Chang et al. BMC Bioinformatics (2020) 21:101
https://doi.org/10.1186/s12859-020-3444-7

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Network hub-node prioritization of gene regulation with intra-network association

Hung-Ching Chang[1], Chiao-Pei Chu[1], Shu-Ju Lin[2] and Chuhsing Kate Hsiao[1,3*]

Abstract

Background: To identify and prioritize the influential hub genes in a gene-set or biological pathway, most analyses rely on calculation of marginal effects or tests of statistical significance. These procedures may be inappropriate since hub nodes are common connection points and therefore may interact with other nodes more often than non-hub nodes do. Such dependence among gene nodes can be conjectured based on the topology of the pathway network or the correlation between them.

Results: Here we develop a pathway activity score incorporating the marginal (local) effects of gene nodes as well as intra-network affinity measures. This score summarizes the expression levels in a gene-set/pathway for each sample, with weights on local and network information, respectively. The score is next used to examine the impact of each node through a leave-one-out evaluation. To illustrate the procedure, two cancer studies, one involving RNA-Seq from breast cancer patients with high-grade ductal carcinoma in situ and one microarray expression data from ovarian cancer patients, are used to assess the performance of the procedure, and to compare with existing methods, both ones that do and do not take into consideration correlation and network information. The hub nodes identified by the proposed procedure in the two cancer studies are known influential genes; some have been included in standard treatments and some are currently considered in clinical trials for target therapy. The results from simulation studies show that when marginal effects are mild or weak, the proposed procedure can still identify causal nodes, whereas methods relying only on marginal effect size cannot.

Conclusions: The NetworkHub procedure proposed in this research can effectively utilize the network information in combination with local effects derived from marker values, and provide a useful and complementary list of recommendations for prioritizing causal hubs.

Keywords: Direction regularization, Intra-network, Neighbor correlation, Pathway activity score, Topology measure

Background

In a disease-associated biological pathway containing nodes such as genes, proteins and other chemical compounds, the detection of nodes that are crucial to this pathway activity may help elucidate the molecular mechanism influencing the response of interest. The prioritization of nodes in this association pathway may provide useful information for follow-up experimental validation and thereby facilitate the search for drug target molecules [1, 2]. In the medical genetics community in recent decades, the prioritization of a set of genes has been an important issue, especially when the research focus is on identification of genes for drug discovery [3, 4].

Current methodology for such prioritization can be grouped into three categories, depending on what information is utilized in the procedures of modeling and ranking. The first group is more traditional; these methods rank the genes in a previously identified gene-set or pathway based

* Correspondence: ckhsiao@ntu.edu.tw
[1]Division of Biostatistics, Institute of Epidemiology and Preventive Medicine, National Taiwan University, No. 17, Xu-Zhou Road, Taipei 10055, Taiwan
[3]Bioinformatics and Biostatistics Core, Center of Genomic Medicine, National Taiwan University, Taipei 10055, Taiwan
Full list of author information is available at the end of the article

only on marginal marker values, such as the fold-change, the t-statistic, or a function of the t-statistic, if gene expression data are involved. These methods are straightforward to carry out and can quite effectively reduce the enormous number of genetic markers to one that is easier to handle. However, these methods may not be proper with the underlying assumption that these genes are independent, as noted in many gene-set analyses [5–9]. Through correlation sharing or through de-correlation such as in the shrinkage correlation-adjusted-t (shrinkage cat) score [10], several modifications to this group of tests have been proposed. Another inappropriate assumption is that of exchangeability, which assumes all candidate genes are equally important a priori, regardless of whether they are hub genes or not. This assumption ignores the fact that some genes, such as hubs, interact with relatively more other genes.

Unlike the above methods which use only expression data in their prioritization analysis, the second group of approaches utilizes information from networks that are established based on research findings from the literature and data from multiple sources to define *seed genes*, calculates similarities between the seed and each of the candidate genes, and then ranks the candidate genes based on the similarities. This concept is usually called guilt-by-association. Examples include Endeavour, ToppGene and GeneDistiller [11–13]. These methods depend heavily on the information available and would be helpful for investigating reproducibility. However, they assume the network to be static, which may inflate false positive results, and some of them do not account for the marginal effects [14].

The third group of analyses integrates both the gene expression data regarding the phenotype of interest and the network information. For instance, PINTA [15, 16] prioritizes candidate genes by combining the protein association network from STRING [17], the disease expression data, and a kernel for distance. The network gene prioritizer (NGP) proposed by Wu et al. [14] uses the correlation under the network rewiring (NR) model or networked differential expression (ND) model to construct a network for each candidate gene, and then carries out gene set enrichment analysis (GSEA) [18] to determine the importance of the genes, called NGP-NR and NGP-ND respectively. These approaches incorporate the network information through correlations. In contrast, with pathways defined in KEGG [19], Lin et al. [20] included expression levels, correlations and degree of the nodes in their Bayesian probabilistic prioritization procedure. The degree they included in their method provides the information about the pathway/network structure and topology.

When a pathway is represented as a network, the structure may provide useful information. For instance, in the network plot, any two nodes having a direct molecular interaction, whether inhibitory or activating, are immediate neighbors of each other; while nodes that have not been found to interact will not be directly connected in the pathway network. A node with a large number of neighbors is considered as a hub gene node. A hub may be more crucial than other nodes, since its absence can disrupt the pathway function and isolate other gene nodes [14, 21–25]. The number of neighboring nodes of a gene is called its degree, which can be derived from the adjacency matrix of the network [26]. In other words, the observed degree distribution or the adjacency matrix of this network can provide a description of the structure of the pathway. Such an affinity measure for the network topology may offer useful information in ranking gene nodes and in summarizing the collective enrichment of a pathway. Similar ideas have been adopted in testing the association of a network/pathway, but not in ranking gene nodes [27–29]. Although these topology-based methods include information about the pathway structure, such an affinity measure has not been utilized in identifying influential hub gene nodes.

In this analysis, we propose to construct the intra-network information that each node carries in the pathway, calculate the local effect from the marginal influence, weight the local and network information separately, and combine them to formulate a pathway activity score. The intra-network information is composed of the pairwise affinity measures of correlation and distance. Based on the adjacency matrix of the network, the shortest distance between two nodes is called the path length, where the length can represent the efficiency of information transmission in this network. The resulting pathway activity score for each sample is next used to prioritize the gene nodes, particularly the hub nodes. We call the method the NetworkHub procedure (Fig. 1).

The rest of the article is organized as follows. The methodology and the underlying rationale will be explained in the Methods section. In Results, applications and simulation studies are conducted to evaluate the performance of the proposed procedure and to compare it with other existing methods, including ones that prioritize with and without correlation, such as the shrinkage cat and t tests, and ones that prioritize with and without network information, such as Endeavour, PINTA, NGP, and Lin's method. We then conclude with the Discussion section.

Results

Breast cancer with DCIS (RNA-Seq)

The first application considered for illustration is a study of high-grade ductal carcinoma in situ (DCIS), a subtype of breast cancer [30]. This study collected RNA-Seq data from 10 unaffected subjects and 25 breast cancer patients, which can be downloaded from the NCBI GEO database (accession number GSE69240). The downloadable data set was examined with procedures for quality

control and normalization, as described in the original report of the study [30]. Four pathways defined in KEGG [19], P53, mTor, Estrogen, and JAK-STAT, are selected here as networks for demonstration. These pathways have been reported to associate with breast cancer [31], and have passed the global test [32], GSEA [18], Fisher's test, SPIA [33] and the Bayesian association test in Lin et al. [20]. For each pathway, the plots of L_j and S_j against path length in Figure S1 show that the local weight of each node neither associates with nor reflects the magnitude of its degree, whereas the topology weight does increase slightly with the degree of the node. The NetworkHub procedure is used to rank the gene nodes inside each pathway network, respectively, with the leave-one-out evaluation.

For the P53 pathway, its network plot is shown in Fig. 2a. In the scree plot in Fig. 2b, the gene nodes on the X-axis are ordered according to their importance with bold red font used to indicate hub nodes, while the grey bars represent the magnitude of negative log-*p*-values. Several interesting findings are indicated when the NetworkHub procedure is implemented. First, note that the top-ranking node *IGF1* in this pathway indeed is known to have cross-talk with estrogens; anti-estrogens such as tamoxifen have served as a routine treatment for breast cancer in many countries [34]. The use of IGF1R inhibitors as molecule targets has been considered in several recent clinical trials, including a phase Ib/II trial (NCT02123823) of the drug BI 836845 and a phase II study of BMS-754807 combined with letrozole (NCT01225172). More reviews can be found in [34]. This top-ranking gene node has a significant marginal effect and therefore it is not surprising that it has been included in several clinical investigations. This gene node is, however, overlooked when some other methods are applied. For instance, as seen in Fig. 2c, *IGF1* is not in the top 25% under Endeavour, PINTA, NGP-ND and NGP-NR; but it ranks second under the shrinkage cat and tenth under the shrinkage t methods. In Fig. 2d, an alternative

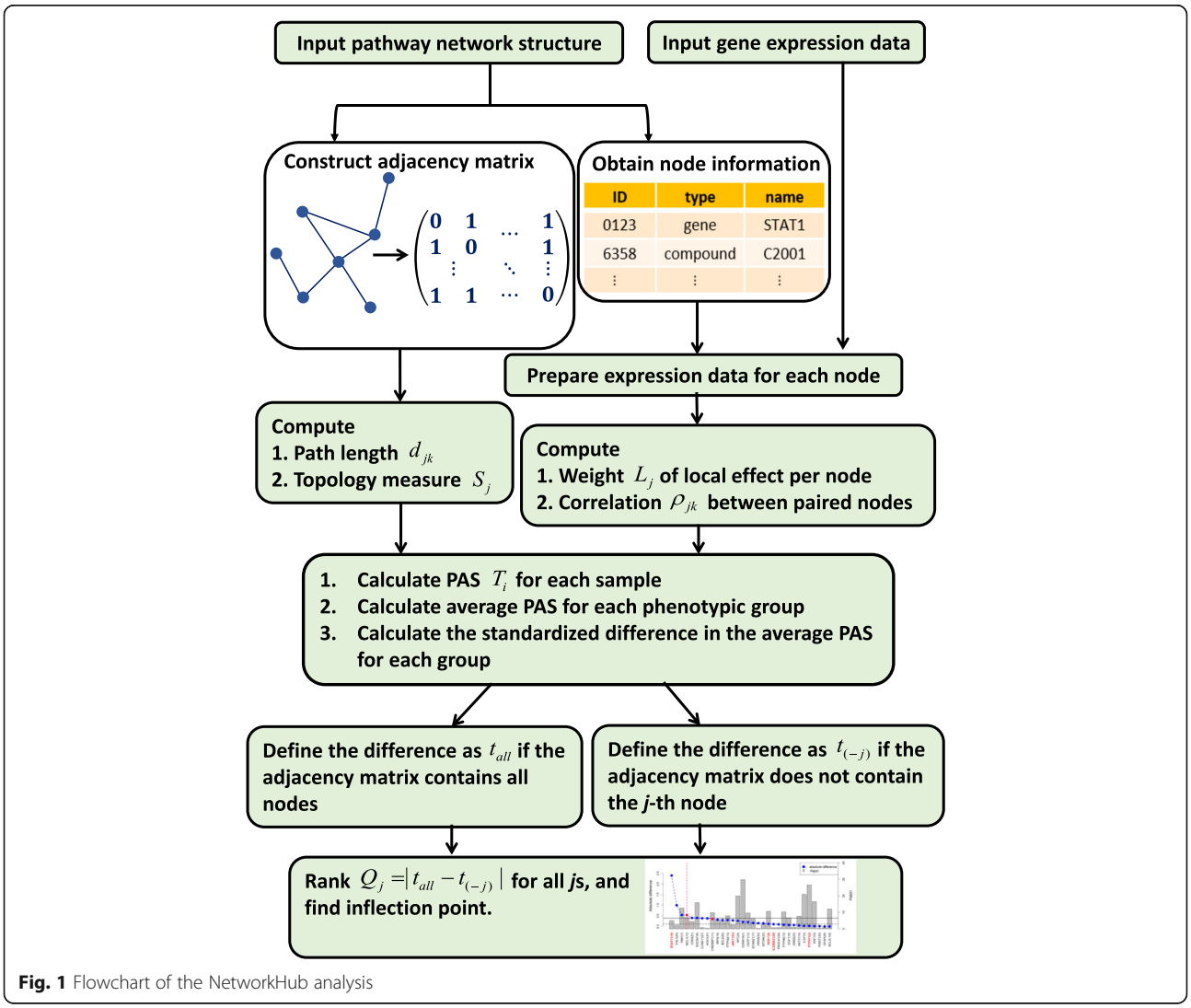
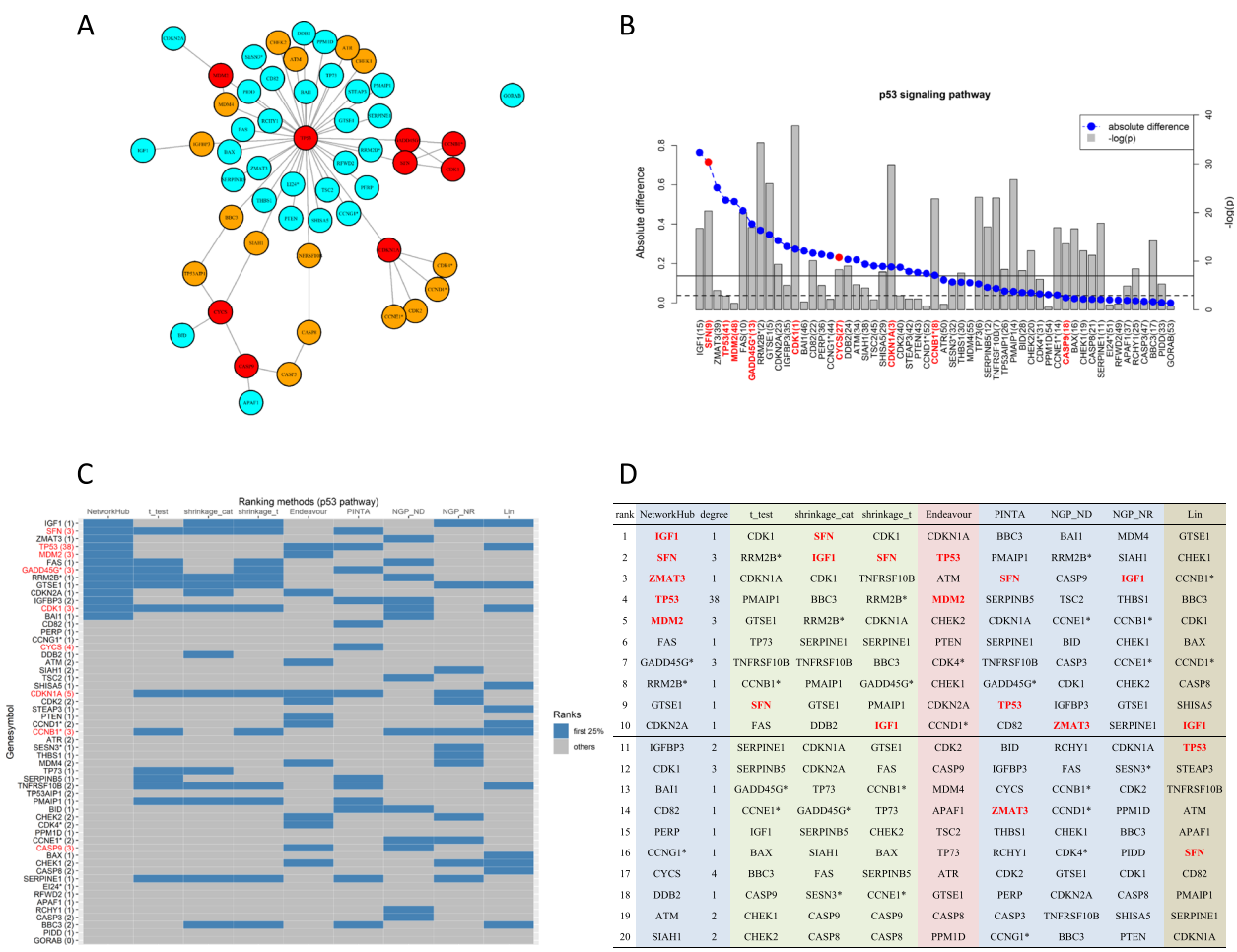


Fig. 1 Flowchart of the NetworkHub analysis



rank	NetworkHub	degree	t_test	shrinkage_cat	shrinkage_t	Endeavour	PINTA	NGP_ND	NGP_NR	Lin
1	**IGF1**	1	CDK1	**SFN**	CDK1	CDKN1A	BBC3	BAI1	MDM4	GTSE1
2	**SFN**	3	RRM2B*	**IGF1**	**SFN**	**TP53**	PMAIP1	RRM2B*	SIAH1	CHEK1
3	**ZMAT3**	1	CDKN1A	CDK1	TNFRSF10B	ATM	**SFN**	CASP9	**IGF1**	CCNB1*
4	**TP53**	38	PMAIP1	BBC3	RRM2B*	**MDM2**	SERPINB5	TSC2	THBS1	BBC3
5	**MDM2**	3	GTSE1	RRM2B*	CDKN1A	CHEK2	CDKN1A	CCNE1*	CCNB1*	CDK1
6	FAS	1	TP73	SERPINE1	SERPINE1	PTEN	SERPINE1	BID	CHEK1	BAX
7	GADD45G*	3	TNFRSF10B	TNFRSF10B	BBC3	CDK4*	TNFRSF10B	CASP3	CCNE1*	CCND1*
8	RRM2B*	1	CCNB1*	PMAIP1	GADD45G*	CHEK1	GADD45G*	CDK1	CHEK2	CASP8
9	GTSE1	1	**SFN**	GTSE1	PMAIP1	CDKN2A	**TP53**	IGFBP3	GTSE1	SHISA5
10	CDKN2A	1	FAS	DDB2	**IGF1**	CCND1*	CD82	**ZMAT3**	SERPINE1	**IGF1**
11	IGFBP3	2	SERPINE1	CDKN1A	GTSE1	CDK2	BID	RCHY1	CDKN1A	**TP53**
12	CDK1	3	SERPINB5	CDKN2A	FAS	CASP9	IGFBP3	FAS	SESN3*	STEAP3
13	BAI1	1	GADD45G*	TP73	CCNB1*	MDM4	CYCS	CCNB1*	CDK2	TNFRSF10B
14	CD82	1	CCNE1*	GADD45G*	TP73	APAF1	**ZMAT3**	CCND1*	PPM1D	ATM
15	PERP	1	IGF1	SERPINB5	CHEK2	TSC2	THBS1	CHEK1	BBC3	APAF1
16	CCNG1*	1	BAX	SIAH1	BAX	TP73	RCHY1	CDK4*	PIDD	**SFN**
17	CYCS	4	BBC3	FAS	SERPINB5	ATR	CDK2	GTSE1	CDK1	CD82
18	DDB2	1	CASP9	SESN3*	CCNE1*	GTSE1	PERP	CDKN2A	CASP8	PMAIP1
19	ATM	2	CHEK1	CASP9	CASP9	CASP8	CASP3	TNFRSF10B	SHISA5	SERPINE1
20	SIAH1	2	CHEK2	CASP8	CASP8	PPM1D	CCNG1*	BBC3	PTEN	CDKN1A

Fig. 2 P53 pathway of the breast cancer study. **a** Network plot. Nodes in red are hubs with degree ≥3, nodes in yellow are of degree 2, and nodes in light blue are of degree ≤1. **b** Scree plot (blue dots) of the relative influence of gene nodes (left Y-axis). The right Y-axis is the negative-logarithm-*p*-value (grey bar). The dashed red line corresponds to $p = 0.05$, and the solid red line to 0.05 divided by the number of nodes inside the pathway. The X-axis label indicates the gene node. A name followed by an * indicates more than one gene is contained inside this node. The number in parentheses is the ranking determined by *p*-value. **c** Heatmap for comparing rankings under different methods. Gene nodes ranked among the first 25% are colored in grey. The nodes considered hubs are denoted in red. The number in parentheses after the gene symbol indicates the degree of the node. **d** Ranking list

representation lists the top 20 gene nodes under each method. The complete list is in Supplementary Table S1.

A similar pattern can be observed in the gene that ranks second, the *SFN* gene. Its marginal effect is statistically significant, as is identified by most t-statistic type methods and PINTA (Fig. 2b-d). Previous studies have reported its ability to increase cell death in breast cancer lines and found that its hypermethylation is related to silencing of the 14–3-3σ protein in epithelial breast cancer tumors [35–37]. Its property as a hub node (Fig. 2a & b) corroborates the importance of the *SFN* gene.

The next ranking gene, *ZMAT3*, also known as *Wig-1*, is identified by NetworkHub as the third-ranking gene node and by NGP-ND as the tenth (Fig. 2d). However, it is not marginally significant and not selected in the top ten by other methods (Fig. 2b). Nevertheless, the chromosome region where it is located is amplified in many tumors including breast cancer [38]. It has been reported to be a direct target of *TP53* [39], to be associated with other targets of *TP53*, such as *FAS* and 14–3-3σ protein, and to regulate the mRNA stability of *TP53* [39, 40].

Another gene worth mentioning is the well-known tumor suppressor *TP53*, the guardian of the genome [41, 42]. Since its discovery in 1979, many studies have been devoted to the investigation of its germline/somatic mutations, its sequence context, and its impact on and association with human cancers [43, 44]. Its role in the etiology of breast cancer is beyond doubt, and yet it was not identified in the top 20 under most methods (Fig. 2c and d) due to its non-differentiability in gene expression levels. The only methods that did identify it this highly were NetworkHub as fourth, Endeavour as second, PINTA as ninth

and Lin's method as ninth. It was ranked fourth by NetworkHub because it is a hub node connecting to most gene nodes in this pathway.

The top-ranked nodes for the other three pathways, mTor, Estrogen, and JAK-STAT, are *IGF1R-INSR*, *ESR*-node (denoted as *ESR**), and *STAT*-node (*STAT**), respectively. In the mTor pathway, *IGF1R-INSR* has similar roles as top-ranked *IGF1* does in the P53 pathway (details in supplementary Figure S2 and Table S2). The top-ranking node *ESR1** in the Estrogen signaling pathway contains the *ESR1* gene (Figure S3). Research indicates that ESR1 mutations emerge during both metastatic breast cancer treatment and tumor evolution, and thus the need for a better-personalized treatment with aromatase inhibitor therapy that sequentially monitors and targets ESR1 mutations has been suggested [45]. *ESR1** also ranks among the top 25% with Lin's (second), NGP-NR (seventh), Endeavor (first), and the t-test (ninth) methods (Figure S3 and Table S3).

The top-ranking node, *STAT*-node, in the JAK-STAT pathway contains the family of STAT genes (Figure S4), which are important for mammary cell survival and tumorigenesis [46]. The expression levels of these genes are associated with breast cancer subtypes and gene *STAT1* is known to transmit information from extracellular signals to the cell nucleus [47]. Interrupted or dysregulated function of *STAT1-STAT3* can cause immune deficiency or development of cancer [48]. Recently, suggestions have been made for using the anti-psychotic drug pimozide to inhibit *STAT3* and *STAT5* in breast cancer patients [49, 50]. The degree of *STAT*-node is 17 and thus it is clearly a hub node. The NGP-ND and NGP-NR methods also recognize this property, ranking it among the top 25%. NGP-ND ranks it fourth and NGP-NR third (Figure S4 and Table S4), whereas with the t-test it ranks 19 out of 32 nodes and is not statistically significant.

Epithelial ovarian carcinoma (array expression)

To demonstrate the utility of the approach using microarray expression levels, we consider here an ovarian serous cystadenocarcinoma study with data retrieved from The Cancer Genome Atlas (TCGA). After data processing and management (quality control, outlier detection, and normalization) and filtering with clinical information (tumor type, cancer stage, and ethnic group), 282 patients with complete node expression values were obtained. Among them, seventy-two did not survive over two years and were categorized in the poor prognosis group.

The same four pathways were examined and only the mTor pathway passed the global, Fisher's, and t_{all} gene-set association tests (Table 1). This pathway also has the highest posterior probability of association among the four, based on the Bayesian approach in Lin et al. [20]. The gene nodes in this pathway were then examined and ranked, as displayed in Fig. 3a and b. The top-ranking gene node is glycogen synthase kinase 3 beta (*GSK3B*), also a hub node, which functions in cellular processes such as proliferation and survival. This gene has been associated with drug resistance in cancer chemotherapy [51]. Higher levels of *GSK3B* are often observed in tumor tissues and overexpression of *GSK3B* can enhance tumorigenicity [52, 53]. An ongoing clinical trial (NCT03678883) is testing a GSK3B inhibitor, 9-ING-41, for treating patients with advanced cancers, including ovarian cancer.

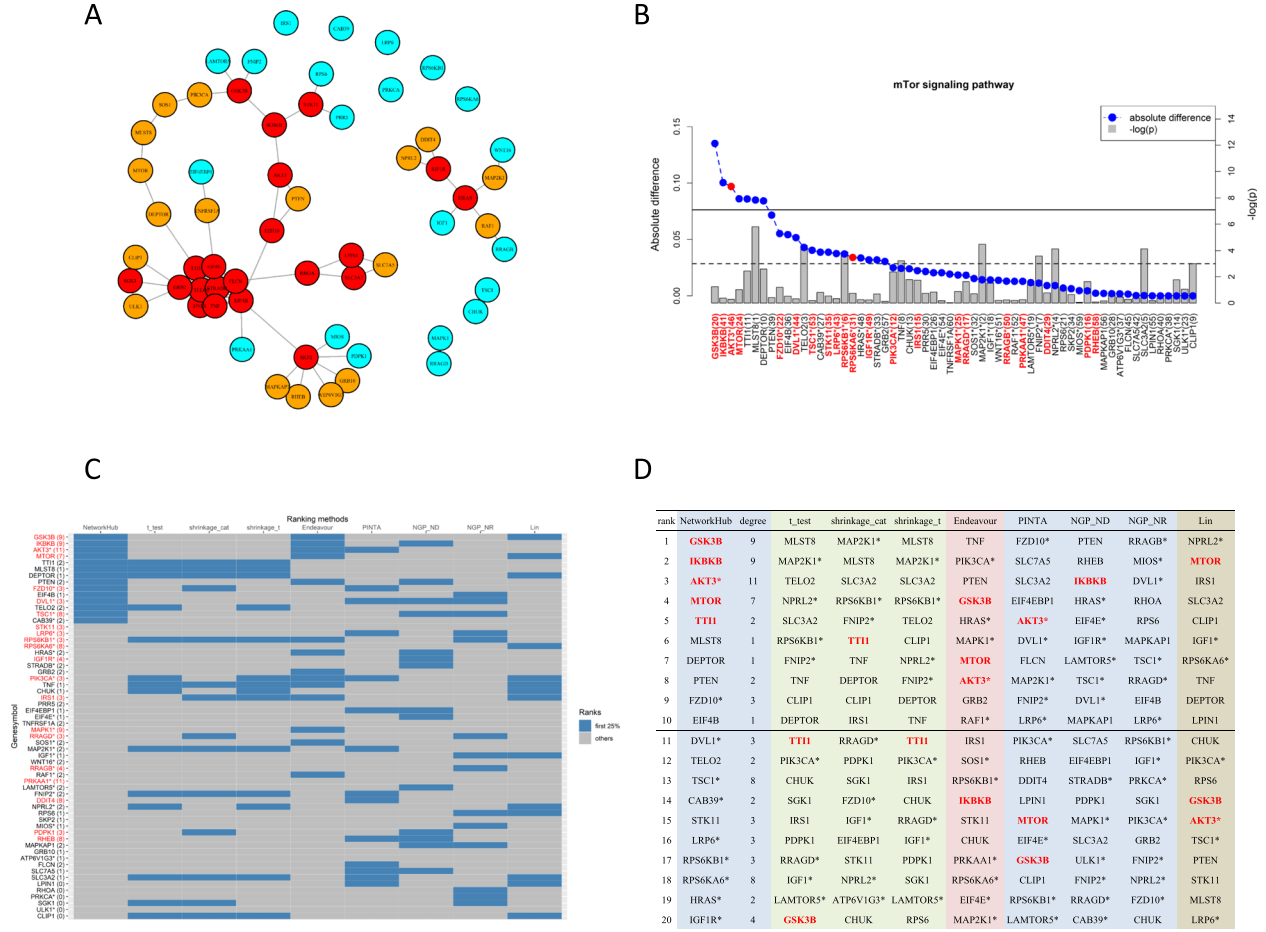
Table 1 *P*-values under the first five gene-set analyses and posterior probability under the Bayesian approach for the ovarian cancer study

	Network-guided	Global	GSEA	Fisher's	SPIA	Bayesian
JAK-STAT	0.76	0.24	0.69	0.07	0.88	0.67
P53	0.53	0.33	0.24	0.36	0.41	0.71
mTor	0.01	0.003	0.66	0.01	0.54	0.79
Estrogen	0.73	0.24	0.43	0.18	0.98	0.73

When compared with other methods, Lin's and Endeavour also rank *GSK3B* among the top 25% (Fig. 3c); the ranks are 14 by Lin's method and 4 by Endeavour (Fig. 3d). However, the ranks obtained by t-test, shrinkage t, shrinkage cat, PINTA, NGP-ND, and NGP-NR are not high, possibly due to the fact that *GSK3B* is not marginally significant. Similar patterns can be seen in the following three gene nodes, *IKBKB*, *AKT3** and *MTOR*, which are all hub nodes and rank higher, but are not differentially expressed genes (more details are in Table S5).

Simulation studies for ranking hub gene nodes

The following simulation studies were designed to evaluate the performance of NetworkHub in ranking the hub gene nodes. The structure of JAK-STAT pathway containing 32 nodes was considered as a prototype of the network, where two hub nodes and eight non-hub nodes were chosen as causal genes of different effect sizes. These causal nodes were then examined to see how well the procedure could prioritize them. In order to maintain the inherent biological relationship among gene nodes, sample individuals were randomly selected from the above ovarian cancer study subjects and their corresponding array expression levels were included for analysis and for generating disease status via a logit link function $\text{logit}(p_i) = \beta_0 + \sum_{j=1}^{10} \beta_j G_{ij}$, where p_i is the disease probability of the binary disease status of this i-th subject. The effect sizes β_1, ..., β_{10} under each of the four scenarios (A-D) are displayed in Table 2, including hub nodes with strong ($\beta_j = 1$), moderate ($\beta_j = 0.5$), or weak effect ($\beta_j = 0.1$). Under each scenario, 1000 replications were performed. In each replication, the number of subjects was 100 with 50 cases and 50 controls. Once the



rank	NetworkHub	degree	t_test	shrinkage_cat	shrinkage_t	Endeavour	PINTA	NGP_ND	NGP_NR	Lin
1	GSK3B	9	MLST8	MAP2K1*	MLST8	TNF	FZD10*	PTEN	RRAGB*	NPRL2*
2	IKBKB	9	MAP2K1*	MLST8	MAP2K1*	PIK3CA*	SLC7A5	RHEB	MIOS*	MTOR
3	AKT3*	11	TELO2	SLC3A2	SLC3A2	PTEN	SLC3A2	IKBKB	DVL1*	IRS1
4	MTOR	7	NPRL2*	RPS6KB1*	RPS6KB1*	GSK3B	EIF4EBP1	HRAS*	RHOA	SLC3A2
5	TTI1	2	SLC3A2	FNIP2*	TELO2	HRAS*	AKT3*	EIF4E*	RPS6	CLIP1
6	MLST8	1	RPS6KB1*	TTI1	CLIP1	MAPK1*	DVL1*	IGF1R*	MAPKAP1	IGF1*
7	DEPTOR	1	FNIP2*	TNF	NPRL2*	MTOR	FLCN	LAMTOR5*	TSC1*	RPS6KA6*
8	PTEN	2	TNF	DEPTOR	FNIP2*	AKT3*	MAP2K1*	TSC1*	RRAGD*	TNF
9	FZD10*	3	CLIP1	CLIP1	DEPTOR	GRB2	FNIP2*	DVL1*	EIF4B	DEPTOR
10	EIF4B	1	DEPTOR	IRS1	TNF	RAF1*	LRP6*	MAPKAP1	LRP6*	LPIN1
11	DVL1*	3	TTI1	RRAGD*	TTI1	IRS1	PIK3CA*	SLC7A5	RPS6KB1*	CHUK
12	TELO2	2	PIK3CA*	PDPK1	PIK3CA*	SOS1*	RHEB	EIF4EBP1	IGF1*	PIK3CA*
13	TSC1*	8	CHUK	SGK1	IRS1	RPS6KB1*	DDIT4	STRADB*	PRKCA*	RPS6
14	CAB39*	2	SGK1	FZD10*	CHUK	IKBKB	LPIN1	PDPK1	SGK1	GSK3B
15	STK11	3	IRS1	IGF1*	RRAGD*	STK11	MTOR	MAPK1*	PIK3CA*	AKT3*
16	LRP6*	3	PDPK1	EIF4EBP1	IGF1*	CHUK	EIF4E*	SLC3A2	GRB2	TSC1*
17	RPS6KB1*	3	RRAGD*	STK11	PDPK1	PRKAA1*	GSK3B	ULK1*	FNIP2*	PTEN
18	RPS6KA6*	8	IGF1*	NPRL2*	SGK1	RPS6KA6*	CLIP1	FNIP2*	NPRL2*	STK11
19	HRAS*	2	LAMTOR5*	ATP6V1G3*	LAMTOR5*	EIF4E*	RPS6KB1*	RRAGD*	FZD10*	MLST8
20	IGF1R*	4	GSK3B	CHUK	RPS6	MAP2K1*	LAMTOR5*	CAB39*	CHUK	LRP6*

Fig. 3 The mTor pathway for the epithelial ovarian carcinoma study. **a** Network plot. Nodes in red are hubs with degree ≥3, nodes in yellow are of degree 2, and nodes in light blue are of degree ≤1. **b** Scree plot (blue dots) of the relative influence of gene nodes (left Y-axis). The right Y-axis is the negative-logarithm-*p*-value (grey bar). The dashed red line corresponds to p = 0.05, and the solid red line to 0.05 divided by the number of nodes inside the pathway. The X-axis label indicates the gene node. A name followed by an * indicates more than one gene is contained inside this node. The number in parenthesis is the ranking determined by *p*-value. **c** Heatmap for comparing rankings under different methods. Gene nodes ranked among the first 25% are colored in grey. The nodes considered hubs are denoted in red. The number in parentheses after the gene symbol indicates the degree of the node. **d** Ranking list

disease status and expression values were available, the NetworkHub procedures were carried out. Since most prioritization methods require literature mining and are not suitable for simulation studies, here we can only compare NetworkHub with three simple methods, the t-test, shrinkage cat, and shrinkage t [54]. This comparison can demonstrate the advantages when network information is included, but does not allow comparison with other network-based methods described in earlier sections.

Two criteria were considered for performance evaluation. The first one focuses on the ability to detect causal hub nodes, the hub ranking rate (HRR). This rate is the proportion of causal hub nodes whose rankings are less than x among the top x gene nodes: HRR = {no. of [rank(hub) ≤ x]} /x, where x is a predetermined decision point representing the number of influential nodes. Figure 4a shows the proportions under Scenarios A-D when $x = 3$. With the inclusion of network information, the resulting HRR is higher than the HRR obtained under other methods that do not allow inclusion of network information. This advantage is consistent across four scenarios. When no node is causal (Scenario D), the proposed method still selects the hub nodes prior to other nodes,

Table 2 Number and effect size of the causal hub nodes and causal non-hub nodes under each scenario

Scenarios	Causal nodes			
	Hub nodes		Non-hub nodes	
	number	effect size	number	effect size
A	2	strong ($\beta_j = 1$)	8	weak ($\beta_j = 0.1$)
B	2	moderate ($\beta_j = 0.5$)	8	weak ($\beta_j = 0.1$)
C	2	weak ($\beta_j = 0.1$)	8	weak ($\beta_j = 0.1$)
D	0	null ($\beta_j = 0$)	0	null ($\beta_j = 0$)

which is expected because it is the network information that now dominates. The other three methods do not have the same tendency; their HRRs are between 16 and 19%, close to the rate expected by chance only (18.1%). Furthermore, when x ranges between 1 and 32, NetworkHub remains advantageous for Scenarios A-C, respectively, for strong, mild, and weak effects (Fig. 4b).

Alternatively, we focus on causal nodes including both hubs and non-hubs and evaluate the false discovery rate for these causal nodes (FDR) and the true detection rate for causal nodes (TDR) among the leading x gene nodes. In Figs. 5a-c, note that when only the top 5 or fewer ($x \le 5$) genes are of interest, NetworkHub does detect the causal ones, with a lower error rate, regardless of the effect size ranging from strong (Scenario A in Fig. 5a) to mild (Scenario B in Fig. 5b), and to weak (Scenario C in Fig. 5c). For TDR, the shrinkage t-test performs the best when the effect size is strong; while others have similar rates (Scenario A in Fig. 5d). This is not surprising because we assumed only marginal effects in simulating the data. Such advantage disappears, however, when the effect size is moderate or weak (Scenarios B and C in Figs. 5e-f). In that case, all four methods performed similarly.

Discussion

Pathways are biological systems connecting genes, proteins, and chemical substances that participate together in a molecular function. It is known that nodes inside a pathway are dependent on each other, and some nodes like hub nodes may serve as gate keepers that can maintain or disrupt this biological function. Such relationships should be considered if the aim is to rank gene nodes in the same pathway. The proposed pathway activity score integrates gene expression values, takes into account their differential status as well as their dependence, and includes available network information. In the breast and ovarian cancer applications we demonstrated that the proposed procedure NetworkHub can identify genes that have been incorporated in current standard treatments or are being evaluated in ongoing clinical trials. This procedure can provide a complementary tool when ranking a set of genes in a network structure where hub nodes are present.

Some details should be noted, however, when applying this procedure. First, when incorporating the network information, the threshold α_S is used to flexibly include or exclude nodes in the pathway. It has been set at the default value 1 in all the analyses. Other choices of this value change the results only slightly. In Figure S5A we calculate the HRR at different values of α_S. Note that the HRR maintains a satisfactory level even when α_S is set at 0.05, remaining above 60% under Scenarios A and B. Second, the proposed ranking procedure can be applied before or after the pathway association has been tested. In Figure S5B we display the HRR based on the same simulations as in Figure S5A, but only for significantly associated pathways. The HRR becomes slightly larger, indicating little gain if a pathway association test is conducted a priori. The third issue is about the network information. The network information included in NetworkHub involves the correlation and path length d_{jk}. The path length is not a measure of Euclidean distance, but rather one that is comparable to the likelihood of molecular interaction between two nodes. A smaller d_{jk} implies a larger chance of interaction.

There are issues involved in this procedure worth further investigation. First, we considered here only one dataset from a breast cancer study and one from an ovarian cancer study to demonstrate the procedure. If one aims at unraveling genetic causality for a specific disease, then the proposed NetworkHub should be applied to other datasets containing comparable diseased subjects or the integrative analysis of multi-omic data should be implemented. Second, the simulation studies

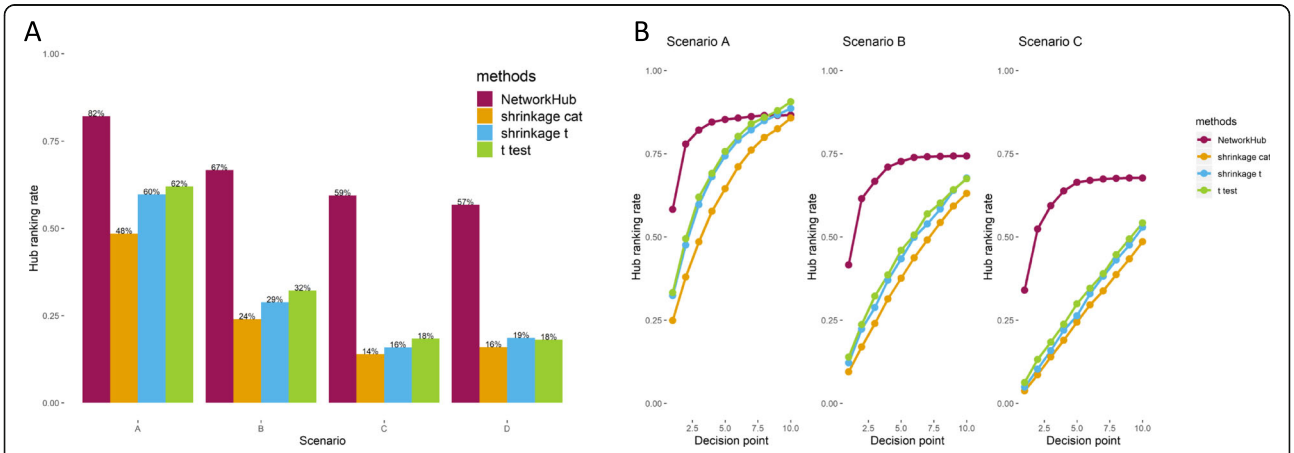


Fig. 4 Hub ranking rates (HRR) in simulation studies. **a** The hub ranking rate of the four ranking methods under Scenarios **a**, **b**, **c**, and **d**, at decision point 3. **b** The HRR of the four ranking methods under Scenarios A-C at different decision points

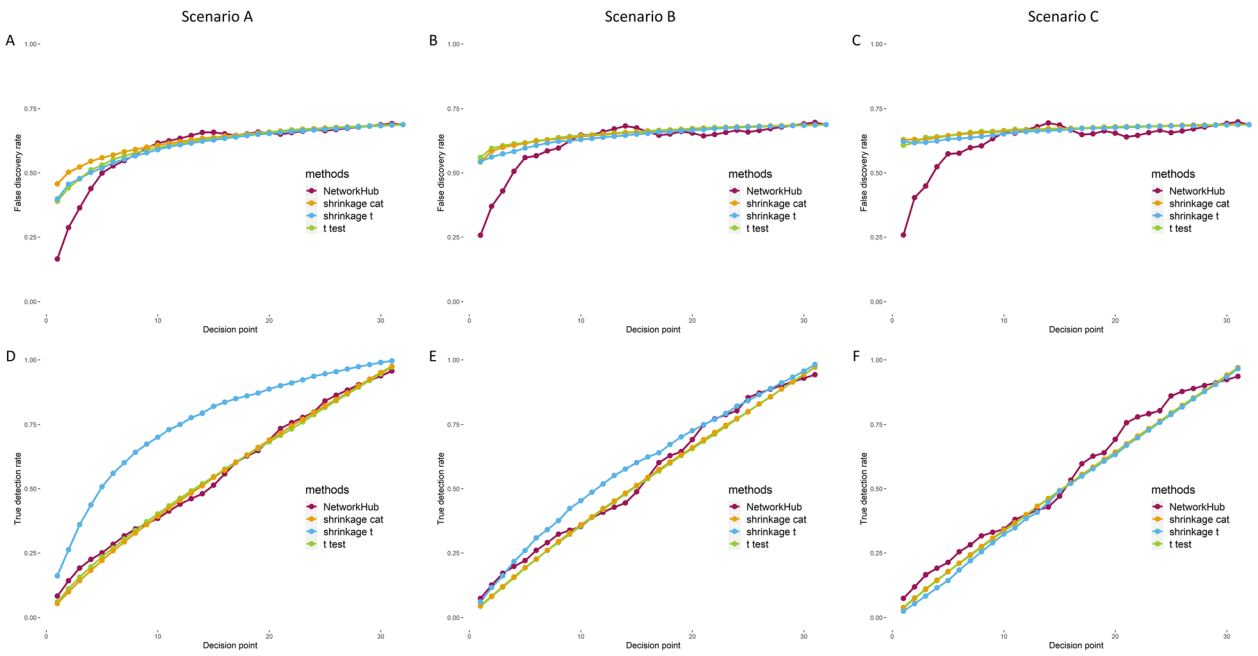


Fig. 5 The false discovery rate (**a-c**) and true detection rate (**d-f**) of causal nodes at different decision points

conducted here only compared NetworkHub with 3 t-statistic-type tests, where sampling variation is accounted for, but no network properties are included, in the ranking procedures. This comparison did not cover other network-based methods because their execution does not include sampling variation. For instance, Endeavour does not incorporate information about gene expression, which leads to the same list of rankings in every replication across all scenarios of simulation. PINTA reads the input expression data from a dialog box online, which requires manual input and thus is not practical for simulation studies. A similar problem exists for NGP-ND and NGP-NR. Therefore, we can compare all nine methods in the two cancer studies, but not in simulation studies. Further studies may focus on modifications of those algorithms, or on other disease datasets, for broader applicability. Third, the definition of pathways in different sources may vary. The use of multiple pathway databases has been suggested because the choice of the database could impact the enrichment analysis and predictive model [55]. The need to name and annotate the pathway with a unique identification number has been called for in a comparison review of twenty-four human cell signaling pathway databases [56]. Here we adopted KEGG [19] simply for demonstration and other platforms such as String with protein-protein interaction [17] or even a user-defined pathway network can be applied straightforwardly. The codes stored in GitHub (https://github.com/Hung-Ching-Chang/NetworkHub) currently work for input from KEGG only and we are working on inclusion of other types of input. This also relates to the fourth issue, namely, when NetworkHub is to be applied in situations where a large user-defined gene-set is provided, instead of a known biological pathway. This case can arise in analysis that explores unknown relationships among gene nodes. Since the network structure is not clear in such a set, a fully connected network where all nodes are linked directly to each other may be considered. Further investigations are warranted to examine if any unnecessary edges between nodes affect the final conclusion.

Conclusions

In summary, we proposed a network-based procedure, NetworkHub, to prioritize the gene nodes, especially the hub nodes, contained in a biological pathway network. This procedure first constructs a pathway activity score based on gene expression value, marginal effect, and network information. The network information, also termed as the intra-network association, is a function of Pearson's correlation and minimum path length between all possible pairs of nodes, subject to a user-defined threshold for nodes to be included in this calculation. This pathway activity score was next used in a leave-one-out evaluation to prioritize the importance of the gene nodes. The application of this procedure to two cancer studies identified several important genes that have now been used in standard treatment or currently considered in clinical trials for target therapy.

Methods

For each sample i ($i = 1, ..., N$), the pathway activity score T_i^{raw} combines the local effect through the weight L_j and the topology information S_j derived from the j th gene node ($j = 1, ..., M$) for all M nodes in the same pathway as

$$T_i^{raw} = \sum_{j=1}^{M} g_{ij} \times \left(L_j + S_j\right).$$

This score summarizes the information contained in this group of genes with three measurements: the expression level of genes g_{ij}, the information L_j about whether the gene is differentially expressed, and the network topology information S_j carried by the j th gene in the network. The details of the last two are as follows.

Summarizing the local effect and network information

Weighting the local effect

The weight of the local effect L_j for the j th gene is a weight based on its negative logarithm of p-values

$$L_j = \frac{-\log p_j \times I\left(p_j < \alpha_L\right)}{\sum_{m=1}^{M}\left[-\log p_m \times I(p_m < \alpha_L)\right]}.$$

This p-value results from a single-marker test such as a t-test on gene expression levels between two phenotypic groups. The indicator function $I(p_j < \alpha_L)$ determines which p-values are included in the weighting system, where α_L is the threshold for differential expression. In other words, this local effect assigns larger weights, on an exponential scale, to genes that are statistically significant at the significance level α_L. The default is $\alpha_L = 0.05$. When $\alpha_L = 1$, the summation of all L_j becomes the scaled test statistic of Fisher's method for gene set analysis [7, 57]. Note that the denominator is included so that the local effect weights sum to 1.

Weighting network information

The topology information measure S_j at the j th node is defined as

$$S_j = \frac{\sum_{k=1,k\neq j}^{M} \frac{|\rho_{jk}|}{d_{jk}} \times I\left(d_{jk} < D, p_j < \alpha_S\right)}{\sum_{m=1}^{M}\left\{\sum_{k=1,k\neq m}^{M} \frac{|\rho_{mk}|}{d_{mk}} \times I(d_{mk} < D, p_m < \alpha_S)\right\}}.$$

This measure contains three elements: (1) the pairwise absolute Pearson's correlation $|\rho_{jk}|$ between the current node j and the rest of the nodes, (2) the path length d_{jk} for the closeness between nodes j and k, and (3) the threshold D which controls the size of the topology influence. The closeness d_{jk} here is defined as the minimum steps/edges between two nodes in the network. It is used to balance the effect of correlation. For example, when two nodes are highly correlated but distant, it implies they are less likely to interact directly, and therefore the correlation is down-weighted by the path length.

The third component, the threshold D, regulates the type and size of the topological influence. It can be a very large number so that all gene nodes contribute to this measure. Alternatively, it can be determined to include only first-degree neighbors such as the immediately adjacent up- and down-stream genes, or genes whose coding protein is being directly regulated.

The number α_S in S_j also serves as a gatekeeper to control the number of genes involved in the topological contribution. When the default value $\alpha_S = 1$ is used, all other nodes, whether differentially expressed or not, will contribute to the topological influence of this gene. If one chooses a small value for α_S, say 0.05, then only differentially expressed genes are included. Again, the denominator in S_j guarantees these weights sum to 1.

In brief, this measure S_j has several features. First, it incorporates the relationship among genes through the use of correlation, which relaxes the traditional assumption of independence. Second, this correlation is balanced with the inclusion of the length d_{jk}. In other words, a large correlation between two genes far apart in the network will be moderated by a correspondingly large d_{jk}; while the correlation between two genes in a direct regulation will not. Third, the threshold D for the minimum distance offers the flexibility to include either all genes or only the nearby ones in the evaluation of network influence. Finally, the number α_S provides the chance to include only genes passing the single-marker association test.

Regularization of effect direction

When combining the gene expressions in a network, caution should be taken if there exist genes negatively correlated with each other. A direct summation of the expression levels without taking into account their interrelation may underestimate the strength of the effect. Therefore, for gene j, if its t-test statistic is negative ($t_j < 0$) in the single-marker test, then the expression values will be regularized by its maximum value O_j, where $O_j = \max\{g_{1j}, g_{2j}, ..., g_{Nj}\}$ is the maximum across all observed gene expressions from N samples at the same gene j, and the expression to be used for calculating the network score becomes

$$g_{ij}^R = \begin{cases} g_{ij}, & \text{if } t_j \geq 0 \\ O_j - g_{ij} + \varepsilon, & \text{if } t_j < 0 \end{cases}.$$

Here the ε is a small positive number, say 0.001, to guard against a value of zero. This standardization leads to our proposed pathway activity score (PAS) for the network

$$T_i = \sum_{j=1}^{M} g_{ij}^R \times (L_j + S_j).$$

This regularization avoids excessive cancellation in summing g_{ij_1} and g_{ij_2} when they are negatively correlated and, instead, gives a relatively large value of PAS to reflect the higher degree of activity level in this pathway for sample i. Here the t statistic t_j is used as a reference for regularization; other alternatives include the difference in the average expression level between two phenotypic groups or the fold-change between the two groups.

Gene-ranking with leave-one-out evaluation

To rank the gene nodes, we first compute the PAS with the procedures described above for each sample, and then calculate the difference in average PAS between two phenotypic groups (grp1 and grp2), t_{all}, standardized by standard errors, *se*, corresponding to each group,

$$t_{all} = \frac{\overline{T}_{grp1} - \overline{T}_{grp2}}{\sqrt{\left[se\left(T_{i,i\in grp1}\right)\right]^2 + \left[se\left(T_{i,i\in grp2}\right)\right]^2}}.$$

Next, we rearrange the network by leaving one gene node out, evaluate again the local weight and topology measure for each node in the new network with the $M - 1$ nodes, derive the corresponding PAS for each sample, denoted as $T_{i(-j)}$, and then calculate the standardized difference in average PAS as $t_{(-j)}$, where j is the index of the node removed.

Once all nodes are visited in turn, and the corresponding standardized differences have been computed, their magnitudes are then compared with the original t_{all}, by taking the absolute difference $Q_j = |t_{all} - t_{(-j)}|$. These values can be ordered and displayed in a scree plot to facilitate analysis. A large value indicates a substantial change when the gene node is removed from the pathway network, while a small value implies little perturbation with this deletion. The scree plot of the sorted absolute differences $Q_{(j)}$—i.e., the order statistics—from largest to smallest, can provide the ranking in terms of importance, indicating which nodes may be useful in target therapy or drug development within this set of interest. This is the NetworkHub procedure. In addition, if a cut-off is needed to select influential gene nodes, the inflection point on the curve, defined as the first ordered $Q_{(j)}$ where $(Q_{(j+1)} - Q_{(j)}) - (Q_{(j)} - Q_{(j-1)})$ becomes negative, can be used where all nodes before $Q_{(j)}$ are considered as influential. The R code for performing the pathway test and gene-ranking, an example, and the document files are freely available online (https://github.com/Hung-Ching-Chang/NetworkHub).

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-3444-7.

Additional file 1: Tables S1-S5. for detailed lists of rankings in each pathway under different methods.

Additional file 2: Figure S1. Boxplots of local weights (left column) and topology weights (right column). Here the thresholds are set at conservative values: 0.05 for a_L, 1 for a_S, and no limit for D so that all nodes are included. **Figure S2:** Analyses of the mTor pathway for the breast cancer study. **Figure S3:** Analyses of the estrogen pathway for the breast cancer study. **Figure S4:** Analyses of the JAK-STAT pathway for the breast cancer study. **Figure S5:** The hub ranking rate of NetworkHub corresponding to different threshold values before and after the pathway association test, under four scenarios. **A:** Before. **B:** After.

Abbreviations

FDR: False discovery rate; GSEA: Gene set enrichment analysis; HRR: Hub ranking rate; KEGG: Kyoto encyclopedia of genes and genomes; ND: Networked differential expression; NetworkHub: Network hub prioritization method; NGP: Network gene prioritizer; NR: Network rewiring; TCGA: The cancer genome atlas; TDR: True detection rate for causal nodes

Acknowledgements

Not applicable.

Authors' contributions

HCC, CPH and CKH conceived the study and designed the method. HCC implemented the simulation studies. HCC and SJL conducted the data analyses. HCC and CKH prepared the draft. All authors read and approved the final manuscript.

Funding

This work was supported by the Ministry of Science and Technology, Taiwan (MOST 106–2314-B-002-097-MY3). The funding agency had no role in the study design, data analysis, writing the manuscript, and decision to publish.

Availability of data and materials

The breast cancer data are publicly available and can be downloaded from NCBI GEO database (GSE69240). The ovarian carcinoma array expression levels are also freely downloadable from The Cancer Genome Atlas (TCGA). The R code for implementation and tutorial documents are freely available and can be downloaded from https://github.com/Hung-Ching-Chang/NetworkHub.

Ethics approval and consent to participate

Not applicable. All analyses were performed either on publicly available data or on simulation data.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Division of Biostatistics, Institute of Epidemiology and Preventive Medicine, National Taiwan University, No. 17, Xu-Zhou Road, Taipei 10055, Taiwan. [2]Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan. [3]Bioinformatics and Biostatistics Core, Center of Genomic Medicine, National Taiwan University, Taipei 10055, Taiwan.

Received: 15 August 2019 Accepted: 6 March 2020
Published online: 12 March 2020

References

1. Laenen G, Thorrez L, Bornigen D, et al. Finding the targets of a drug by integration of gene expression data with a protein interaction network. Mol BioSyst. 2013;9:1676.
2. Isik Z, Baldow C, Cannistraci CV, et al. Drug target prioritization by perturbed gene expression and network information. Sci Rep. 2015;5:17417.
3. Moreau Y, Tranchevent LC. Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat. Rev. Genet. 2012;13:523–36.
4. Piro RM, Di Cunto F. Computational approaches to disease-gene prediction: rationale, classification and successes. FEBS J. 2012;279:678–96.
5. Goeman JJ, Buhlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics. 2007;23:980–7.
6. Gatti DM, Barry WT, Nobel AB, et al. Heading down the wrong pathway: on the influence of correlation within gene sets. BMC Genomics. 2010;11:574.
7. Maciejewski H. Gene set analysis methods: statistical models and methodological differences. Brief Bioinform. 2013;15:504–18.
8. de Leeuw CA, Neale BM, Heskes T, et al. The statistical properties of gene-set analysis. Nat Rev Genet. 2016;17:353–64.
9. Tamayo P, Steinhardt G, Liberzon A, et al. The limitations of simple gene set enrichment analysis assuming gene independence. Stat Methods Med Res. 2016;25:472–87.
10. Zuber V, Strimmer K. Gene ranking and biomarker discovery under correlation. Bioinformatics. 2009;25:2700–7.
11. Tranchevent L, Ardeshirdavan A, EIShal S, et al. Candidate gene prioritization with Endeavour. Nucleic Acids Res. 2016;44:W117–21.
12. Chen J, Bardes EE, Aronow BJ, et al. ToppGene suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 2009;37:W305–411.
13. Seelow D, Schwarz JM, Schuelke M. GeneDistiller-distilling candidate genes from linkage intervals. PLoS One. 2018;3:e3874.
14. Wu C, Zhu J, Zhang X. Integrating gene expression and protein-protein interaction network to prioritize cancer-associated genes. BMC Bioinformatics. 2012;13:182.
15. Nitsch D, Tranchevent LC, Thienpont B, et al. Network analysis of differential expression for the identification of disease-causing genes. PLoS One. 2009;4: e5526.
16. Nitsch D, Tranchevent LC, Goncalves JP, et al. PINTA: a web server for network-based gene prioritization from expression data. Nucleic Acids Res. 2011;39:W334–8.
17. Jensen LJ, Kuhn M, Stark M, et al. STRING 8-a global view on protein and their functional interactions in 630 organisms. Nucleic Acids Res. 2009;37: D412–5.
18. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50.
19. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.
20. Lin SJ, Lu TP, Yu QY, et al. Probabilistic prioritization of candidate pathway association with pathway score. BMC Bioinformatics. 2018;19:391.
21. Albert R, Barabasi A. Statistical mechanics of complex networks. Rev Mod Phys. 2002;74:47–97.
22. Barabasi A, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5:101–13.
23. Han J, Bertin N, Hao T, et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature. 2004;430:88–93.
24. Almaas E. Biological impacts and context of network theory. J Exp Biol. 2007;210:1548–58.
25. Langfelder P, Mischel PS, Horvath S. When is hub gene selection better than standard meta-analysis? PLoS One. 2013;8:e61505.
26. Estrada E, Higham DJ. Network properties revealed through matrix functions. SIAM Rev. 2010;52:696–714.
27. Glaab E, Baudot A, Krasnogor N, et al. TopoGSA: network topological gene set analysis. Bioinformatics. 2010;26:1271–2.
28. Chang CW, Lu TP, She CX, et al. Gene-set analysis with CGI information for differential DNA methylation profiling. Sci Rep. 2016;6:24666.
29. Chu CP. A topology-based pathway analysis under self-contained hypothesis. Master thesis: National Taiwan University, Taiwan; 2017.
30. Abba MC, Gong T, Lu Y, Lee Y, et al. A molecular portrait of high-grade ductal carcinoma in situ. Cancer Res. 2015;75:3980–90.
31. Slattery ML, Lundgreen A, Hines LM, et al. Genetic variation in the JAK/STAT/SOCS signaling pathway influences breast cancer-specific mortality through interaction with cigarette smoking and use of aspirin/NSAIDs: the breast Cancer health disparities study. Breast Cancer Res Treat. 2014;147:145–58.
32. Goeman JJ, van de Geer AA, de Kort F, et al. A global test for group of genes: testing association with a clinical outcome. Bioinformatics. 2004;20: 93–9.
33. Draghici S, Khatri P, Tarca AL, et al. A systems biology approach for pathway level analysis. Genome Res. 2007;17:1537–45.
34. Christopoulos PF, Msaouel P, Koutsilieris M. The role of the insulin-like growth factor-1 system in breast cancer. Mol Cancer. 2015;14:43.
35. Ferguson AT, Evron E, Umbricht C, et al. High frequency of hypermethylation at the 14-3-3 sigma locus leads to gene silencing in breast cancer. Proc Natl Acad Sci U S A. 2000;97:6049–54.
36. Phan L, Chou PC, Velazquez-Torres G, et al. The cell cycle regulator 14-3-3σ opposes and reverses cancer metabolic reprogramming. Nat Commun. 2015;6:7530.
37. Shiba-Ishii A, Noguchi M. Aberrant stratifin overexpression is regulated by tumor-associated CpG demethylation in lung adenocarcinoma. Am J Pathol. 2012;180:1653–62.
38. Vilborg A, Bersani C, Wilhelm MT, et al. The p53 target Wig-1: a regulator of mRNA stability and stem cell fate? Cell Death Differ. 2011;18:1434–40.
39. Vilborg A, Glahder JA, Wilhelm MT, et al. The p53 target Wig-1 regulates p53 mRNA stability through an AU-rich element. Proc Natl Acad Sci U S A. 2009;106:15756–61.
40. Bersani C, LD XU, Vilborg A, et al. Wig-1 regulates cell cycle arrest and cell death through the p53 targets FAS and 14-3-3σ. Oncogene. 2014;33:4407–17.
41. Lane DP. P53, guardian of the genome. Nature. 1992;385:15–6.
42. Kastenhuber ER, Lowe SW. Putting p53 in context. Cell. 2017;170:1062–78.
43. Levine AJ, Oren M. The first 30 years of p53: growing ever more complex. Nat Rev Cancer. 2009;9:749–58.
44. Bouaoun I, Sonkin D, Ardin M, et al. TP53 variations in human cancers: new lessons from the IARC TP53 database and genomics data. Hum Mutat. 2016; 37:865–76.
45. Pejerrey SM, Dustin D, Kim JA, et al. The impact of ESR1 mutations on the treatment of metastatic breast cancer. Horm Cancer. 2018;9:215–28.
46. Haricharan S, Li Y. STAT signaling in mammary gland differentiation, cell survival and tumorigenesis. Mol Cell Endocrinol. 2014;382:560–9.
47. Furth PA. STAT signaling in different breast cancer subtypes. Mol Cell Endocrinol. 2014;382:612–5.
48. Aaronson DS, Horvath CM. A road map for those who don't know JAK-STAT. Science. 2002;296:1653–5.
49. Walker SR, Xiang M, Frank DA. Distinct roles of STAT3 and STAT5 in the pathogenesis and targeted therapy of breast cancer. Mol Cell Endocrinol. 2014;382:616–21.
50. Dakir EH, Pickard A, Srivastava K, et al. The anti-psychotic drug pimozide is a novel chemotherapeutic for breast cancer. Oncotarget. 2018;9:34889–910.
51. Luo J. Clycogen synthase kinase 3β (GSK3β) in tumorigenesis and cancer chemotherapy. Cancer Lett. 2009;273:194–200.
52. Cao Q, Lu X, Feng YJ. Glycogen synthase jubase-3β positively regulates the proliferation of human ovarian cancer cells. Cell Res. 2006;16:671–7.
53. McCubrey JA, Steelman LS, Bertrand FE, et al. GSK-3 as potential target for therapeutic intervention in cancer. Oncotarget. 2014;5:2881–911.
54. Opgen-Rhein R, Strimmer K. Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. Stat Appl Genet Mol Biol. 2007;6:9.
55. Mubeen S, Hoyt CT, Gemünd A, et al. The impact of pathway database choice on statistical enrichment analysis and predictive modeling. Front Genet. 2019;10:1203.
56. Chowdhury S, Sarkar RR. Comparison of human cell signaling pathway databases-evolution, drawbacks and challenges. Database (Oxford). 2015. https://doi.org/10.1093/database/bau126.
57. Fridley BL, Jenkins GD, Biernacka JM. Self-contained gene-set analysis of expression data: an evaluation of existing and novel methods. PLoS One. 2010;5:212693.

Publisher's Note