas with a null proportion ≥ 0.8 as estimated by the GMM. We ran NetMix, jActiveModules*, and heinz on these 157 experiments with the HINT+HI network. For heinz, we used three FDR values: FDR = 0.1, FDR = 0.001, and the FDR value such that $|\hat{A}_{\text{NetMix}}|$ genes have a positive weight in the heinz scoring. These choices demonstrate how users might ''*p*-hack'' the FDR value to achieve desirable results. We also compared these with a method that ignores network topology, selecting the $|\hat{A}_{\text{NetMix}}|$ genes with the lowest *p*-values; we call this method ''top *p*-values.'' See the Supplementary Data for specific details on these methods.

Both NetMix and heinz identified subnetworks that were significantly smaller than jActiveModules* (Fig. 5A), which is consistent with previous observations (Nikolayeva et al., 2018) that jActiveModules estimates overly large subnetworks. At the same time, NetMix identified subnetworks with significant overlap (FDR ≤ 0.01) with more biological process gene ontology (GO) terms than heinz ($p = 3.3 \times 10^{-12}$, *t*-test; Fig. 5B) or top *p*-values ($p < 2.2 \times 10^{-16}$, *t*-test; Fig. 5B). We also found that subnetworks identified by NetMix had greater overlap (as quantified by *F*-measure) with genes in the top *k* GO terms (Fig. 5C). These results suggest that NetMix identifies subnetworks that are more relevant to differential expression experiments than other methods.

We examined the experiment E-GEOD-11199 in more detail. This experiment compared *Mtb*-stimulated and unstimulated macrophages (Thuong et al., 2008). NetMix identified a subnetwork containing 706 genes, half the size of the jActiveModules* subnetwork containing 1450 genes. Both of these subnetworks contained 37 of the 42 genes whose differential expression was experimentally validated by reverse-transcription polymerase chain reaction (RT-PCR) (Thuong et al., 2008). Although the NetMix subnetwork was less than half the size of the jActiveModules* subnetwork, the NetMix subnetwork overlapped more GO terms (445 vs. 179). In contrast, heinz (using FDR = 0.27) identified a subnetwork of 382 genes containing only 25 RT-PCR validated genes. Finally, the 692 genes with the smallest *p*-values include only 7 validated genes. These results show that the NetMix subnetwork contains many biologically relevant genes, including most of the RT-PCR validated genes, without being overly large.

4.3. Somatic mutations in cancer

We compared the performance of NetMix, jActiveModules* (Arias-Castro et al., 2008; Addario-Berry et al., 2010), jActiveModules (Ideker et al., 2002), heinz (Dittrich et al., 2008), and Hierarchical HotNet (Reyna et al., 2018) in identifying cancer driver genes, using the MutSig2CV driver *p*-values (Lawrence

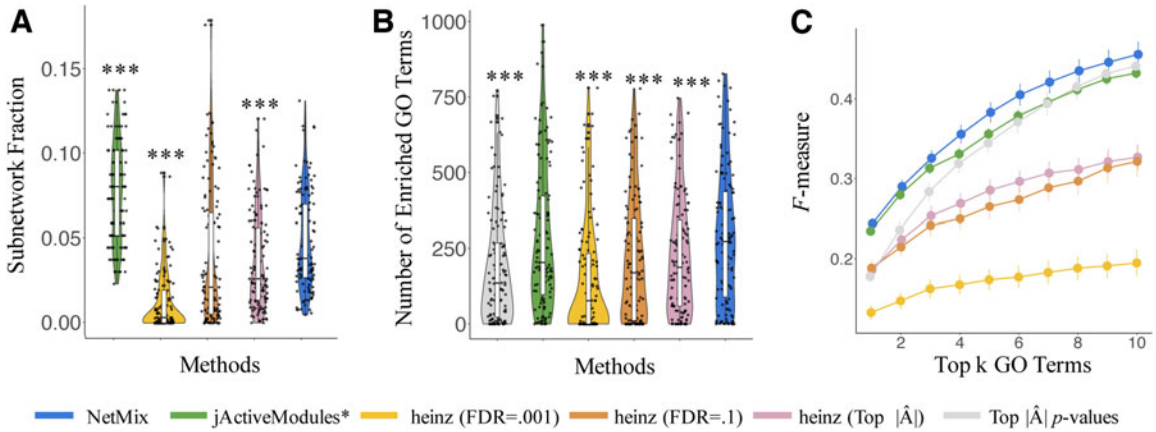

FIG. 5. **(A)** Fraction of genes in the HINT+HI interaction network that are in the subnetwork identified by each method. $^{***}p \leq 10^{-4}$ indicate significant *p*-values in paired *t*-test between and other methods. **(B)** Number of enriched GO biological process terms for altered subsets identified by each method. **(C)** *F*-measure of the *k* most enriched GO terms. GO, gene ontology.

Table 1. Results of Network Methods on Cancer Driver Gene Prediction Using MutSig2CV Driver *p*-Values from the TCGA PanCanAtlas Project and Multiple Interaction Networks

Method	*Network*			
	None	*HINT+HI*	*iRefIndex*	*ReactomeFI*
jActiveModules*	2,136/0.155	1,575/0.191	1,815/0.174	557/0.261
jActiveModules (greedy search)	NA/NA	NA/NA	NA/NA	NA/NA
jActiveModules (simulated annealing)	NA/NA	12,284/0.086	15,046/0.074	8,329/0.118
heinz (FDR = 0.001)	115/0.205	119/0.216	109/0.217	114/0.215
heinz (FDR = 0.1)	259/0.244	249/0.264	259/0.255	253/0.215
Hierarchical HotNet	NA/NA	228/0.214	297/0.215	228/0.214
NetMix	307/**0.254**	263/**0.277**	296/**0.270**	264/**0.270**

Bold values indicate the highest score.

Each entry reports the size/*F*-measure of the altered subnetwork identified by each method.

FDR, false discovery rate; HI, Human Interactone; HINT, High-quality INTeractone.

et al., 2014) from the TCGA PanCanAtlas project (Bailey et al., 2018). We ran all methods on the HINT+HI interaction network described earlier, as well as the iRefIndex 15.0 (Razick et al., 2008) and ReactomeFI 2016 (Croft et al., 2014; Fabregat et al., 2016) interaction networks. See the Supplementary Data for more details on the datasets.

We evaluated the quality of the subnetwork $\hat{A}$ reported by each method by computing the overlap with the curated list of cancer genes from the COSMIC Cancer Gene Census (CGC) (Futreal et al., 2004; Forbes et al., 2016) (Table 1). We found that NetMix outperforms all other methods in *F*-measure across all interaction networks. For example, using the HINT+HI network, NetMix achieved an *F*-measure of 0.277, compared with *F*-measures of 0.191 for jActiveModules*, 0.216 for heinz (FDR = 0.001), 0.264 for heinz (FDR = 0.1), and 0.214 for Hierarchical HotNet.[‡] Both the NetMix and Hierarchical HotNet results were statistically significant ($p < 0.01$) on all three interaction networks according to permutation tests from Reyna et al. (2018). The modest *F*-measures for all methods are not surprising; the genes in CGC have diverse alterations across cancer types and, thus, high recall is not expected by this restricted analysis of single-nucleotide mutations in a subset of cancer types. Nevertheless, the higher performance of NetMix across all networks is encouraging.

5. DISCUSSION

In this article, we revisit the classic problem of identifying altered subnetworks in high-throughput biological data. We formalize the Altered Subnetwork Problem as the estimation of the parameters of the ASD. We show that the seminal algorithm for this problem, jActiveModules (Ideker et al., 2002), is equivalent to an MLE of the ASD. At the same time, we show that the MLE is a biased estimator of the altered subnetwork, with especially a large positive bias for small altered subnetworks. This bias explains previous reports that jActiveModules tends to output large subnetworks (Nikolayeva et al., 2018).

We leverage these observations to design NetMix, a new algorithm for the Altered Subnetwork Problem. We show that NetMix outperforms existing methods on simulated and real data. NetMix fits a GMM to observed vertex scores and then finds a maximum-weighted connected subgraph by using vertex weights derived from the GMM. heinz (Dittrich et al., 2008), another widely used method for altered subnetwork identification that also derives vertex weights from a mixture model (a BUM of *p*-values) and finds a maximum weighted connected subgraph. However, heinz does not solve the Altered Subnetwork Problem in a strict sense; rather, heinz requires users to choose a parameter (an FDR estimate for the mixture fit) that implicitly constrains the size of the identified subnetwork. This user-defined parameter may encourage *p*-hacking (Ioannidis, 2005; Head et al., 2015; Nuzzo, 2015), and we find that nearly all values of this parameter lead to biased estimates of the size of the altered subnetwork.

[‡]The jActiveModules greedy search algorithm failed to complete within 100 hours, whereas the jActiveModules simulated annealing algorithm yielded an *F*-measure of 0.086.

We note a number of directions for future work. The first is to generalize our theoretical contributions to the identification of *multiple* altered subnetworks, a situation that is common in biological applications where multiple biological processes may be perturbed (Menche et al., 2015). Although it is straightforward to run NetMix iteratively to identify multiple subnetworks—as jActiveModules does—a rigorous assessment of the identification of multiple altered subnetworks would be of interest.

Second, our results on simulated data (Section 4.1) show that altered subnetwork methods have only marginal gains over simpler methods that rank vertices without information from network interactions. We hypothesize that this is because connectivity is not a strong constraint for biological networks; indeed, the biological interaction networks that we use have both small diameter and small average shortest path between vertices (see the Supplementary Data for specific statistics). Specifically, we suspect that most subsets of vertices are ''close'' to a connected subnetwork in such biological networks, and thus the MLE of the connected ASD has similar bias as the MLE of the unstructured ASD. In contrast, for other network topologies such as the line graph, connectivity is a much stronger topological constraint (see the Supplementary Data for a brief comparison of different topologies). It would be useful to investigate this hypothesis and characterize the conditions when networks provide benefit for finding altered subnetworks. In particular, other topological constraints such as dense subgraphs (Guo et al., 2007; Ayati et al., 2015), cliques (Amgalan and Lee, 2014), and subgraphs resulting from heat diffusion and network propagation processes (Vanunu et al., 2010; Vandin et al., 2011; Leiserson et al., 2015; Cowen et al., 2017) have been used to model altered subnetworks in biological data. Generalizing the theoretical results in this article to these other topological constraints may be helpful for understanding the parameter regimes where these topological constraints provide a signal for the identification of altered subnetworks. In general, which topological constraints best model altered subnetworks remains an open question.

Finally, we note that biological networks often have substantial ascertainment bias: More interactions are annotated for well-studied genes (Rolland et al., 2014; Horn et al., 2018), and so these well-studied genes, in turn, may also be more likely to have outlier measurements/scores. Thus, any network method should carefully quantify the regime where it outperforms straightforward approaches—for example, methods based on ranking vertices by gene scores or degree—both on well-calibrated simulations and on real data.

DATA AVAILABILITY

NetMix is available online at https://github.com/raphael-group/netmix

ACKNOWLEDGMENTS

The authors thank Mohammed El-Kebir for assistance with implementing jActiveModules* by modifying the ILP in heinz. They also thank David Tse for directing them to the network anomaly literature.

AUTHOR DISCLOSURE STATEMENT

B.J.R. is a cofounder of, and consultant to, Medley Genomics.

FUNDING INFORMATION

M.A.R. was supported in part by the National Cancer Institute of the NIH (Cancer Target Discovery and Development Network grant U01CA217875). B.J.R. was supported by US National Institutes of Health (NIH) grants R01HG007069 and U24CA211000.

SUPPLEMENTARY MATERIAL

Supplementary Data

REFERENCES

Addario-Berry, L., Broutin, N., Devroye, L., et al. 2010. On combinatorial testing problems. *Ann. Stat.* 38, 3063–3092.

Amgalan, B. and Lee, H. 2014. Wmaxc: A weighted maximum clique method for identifying condition-specific subnetwork. *PLoS One* 9, e104993.

Arias-Castro, E., Candès, E.J., and Durand, A. 2011. Detection of an anomalous cluster in a network. *Ann. Stat.* 39, 278–304.

Arias-Castro, E., Candès, E.J., Helgason, H., et al. 2008. Searching for a trail of evidence in a maze. *Ann. Stat.* 36, 1726–1757.

Arias-Castro, E., Castro, R.M., Tánczos, E., et al. 2018. Distribution-free detection of structured anomalies: Permutation and rank-based scans. *J. Am. Stat. Assoc.* 113, 789–801.

Arias-Castro, E., Donoho, D L., and Huo, X. 2006. Adaptive multiscale detection of filamentary structures in a background of uniform random points. *Ann. Stat.* 34, 326–349.

Ayati, M., Erten, S., Chance, M.R., et al. 2015. Mobas: Identification of disease-associated protein subnetworks using modularity-based scoring. *EURASIP J. Bioinform. Syst. Biol.* 2015, 7.

Bailey, M.H., Tokheim, C., Porta-Pardo, E., et al. 2018. Comprehensive characterization of cancer driver genes and mutations. *Cell* 173, 371–385.e18.

Batra, R., Alcaraz, N., Gitzhofer, K., et al. 2017. On the performance of de novo pathway enrichment. *NPJ Syst. Biol. Appl.* 3, 6.

Berger, B., Peng, J., and Singh, M. 2013. Computational solutions for omics data. *Nat. Rev. Genet.* 14, 333.

Bhalla, U.S., and Iyengar, R. 1999. Emergent properties of networks of biological signaling pathways. *Science* 283, 381–387.

Califano, A., Butte, A.J., Friend, S., et al. 2012. Leveraging models of cell regulation and gwas data in integrative network-based association studies. *Nat. Genet.* 44, 841–847.

Chasman, D., Siahpirani, A.F., and Roy, S. 2016. Network-based approaches for analysis of complex biological systems. *Curr. Opin. Biotechnol.* 39, 157–166.

Chen, J. 2017. Consistency of the mle under mixture models. *Stat. Sci.* 32, 47–63.

Cho, A., Shim, J.E., Kim, E., et al. 2016. Muffinn: Cancer gene discovery via network analysis of somatic mutation data. *Genome Biol.* 17, 129.

Cho, D.-Y., Kim, Y.-A., and Przytycka, T.M. 2012. Chapter 5: Network biology approach to complex diseases. *PLoS Comput. Biol.* 8, 1–11.

Choi, J., Shooshtari, P., Samocha, K.E., et al. 2016. Network analysis of genome-wide selective constraint reveals a gene network active in early fetal brain intolerant of mutation. *PLoS Genet.* 12, e1006121.

Chua, H.N., Sung, W.-K., and Wong, L. 2006. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics* 22, 1623–1630.

Cline, M.S., Smoot, M., Cerami, E., et al. 2007. Integration of biological networks and gene expression data using cytoscape. *Nat. Protoc.* 2, 2366.

Cowen, L., Ideker, T., Raphael, B.J., et al. 2017. Network propagation: A universal amplifier of genetic associations. *Nat. Rev. Genet.* 18, 551–562.

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

Croft, D., Mundo, A.F., Haw, R., et al. 2014. The reactome pathway knowledgebase. *Nucleic Acids Res.* 42, D472–D477.

Das, J. and Yu, H. 2012. Hint: High-quality protein interactomes and their applications in understanding human disease. *BMC Syst. Biol.* 6, 92.

Daskalakis, C., Tzamos, C., and Zampetakis, M. 2017. Ten steps of EM suffice for mixtures of two gaussians. Presented at the Proceedings of the 2017 Conference on Learning Theory, Amsterdam, The Netherlands. pp. 704–710.

de la Fuente, A. 2010. From 'differential expression' to 'differential networking'—identification of dysfunctional regulatory networks in diseases. *Trends Genet.* 26, 326–333.

Dempster, A.P., Laird, N.M., and Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 39, 1–38.

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

Dimitrakopoulos, C.M., and Beerenwinkel, N. 2017. Computational approaches for the identification of cancer genes and pathways. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 9, e1364.

Dittrich, M.T., Klau, G., Rosenwald, A., et al. 2008. Identifying functional modules in protein–protein interaction networks: An integrated exact approach. *Bioinformatics* 24, i223–i231.

Fabregat, A., Sidiropoulos, K., Garapati, P., et al. 2016. The reactome pathway knowledgebase. *Nucleic Acids Res.* 44, D481–D487.

Ferguson, T.S. 1982. An inconsistent maximum likelihood estimate. *J. Am. Stat. Assoc.* 77, 831–834.
Firth, D. 1993. Bias reduction of maximum likelihood estimates. *Biometrika* 80, 27–38.
Forbes, S.A., Beare, D., Boutselakis, H., et al. 2016. Cosmic: Somatic cancer genetics at high-resolution. *Nucleic Acids Res.* 45, D777–D783.
Futreal, P.A., Coin, L., Marshall, M., et al. 2004. A census of human cancer genes. *Nat. Rev. Cancer* 4, 177–183.
Glaz, J., Naus, J., and Wallenstein, S. 2001. *Scan Statistics*. Springer-Verlag, New York.
Gligorijević, V. and Pržulj, N. 2015. Methods for biological data integration: Perspectives and challenges. *J. R. Soc. Interface* 12, 20150571.
Gulsuner, S., Walsh, T., Watts, A.C., et al. 2013. Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* 154, 518–529.
Guo, M., Bao, E.L., Wagner, M., et al. 2016. SLICE: Determining cell differentiation and lineage based on single cell entropy. *Nucleic Acids Res.* 45, e54–e54.
Guo, Z., Li, Y., Gong, X., et al. 2007. Edge-based scoring and searching method for identifying condition-responsive protein–protein interaction sub-network. *Bioinformatics* 23, 2121–2128.
Halldórsson, B.V., and Sharan, R. 2013. Network-based interpretation of genomic variation data. *J. Mol. Biol.* 425, 3964–3969.
He, H., Lin, D., Zhang, J., et al. 2017. Comparison of statistical methods for subnetwork detection in the integration of gene expression and protein interaction network. *BMC Bioinf.* 18, 149.
Head, M.L., Holman, L., Lanfear, R., et al. 2015. The extent and consequences of p-hacking in science. *PLoS Biol.* 13, e1002106.
Hofree, M., Shen, J.P., Carter, H., et al. 2013. Network-based stratification of tumor mutations. *Nat. Methods* 10, 1108–1115.
Hormozdiari, F., Penn, O., Borenstein, E., et al. 2015. The discovery of integrated gene networks for autism and related disorders. *Genome Res.* 25, 142–154.
Horn, H., Lawrence, M.S., Chouinard, C.R., et al. 2018. Netsig: Network-based discovery from cancer genomes. *Nat. Methods* 15, 61–66.
Huang, J.K., Carlin, D.E., Yu, M.K., et al. 2018. Systematic evaluation of molecular networks for discovery of disease genes. *Cell Syst.* 6, 484–495.
Hung, H.M.J., O'Neill, R.T., Bauer, P., et al. 1997. The behavior of the p-value when the alternative hypothesis is true. *Biometrics* 53, 11–22.
Hung, J.-H., Yang, T.-H., Hu, Z., et al. 2011. Gene set enrichment analysis: Performance evaluation and usage guidelines. *Brief Bioinform.* 13, 281–291.
Ideker, T., Ozier, O., Schwikowski, B., et al. 2002. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18, S233–S240.
Ioannidis, J.P. 2005. Why most published research findings are false. *PLoS Med.* 2, e124.
Kelley, B.P., Yuan, B., Lewitter, F., et al. 2004. Pathblast: A tool for alignment of protein interaction networks. *Nucleic Acids Res.* 32, W83–W88.
Kim, M., and Hwang, D. 2016. Network-based protein biomarker discovery platforms. *Genomics Inf.* 14, 2–11.
Klimm, F., Toledo, E.M., Monfeuga, T., et al. 2019. Functional module detection through integration of single-cell rna sequencing data with protein–protein interaction networks. *bioRxiv* 698647.
Kulldorff, M. 1997. A spatial scan statistic. *Commun. Stat. Theor. Methods* 26, 1481–1496.
Lawrence, M.S., Stojanov, P., Mermel, C.H., et al. 2014. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 505, 495.
Lee, I., Blom, U.M., Wang, P.I., et al. 2011. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.* 21, 1109–1121.
Leiserson, M.D., Eldridge, J.V., Ramachandran, S., et al. 2013. Network analysis of gwas data. *Curr. Opin. Genet. Dev.* 23, 602–610.
Leiserson, M.D.M., Vandin, F., Wu, H.-T., et al. 2015. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* 47, 106–114.
Liang, D., Han, G., Feng, X., et al. 2012. Concerted perturbation observed in a hub network in alzheimer's disease. *PLoS One* 7, 1–17.
Liu, J.J., Sharma, K., Zangrandi, L., et al. 2018. In vivo brain gpcr signaling elucidated by phosphoproteomics. *Science* 360, eaao4927.
Lu, X., and Bressan, S. 2012. Sampling connected induced subgraphs uniformly at random. Presented at the Proceedings of the 24th International Conference on Scientific and Statistical Database Management, SSDBM'12. Springer, Berlin-Heidelberg. pp. 195–212.
Luo, Y., Zhao, X., Zhou, J., et al. 2017. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat. Commun.* 8, 573.

McLachlan, G., Bean, R.W., and Jones, L.B.-T. 2006. A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics* 22, 1608–1615.

Menche, J., Sharma, A., Kitsak, M., et al. 2015. Uncovering disease-disease relationships through the incomplete human interactome. *Science* 347, 1257601–1257601.

Mitra, K., Carvunis, A.-R., Ramesh, S.K., et al. 2013. Integrative approaches for finding modular structure in biological networks. *Nat. Rev. Genet.* 14, 719–732.

modENCODE Consortium, Roy, S., Ernst, J., et al. 2010. Identification of functional elements and regulatory circuits by drosophila modencode. *Science* 330, 1787–1797.

Nabieva, E., Jim, K., Agarwal, A., et al. 2005. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics* 21, i302–i310.

Nibbe, R.K., Koyutürk, M., and Chance, M.R. 2010. An integrative-omics approach to identify functional sub-networks in human colorectal cancer. *PLoS Comput. Biol.* 6, e1000639.

Nikolayeva, I., Pla, O.G., and Schwikowski, B. 2018. Network module identification—A widespread theoretical bias and best practices. *Methods* 132, 19–25.

Nuzzo, R. 2015. How scientists fool themselves—And how they can stop. *Nat. News* 526, 182.

Pan, W., Lin, J., and Le, C.T. 2003. A mixture model approach to detecting differentially expressed genes with microarray data. *Funct. Integr. Genomics* 3, 117–124.

Petryszak, R., Keays, M., Tang, Y.A., et al. 2015. Expression atlas update: An integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res.* 44, D746–D752.

Pounds, S., and Cheng, C. 2004. Improving false discovery rate estimation. *Bioinformatics* 20, 1737–1745.

Pounds, S., and Morris, S.W. 2003. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics* 19, 1236–1242.

Radivojac, P., Clark, W.T., Oron, T.R., et al. 2013. A large-scale evaluation of computational protein function prediction. *Nat. Methods* 10, 221–227.

Razick, S., Magklaras, G., and Donaldson, I.M. 2008. irefindex: A consolidated protein interaction database with provenance. *BMC Bioinformatics* 9, 1.

Reyna, M. A., Leiserson, M. D., and Raphael, B. J. 2018. Hierarchical hotnet: Identifying hierarchies of altered subnetworks. *Bioinformatics* 34, i972–i980.

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

Shannon, P., Markiel, A., Ozier, O., et al. 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504.

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

Sharpnack, J., Krishnamurthy, A., and Singh, A. 2013a. 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, pp. 1959–1967.

Sharpnack, J., Rinaldo, A., and Singh, A. 2016. Detecting anomalous activity on networks with the graph fourier scan statistic. *IEEE Trans. Signal Process.* 64, 364–379.

Sharpnack, J., and Singh, A. 2013. Near-optimal and computationally efficient detectors for weak and sparse graph-structured patterns. Presented at the 2013 IEEE Global Conference on Signal and Information Processing. Austin, Texas. pp. 443–446.

Sharpnack, J., Singh, A., and Rinaldo, A. 2013b. Changepoint detection over graphs with the spectral scan statistic. *Artif. Intell. Stat.* 31, 545–553.

Shrestha, R., Hodzic, E., Sauerwald, T., et al. 2017. Hit'ndrive: Patient-specific multidriver gene prioritization for precision oncology. *Genome Res.* 27, 1573–1588.

Soul, J., Hardingham, T.E., Boot-Handford, R.P., et al. 2015. Phenomeexpress: A refined network analysis of expression datasets by inclusion of known disease phenotypes. *Sci. Rep.* 5, 8117.

Thuong, N.T.T., Dunstan, S.J., Chau, T.T.H., et al. 2008. Identification of tuberculosis susceptibility genes with human macrophage gene expression profiles. *PLoS Pathog.* 4, e1000229–e1000229.

Vandin, F., Upfal, E., and Raphael, B.J. 2011. Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.* 18, 507–522.

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

Wang, X., Terfve, C., Rose, J.C., et al. 2011. HTSanalyzeR: An R/Bioconductor package for integrated network analysis of high-throughput screens. *Bioinformatics* 27, 879–880.

Wang, Y.H., Bower, N.I., Reverter, A., et al. 2009. Gene expression patterns during intramuscular fat development in cattle. *J. Anim. Sci.* 87, 119–130.

Xia, J., Gill, E.E., and Hancock, R.E.W. 2015. Networkanalyst for statistical, visual and network-based meta-analysis of gene expression data. *Nat. Protoc.* 10, 823–844.

Xu, J., Hsu, D., and Maleki, A. 2016. Global analysis of expectation maximization for mixtures of two gaussians. Presented at the Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16. Barcelona, Spain. pp. 2684–2692.

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

E-mail: braphael@princeton.edu