ORIGINAL RESEARCH
published: 22 April 2021
doi: 10.3389/fgene.2021.608042

NEM-Tar: A Probabilistic Graphical Model for Cancer Regulatory Network Inference and Prioritization of Potential Therapeutic Targets From Multi-Omics Data

Yuchen Zhang[1], *Lina Zhu*[1] and *Xin Wang*[1,2]*

[1] Department of Biomedical Sciences, City University of Hong Kong, Hong Kong, China, [2] Key Laboratory of Biochip Technology, Biotech and Health Centre, Shenzhen Research Institute, City University of Hong Kong, Shenzhen, China

Edited by:
Bairong Shen,
Sichuan University, China

Reviewed by:
Yuedong Yang,
Sun Yat-sen University, China
Rodrigo Juliani Siqueira Dalmolin,
Federal University of Rio Grande do Norte, Brazil
Yuxin Lin,
Soochow University, China
Wenying Yan,
Soochow University, China

***Correspondence:**
Xin Wang
xin.wang@cityu.edu.hk

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Genetics

Received: *18 September 2020*
Accepted: *22 March 2021*
Published: *22 April 2021*

Citation:
Zhang Y, Zhu L and Wang X (2021) NEM-Tar: A Probabilistic Graphical Model for Cancer Regulatory Network Inference and Prioritization of Potential Therapeutic Targets From Multi-Omics Data. Front. Genet. 12:608042. doi: 10.3389/fgene.2021.608042

Targeted therapy has been widely adopted as an effective treatment strategy to battle against cancer. However, cancers are not single disease entities, but comprising multiple molecularly distinct subtypes, and the heterogeneity nature prevents precise selection of patients for optimized therapy. Dissecting cancer subtype-specific signaling pathways is crucial to pinpointing dysregulated genes for the prioritization of novel therapeutic targets. Nested effects models (NEMs) are a group of graphical models that encode subset relations between observed downstream effects under perturbations to upstream signaling genes, providing a prototype for mapping the inner workings of the cell. In this study, we developed NEM-Tar, which extends the original NEMs to predict drug targets by incorporating causal information of (epi)genetic aberrations for signaling pathway inference. An information theory-based score, weighted information gain (WIG), was proposed to assess the impact of signaling genes on a specific downstream biological process of interest. Subsequently, we conducted simulation studies to compare three inference methods and found that the greedy hill-climbing algorithm demonstrated the highest accuracy and robustness to noise. Furthermore, two case studies were conducted using multi-omics data for colorectal cancer (CRC) and gastric cancer (GC) in the TCGA database. Using NEM-Tar, we inferred signaling networks driving the poor-prognosis subtypes of CRC and GC, respectively. Our model prioritized not only potential individual drug targets such as HER2, for which FDA-approved inhibitors are available but also the combinations of multiple targets potentially useful for the design of combination therapies.

Keywords: nested effects model, molecular subtype, regulatory network, drug targets, combination therapy, cancer

INTRODUCTION

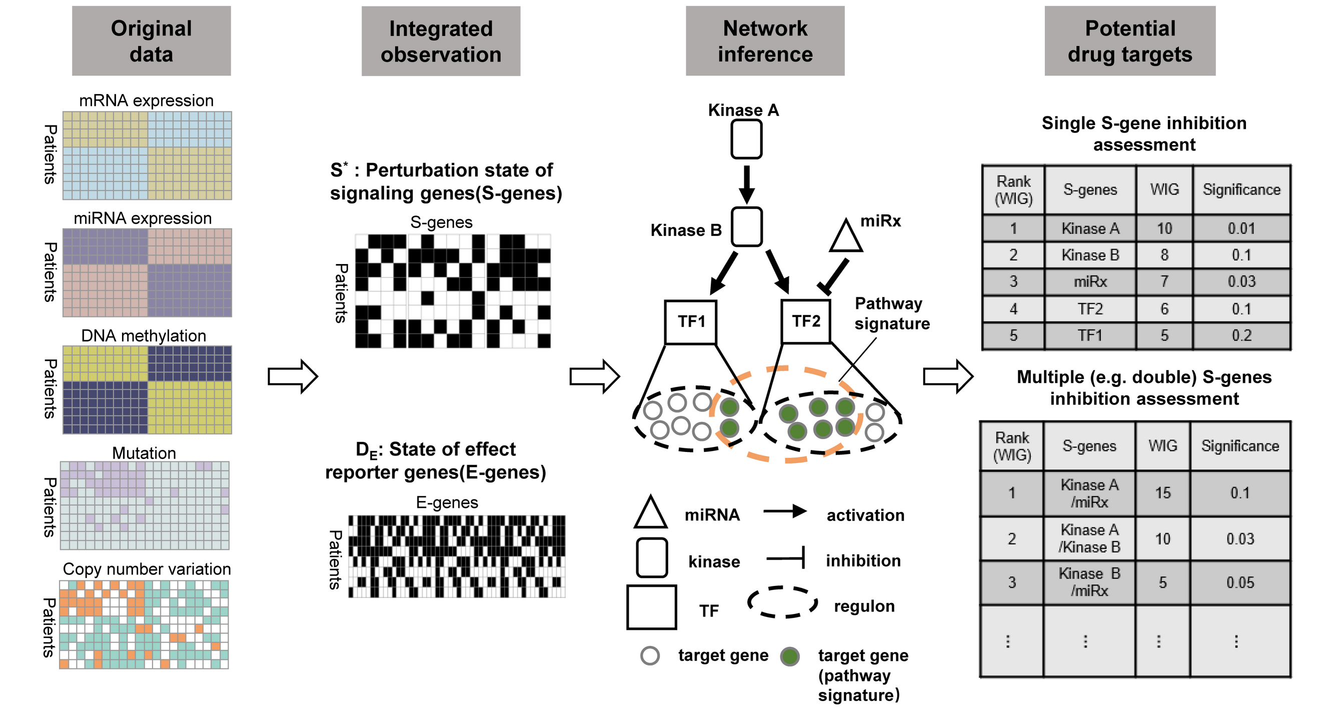
Cancers are always discovered with diverse molecular properties and heterogeneous clinical outcomes, even when occurring in the same tissues or organs. The last decade has witnessed tremendous progress in the emerging field of precision medicine for more accurate patient stratification for more optimized therapeutic treatment. However, it remains challenging to dissect the mechanism underlying cancer heterogeneity to identify novel drug targets for further development of targeted therapies. Targeted cancer therapy has been accepted as an effective weapon to conquer cancer (Green, 2004; Polyak and Garber, 2011), aiming to inhibit or reverse the activation patterns of particular cancer signaling pathways. Unfortunately, pathway redundancies, complex feedback, and crosstalk present in cancer cells often result in drug resistance, leading to treatment failure (Bernards, 2012; Yamaguchi et al., 2014). Therefore, a key task of precision medicine is excavating the causally wired relationship among the regulatory elements contributing to specific cancer molecular subtypes.

The identification of cancer therapeutic targets has long been based on biological knowledge and experience, which lacks a global functional overview and efficiency. Mathematical modeling could be established to predict potential drug targets in a more systematic and efficient way (**Supplementary Table 1**). Studies like iODA (Yu et al., 2020) integrated basic bioinformatic analysis and statistical methods to prioritize consistent molecular signatures at the pathway level for further investigation of cancer pathogenesis. Methods such as MiRNA-BD (Lin et al., 2018) focused on the discovery of novel miRNA biomarkers in diseases such as cancers without training or prior knowledge. Graphical models (e.g., Mezlini and Goldenberg, 2017; Manatakis et al., 2018; Kotiang and Eslami, 2020) were also proposed to infer the regulatory relationship and key driver genes, but the networks mainly encode gene expression associations, without support of multi-omics input. Other methods such as the miRNA-TF-mRNA network (Pham et al., 2019) and bipartite graphs (Bashashati et al., 2012) employed complex structures and multi-omics data to identify cancer driver genes as potential therapeutic targets. Furthermore, computational models were also proposed for the personalized prediction of potential target genes (Hou and Ma, 2014; Guo et al., 2018). All the previous methods have demonstrated their usefulness in various applications, very few of them infer causal regulatory relationships. To study the dysregulation of pathways and discover causal regulation relationships, typical approaches are Bayesian Networks, which encode conditional independence between genes on edges [e.g., (Sachs et al., 2005)]. However, the major limitation of Bayesian networks lies in their requirement of direct observations (e.g., protein activities) of perturbation effects on other pathway components, which are often not available. Besides, these methods require a large sample size to distinguish signal from noise and only capture parts of biologically relevant networks (Markowetz and Spang, 2007). Nested effects models (NEMs) (Markowetz et al., 2005, 2007) are specifically tailored to reconstruct signaling networks from indirect observations of experimental interventions. In each experiment, one component (e.g., kinase, transcription factor) in the pathway is perturbed, and multi-dimensional downstream effects are observed (e.g., gene expression or cell imaging data) (Siebourg-Polster et al., 2015). Different from other graphical models, NEMs encode subset relations between the observed downstream effects reporter genes under perturbations to signaling genes.

Nested effects models have been successfully applied to various biological scenarios to infer the causal network of signaling components (Markowetz et al., 2005; Fröhlich et al., 2009; MacNeil et al., 2015). Several extensions of NEMs have been proposed to adapt to different experimental designs or data types. For instance, Boolean NEMs (Pirkl et al., 2016) creatively model the data observed from arbitrary experimental combinations (excitation or inhibition) to infer a full Boolean network and further integrate the information from the literature. Epistatic NEMs (Pirkl et al., 2017) infer epistasis from phenotyping screens of double knock-downs systematically to test the hypothesis that complex relationships between a gene pair can be explained by the action of a third gene that modulates the interaction. Dynamic NEMs (Anchang et al., 2009; Fröhlich et al., 2011) infer the rate of the signal flow within the network from time-series data, while Hidden Markov NEMs (Wang et al., 2014) model the evolution of the network itself over time. Motivated by a recent experiment investigating epithelial-mesenchymal transition (EMT) in murine mammary gland cells, a method for mapping a non-interventional time series onto a static NEM has been proposed (Cardner et al., 2019). Furthermore, with the rapid development of single-cell sequencing technologies, a mixture of NEMs (M&NEM) tailored explicitly for single-cell data has been proposed (Pirkl and Beerenwinkel, 2018), which is capable of identifying different cellular subpopulations and inferring their corresponding causal networks simultaneously.

To prioritize potential therapeutic targets based on tissue-derived multi-omics profiles from cancer patients, we extended the classic NEMs to model the causal effects of genetic and epigenetic aberrations of various regulatory components (kinases, transcriptional factors, and miRNAs) on downstream genes. Importantly, the computational evaluation was conducted on the regulatory components (mainly on kinases) to prioritize potential therapeutic targets. **Figure 1** illustrated the framework and major steps of NEM-Tar, which is featured with the following highlights: (1) Different from pre-existing NEMs developed for phenotyping screens derived from experimental perturbations, NEM-Tar integrates natural perturbations (e.g., somatic mutations, DNA hyper- or hypo-methylation, copy number alterations) at multiple levels of gene regulations for cancer-related signaling network inference; (2) We proposed a scoring method based on information theory, named weighted information gain (WIG), which could prioritize not only individual therapeutic targets but also evaluate potential combination therapies; (3) NEM-Tar is a versatile framework for dissecting the cancer molecular heterogeneity by inferring cancer subtype-specific signaling network. In our case studies, we specifically focused on the 'EMT' subtype in gastric cancer and the CMS4-mesenchymal subtype in colorectal cancer (Cristescu et al., 2015; Guinney et al., 2015), which are associated with a higher risk of recurrence and poor prognosis. Potential drug targets are evaluated specifically on the epithelial-mesenchymal transition (EMT) pathway, which is directly associated with cancer metastasis.

In the 'Methods and Materials' section, we introduce the design of NEM-Tar and the inference strategies in detail. Subsequently, we test the effectiveness of NEM-Tar in a simulation study ('Results on Simulated Data') and demonstrate



Single S-gene inhibition assessment

Rank (WIG)	S-genes	WIG	Significance
1	Kinase A	10	0.01
2	Kinase B	8	0.1
3	miRx	7	0.03
4	TF2	6	0.1
5	TF1	5	0.2

Multiple (e.g. double) S-genes inhibition assessment

Rank (WIG)	S-genes	WIG	Significance
1	Kinase A /miRx	15	0.1
2	Kinase A /Kinase B	10	0.03
3	Kinase B /miRx	5	0.05
⋮	⋮	⋮	⋮

FIGURE 1 | The workflow of NEM-Tar for cancer regulatory network inference and potential drug targets prioritization. Observations of the states of S-genes and E-genes could be obtained after the preprocessing of multi-omics data. The signaling network regulating a specific cancer subtype will subsequently be inferred. Finally, based on quantification of the causal impact and specificity to downstream genes using WIG, potential drug targets could be prioritized for single and double perturbations.

its potential by real case studies on colorectal cancer and gastric cancer ('Results on Case Studies').

MATERIALS AND METHODS

The Original Nested Effects Model (NEM)

We first review the original nested effect model (NEM), before we explain in detail how we extend the original model design to fit multi-omics high-throughput profiles of cancer samples.

The structure of a NEM is illustrated in **Figure 2A**. The goal is to infer a signaling network *G*, represented as a directed acyclic graph involving the regulators, also referred to as signaling genes (S-genes), denoted S_j for j∈{1,2,....,m}. In the initial phenotypic screening experiments, the S-genes are individually perturbed during RNAi experiments, but their effects are indirectly measured by the expression level of effect reporter genes (E-genes) denoted E_i for i∈{1,2,....,n}. The attachment of E-genes to S-genes is denoted by Θ, within which $\theta_{ij} = 1$, if E-gene *i* is attached to S-gene *j*. The initial NEMs assumed that each E-gene can be attached to at most one S-gene, but this constraint has been relaxed thereafter. Tresch and Markowetz have proposed to add a null S-gene, which predicts no effects to account for uninformative features (Tresch and Markowetz, 2008). Due to the nested effects, it is assumed that the signaling network *G* is transitively closed; for instance, in **Figure 2A**, the signaling information flow is S2→S3→S4, then S2→S4 also exists.

We calculate the expected E-gene profiles for a given model (*G*;Θ) as the matrix product with E_{ij} the predicted state of E-gene *i* under knock-down of S-gene *j*, namely $E_{\text{exp}} = G\Theta$. In practice, we cannot neglect potential noise in the data, which requires probabilistic modeling to infer an optimal *G* to interpret the observation of E-genes. Suppose that we have a candidate network structure *G*, which is a directed acyclic graph (DAG) of S-genes. What matters ultimately is the posterior probability of the model:

$$P(G|E) = \frac{P(E|G)P(G)}{P(E)} \tag{1}$$

where the denominator does not depend on G and cannot be taken into consideration for model comparison. Since almost nothing is known about the signaling network without reliable knowledge, we use a uniform prior P(G). Thus, we focus entirely on the likelihood P(E|G). It can be computed by marginalizing over E-gene attachments Θ, or by employing the maximum *a posteriori* (MAP) estimate of Θ (Tresch and Markowetz, 2008). The former choice is more intuitive, and the marginal likelihood can be deduced as:

$$P(G|E) = \int P(E|G, \Theta)P(\Theta, G)d\Theta$$

$$= \frac{1}{m^n}\prod_{i=1}^{n}\sum_{j=1}^{m}\prod_{k=1}^{l} P(e_{ik}|G, \theta_i = j) \tag{2}$$

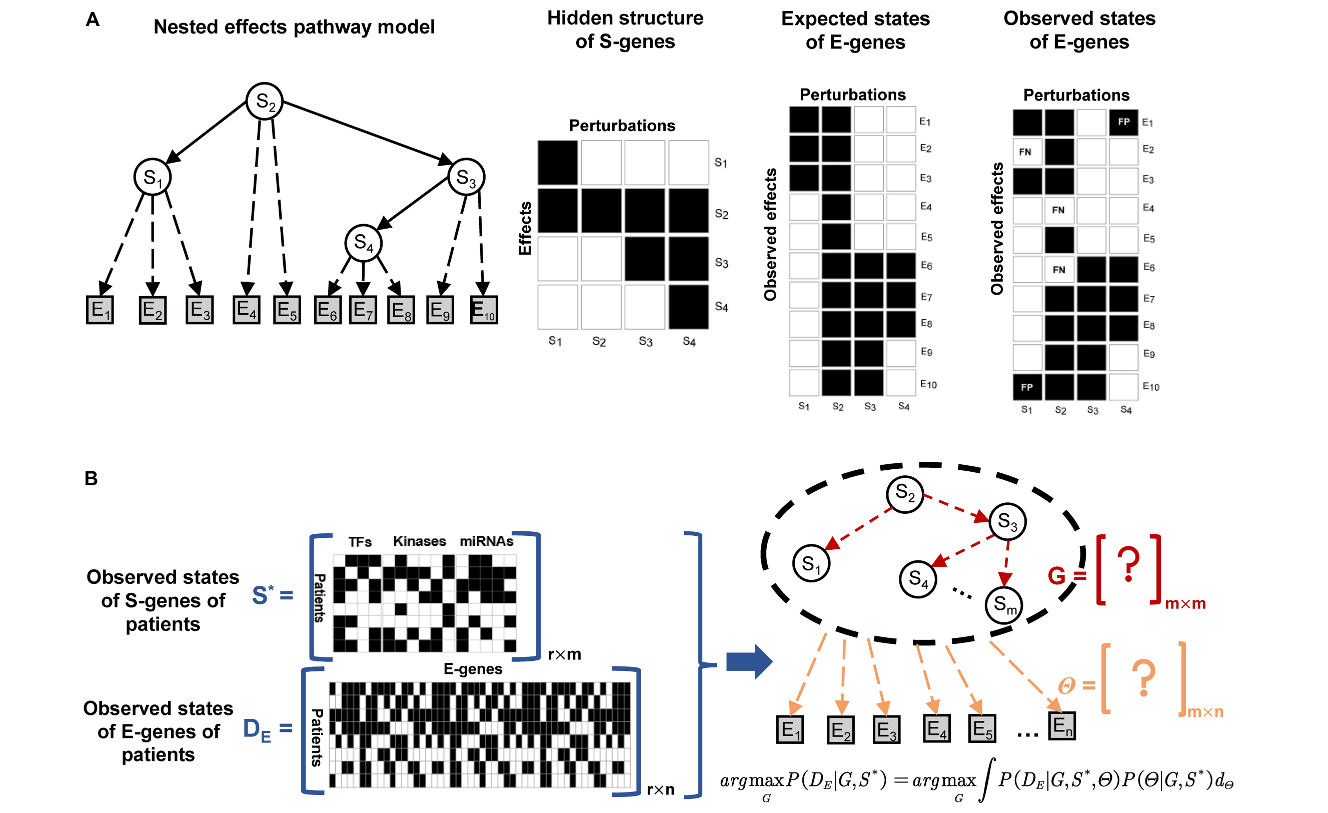

A
Nested effects pathway model
S_2
S_1
S_3
S_4
E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10}
Hidden structure of S-genes
Perturbations
Effects
S1 S2 S3 S4
Expected states of E-genes
Perturbations
Observed effects
E1 E2 E3 E4 E5 E6 E7 E8 E9 E10
Observed states of E-genes
Perturbations
Observed effects
FP
FN
FN
FN
FP
B
Observed states of S-genes of patients
$S^* =$
TFs Kinases miRNAs
Patients
r×m
Observed states of E-genes of patients
$D_E =$
E-genes
Patients
r×n
S_1 S_2 S_3 S_4 S_m
G = [?] m×m
Θ = [?] m×n
E_1 E_2 E_3 E_4 E_5 … E_n
$\arg\max_G P(D_E|G,S^*) = \arg\max_G \int P(D_E|G,S^*,\Theta)P(\Theta|G,S^*)d_\Theta$

FIGURE 2 | Illustration of the nested effects model and NEM-Tar for real cancer samples. **(A)** The S-genes are modeled as hidden variables, and their signaling interaction graph *G* (solid arrows) is the target to infer. In experiments with perturbations to individual S-genes, differential expression of downstream genes could be observed and considered as effect reporter genes (E-genes). Assuming that each E-gene is directly regulated by at most one S-gene in *G*, the maximum *a posteriori* attachment Θ (dashed arrows) of effect genes to S-genes could be computed. The goal is to search for the signaling graph *G*, which yields the most likely probabilistic nested effects. **(B)** For an extra observational dimension (the real patients), the necessary adjustment should be conducted on the design and inference strategies of classic NEM. However, the information that needs to be inferred is also the hidden interaction between S-genes and the attachment relationship of E-genes to S-genes.

The term $P(e_{ik}|G, \theta_i = j)$ in Eq. 2 reflects the noise rate of the real binary observation e_{ik}. The distribution of e_{ik} is determined by the network structure G and the error probabilities α and β. For all E-genes and targets of perturbation, the conditional probability of the E-gene state e_{ik} given the network structure G can then be written as:

$$P(e_{ik}|G, \theta_i = j) = \begin{cases} \overset{e_{ik}=1}{\alpha} & \overset{e_{ik=0}}{1-\alpha} \\ 1-\beta & \beta \end{cases} \tag{3}$$

Equation 3 means that if E_i is not an influenced target of the S-gene perturbed in experiment k, the probability of observing $e_{ik} = 1$ is α (probability of false alarm, type-I error); the probability to miss an effect and observe $e_{ik} = 0$ even though E_i is an influenced target is β (type-II error).

NEM-Tar for Multi-Omics Data

Figure 1 illustrated the NEM-Tar framework and the major steps involved to infer a signaling network using a toy example, and a comparison was made with a classic NEM in observation (**Figure 2B**). We model copy number variations or mutations (e.g., copy number gain/mutation in kinase A/B, mutation in transcription factor TF1), hyper/hypo methylation (e.g., hypermethylation of miRx) as 'natural' perturbations in tumors, which are different from experimental perturbations such as RNA interference and CRISPR-Cas9 knockout modeled in the classic NEMs. Regulators considered in the network are master regulators (TFs and miRNAs) and modulators (kinases) resulting from the reported literature (Fessler et al., 2016; Kiyozumi et al., 2018; Xie et al., 2020) and our prioritized candidates. Let $S^* = [s_{kj}]$ denote the state matrix of regulators, where s_{kj} represents whether regulator j is aberrant in sample k or not. Let G represent the signaling network of interactions between kinases, TFs, and miRNAs, and Θ be the set of interactions between regulators and their target genes. Let $D = [e_{ki}]$ be the observed data, where e_{ki} denotes whether the E-gene i is differentially expressed in patient k ($e_{ki} = 1$) or not ($e_{ki} = 0$). Our goal is to infer the optimal G that maximizes the following marginal likelihood:

$$\arg\max_G P(D|G, S^*) = \arg\max_G \int P(D_E|G, S^*, \Theta)P(\Theta|G, S^*)d\Theta \tag{4}$$

It should be noted that Eq 4 is similar to the original likelihood function of NEMs (Eq. 2), except the state matrix of regulators (S-genes) in our model.

When the optimal S-genes structure G^* is determined, we could compute the posterior probability for the edge between S_j and E_i.

$$P(\theta_i = j|G^*, S^*, D) = \frac{1}{Z}\prod_{k=1}^{l} P(e_{ki}|G^*, S^*, \theta_i = j) \qquad (5)$$

where Z is a constant and does not rely on G^*.

When using NEM-Tar in real-world applications, we recommend the following criteria to select E-genes and S-genes. E-genes can be prioritized based on genes that are significantly upregulated ($\log_2$FC > 1, FDR < 0.01) in a specific cancer subtype of interest. If the selected E-genes are too few (e.g., only 238 E-genes for the EMT subtype of GC based on the above criteria), the cutoff on $\log_2$FC may be relaxed to 0.5. The prioritization of S-genes can be based on the following criteria. First, subtype-specific miRNAs and TFs can be prioritized based on differentially expressed genes. By default, we recommend selecting TFs that are significantly upregulated ($\log_2$FC > 1 and FDR < 0.01) and miRNAs that are significantly downregulated ($\log_2$FC < -1 and FDR < 0.01). However, due to the heterogeneity between different cancer subtypes, the number of candidate miRNAs or TFs may be limited. In the situation, the cutoff on $\log_2$FC may also be relaxed to 0.5. Second, the selection of S-genes should also satisfy the following perturbation criteria: (1) Mutation: the cutoff on mutation frequency in kinases/TFs should be $>5\%$. When the overall mutation frequency of candidate S-genes is lower than 5%, the cutoff might also be relaxed appropriately. (2) copy number variations (CNVs): kinases and membrane proteins with $>5\%$ frequency of copy number gains. (3) DNA methylation: miRNAs with significant hypermethylation (delta-beta >0.1, BH-adjusted $P < 0.001$).

Inference Methods of NEM-Tar

The original NEM performs an exhaustive search over all transitively closed graphs to identify the optimal graph by the maximum likelihood estimation (Markowetz et al., 2005). Since the number of candidate network structure G grows exponentially with the number of nodes, an exhaustive enumeration is not feasible for signaling networks with more than five S-genes. In real applications, it is always necessary to search for a larger network, where heuristics are more appropriate to explore the network space. Many heuristic inference methods have been proposed, with respective advantages as well as limitations. To determine the optimal inference strategy for NEM-Tar, we investigated the triple relations, greedy hill-climbing, and MCMC sampling methods.

Instead of scoring the whole network, the model could be learned using a pairwise method (Markowetz et al., 2007). For a pair of genes A and B, their relationship could be determined by maximum *a posteriori* (MAP) from four possible models: $A{\cdot}B$ (unconnected), $A{\rightarrow}B$ (effects of A are a superset of effects of B), $A \leftarrow B$ (subset), and $A \leftrightarrow B$ (undistinguishable effects). However, the pairwise learning assumes independence of edges, which is not true in transitively closed graphs. Hence, the natural extension of pairwise learning is the inference from the triples of nodes (Markowetz et al., 2007), which comprises two steps. First, for each triple (x,y,z) in the graph with n nodes, all 29 possible quasi-orders are scored, and the MAP model is selected. Edgewise model averaging was subsequently employed to combine all models into the final graph.

Greedy hill-climbing is a more straightforward optimization strategy known from the literature (Russell and Norvig, 2016). Given an initial network hypothesis (usually an empty graph), a local maximum of the likelihood function could be reached by successively adding an edge. This procedure is continued until no improving edge can be found anymore. We also evaluated the performance of greedy hill-climbing for the benchmark in our simulation study.

Furthermore, Markov chain Monte Carlo (MCMC) methods are a class of algorithms for sampling from a probability distribution. Niederberger et al. (2012) proposed an inference method by combining MCMC sampling with an Expectation-Maximization (EM) algorithm. For reconstructing evolving signaling networks, MCMC sampling was also an important procedure in HM-NEM (Wang et al., 2014). In our simulation study, we also examined MCMC sampling, and the detailed pseudocode is in **Supplementary Figure 1**.

Weighted Information Gain (WIG) for Evaluation of the Causal Impact of S-Genes on Downstream Reporter Genes

Given the inferred optimal network G^* and interactions between regulators and target genes Θ^*, we sought to quantify the causal impact that a regulator has on downstream reporter genes, especially signature genes for a particular biological process of interest such as epithelial–mesenchymal transition (or EMT) (Nieto et al., 2016; Lambert et al., 2017). The fundamental assumptions of the assessment criteria for the impact should satisfy: (1) The more E-genes related to a particular pathway are affected by an S-gene, the more significant the influence is; (2) The more likely a particular E-gene is attached to an S-gene, the higher the global influence of the S-gene is. On the basis of the above assumptions, we defined a score called **Weighted Information Gain (WIG)** on every E-gene within the regulons of S-genes based on KL divergence (Kullback and Leibler, 1951) in information theory, which measures the information gain after network inference.

$$\begin{aligned} WIG(S_j) &= \textstyle\sum_{i=1}^{r} WIG(S_j \rightarrow E_i) \\ &= \textstyle\sum_{i=1}^{r} P(S_j \rightarrow E_i)\log[(m+1)P(S_j \rightarrow E_i)] \end{aligned} \qquad (6)$$

As shown in **Figure 3**, before the network inference, for every E-gene, we assume that the probability of an E-gene attached to an S-gene is uniformly distributed, which could be denoted as$P(\theta_i = j|G) = 1/m + 1$, if we set a ‘null’ S-gene and no particular prior knowledge is involved. While after the inference, the posterior distribution of nested effect positions of E-genes changes into$P(S_j \rightarrow E_i) = p(\theta_i = j|G^*, S^*, D)$. According to the original definition of KL divergence, the increase of the information of the attachment of an E-gene could be computed, like the highlighted $WIG(S_3 \rightarrow E_3)$ and $WIG(S_3 \rightarrow E_{14})$. As for an S-gene, the global causal impact over all the E-genes or some signature genes of key pathways could be obtained by summing up the WIG of related E-genes, as shown in Eq. 6. The statistical significance for the specificity of WIG on key pathways could be estimated by the bootstrap of the same number as the pathway signature genes of arbitrary E-genes within the regulon of a S-gene.

Ultimately, kinases/TFs/miRNAs with top causal WIG and/or enough significance will be prioritized as potential drug targets. For more convenient drug design, kinases, or membrane proteins are preferred.

Data Source

In our case studies, we analyzed multi-omics data for colorectal cancer (CRC) and gastric cancer (GC) patients from TCGA, including the following data types: (1) whole-genome gene expression data for 382 CRC and 415 GC patients based on RNA sequencing platform; (2) copy number variation data (scores on gene level) for 374 CRC patients and 268 GC patients; (3) somatic mutations profiles for 423 CRC patients and 433 GC patients; (4) miRNA expression data for 297 CRC and 446 gastric tumors based on Illumina sequencing platform; (5) DNA methylation data for 396 CRC and 395 GC tumor samples based on Infinium Methylation 450K platform.

RESULTS

Results on Simulated Data

Generation of *in silico* Data

The simulations evaluating the inference strategies of NEM-Tar were performed on datasets generated with varying network sizes and noise levels. The generation of simulated data is described in detail as follows.

(1) **S-gene graph generation**: We first randomly generated a graph of m S-genes, m$\in$ {6,8,10,12,15,20,30}. These graphs of S-genes were transformed to transitively closed graphs.

(2) **S-gene state generation**: For each S-gene graph generated, we simulated patient samples with a random fraction of S-genes perturbed according to the real proportions of S-genes with genetic and epigenetic alterations in the gastric cancer case study. An S-gene state matrix was subsequently generated according to the S-gene graph and simulated perturbations.

(3) **Attachment of E-genes to S-genes**: In each S-gene graph simulated, we attached effect reporter genes (or E-genes) to each S-gene, and the number of E-genes per S-gene was roughly equivalent to the average number of E-genes in the gastric cancer case study.

(4) **Generation of E-gene observations**: For each simulated graph, with the corresponding S-gene state matrix and E-gene attachment, we next generated the corresponding E-gene observation matrix. For E-genes without downstream effects expected, observations were sampled from a null distribution, or otherwise from an alternative distribution. In the simplest case, we only sampled binary data, where 1 indicated an effect and 0 no effect, according to the Type-I error α_{sim} (FP) and Type-II error β_{sim} (FN).

Using the simulation strategy, we generated data to test the performance of NEM-Tar:

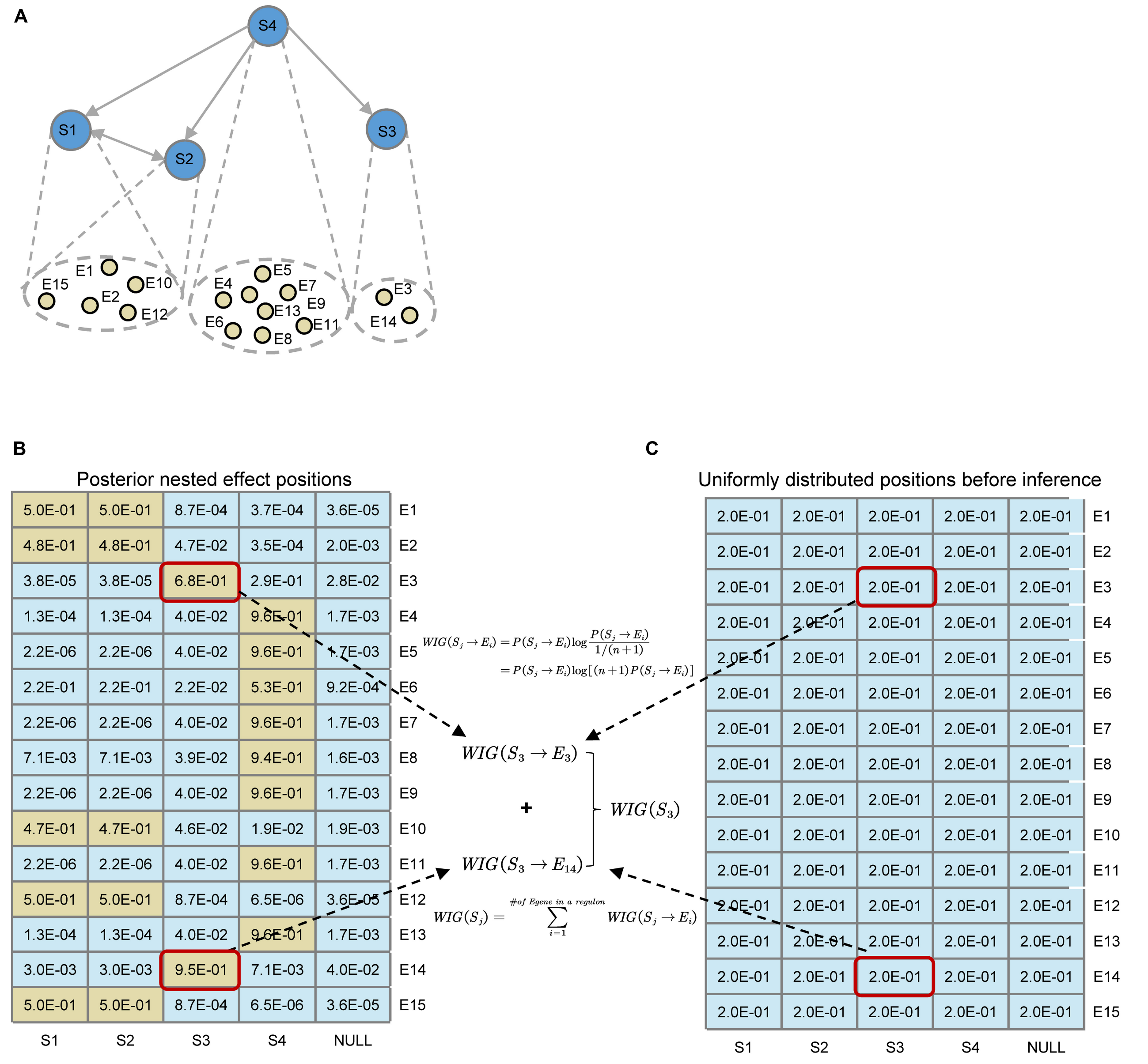
(1) Scalability. Fix $\alpha_{sim} = \beta_{sim} = 0.05$ and vary the number of S-genes from 6 to 30, representing the size of a typical signaling pathway. For each number of S-genes, 200 different random S-gene networks were generated, and the simulated S-gene network structures were inferred using MCMC sampling, triple relations, and greedy hill-climbing, respectively;

(2) Robustness to noise. Fix $\beta_{sim} = 0.05$ and the number of S-genes $m = 12$ (medium size) and vary α_{sim} from 0.05 to 0.5. For the inference of S-gene network, we set $\alpha = 0.2$, $\beta = 0.1$, which were arbitrarily chosen and different from the α_{sim} and β_{sim} used for the generation of E-gene data. The evaluation criteria of their performance were TPR = TP/(TP + FN), TNR = TN/(TN + FP), Accuracy = (TP + TN)/(TP + FN + TN + FP) and Precision = TP/(TP + FP).

Benchmark the Performance of Inference Methods

The simulation results are shown in **Figure 4**. Using MCMC sampling (**Figure 4A**), although the performance showed a decreasing trend due to the increase of the size of the network, the magnitude of decrease was quite significant (e.g., the averaged TPR of 200 networks decreased from 0.867 to 0.136). Especially, the most concerned measure ‘Precision’ was unacceptable in real applications, no matter for smaller networks (S-genes $\leq$ 10) or larger networks (S-genes $>$ 10). Even for smaller networks with only six S-genes the instability of MCMC was evident, as the median of Precision (0.845) was much larger than the mean (0.770). For relatively large networks (e.g., 20 S-genes), the averaged Precision was too low (0.328) to accept. Using the triple relations inference, the result was slightly better than MCMC sampling (**Figure 4B**), but a dramatic decrease of Precision was also observed for networks with $\geq$ 10 S-genes. A special observation on the triple relations is that the performance on the networks with a medium size (10-15) showed fluctuating TPR, TNR, and Accuracy rather than a steady decrease, suggesting that the inference based on triple relations was also unstable.

Compared to MCMC sampling and triple relations methods, greedy hill-climbing showed much higher performance (**Figure 4C-D**). For small and medium networks (6-15 S-genes), the median of all the evaluation metrics were close to 1. Even for relatively large networks, the TPR and Precision were still reliable. Though, in essence, the greedy hill-climbing is likely to be trapped in a local optimum, at least for the graphs with less than 30 S-genes, the performance is reasonably good. The robustness for the inference with varying α_{sim} based on greedy hill-climbing is also stable and acceptable. Even for the very noisy condition (α_{sim} = 0.5), the averaged TPR and Precision could still reach 0.947 and 0.766, respectively. Furthermore, compared to the other two methods, the greedy hill-climbing algorithm was not only superior in the performance, but also less time consuming (**Supplementary Table 2**).



Posterior nested effect positions

S1	S2	S3	S4	NULL	
5.0E-01	5.0E-01	8.7E-04	3.7E-04	3.6E-05	E1
4.8E-01	4.8E-01	4.7E-02	3.5E-04	2.0E-03	E2
3.8E-05	3.8E-05	6.8E-01	2.9E-01	2.8E-02	E3
1.3E-04	1.3E-04	4.0E-02	9.6E-01	1.7E-03	E4
2.2E-06	2.2E-06	4.0E-02	9.6E-01	1.7E-03	E5
2.2E-01	2.2E-01	2.2E-02	5.3E-01	9.2E-04	E6
2.2E-06	2.2E-06	4.0E-02	9.6E-01	1.7E-03	E7
7.1E-03	7.1E-03	3.9E-02	9.4E-01	1.6E-03	E8
2.2E-06	2.2E-06	4.0E-02	9.6E-01	1.7E-03	E9
4.7E-01	4.7E-01	4.6E-02	1.9E-02	1.9E-03	E10
2.2E-06	2.2E-06	4.0E-02	9.6E-01	1.7E-03	E11
5.0E-01	5.0E-01	8.7E-04	6.5E-06	3.6E-05	E12
1.3E-04	1.3E-04	4.0E-02	9.6E-01	1.7E-03	E13
3.0E-03	3.0E-03	9.5E-01	7.1E-03	4.0E-02	E14
5.0E-01	5.0E-01	8.7E-04	6.5E-06	3.6E-05	E15

Uniformly distributed positions before inference

S1	S2	S3	S4	NULL	
2.0E-01	2.0E-01	2.0E-01	2.0E-01	2.0E-01	E1
2.0E-01	2.0E-01	2.0E-01	2.0E-01	2.0E-01	E2
2.0E-01	2.0E-01	2.0E-01	2.0E-01	2.0E-01	E3
2.0E-01	2.0E-01	2.0E-01	2.0E-01	2.0E-01	E4
2.0E-01	2.0E-01	2.0E-01	2.0E-01	2.0E-01	E5
2.0E-01	2.0E-01	2.0E-01	2.0E-01	2.0E-01	E6
2.0E-01	2.0E-01	2.0E-01	2.0E-01	2.0E-01	E7
2.0E-01	2.0E-01	2.0E-01	2.0E-01	2.0E-01	E8
2.0E-01	2.0E-01	2.0E-01	2.0E-01	2.0E-01	E9
2.0E-01	2.0E-01	2.0E-01	2.0E-01	2.0E-01	E10
2.0E-01	2.0E-01	2.0E-01	2.0E-01	2.0E-01	E11
2.0E-01	2.0E-01	2.0E-01	2.0E-01	2.0E-01	E12
2.0E-01	2.0E-01	2.0E-01	2.0E-01	2.0E-01	E13
2.0E-01	2.0E-01	2.0E-01	2.0E-01	2.0E-01	E14
2.0E-01	2.0E-01	2.0E-01	2.0E-01	2.0E-01	E15

$$\begin{aligned} WIG(S_j \rightarrow E_i) &= P(S_j \rightarrow E_i)\log\frac{P(S_j \rightarrow E_i)}{1/(n+1)} \\ &= P(S_j \rightarrow E_i)\log[(n+1)P(S_j \rightarrow E_i)] \end{aligned}$$

$$\left.\begin{aligned} &WIG(S_3 \rightarrow E_3) \\ &+ \\ &WIG(S_3 \rightarrow E_{14}) \end{aligned}\right\} WIG(S_3)$$

$$WIG(S_j) = \sum_{i=1}^{\#of\ Egene\ in\ a\ regulon} WIG(S_j \rightarrow E_i)$$

FIGURE 3 | Illustration for the definition of Weighted Information Gain (WIG) using a toy example. **(A)** A toy network containing four S-genes with their corresponding E-gene attachment. Note that the hierarchies of S1 and S2 genes cannot be distinguished. **(B)** Posterior effect positions obtained after network inference; **(C)** Uniformly distributed effect positions before inference. Suppose that the attached E-genes to a S-gene are all signature genes related to a pathway of interest (e.g., EMT), it could be easily calculated that S_4 has the highest causal impact on the particular downstream pathway, and S_1 and S_2 have the same impact. As an example, we illustrated the calculation of *WIG*(S_3).

Therefore, based on the simulation study, we employed greedy hill-climbing as the inference method for the following case studies.

Results on Case Studies

To exemplify NEM-Tar for inference of cancer subtype-specific signaling network and prioritization of potential therapeutic targets, we did two case studies using multi-omics data in gastric cancer and colorectal cancer, respectively.

Inferring the Signaling Network Driving the EMT Subtype of Gastric Cancer and Prioritization of Potential Drug Targets

Gastric cancer (GC), a leading cause of cancer-related deaths, is known to be a heterogeneous disease. The presence of molecular heterogeneity in GC has been shown through the existence of subtypes with distinct genetic/epigenetic aberrations associated with clinical outcomes. Based on 300 primary gastric cancer tumor specimens, the Asian Cancer Research Group (ACRG) identified four molecular

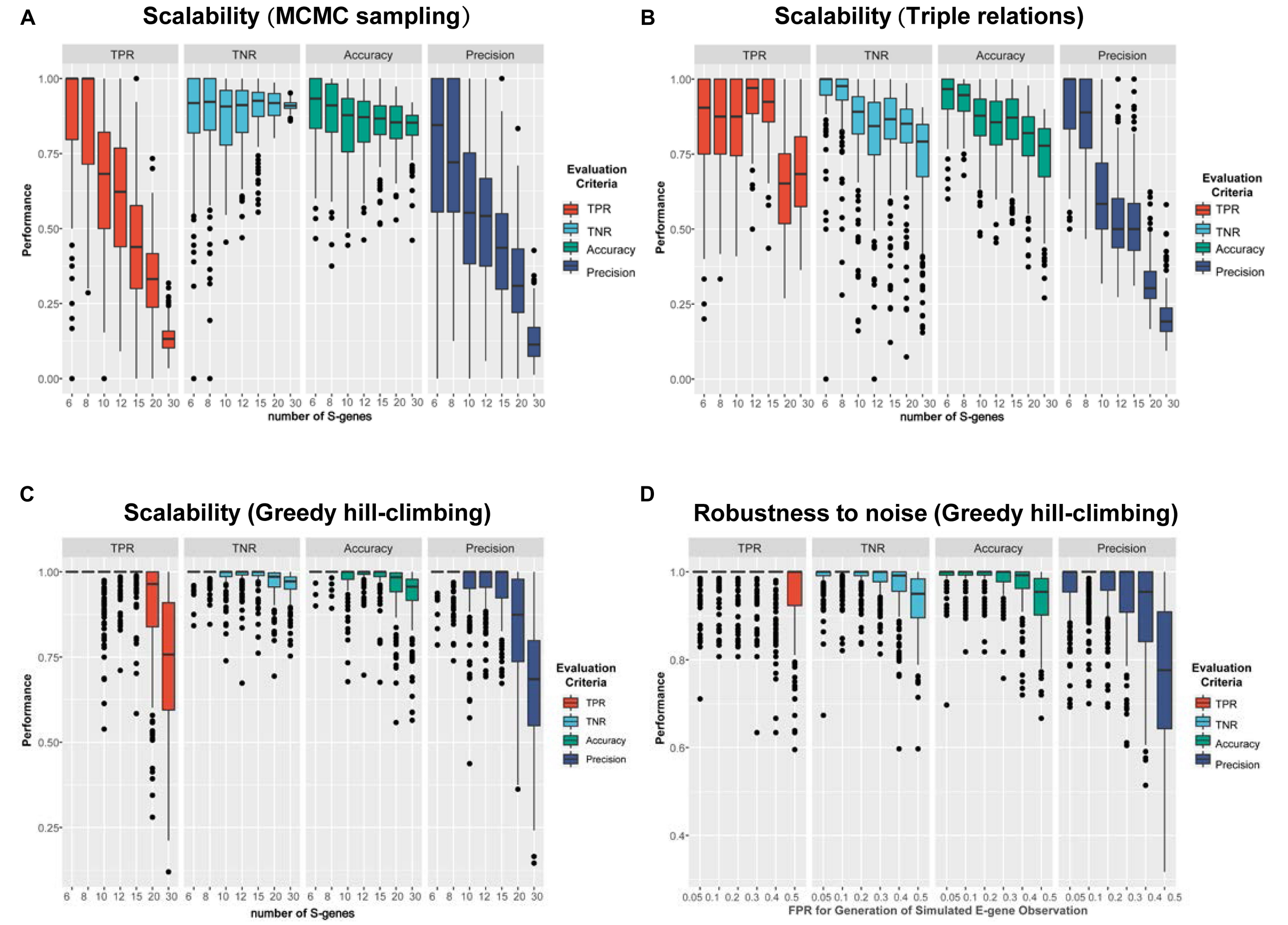


FIGURE 4 | A comparison of the performance of three representative network inference strategies. **(A–C)** The performance of NEM-Tar based on **(A)** MCMC sampling, **(B)** triple relations, and **(C)** greedy hill-climbing, respectively, on simulated data for varying numbers of S-genes. For each method, we generated 200 random signaling networks and inferred their structures using NEM-Tar from the simulated E-gene data. **(D)** The performance of NEM-Tar based on greedy hill-climbing testing its robustness to simulated data with different levels of noise.

subtypes with distinct patterns of molecular alterations and clinical outcomes (Cristescu et al., 2015). Among these four subtypes, patients classified to the MSS/EMT (in short, EMT) subtype showed the worst prognosis. Despite the extensive subtyping studies published, the regulatory mechanism underlying specific molecular subtypes has not been fully explored explicitly. Here, we employed NEM-Tar to infer the signaling network driving the EMT gastric cancer and quantitatively evaluate single and double perturbations to prioritize potential drug targets.

For the choice of the regulatory elements, we focused on the signature genes of the MAP-kinase pathway (KRAS, BRAF), frequently mutated kinases/TFs (TP53, ARID1A, CDH1, and ERBB2) (Gastric Adenocarcinoma - My Cancer Genome) and significantly upregulated TFs ($\log_2$FC > 0.5, BH-adjusted $P < 0.01$) as well as downregulated miRNAs ($\log_2$FC < -1, BH-adjusted $P < 0.01$) in the EMT subtype. The regulatory elements were filtered through the integration with the somatic mutation profiles. More specifically, we kept the kinases and TFs with mutation frequency > 5% and ZEB2 (Dai et al., 2012) and KRAS (Yoon et al., 2019), which were well characterized before for their roles in EMT regulation. As a result, we included nine kinases/TFs in 177 patient samples for the following analysis. The perturbations to miRNAs were measured by DNA methylation in the promoters, and six miRNAs were selected with highly significant hypermethylation (delta-beta > 0.1, BH-adjusted $P < 0.001$) in the samples of the EMT subtype. Since copy number variations (CNVs) were frequently found in kinases and membrane proteins in many cancer types, we also incorporated copy number gains as a type of perturbation in the case study. Furthermore, 1194 genes significantly upregulated in the EMT subtype ($\log_2$FC > 0.5, BH-adjusted $P < 0.01$) were selected as E-genes for the following analysis.

In the classic NEMs, E-genes' states are the production of individually perturbed S-genes, while for NEM-Tar E-genes' states can be the production of multiple S-genes with genetic and/or epigenetic perturbations in a tumor sample. Therefore, the concepts of positive and negative controls for the discretization of E-genes' states should be revised accordingly. A positive control referred to the patients belonging to a particular subtype (e.g., EMT subtype) but without any (epi)genetic aberrations in the S-genes. In contrast, a negative control referred to patients not assigned to a particular subtype (e.g., Non-EMT subtypes) and had no aberrations in any S-genes. Using the strategy, we transformed the continuous gene expression data into binary observations. Denote C_{ik} as the continuous expression level of E_i of patient k. Let $\mu_i{}^+$ be the mean of positive controls for $E_{i,}$ and $\mu_i{}^-$ the mean of negative controls. To derive binary data $E_{ik,}$ we defined individual cutoffs for every gene E_i by:

$$E_{ik} = \begin{cases} 1 \ \ if \ C_{ik} < \sigma \cdot \mu_i^+ + (1 - \sigma) \cdot \mu_i^- \\ 0 \ else \end{cases} \tag{7}$$

Based on the method introduced in Markowetz et al. (2005), to make a balance between Type-I and Type-II errors, we set σ = 0.5 for the discretization. As a result, we obtained the estimated error rates α = 0.07, β = 0.08.

Using the discretized E-gene data, we inferred the S-gene network regulating the EMT subtype of GC using NEM-Tar (**Figure 5A**). Interestingly, CDH1, ERBB2 (HER2), and KRAS were predicted to be sitting at the top hierarchies in the signaling network. Indeed, Trastuzumab, a monoclonal antibody for human epidermal growth factor receptor 2 (HER2), has already been established with chemotherapy as a first-line treatment for HER2-positive metastatic advanced GC patients (Bang et al., 2010). Besides, CDH1, coding for the E-cadherin protein, was reported to be linked to GC susceptibility and tumor invasion, and preliminary studies indicated the potential clinical value to employ CDH1 haplotypes in metastatic GC to stratify patients that will benefit from Trastuzumab-based treatments (Caggiari et al., 2017). NEM-Tar further supports the important discovery by computationally predicting and statistically evaluating the potential drug targets. Summarizing the single and double S-gene perturbations (to kinases only) with top WIGs, we found that both CDH1 and HER2 had a strong causal impact on the signature genes of epithelial-mesenchymal transition (EMT) (Zhao et al., 2019). More importantly, the causal effect was statistically significant and specific to the EMT pathway only (**Table 1**), as quantified by permutation tests, i.e., random sampling of E-genes with the same number of EMT signature genes in the regulon of a S-gene, and calculating the frequency of observing a same or higher WIG from the sampled E-gene sequences. Moreover, the combinatorial perturbations (**Table 2**) to CDH1 and ERBB2, CDH1 and KRAS or CDH1 and BRAF had the strongest and specific causal effect on the EMT pathway among all possible combinations.

Inferring the Signaling Network Driving the CMS4-Mesenchymal Subtype of Colorectal Cancer and Prioritization of Potential Drug Targets

Similar to gastric cancer, colorectal cancer (CRC) is also a heterogeneous disease posing a challenge for accurate classification and treatment of this malignancy. Recently, CRC patients have been categorized using unsupervised classification of gene expression profiling, which resulted in distinct CRC subtypes. In order to generate unified subtyping of CRC, based on a large panel of CRC patients (n = 4151), the CRC Subtyping Consortium identified four consensus molecular subtypes (CMSs) (Guinney et al., 2015). Linking the subtypes to disease outcomes revealed that the mesenchymal subtype CMS4 displayed a worse prognosis, highlighting the clinical relevance of the CMS taxonomy. As another case study, we employed NEM-Tar to infer the signaling network driving the CMS4 CRC and

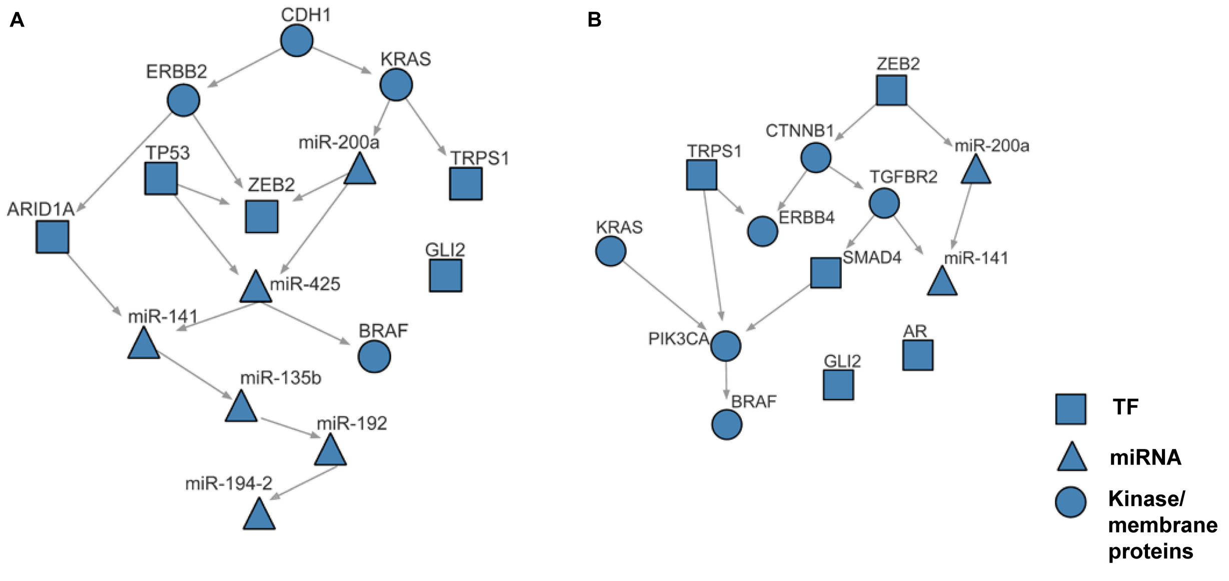


FIGURE 5 | The case studies of NEM-Tar on gastric cancer and colorectal cancer. **(A)** Reconstructed signaling network for the EMT subtype of gastric cancer. **(B)** Reconstructed signaling network for the CMS4 subtype of colorectal cancer.

TABLE 1 | WIGs assessing the impact of single perturbations (kinase only) on EMT in GC.

S-genes	Total No. of downstream E-genes within the regulon	No. of E-genes (EMT related) within the regulon	WIG	Significance of WIG (100,000 sampling, BH-adjusted *P*)
CDH1	591	57	66.15	<1e-05
ERBB2	491	44	51.65	<1e-05
KRAS	229	20	20.05	<1e-05
BRAF	14	1	2.71	3.21e-01

TABLE 2 | Double perturbations (kinase only) with top WIGs in GC.

S-genes	Total No. of downstream E-genes within the regulon	No. of E-genes (EMT related) within the regulon	WIG	Significance of WIG (50,000 sampling, BH-adjusted *P*)
CDH1/ERBB2	591	57	66.15	<5e-04
CDH1/KRAS	591	57	66.15	<5e-04
BRAF/CDH1	591	57	66.15	<5e-04
KRAS/ERBB2	558	51	59.12	<5e-04
BRAF/ERBB2	505	45	54.36	<5e-04
KRAS/BRAF	229	20	20.05	<5e-04

calculated WIGs for single and double perturbations to signaling elements in order to prioritize potential drug targets.

To select regulatory elements, we incorporated the signature genes of the MAP-kinase pathway (KRAS, BRAF, PIK3CA), and the TFs significantly upregulated in CMS4 ($\log_2$FC > 1, BH-adjusted $P < 0.01$) as well as the miRNAs significantly downregulated in CMS4 ($\log_2$FC < -0.5, BH-adjusted $P < 0.01$). The regulatory elements were filtered through the integration with the somatic mutation profiles. More specifically, the kinases and TFs with the mutation frequency $> 5\%$ were left, resulting in 11 kinases/TFs in 212 patient samples for analysis. The perturbations to miRNAs were measured by DNA methylation in the promoters, and two miRNAs were selected with highly significant hypermethylation (delta-beta > 0.1, BH-adjusted $P < 0.001$) in the samples of the CMS4 subtype. The copy number variations (CNVs) profiles were also preprocessed, but the frequency of copy number gain was too low (less than 5%) to integrate. Finally, after integration with downstream E-genes ($\log_2$FC = 1, BH-adjusted $P = 0.01$) that are differentially expressed between CMS4 and non-CMS4 samples, we obtained a 212 × 1337 E-gene observation matrix for the following analysis.

The whole discretization analysis of E-genes is similar to what we did in gastric cancer. In CRC, the positive controls are patients belonging to the CMS4 subtype without any aberrations in any S-genes, while the negative controls are patients assigned to Non-CMS4 subtypes without aberrations in any S-genes. We set $\sigma = 0.6$ for the discretization, and the estimated error rates were $\alpha = 0.22$ and $\beta = 0.18$. Using the discretized E-gene data, we inferred the S-genes network regulating the CMS4 subtype of CRC (**Figure 5B**). Based on the WIG calculation (**Table 3**, **4**), we found that the perturbation on KRAS has the highest impact on the EMT pathway, though the influence is not specific to EMT, which is reasonable as KRAS is a frequently mutated oncogene in cancer. Currently, a variety of methods to inhibit KRAS for the treatment of metastatic CRC have been proposed (Porru et al., 2018). Besides, CTNNB1, which encodes β-catenin, has the second highest impact on the EMT pathway. CTNNB1 is involved in the Wnt-β-catenin signaling pathway, which often drives a transcriptional program that is reminiscent of EMT (Anastas and Moon, 2013). Particularly, the role of Wnt-β-catenin signaling in CRC and its potential as a therapeutic target for CRC has been extensively explored. Existing drugs targeting β-catenin, such as Aspirin, are already available, and several small molecules are under clinical trials (Cheng et al., 2019). Furthermore, the combinatorial perturbations to KRAS and CTNNB1, as well as KRAS and TGFBR2, enhanced the causal impact on the EMT pathway compared to their single perturbations, suggesting potential combination therapies for the specific CMS4 subtype of CRC.

DISCUSSION

Although quite a few computational approaches have been developed for the identification of cancer therapeutic targets, they differ in the types of input data, the design of models/algorithms, the output of the results and the angles of biological interpretations. The unique strength of our NEM-Tar lies in its capability to prioritize not only individual therapeutic targets but also combinational therapies, which has not been realized before as far as we know. As a result, it is very difficult to quantitatively compare NEM-Tar with other computational approaches directly. However, we tried to make a rough comparison with two widely used methods, DawnRank (Hou and Ma, 2014) and DriverNet (Bashashati et al., 2012), which were proposed to discover cancer driver genes. Using DawnRank, we found that for the CMS4 subtype in CRC, AR, and GLI2, two TFs in our regulatory network, were also ranked among the top 5% (**Supplementary Table 3**). More excitingly, CDH1 and TP53 were ranked as the top two drivers for the EMT subtype in GC (**Supplementary Table 3**). When it comes to the result of DriverNet, only CDH1 and TP53 were prioritized as the 2nd and 3rd for EMT subtype in GC (**Supplementary Table 4**). However,

TABLE 3 | WIGs assessing the impact of single perturbations (kinase only) on EMT in CRC.

S-genes	Total No. of downstream E-genes within the regulon	No. of E-genes (EMT related) within the regulon	WIG	Significance of WIG (100,000 sampling, BH-adjusted *P*)
KRAS	525	49	38.61	1.48e-01
CTNNB1	151	14	23.26	< 1e-05
TGFBR2	85	10	16.67	< 1e-05
PIK3CA	26	3	5.98	3.14e-02
BRAF	15	2	3.94	5.21e-01
ERBB4	23	2	2.95	5.25e-01

TABLE 4 | Double perturbations (kinase only) with top WIGs in CRC.

S-genes	Total No. of downstream E-genes within the regulon	No. of E-genes (EMT related) within the regulon	WIG	Significance of WIG (50,000 sampling, BH-adjusted *P*)
KRAS/CTNNB1	650	60	55.88	<5e-04
KRAS/TGFBR2	584	56	49.29	2.73e-04
KRAS/ERBB4	548	51	41.56	1.06e-01
KRAS/BRAF	525	49	38.61	2.62e-01
KRAS/PIK3CA	525	49	38.61	1.71e-01
BRAF/CTNNB1	151	14	23.26	<5e-04
PIK3CA/CTNNB1	151	14	23.26	<5e-04
TGFBR2/CTNNB1	151	14	23.26	<5e-04
ERBB4/CTNNB1	151	14	23.26	< 5e-04
TGFBR2/ERBB4	108	12	19.62	4.20e-05

no driver genes were found consistent between NEM-Tar and DriverNet for the CMS4 subtype of CRC (**Supplementary Table 5**). It should be noted that DawnRank and DriverNet could dissect the driver genes only based on the modeling of association networks, which lack the inference of causal relationships and cannot measure double or multiple therapeutic targets. Furthermore, neither DriverNet nor DawnRank were designed to distinguish TFs and kinases and could not incorporate perturbation information at other levels of gene expression regulations except for gene mutations. Instead, NEM-Tar was developed to prioritize potential therapeutic targets using regulatory network inference based on nested effects models.

The hierarchical causal relationship between signaling components is not only central for understanding the regulatory mechanism of cancers but also critical for developing potential drug targets to overcome the pervasive genetic redundancies. Inspired by NEMs encoding subset relations between observed downstream effects of experimental perturbations in signaling genes, we proposed NEM-Tar to infer signaling networks from various genetic and epigenetic perturbations to regulatory elements such as kinases, transcriptional factors, and miRNAs. The marginal likelihood function of NEM-Tar is similar to the original likelihood function of NEM, except the state matrix of regulators (S-genes) in our model. Based on NEM-Tar, a new score named weighted information gain (WIG) was defined to assess the causal impact of S-genes on downstream reporter genes.

Colorectal cancer and GC are two major malignancies of the gastrointestinal tract, for which molecular subtyping has been well studied. To exemplify the usefulness of NEM-Tar, we performed two case studies to infer signaling networks that drive the poor prognosis subtypes of GC and CRC, respectively. In GC, we found that among the top significant signaling genes with high WIGs, CDH1, and ERBB2 are particularly attractive. Indeed, the FDA-approved drug Trastuzumab targeting ERBB2 has already been established with chemotherapy as a first-line treatment for HER2-positive metastatic advanced GC patients. Our further evaluation of combinatorial perturbations suggested that simultaneous inhibition of CDH1 and ERBB2/KRAS/BRAF, ERBB2, and KRAS/BRAF, as well as KRAS and BRAF may be potential combination therapies. For CMS4 CRC, except for KRAS, a representative oncogene employed as a therapeutic target, the kinase CTNNB1 with the second highest WIG may be a potential alternative therapeutic target to CRC, and combinatorial inhibition of KRAS and CTNNB1 may provide a potential combination therapy.

Within the inferred signaling networks, we noticed many interesting interactions between the S-gene regulators. First, in the signaling network inferred for the EMT subtype in GC (**Figure 5A**), CDH1 and ERBB2 were prioritized as potential therapeutic targets (**Table 1**). A signal flow was inferred between them, which could be explained by the direct interaction (PPI) between them (Guo et al., 2014) or their PPIs via β-catenin (CTNNB1) (Schroeder et al., 2002; Tang et al., 2008). The signal flow miR-200a→ZEB2 could be strongly supported by the previous finding that miR-200a can regulate the expression of ZEB2 by directly binding the 3′UTR (Cong et al., 2013). Furthermore, the signal flow KRAS→miR-200a was also supported by the previous finding that oncogenic KRAS activation can suppress the expression of miR-200s

(Zhong et al., 2016), and TP53→ZEB2 could be verified by their interactions with the miR-200 family (Rokavec et al., 2014). Second, in the signaling network inferred for the CMS4 subtype in CRC (**Figure 5B**), the advantages of our work were demonstrated more explicitly. The signal flow KRAS→PIK3CA→BRAF, supported by the MAP-kinase pathway (Dhillon et al., 2007), i.e., the PPIs between KRAS and PIK3CA (Hart et al., 2015) and between PIK3CA and BRAF (Shen et al., 2017), which is known as a typical signaling pathway driving EMT. The interaction between TGFBR2 and SMAD4 is involved in the TGFβ signaling pathway (Zhang et al., 1996). The signal flow CTNNB1→TGFBR2 is involved in the crosstalk between Wnt/β-catenin and TGFβ signaling pathways (Tian and Phillips, 2002). Together, the literature supports the effectiveness of NEM-Tar in predicting the regulatory hierarchy involving multiple redundant pathways driving EMT. Moreover, we also found signal flows between miRNAs, like the links miR-200a→miR-425, miR-141→miR-135b (**Figure 5A**) and miR-200a→miR-141 (**Figure 5B**), which are interesting but have not been previously reported yet. The miRNAs may interact indirectly via intermediate regulators, which were not included in the regulatory network inference based on our criteria for the selection of S-genes. The crosstalk between the miRNAs might also indicate their synergistic relationship on co-regulating downstream targets, which is frequently reported in the literature [reviewed in Xu et al. (2016)]. Integrating the computational prediction with experimental validation will be more convincing in revealing the crosstalks between the miRNAs, which will be an interesting direction to explore in our future work.

NEM-Tar can be improved in multiple ways in our future work. First, known signaling pathway structures can be incorporated into the model as prior knowledge to strengthen the accuracy of inference. Second, NEM-Tar proposed in this article is designed for binary effects and treats E-genes as independent random variables. However, we can possibly model log odds ratios like the methods in Tresch and Markowetz (2008), where alternative and null distribution are both normal, to decrease the information loss. Third, in this work, we focused on the S-genes with subtype-specificity or with functional relations reported to key pathways (e.g., MAP-kinase) or biological processes (e.g., EMT), and therefore the number of S-genes was limited. The limitation of scalability to a larger perturbation scale could be one future direction to improve our method. In our simulation study, greedy hill-climbing demonstrated high and robust performance in signaling networks with up to 30 S-genes, which meets the regular need for signaling network inference and drug targets prioritization. Many techniques may improve the performance of MCMC sampling (Andrieu et al., 2003), which warrants further exploration in our future work. Last but not least, we can also change the modeling framework radically using graph embedding based methods (Yue et al., 2020), as the observation of S-genes and E-genes are all high-dimensional vectors. However, the question of how to preserve the assumption of nested subset structures in the embedding space needs to be conquered tactfully.

In conclusion, NEM-Tar presents a useful computational framework for dissecting the regulatory architecture underlying specific cancer subtypes and prioritizing potential drug targets. With the explosive increase of high-throughput sequencing data, NEM-Tar warrants further evaluation using large-scale multi-omics data cohorts in the future.

DATA AVAILABILITY STATEMENT

The datasets analyzed during the current study are available in the TCGA repository (https://cancergenome.nih.gov/).

AUTHOR CONTRIBUTIONS

XW contributed to study concept and design. YZ and LZ contributed to data collection, analysis, and interpretation. XW contributed to critical revision of the manuscript for important intellectual content. LZ provided important advice and assistance for manuscript drafting. XW supervised the study. All authors read and approved the final manuscript.

FUNDING

This work was supported by a grant from Guangdong Basic and Applied Basic Research Foundation (Project No. 2019B030302012), a grant supported by the Young Scientists Fund of the National Natural Science Foundation of China (81802384), a grant by the Science Technology and Innovation Commission of Shenzhen Municipality (Project No. JCYJ20200109120425045), and grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 21101115, 11102317, 11103718, 11103619, R4017-18, C4041-17GF) awarded to XW.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.608042/full#supplementary-material

REFERENCES

Anastas, J. N., and Moon, R. T. (2013). WNT signalling pathways as therapeutic targets in cancer. *Nat. Rev. Cancer* 13, 11–26. doi: 10.1038/nrc3419

Anchang, B., Sadeh, M. J., Jacob, J., Tresch, A., Vlad, M. O., Oefner, P. J., et al. (2009). Modeling the temporal interplay of molecular signaling and gene expression by using dynamic nested effects models. *Proc. Natl. Acad. Sci. U. S. A.* 106, 6447–6452. doi: 10.1073/pnas.0809822106

Andrieu, C., de Freitas, N., Doucet, A., and Jordan, M. I. (2003). An Introduction to MCMC for Machine Learning. *Mach. Learn.* 50, 5–43. doi: 10.1023/a:1020281327116

Bang, Y.-J., Van Cutsem, E., Feyereislova, A., Chung, H. C., Shen, L., Sawaki, A., et al. (2010). Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): a phase 3, open-label, randomised controlled trial. *Lancet* 376, 687–697. doi: 10.1016/S0140-6736(10)61121-X

Bashashati, A., Haffari, G., Ding, J., Ha, G., Lui, K., Rosner, J., et al. (2012). DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol.* 13:R124. doi: 10.1186/gb-2012-13-12-r124

Bernards, R. (2012). A missing link in genotype-directed cancer therapy. *Cell* 151, 465–468. doi: 10.1016/j.cell.2012.10.014

Caggiari, L., Miolo, G., Buonadonna, A., Basile, D., Santeufemia, D. A., Cossu, A., et al. (2017). Characterizing metastatic HER2-positive gastric cancer at the CDH1 haplotype. *Int. J. Mol. Sci.* 19:19010047. doi: 10.3390/ijms19010047

Cardner, M., Meyer-Schaller, N., Christofori, G., and Beerenwinkel, N. (2019). Inferring signalling dynamics by integrating interventional with observational data. *Bioinformatics* 35, i577–i585. doi: 10.1093/bioinformatics/btz325

Cheng, X., Xu, X., Chen, D., Zhao, F., and Wang, W. (2019). Therapeutic potential of targeting the Wnt/β-catenin signaling pathway in colorectal cancer. *Biomed. Pharmacother.* 110, 473–481. doi: 10.1016/j.biopha.2018.11.082

Cong, N., Du, P., Zhang, A., Shen, F., Su, J., Pu, P., et al. (2013). Downregulated microRNA-200a promotes EMT and tumor growth through the wnt/β-catenin pathway by targeting the E-cadherin repressors ZEB1/ZEB2 in gastric adenocarcinoma. *Oncol. Rep.* 29, 1579–1587. doi: 10.3892/or.2013.2267

Cristescu, R., Lee, J., Nebozhyn, M., Kim, K.-M., Ting, J. C., Wong, S. S., et al. (2015). Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. *Nat. Med.* 21, 449–456. doi: 10.1038/nm.3850

Dai, Y.-H., Tang, Y.-P., Zhu, H.-Y., Lv, L., Chu, Y., Zhou, Y.-Q., et al. (2012). ZEB2 promotes the metastasis of gastric cancer and modulates epithelial mesenchymal transition of gastric cancer cells. *Dig. Dis. Sci.* 57, 1253–1260. doi: 10.1007/s10620-012-2042-6

Dhillon, A. S., Hagan, S., Rath, O., and Kolch, W. (2007). MAP kinase signalling pathways in cancer. *Oncogene* 26, 3279–3290. doi: 10.1038/sj.onc.1210421

Fessler, E., Jansen, M., De Sousa, E., Melo, F., Zhao, L., Prasetyanti, P. R., et al. (2016). A multidimensional network approach reveals microRNAs as determinants of the mesenchymal colorectal cancer subtype. *Oncogene* 35, 6026–6037. doi: 10.1038/onc.2016.134

Fröhlich, H., Praveen, P., and Tresch, A. (2011). Fast and efficient dynamic nested effects models. *Bioinformatics* 27, 238–244. doi: 10.1093/bioinformatics/btq631

Fröhlich, H., Sahin, O., Arlt, D., Bender, C., and Beissbarth, T. (2009). Deterministic Effects Propagation Networks for reconstructing protein signaling networks from multiple interventions. *BMC Bioinformatics* 10:322. doi: 10.1186/1471-2105-10-322

Gastric Adenocarcinoma (2020). *My Cancer Genome*. Available online at: https://www.mycancergenome.org/content/disease/gastric-adenocarcinoma/ (accessed September 18, 2020).

Green, M. R. (2004). Targeting targeted therapy. *N. Engl. J. Med.* 350, 2191–2193. doi: 10.1056/NEJMe048101

Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., et al. (2015). The consensus molecular subtypes of colorectal cancer. *Nat. Med.* 21, 1350–1356. doi: 10.1038/nm.3967

Guo, W.-F., Zhang, S.-W., Liu, L.-L., Liu, F., Shi, Q.-Q., Zhang, L., et al. (2018). Discovering personalized driver mutation profiles of single samples in cancer by network control strategy. *Bioinformatics* 34, 1893–1903. doi: 10.1093/bioinformatics/bty006

Guo, Z., Neilson, L. J., Zhong, H., Murray, P. S., Zanivan, S., and Zaidel-Bar, R. (2014). E-cadherin interactome complexity and robustness resolved by quantitative proteomics. *Sci. Signal.* 7:rs7. doi: 10.1126/scisignal.2005473

Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K. R., MacLeod, G., et al. (2015). High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell* 163, 1515–1526. doi: 10.1016/j.cell.2015.11.015

Hou, J. P., and Ma, J. (2014). DawnRank: discovering personalized driver genes in cancer. *Genome Med.* 6:56. doi: 10.1186/s13073-014-0056-8

Kiyozumi, Y., Iwatsuki, M., Yamashita, K., Koga, Y., Yoshida, N., and Baba, H. (2018). Update on targeted therapy and immune therapy for gastric cancer, 2018. *J. Cancer Metastasis. Treat* 4:31.

Kotiang, S., and Eslami, A. (2020). A probabilistic graphical model for system-wide analysis of gene regulatory networks. *Bioinformatics* 36, 3192–3199. doi: 10.1093/bioinformatics/btaa122

Kullback, S., and Leibler, R. A. (1951). On Information and Sufficiency. *Ann. Math. Stat.* 22, 79–86.

Lambert, A. W., Pattabiraman, D. R., and Weinberg, R. A. (2017). Emerging Biological Principles of Metastasis. *Cell* 168, 670–691. doi: 10.1016/j.cell.2016.11.037

Lin, Y., Wu, W., Sun, Z., Shen, L., and Shen, B. (2018). MiRNA-BD: an evidence-based bioinformatics model and software tool for microRNA biomarker discovery. *RNA Biol.* 15, 1093–1105. doi: 10.1080/15476286.2018.1502590

MacNeil, L. T., Pons, C., Arda, H. E., Giese, G. E., Myers, C. L., and Walhout, A. J. M. (2015). Transcription Factor Activity Mapping of a Tissue-Specific in vivo Gene Regulatory Network. *Cell Syst* 1, 152–162. doi: 10.1016/j.cels.2015.08.003

Manatakis, D. V., Raghu, V. K., and Benos, P. V. (2018). piMGM: incorporating multi-source priors in mixed graphical models for learning disease networks. *Bioinformatics* 34, i848–i856. doi: 10.1093/bioinformatics/bty591

Markowetz, F., Bloch, J., and Spang, R. (2005). Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. *Bioinformatics* 21, 4026–4032. doi: 10.1093/bioinformatics/bti662

Markowetz, F., Kostka, D., Troyanskaya, O. G., and Spang, R. (2007). Nested effects models for high-dimensional phenotyping screens. *Bioinformatics* 23, i305–i312. doi: 10.1093/bioinformatics/btm178

Markowetz, F., and Spang, R. (2007). Inferring cellular networks–a review. *BMC Bioinformatics* 8:S5. doi: 10.1186/1471-2105-8-S6-S5

Mezlini, A. M., and Goldenberg, A. (2017). Incorporating networks in a probabilistic graphical model to find drivers for complex human diseases. *PLoS Comput. Biol.* 13:e1005580. doi: 10.1371/journal.pcbi.1005580

Niederberger, T., Etzold, S., Lidschreiber, M., Maier, K. C., Martin, D. E., Fröhlich, H., et al. (2012). MC EMiNEM maps the interaction landscape of the Mediator. *PLoS Comput. Biol.* 8:e1002568. doi: 10.1371/journal.pcbi.1002568

Nieto, M. A., Huang, R. Y.-J., Jackson, R. A., and Thiery, J. P. (2016). EMT: 2016. *Cell* 166, 21–45. doi: 10.1016/j.cell.2016.06.028

Pham, V. V. H., Liu, L., Bracken, C. P., Goodall, G. J., Long, Q., Li, J., et al. (2019). CBNA: A control theory based method for identifying coding and non-coding cancer drivers. *PLoS Comput. Biol.* 15:e1007538. doi: 10.1371/journal.pcbi.1007538

Pirkl, M., and Beerenwinkel, N. (2018). Single cell network analysis with a mixture of Nested Effects Models. *Bioinformatics* 34, i964–i971. doi: 10.1093/bioinformatics/bty602

Pirkl, M., Diekmann, M., van der Wees, M., Beerenwinkel, N., Fröhlich, H., and Markowetz, F. (2017). Inferring modulators of genetic interactions with epistatic nested effects models. *PLoS Comput. Biol.* 13:e1005496. doi: 10.1371/journal.pcbi.1005496

Pirkl, M., Hand, E., Kube, D., and Spang, R. (2016). Analyzing synergistic and non-synergistic interactions in signalling pathways using Boolean Nested Effect Models. *Bioinformatics* 32, 893–900. doi: 10.1093/bioinformatics/btv680

Polyak, K., and Garber, J. (2011). Targeting the missing links for cancer therapy. *Nat. Med.* 17, 283–284. doi: 10.1038/nm0311-283

Porru, M., Pompili, L., Caruso, C., Biroccio, A., and Leonetti, C. (2018). Targeting KRAS in metastatic colorectal cancer: current strategies and emerging opportunities. *J. Exp. Clin. Cancer Res.* 37, 719–711. doi: 10.1186/s13046-018-0719-1

Rokavec, M., Li, H., Jiang, L., and Hermeking, H. (2014). The p53/microRNA connection in gastrointestinal cancer. *Clin. Exp. Gastroenterol.* 7, 395–413. doi: 10.2147/CEG.S43738

Russell, S., and Norvig, P. (2016). *Artificial intelligence: A modern approach, global edition*, 3rd Edn. London: Pearson Education.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 308, 523–529. doi: 10.1126/science.1105809

Schroeder, J. A., Adriance, M. C., McConnell, E. J., Thompson, M. C., Pockaj, B., and Gendler, S. J. (2002). ErbB-beta-catenin complexes are associated with human infiltrating ductal breast and murine mammary tumor virus (MMTV)-Wnt-1 and MMTV-c-Neu transgenic carcinomas. *J. Biol. Chem.* 277, 22692–22698. doi: 10.1074/jbc.M201975200

Shen, J. P., Zhao, D., Sasik, R., Luebeck, J., Birmingham, A., Bojorquez-Gomez, A., et al. (2017). Combinatorial CRISPR-Cas9 screens for de novo mapping of genetic interactions. *Nat. Methods* 14, 573–576. doi: 10.1038/nmeth.4225

Siebourg-Polster, J., Mudrak, D., Emmenlauer, M., Rämö, P., Dehio, C., Greber, U., et al. (2015). NEMix: single-cell nested effects models for probabilistic pathway stimulation. *PLoS Comput. Biol.* 11:e1004078. doi: 10.1371/journal.pcbi.1004078

Tang, Y., Liu, Z., Zhao, L., Clemens, T. L., and Cao, X. (2008). Smad7 stabilizes beta-catenin binding to E-cadherin complex and promotes cell-cell adhesion. *J. Biol. Chem.* 283, 23956–23963. doi: 10.1074/jbc.M800351200

Tian, Y. C., and Phillips, A. O. (2002). Interaction between the transforming growth factor-beta type II receptor/Smad pathway and beta-catenin during transforming growth factor-beta1-mediated adherens junction disassembly. *Am. J. Pathol.* 160, 1619–1628. doi: 10.1016/s0002-9440(10)61109-1

Tresch, A., and Markowetz, F. (2008). Structure learning in Nested Effects Models. *Stat. Appl. Genet. Mol. Biol.* 7:Article9. doi: 10.2202/1544-6115.1332

Wang, X., Yuan, K., Hellmayr, C., Liu, W., and Markowetz, F. (2014). Reconstructing evolving signalling networks by hidden Markov nested effects models. *Ann. Appl. Stat.* 8, 448–480. doi: 10.1214/13-AOAS696

Xie, Y.-H., Chen, Y.-X., and Fang, J.-Y. (2020). Comprehensive review of targeted therapy for colorectal cancer. *Signal Transduct. Target. Ther.* 5:22. doi: 10.1038/s41392-020-0116-z

Xu, J., Shao, T., Ding, N., Li, Y., and Li, X. (2016). miRNA–miRNA crosstalk: from genomics to phenomics. *Brief. Bioinform.* 2016:bbw073. doi: 10.1093/bib/bbw073

Yamaguchi, H., Chang, S.-S., Hsu, J. L., and Hung, M.-C. (2014). Signaling cross-talk in the resistance to HER family receptor targeted therapy. *Oncogene* 33, 1073–1081. doi: 10.1038/onc.2013.74

Yoon, C., Till, J., Cho, S.-J., Chang, K. K., Lin, J.-X., Huang, C.-M., et al. (2019). KRAS activation in gastric adenocarcinoma stimulates epithelial-to-mesenchymal transition to cancer stem-like cells and promotes metastasis. *Mol. Cancer Res.* 17, 1945–1957. doi: 10.1158/1541-7786.MCR-19-0077

Yu, C., Qi, X., Lin, Y., Li, Y., and Shen, B. (2020). iODA: An integrated tool for analysis of cancer pathway consistency from heterogeneous multi-omics data. *J. Biomed. Inform.* 112, 103605. doi: 10.1016/j.jbi.2020.103605

Yue, X., Wang, Z., Huang, J., Parthasarathy, S., Moosavinasab, S., Huang, Y., et al. (2020). Graph embedding on biomedical networks: methods, applications and evaluations. *Bioinformatics* 36, 1241–1251. doi: 10.1093/bioinformatics/btz718

Zhang, Y., Feng, X., We, R., and Derynck, R. (1996). Receptor-associated Mad homologues synergize as effectors of the TGF-beta response. *Nature* 383, 168–172. doi: 10.1038/383168a0

Zhao, M., Liu, Y., Zheng, C., and Qu, H. (2019). dbEMT 2.0: An updated database for epithelial-mesenchymal transition genes with experimentally verified information and precalculated regulation information for cancer metastasis. *J. Genet. Genomics* 46, 595–597. doi: 10.1016/j.jgg.2019.11.010

Zhong, X., Zheng, L., Shen, J., Zhang, D., Xiong, M., Zhang, Y., et al. (2016). Suppression of MicroRNA 200 family expression by oncogenic KRAS activation promotes cell survival and epithelial-mesenchymal transition in KRAS-driven cancer. *Mol. Cell. Biol.* 36, 2742–2754. doi: 10.1128/mcb.00079-16

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.