WINNER: A network biology tool for biomolecular characterization and prioritization

Thanh Nguyen[1,2], Zongliang Yue[1], Radomir Slominski[1], Robert Welner[3], Jianyi Zhang[2] and Jake Y. Chen[1]*

[1]Informatics Institute in School of Medicine, The University of Alabama at Birmingham, Birmingham, AL, United States, [2]Department of Biomedical Engineering, The University of Alabama at Birmingham, Birmingham, AL, United States, [3]Comprehensive Arthritis, Musculoskeletal, Bone and Autoimmunity Center (CAMBAC), School of Medicine, The University of Alabama at Birmingham, Birmingham, AL, United States

Background and contribution: In network biology, molecular functions can be characterized by network-based inference, or "guilt-by-associations." PageRank-like tools have been applied in the study of biomolecular interaction networks to obtain further the relative significance of all molecules in the network. However, there is a great deal of inherent noise in widely accessible data sets for gene-to-gene associations or protein-protein interactions. How to develop robust tests to expand, filter, and rank molecular entities in disease-specific networks remains an ad hoc data analysis process.

Results: We describe a new biomolecular characterization and prioritization tool called Weighted In-Network Node Expansion and Ranking (WINNER). It takes the input of any molecular interaction network data and generates an optionally expanded network with all the nodes ranked according to their relevance to one another in the network. To help users assess the robustness of results, WINNER provides two different types of statistics. The first type is a node-expansion *p*-value, which helps evaluate the statistical significance of adding "non-seed" molecules to the original biomolecular interaction network consisting of "seed" molecules and molecular interactions. The second type is a node-ranking *p*-value, which helps evaluate the relative statistical significance of the contribution of each node to the overall network architecture. We validated the robustness of WINNER in ranking top molecules by spiking noises in several network permutation experiments. We have found that node degree−preservation randomization of the gene network produced normally distributed ranking scores, which outperform those made with other gene network randomization techniques. Furthermore, we validated that a more significant proportion of the WINNER-ranked genes was associated with disease biology than existing methods such as PageRank. We demonstrated the performance of WINNER with a few case studies, including Alzheimer's disease, breast cancer, myocardial infarctions, and Triple negative breast cancer (TNBC). In all these case studies, the expanded and top-ranked genes identified by WINNER reveal disease biology more significantly than those identified by other gene prioritizing software tools, including Ingenuity Pathway Analysis (IPA) and DiAMOND.

Conclusion: WINNER ranking strongly correlates to other ranking methods when the network covers sufficient node and edge information, indicating a high network quality. WINNER users can use this new tool to robustly evaluate a list of candidate genes, proteins, or metabolites produced from high-throughput biology experiments, as long as there is available gene/protein/metabolic network information.

Introduction

Gene prioritization from large-scale omics projects is a central topic in disease biology (Huang H. et al., 2009). Manual searches of the literature and publicly annotated databases (Gene Ontology et al., 2013; Kanehisa et al., 2017; Tyner et al., 2017) for genes associated with a particular disease or biological process can be biased, because they are limited to existing knowledge. Sifting hundreds and thousands of gene or genetic variations associated with genes from genomic studies can also be daunting (Moreau and Tranchevent, 2012), e.g., even for a user to search for genes associated with cardiac arrhythmia (Rajab et al., 2010) within a 2-Mb region of chromosome 17 may return 77 candidate genes. For many biologists, the lack of ranking of genes based on biological relevance of disease context is an experience analogous to the pre-Google days of Internet search of web content. With influx of data from large-scale sequencing projects (Schlotterer et al., 2014), bioinformatics users increasingly count on good gene prioritization to help them generate biological hypotheses (Chen et al., 2006a; Hale et al., 2012), find potential disease biomarkers (Saha et al., 2008; Zhang and Chen, 2010, 2013), and identify candidate drug targets (Chen et al., 2006b, 2013; Li et al., 2009; Muhammad et al., 2017). However, as datasets continue to become larger and more heterogeneous, statistical (Subramanian et al., 2005; Aerts et al., 2006; Cantor et al., 2010) and text-mining (Krallinger et al., 2008; Liu et al., 2015; ElShal et al., 2016) approaches to gene prioritization lack sufficient precision in the biological knowledge context. For example, surveys of PAGER (Yue et al., 2018) for genes associated with the response of breast cancer to doxorubicin treatment may retrieve more than 2,000 statistically significant genes with MSigDB (Liberzon et al., 2015), or 234 candidate genes with the online text-mining platform Beegle (ElShal et al., 2016). The use of statistical p-values to prioritize retrieved genes can mislead biology users who assume statistical significance in samples equate the gene's true biological significance against one another in the experiment (Kim and Bang, 2016).

To overcome the limitations gene prioritization in practice, bioinformatics researchers have developed gene network models with which they perform knowledge-based gene prioritization and novel candidate genes identification (Chen et al., 2006a; Cowen et al., 2017). A molecular network consists of nodes (e.g., proteins) linked by edges that represent the pairwise interactions between nodes, forming a convenient computational model that is easy to interpret and has been widely used to discover (and rediscover) disease-specific genes and potential targets for treatment (Chen et al., 2009; Wu et al., 2009; Erten et al., 2011; Gottlieb et al., 2011; Guney and Oliva, 2012; Singh-Blom et al., 2013; Smedley et al., 2014; Peters et al., 2017; do Valle et al., 2018). Network-based methods also enable researchers to integrate data from a wide variety of sources, including analyses of gene-gene similarity (Alvarez-Ponce et al., 2013), proteomic interactions (Rolland et al., 2014), and regulatory pathways (Li and Campos, 2015); however, the results of prioritization strongly depend on the input gene list (Antanaviciute et al., 2015), and the list is often derived from existing databases that may lack important genes because of statistical errors or human errors during annotation. For example, acetylcholinesterase (ACHE), which is commonly associated with β-amyloid plaques and neurofibrillary tangles in the brains of patients with Alzheimer's Disease (AD; Talesa, 2001), is not among the annotated genes for AD in the KEGG database (Kanehisa et al., 2017). Input lists may also be compromised by redundancy, which can be generated from at least two sources: (1) the inclusion of genes that were falsely identified during the statistical analysis of an experiment (Yu et al., 2017), and (2) when, in an attempt to increase comprehensiveness, the list is expanded to include the gene for a "hub" protein that interacts with dozens, or even hundreds, of other proteins [e.g., ubiquitin C binds to 4,658 other molecules (Chen et al., 2017)] and, consequently is unlikely to be specific for the phenotype of interest. Furthermore, the statistical significance of a ranking is typically calculated *via* comparison to the rankings from a randomized version of the original network, but since the randomized network is often created by adding or deleting a small number of gene-gene interactions (i.e., increasing noise), or *via* total network permutation (Xie et al., 2015; Guala and Sonnhammer, 2017), much of the topology of the original network may be lost.

Related works

According to Bromberg (2013), molecular-interaction-based disease gene prioritization started in the early 2000's by pioneering techniques such as G2D (Perez-Iratxeta et al., 2002). In principle, statistical analysis of the patients' genetic data yields 100's of disease-associated genes. These genes often belong to an interaction network (Sun and Zhao, 2010), which is also called a "disease pathway." Assume that the disease phenotypes occur due to a disturbance at any point of the pathway, then disturbing the "most influential" genes is the most likely reason leading to the disease. Then, having a good disease pathway, network ranking algorithms, especially the eigenvector-based [Random Walk (Smedley et al., 2014) and PageRank (Page et al., 1999)] and centric-based [betweenness centrality (Newman, 2005)] can be used to prioritize the genes. Also, this idea can be applied to analyze key regulators in non-disease-specific biological processes. However, the pathways are usually incompleted: new disease regulators are still not discovered or some interaction among disease-associated genes are not yet shown (Bromberg, 2013). Therefore, the ranking techniques are required to extend the interaction network beyond the known disease-associated genes. Recent gene prioritization techniques have this ability. For example, DIAMOnD (Ghiassian et al., 2015) built a large network comprising genes related to 70 diseases, clustered the large network into multiple network modules, then assigned the network module to a disease; here, in the same module, genes not related to the disease module are added (extended) into the disease-specific network-module for prioritization. Ingenuity Pathway Analysis (Kramer et al., 2014) extended the disease-specific pathway by statistically estimating the likelihood of how a new gene interacts with the known disease-related gene. In Node2Vec (Grover and Leskovec, 2016; Peng et al., 2019), a "global gene network," which includes the known disease-specific genes, their direct interacting genes, and indirect interacting ones (optionally) was constructed; then, each gene is represented by a numerical vector having a fixed-length dimension to allow computing the cosine similarity between a known disease-specific gene and another gene; so, the extension can be made by choosing the genes having high cosine similarity to any of the disease-specific ones. Or, in GenePANDA (Yin et al., 2017), given a "global gene network" (similar to Node2Vec), for a specific gene, the average distance between itself and any other gene in the "global" network was subtracted by the average distance between itself and the known disease-specific genes; then, this difference was used to rank the genes.

Besides the network-based approach, gene prioritization could be performed using text mining and similarity profiling approaches (Yin et al., 2017). In the text mining approach, it is hypothesized that important genes are more likely to be mentioned in an article than non-important ones. Therefore, text mining tools, such as aBandApart (Van Vooren et al.,

2007) and Gene Prospector (Yu et al., 2008), emphasize efficient queries in MEDLINE and other large literature collections to find important disease-specific genes. However, these approaches may not find important genes when the disease is not yet well-researched or when a new disease model (i.e., a new cell line or new organoid) is built to represent the disease. On the other hand, similarity profiling defines the similarity among the genes according to the disease-related information; then, if a novel gene shares a high similarity with genes that are known to be important, the novel gene will be ranked highly. For example, Endeavor (Aerts et al., 2006) and ToppGene (Chen et al., 2009) integrated multiple disease-omic databases by a machine-learning model; the model was trained to classify between the known-important genes and non-important genes; the model will produce a ranking score reflecting how important a novel gene is, respecting the already known ones. Meanwhile, the disease-specific gene expression and correlation matrix can be clustered or latent-based represented, such as in Pinta (Nitsch et al., 2011), Maxlink (Guala et al., 2014), and Genefriends (van Dam et al., 2012), where the well-known disease-specific genes are expected to concentrate in one or a few clusters/latent modules, and the novel genes in these clusters or modules would be ranked highly.

Here, we introduce a new ranking method, W̲eighted I̲n-N̲etwork N̲ode E̲xpansion and R̲anking (WINNER), that addresses many of the current limitations of network-based gene prioritization methods. As with PageRank (Winter et al., 2012) and many other gene prioritization techniques, the ranking engine of WINNER uses random-walk principles (Zhao et al., 2015). However, WINNER was designed to address the following three specific network biology tasks: (1) perform gene prioritization in a weighted biomolecular association network, (2) identify upstream regulators and targeted genes (i.e., "upstream" ranking), or (3) identifying downstream effector molecules that are specific for a particular disease or phenotype ("downstream" ranking). WINNER can generate a ranking score for each input gene, derive optional genes that are "expanded" from the original seed gene lists, and provide two different statistic for users (1) the gene expansion p-value (p_e) for adding a gene to the network, which addresses both incomprehensiveness and redundancy; and (2) the gene ranking p-value (p_r), which represents the significance of the ranking when compared to the randomized network. Furthermore, we found that compared to total network permutation (Xie et al., 2015; Guala and Sonnhammer, 2017), preserving the modularity randomization (Cowen et al., 2017) produces a randomized network that is topologically similar to the original network and yields a more normal distribution of ranks (Espinoza, 2012). We further demonstrated the benefit of WINNER in omics study result interpretations with the following case studies: (1) ranking genes that are genetically associated with Alzheimer's disease (AD); (2) ranking breast-cancer survival-related genes (Lanczky et al., 2016); (3) ranking differentially expressed genes involved in

myocardial injury in pigs for their potential roles in myocardial regeneration (Eschenhagen et al., 2017). In all these studies, we discuss how our prioritization score and statistic associated with high-ranked genes enable biology users to derive new insights and hypotheses worth further experimental investigations.

Methods

For this work, we postulated (1) that the seeded (i.e., input) genes consist of (but are not limited to) differentially expressed genes identified in a wet-lab experiment, genes in a well-curated pathway, and phenotype-associated genes mined from the literature; and (2) that genes added to the expanded network (i.e., "expansion genes") would have significantly more interactions with seeded genes (i.e., "seeded interactions") than with non-seeded genes. WINNER begins with the set of seeded genes and a collection of gene-gene interactions, iteratively applies network ranking for gene prioritization, and expands the ranked list of genes one gene at a time (Supplementary Figure 1). Each gene-gene interaction has a confidence score (scaled between 0 and 1), which is commonly included in interactome databases (Chatr-Aryamontri et al., 2013; Szklarczyk et al., 2015); however, if a confidence score is not available, then the confidence score is set to 1 for all interactions. Network ranking is first applied to the seeded genes and the interactions among them (S_0 metric, Equation 1); then, genes adjacent to the seeded genes are filtered for significant interactions with the seeded genes (p_e) to identify candidates for the expanded network. The identified candidate is added to the ranked list, and network ranking is re-applied to initiate the next iteration of the cycle. A more detailed description of each step is provided below.

Ranking genes in the network by WINNER

Undirected networks

Given a gene-gene association network, the genes are ranked as in Supplementary Video 1. First, WINNER assigns an initial score (S_0) to the genes, according to Yue et al. (2017):

$$S_0(i) = e^{2\ln(w(i)) - ln(I(i))} \tag{1}$$

where i represents the gene index, $w(i)$ is the sum of the confidence scores (normalized to between 0 and 1) for all gene-gene interactions associated with i, and $I(i)$ is the number of gene-gene interactions associated with i. Here, larger confidence scores imply stronger associations. Second, WINNER iteratively updates the gene score by applying the Random Walk technique (Page et al., 1999):

$$S_t(i) = (1 - \sigma) \times S_0(i) + \sigma \times \sum_{\forall j} \frac{c(j, i) \times S_{t-1}(j)}{w(j)} \tag{2}$$

where s is the random walk damping parameter [set to $s = 0.85$ as described (Page et al., 1999)], $c(j, i)$ represents the confidence score of the interaction between gene i and gene j, and t is the index of iteration (starting at 1); $S = 0$ for genes that are outside the network but appear in the collection of gene-gene interactions. PageRank theory (Page et al., 1999) demonstrates that S_t converges ($|S_t - S_{t-1}|^{®} 0$) if t is large enough, so the iterative cycle was continued until $|S_t - S_{t-1}| < 0.001$.

Directed networks

Directed networks, such as networks of regulatory pathways, include more annotation than undirected networks. Thus, we adapted the definitions of terms in Equations 1, 2 so that WINNER could be used to (for example) infer upstream regulatory and downstream effector genes (Kramer et al., 2014). For "upstream" ranking, i is the regulatory gene and j is the gene regulated by i; thus, $w(i)$ is the sum of the confidence scores for all gene-gene relationships that i regulates, $I(i)$ is the number of gene-gene relationships regulated by i, and $c(j, i)$ is the confidence score for the regulation of j by i. For "downstream" ranking, i is the regulated gene and j is the gene that regulates i; thus, $w(i)$ is the sum of the confidence scores for all gene-gene relationships in which i is regulated, $I(i)$ is the number of gene-gene relationships in which i is regulated, and $c(j, i)$ is the confidence score for the regulation of i by j.

Statistical significance of gene ranking

To evaluate the statistical significance (p-value) of the gene ranking, we determined how likely the converging result of S (by default, S_{200}) in Equations 1, 2 is higher than in random networks. Randomization was performed in Matlab with degree-preservation (Espinoza, 2012; Tiong and Yeang, 2019) to maintain the topological characteristics of the original gene-gene network; however, the technique only generates unweighted relationships, so weights were randomly assigned from the distribution of relationship weights in the original network. One thousand random networks were generated, and the ranking scores (S_{200}) of the genes in the random networks were normally distributed (as validated *via* the Chi-square goodness-of-fit test). Thus, the ranking p-value (p_r) for each gene i was calculated by using the normal distribution [$m(i)$, $s(i)$] parameter estimation (Bowman and Azzalini, 1997):

$$p_r(i) = \begin{cases} \int_{-\infty}^{S_{200}(i)} \frac{1}{\sigma(i)\sqrt{2\pi}} e^{-\frac{(x-\mu(i))^2}{2\sigma^2}} dx & if \ S_{200}(i) < \mu(i) \\ \int_{S_{200}(i)}^{\infty} \frac{1}{\sigma(i)\sqrt{2\pi}} e^{-\frac{(x-\mu(i))^2}{2\sigma^2}} dx & if \ S_{200}(i) > \mu(i) \end{cases} \tag{3}$$

which is equivalent to computing the two-tailed p-value for a normal distribution.

Filtering candidates for expansion

We chose two hypergeometric tests that are common practice in annotation (Huang et al., 2009). First, we tested the likelihood of the candidate expansion gene having a seeded interaction relative to its total number of interactions. Second, we tested the likelihood of the candidate expansion gene having seeded interactions relative to the seeded interactions of its most similar seeded gene, with similarity determined by node degree. Thus, we calculated two p-values for each expansion gene j from the "overrepresented" point of view (Beissbarth and Speed, 2004; terms are defined in Supplementary Figure 2):

Test 1:

$$p_{1e}(j) = \sum_{l=k(j)}^{\min(n,K)} \frac{\binom{K}{l}\binom{N-K}{n-l}}{\binom{N}{K}} \quad (4)$$

Test 2:

$$\begin{cases} p_{2e}(j) = \sum_{l=k(j)}^{\min(n,K)} \frac{\binom{K}{l}\binom{N-K}{n-l}}{\binom{N}{K}} & \text{if } N > K \\ \\ p_{2e}(j) = 1 - \sum_{l=0}^{\min(n,K)} \frac{\binom{N}{l}\binom{K-N}{k-l}}{\binom{K}{N}} & \text{if } N < K \end{cases} \quad (5)$$

in which the double-line bracket operator represents the combination operator:

$$\binom{N}{K} = \frac{N(N-1)(N-2)\ldots(N-K+1)}{K(K-1)(K-2)\ldots 1} \quad (6)$$

Genes for which both $p_{1e}(j) < 0.05$ and $p_{2e}(j) < 0.05$ were chosen as candidates for expansion. Thus, the expansion p-value (p_e) for each gene j is defined by the equation $p_e(j) = \max[p_{1e}(j), p_{2e}(j)]$.

Selecting one candidate for expanded ranking

Since there will likely be more than one candidate expansion gene remaining after filtration, WINNER estimates which of the candidates should be added to the network by calculating an expansion score (e) from the confidence score of the interaction between the candidate gene and the ranked genes, and the ranking score (S) of the ranked genes:

$$e(i) = \sum \frac{c(i,j)S(j)}{W(j)} \quad (7)$$

Where i is the candidate expansion gene, j represents all seeded genes that interact with the candidate expansion gene, and $W(j)$ is the sum of the confidence scores for all interactions involving all seeded genes. Note that $W(j)$ differs from $w(j)$ in Equation 2, because $w(j)$ is restricted to interactions among ranked genes.

Informatics databases and benchmarking metrics

Correlations among WINNER, PageRank (Winter et al., 2012), dual node-edge ranking (Wang et al., 2015), eigenvector centrality, betweenness centrality, node degree, and clustering coefficient (Newman, 2008) were evaluated by computing the linear correlation coefficients and p-values with Matlab (Neupane and Kiser, 2018).

For analyses of upstream and downstream genes (directed network), genes were distributed into layers *via* the breadth-first-search approach, and groups of genes that formed a self-contained cycle were treated as a single node. Results were visualized with boxplots. In each pathway, the gene rank numbers were converted into percentile format: the first rank (number 1) was converted to 100% percentile, while the last rank was converted to 0% percentile. The percentile format allowed boxplot aggregation from multiple pathways, where the different pathways had different number of genes.

Experiments demonstrating the general topological and biological significance of the WINNER ranking were conducted with the small gene set associated with AD from KEGG release 50 (2009) (Kanehisa et al., 2010) and with undirected gene-gene interactions from HAPPI version 1.0 (Chen J. Y. et al., 2009). Rankings of upstream regulators and downstream effectors were conducted with all cancer disease pathways in KEGG release 85 (Kanehisa et al., 2017; Tessier et al., 2018) and gene-gene regulatory relationships from STRING v.10.5 (Szklarczyk et al., 2017).

The effectiveness of WINNER for identifying network-expansion genes was evaluated by using KEGG release 50 [stored in PAGER 1.0 (Yue et al., 2015)] as the input with interactions of all types (without directionality) from HAPPI v.2.0 whose confidence scores exceeded 0.75 (Chen et al., 2017), and then determining how closely the expanded network matched the updated KEGG release 85 (Kanehisa et al., 2017). An analogous experiment was conducted with Ingenuity Pathway Analysis (IPA), which (in theory) can be used for both upstream and downstream expansion and HAPPI v.2.0 (Kramer et al., 2014) for comparison. Precision, recall, and F1 scores were calculated *via* the following equations:

$$precision = \frac{|E \cap U|}{|E|} \quad (8)$$

$$recall = \frac{|E \cap U|}{|E|} \quad (9)$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (10)$$

where E is the set of expansion genes determined by Winner or IPA and U is the set of genes present in KEGG release 85 but not in KEGG release 50.

The biological relevance of our rankings was evaluated by (1) determining whether the top-ranked genes from WINNER ranking of the KEGG breast cancer pathway (Kanehisa et al., 2017; https://www.genome.jp/kegg-bin/show_pathway? hsa05224) were included among the genes correlated with survival in 3951 Breast Cancer patients (Gyorffy et al., 2010); and (2) by ranking the set of differentially expressed genes from a study of myocardial regeneration in neonatal pigs (Zhu et al., 2018) with WINNER and determining whether the top-ranked genes could contribute to cardiac repair and cardiomyocyte proliferation. For the analysis of breast-cancer survival genes, we calculated the ratio of the number of genes that were both significant (survival p-value < 0.05) in the breast cancer study (Gyorffy et al., 2010) and highly ranked by WINNER (i.e., scored above a defined threshold) to the number of highly-ranked genes.

Network randomization and testing for ranking normal distribution in random networks

In WINNER, given a network (also called the original network), we examined the following network randomization approaches to evaluate which network randomization approach was the most suitable for computing the ranking p-value for each gene:

- Total rewiring (also called total network permutation; Waksman, 1968). To implement this approach, for each interaction (edge) in the original network, we randomly changed the two genes (node) connecting through this edge. Therefore, this approach preserves the number of interactions, yet it totally changes the network and gene topology.
- Randomly drawing a new network such that each gene's degree is the same to what it is in the original network (also called preserving degree; Rao et al., 1996). A gene degree, in simple description, is the number of other genes connecting to the gene in the network.
- Randomly drawing a new network with the same modularity to the original network (also called preserving modularity). We implemented this strategy according to the network modularity definition in Newman (2006). Modularity measures likely the network can be partitioned into clusters of interacting genes.
- Randomly adding 5% new interactions into the original network. These interactions were

not reported in the gene-gene interaction databases.
- Randomly removing 5% of the interactions from the original network.

For each network randomization approach, starting from the same original network, we repeated 10,000 times, yielding 10,000 different random networks. Then, applying WINNER (and other ranking algorithms) yielded 10,000 random ranking results for each gene. We tested whether these random rankings followed a normal distribution using chi-square goodness of fit test (chi2gof)[1] in Matlab. In this test, the smaller chi-square (chi2) indicates that the rankings are more naturally distributed.

Literature validation using co-citations from PubMed

Important disease-specific genes are often co-mentioned in a research article. Therefore, to demonstrate the significance of the genes related to a disease, we applied a co-citations from the NCBI e-utils application programming interface (API; Sayers, 2008) that implements semantic searches of PubMed abstracts to report biomedical literature citations (https://eutils.ncbi.nlm. nih.gov/entrez/eutils/esearch.fcgi?). We applied "pubmed" as input of database and the concatenated string of the candidate gene and the disease name as input of terms. To identify the co-citation support for the winner scores, we separated the genes into two categories, with literature co-citation ($k = 0$) or without literature co-citation ($k > 0$) to find the differences between the winner scores. We applied the Kruskal-Wallis test to report p-values.

Biomedical case studies, data, and preprocessing

Cardiac regeneration dataset

For the cardiac regeneration case study, the bulk-RNA expression dataset was obtained from Zhang et al. (2020). Briefly, two groups of pig hearts were sent for sequencing when they reached postnatal days (P) 7, 14, and 28. In the first group, the pigs underwent myocardial infarction (a heart attack model) on the postnatal day 1, then their heart fully recovered to normal cardiac functionality with no scar. In the second group, the pig did not undergo injury (sham control). For each group at each day (P7, P14, or P28), three pigs were sequenced. The bulk-RNA data were processed by applying trim-galore (Krueger, 2015) for trimming the fastQ read, then STAR package v2.5.2 for mapping to Pig genome (Dobin et al., 2013), then the RNA transcripts

---

1 chi2gof: Chi-square goodness-of-fit test [https://www.mathworks. com/help/stats/chi2gof.html].

were counted using HtSeq version 0.6.1 (Anders et al., 2015). The gene expression was normalized, and fold-change was calculated using Deseq2 software (Love et al., 2014). Due to the small sample size ($n = 3$), the p-values for differentially expressed genes, compared between two groups at P7, P14, and P21, were calculated using the approach in Bian et al. (2021). After calculating and comparing two groups at these three different postnatal time points, this process yielded 276 seed genes as input for WINNER. Then, these genes were queried in HAPPI v2 database (Chen et al., 2017) to build their interacting network. These gene lists, their interaction, and WINNER results were summarized in Supplementary Tables 1, 2.

Data processing of triple negative breast cancer (TNBC)

Triple negative breast cancer (TNBC) has been found in 15% of breast cancer cases and is characterized by the tumor cells lacking the expression of the following: epidermal growth factor 2 (HER2), estrogen receptor (ER), and progesterone receptor (PR; Liu et al., 2014; Ueda et al., 2019). Unfortunately, because of its nature, TNBC has a poorer prognosis than other types of breast cancers and treatment options are limited (Xia et al., 2014; Eltohamy et al., 2018; Lu et al., 2020). While TNBC markers are already well-studied, finding the key disease regulators and promising targeted genes is still challenging (Nedeljkovic and Damjanovic, 2019). Therefore, we applied WINNER to explore novel answers for this question.

We took the triple negative breast cancer candidate genes from the University of Alabama at Birmingham Cancer data analysis Portal (UALCAN) database (Chandrashekar et al., 2022). In the comparison between the 116 triple negative breast cancer samples and 114 normal samples, UALCAN provided the top 250 up-regulated genes and 250 down-regulated genes selected by the t-test p-value. Next, we retrieved the Protein-Protein Interaction (PPI) using the medium confidence (score $\geq$ 0.4) and extended 100 genes using the STRING database. We performed WINNER and generated the gene ranking and p-values (Supplementary Tables 3, 4).

PubMed co-citation analysis of the WINNER ranked genes

We hypothesize that important disease-specific genes are often co-mentioned in a research article (Olsen et al., 2014); if so, WINNER high-ranking genes tend to be more co-cited in the literature than the low-ranking ones. Therefore, to demonstrate the significance of the genes related to a disease, we applied co-citations from the NCBI e-utils application programming interface (API; Sayers, 2008) that implements semantic searches of PubMed abstracts to report biomedical literature citations (https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?). We applied "pubmed" as an input of the database and the concatenated string of the candidate gene and the

disease name as input of terms. To identify the co-citation support for the winner scores, we separated the genes into two categories, the WINNER significant ranked genes (p-value $\leq$ 0.05) or WINNER non-significant ranked genes (p-value > 0.05) to find the differences between the co-citations. We applied the Kruskal-Wallis test to report p-values to test differences of co-citations between significant and non-significant genes.

Pathway level assignment

We retrieved significantly enriched pathways from PAGER 2.0 database (Yue et al., 2018) using WINNER highly ranked genes with p-values $\leq$ 0.05. We applied the parameter set as follows. The data sources were KEGG, WikiPathway, BioCarta, NCI-Nature Curated, Reactome, Protein Lounge, and Spike, the similarity was set to be 0.05, and FDR was set to be 0.01. We constructed the regulatory (r-type) PAG-to-PAG network using the default r-type relationship score cutoff (=1). We performed a 5-step procedure in the pathway level assignment. Firstly, we calculated shortest paths among the pairwise r-type PAG-PAG relationships. Secondly, we extracted the longest shortest path and assigned levels of pathway from the upstream to the downstream pathway using 1 to n. Thirdly, we expanded the level assignment to the using shortest distances, such as the current pathway is level m, the shortest distance between the expanded pathway in the upstream to the current pathway is 2, the expanded pathway level will be assigned by m-2. Fourthly, we took the average of the levels assigned to pathways. Fifthly, we repeated the steps three and four until all the pathways had been assigned.

The correlation analysis of WINNER ranking and the enriched pathways using the exponential scale of top gene bins

Firstly, we segregated the WINNER significant genes into 2^x bins. Secondly, we took the top 2^x bins (x is [1, X]) and merge the genes to perform the enrichment analysis. Thirdly, we had the pathways enriched in the top 2^x gene bins minus the pathways enriched in $2^1, \ldots, 2^{x-1}$ to seek the add-on pathways enriched in the top 2^x gene bins. Fourthly, we mapped the levels from the r-type pathway-to-pathway relationships to the add-on enriched pathways in each top 2^x gene bins, and plotted the curve of pathway levels vs. the gene bins. Meanwhile, we performed the Pearson correlation analysis to report the correlation coefficient between the pathways' levels and gene bins.

Results

Characteristics of WINNER ranking

WINNER ranking of undirected networks

When genes in the KEGG [release 50, stored in the PAGER 1.0 database (Yue et al., 2015)] AD pathway

FIGURE 1
WINNER gene prioritization is well-correlated with other ranking techniques and most network topological metrics. Genes in the KEGG AD pathway were ranked *via* WINNER (WN), PageRank (PG), Dual Node-edge Rank (DR), Betweenness Centrality (BC), clustering coefficient (CC), eigenvector centrality (EV), and node degree (ND); then, the correlation coefficients for all pairwise comparisons between ranking methods were calculated *via* Pearson's correlation.

(Supplementary Figure 3) were ranked *via* WINNER gene prioritization, our results were strongly correlated with those obtained *via* analyses of both eigenvector (Newman, 2008; $p = 1.45 \times 10^{-39}$) and node-betweenness (Newman, 2008; $p = 1.67 \times 10^{-11}$) centrality, but not with the clustering coefficient (Newman, 2008; $p = 0.22$). Similar patterns of correlation were obtained with two other state-of-the-art network-based ranking techniques, PageRank (Winter et al., 2012), eigenvector (Newman, 2008), betweenness centrality (Newman, 2005), and dual node-edge ranking (dual rank; Wang et al., 2015) (Figure 1), and all three ranking techniques were strongly correlated with node degree. Notably, the clustering coefficient, but no other metric or technique, failed to identify some of the most important markers for Alzheimer's, including Amyloid Beta Precursor Protein (A4 or APP; Jonsson et al., 2012), Caspase 8 (CASP8; Wei et al., 2002), Caspase 3 (CASP3; D'Amelio et al., 2011), and Presenilin 1 (PSN1; La Bella et al., 2004). Thus, WINNER was at least equivalent to other network topological metrics and well-established prioritization techniques for ranking genes in undirected biological networks.

The strong correlation between the WINNER and node-degree rankings prompted us to preserve the node degree and modularity during randomization. Examining the AD-associated genes network, the pairwise rank differences between the original network and the total-permutation random network were significantly large (Figure 2A). When the difference between the random ranking and the original ranking is too large, the random network topology would be too different from the original network topology; thus,
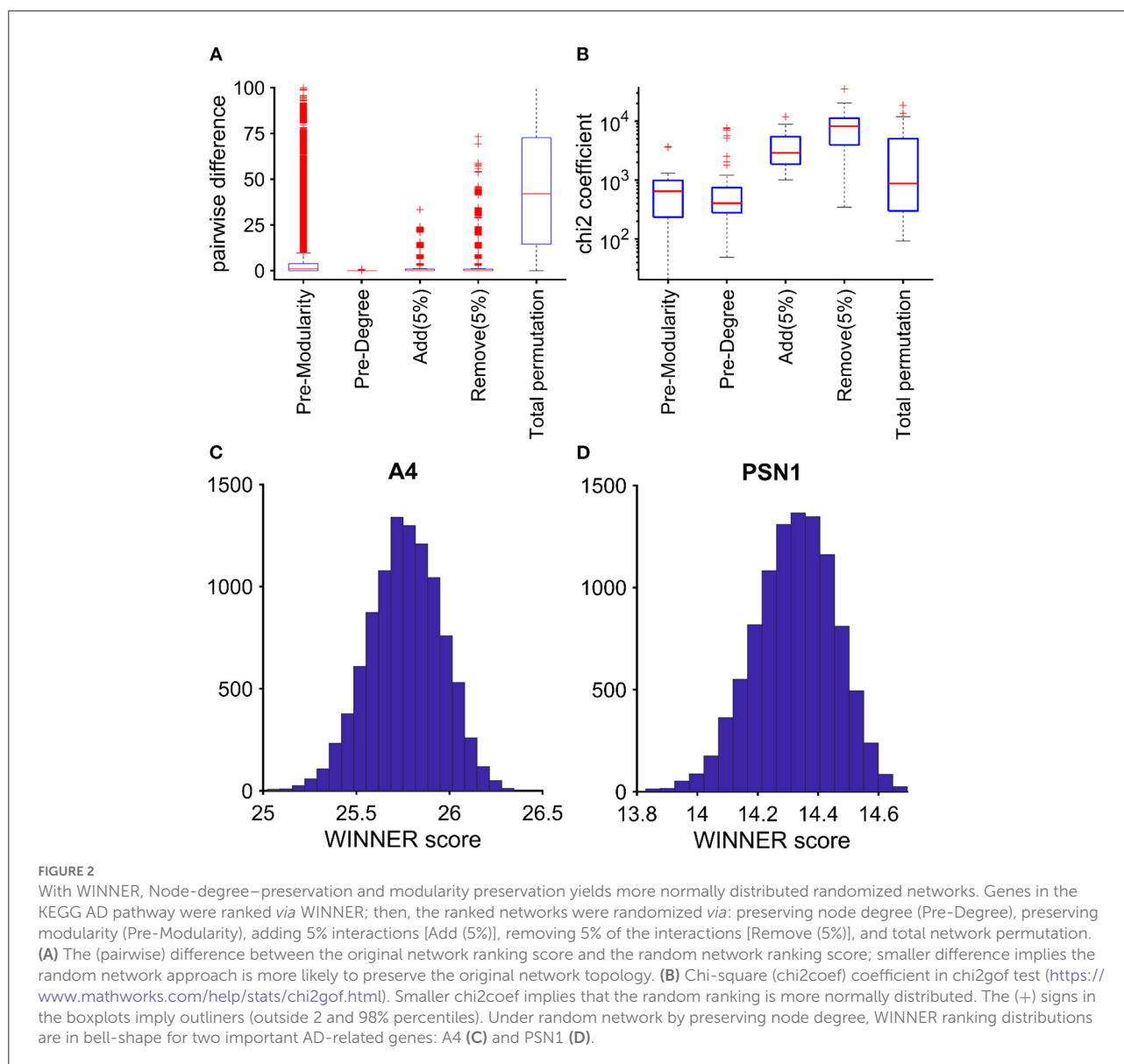
the random ranking may not be suitable to test statistical significance of the original ranking. Besides, when compared to other randomization techniques (total network permutation, preserving modularity, or adding/removing 5% of edges), the distribution of rankings of AD-associated genes in the degree-preserved randomized network was significantly more normally-distributed (Figure 2B). Furthermore, when examining the ranking distributions of two important AD-associated genes A4 and Presenilin 1 (PSN1; Figures 2C,D), it was clear that their distributions had the bell-shape. Thus, rather than relying on the empirical *p*-value (Cornish et al., 2018) for gene rankings, we generated 1,000 node-preserved randomized networks and calculated a ranking *p*-value (p_r) for all genes in all KEGG pathways. Notably, the rankings were much less likely to change in response to the addition of noise for genes with $p_r < 0.05$ than for genes with $p_r \geq 0.05$, especially as the amount of noise increased (Figure 3). These observations suggest that when randomized networks are generated with node-degree preservation, fewer randomizations may be required to achieve adequate precision, and fewer noise simulation may be necessary to evaluate the robustness of the rankings.

The accuracy of WINNER gene prioritization was evaluated by ranking genes in the KEGG breast cancer pathway (https://www.genome.jp/kegg-bin/show_pathway?hsa05224) and then determining whether the top-ranked genes correlated with the genes' effect on survival for patients with breast cancer, as estimated with an online Kaplan-Meier (Bland and Altman, 1998) tool that calculates the breast-cancer survival rates associated with more than 6,000 genes (Gyorffy et al., 2010). The KEGG breast cancer pathway contains 146 genes [annotated by UniProt Consortium (2018)], 62% of which significantly influenced patient survival, and a greater proportion of the most highly ranked genes were significantly associated with breast-cancer survival when prioritized with WINNER than with other gene prioritization techniques (PageRank and dual node-edge ranking; Figure 4). Furthermore, the precision of WINNER for retrieving survival-related genes (i.e., the proportion of retrieved genes that were significantly related to breast cancer survival) was even greater when restricted to genes with a ranking *p*-value of $p_r < 0.05$.

WINNER ranking of directed networks

WINNER ranking of directed networks was evaluated *via* WINNER upstream prioritization with all cancer disease pathways in KEGG release 85 (Kanehisa et al., 2017; KEGG, 2022) and the gene-gene regulatory relationships in STRING v.10.5 (Szklarczyk et al., 2017). Genes were distributed into layers using the breadth-first search approach (Wang et al., 2012) with genes coding for proteins that function further upstream in the pathways assigned to the lower-numbered layers. Thus, genes in the lowest-numbered layers tend to encode master regulatory molecules/receptors and first/second messengers, which are

FIGURE 2
With WINNER, Node-degree−preservation and modularity preservation yields more normally distributed randomized networks. Genes in the KEGG AD pathway were ranked *via* WINNER; then, the ranked networks were randomized *via*: preserving node degree (Pre-Degree), preserving modularity (Pre-Modularity), adding 5% interactions [Add (5%)], removing 5% of the interactions [Remove (5%)], and total network permutation. **(A)** The (pairwise) difference between the original network ranking score and the random network ranking score; smaller difference implies the random network approach is more likely to preserve the original network topology. **(B)** Chi-square (chi2coef) coefficient in chi2gof test (https://www.mathworks.com/help/stats/chi2gof.html). Smaller chi2coef implies that the random ranking is more normally distributed. The (+) signs in the boxplots imply outliners (outside 2 and 98% percentiles). Under random network by preserving node degree, WINNER ranking distributions are in bell-shape for two important AD-related genes: A4 **(C)** and PSN1 **(D)**.
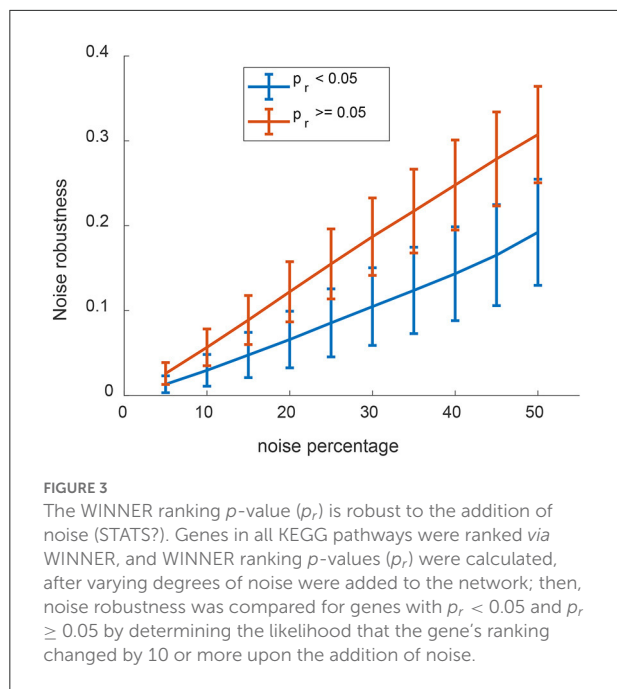
located where the signaling cascade originates (e.g., near the cell membrane; Koschmann et al., 2015), while genes with the highest layer numbers tend to encode downstream effector molecules that are closely associated with a specific disease phenotype, such as drug resistance in breast cancer (Johnston, 2006). Our results indicated that using WINNER, layer 1–3 genes, which were the upstream layers in the pathways, were consistently ranked at higher percentiles than genes at other layers (more downstream; Figure 5). But this consistency was not observed when the genes were prioritized *via* equivalent (directed-network ranking) analyses with PageRank (Winter et al., 2012) and dual node-edge ranking (Wang et al., 2015). WINNER upstream overestimated the ranking of genes in layer 8, but this can likely be attributed to noise, because the layer contained only 12 ranked genes.

WINNER network expansion and ranking upstream regulators

We demonstrated how WINNER could identify upstream regulators of two cancer pathways, Chronic Myeloid Leukemia (CML; https://www.genome.jp/kegg-bin/show_pathway?hsa05220) and hepatocellular carcinoma (https://www.genome.jp/pathway/hsa05225), that were missing from the existing pathways in KEGG but were present in the KEGG database itself. WINNER upstream prioritization distributed genes into five different layers for each pathway, and WINNER expansion added several highly ranked genes to both networks. Additions to the CML network (Figure 6) included JAK1/2/3 and proteins that participate in IL-2 (IL2, IL2RA, and IL2RB), IL-3 (IL-3, IL-3RA, and IL-3RB), and GM-CSF (CSF2) signaling, which is consistent with the JAK2/STAT5 pathway's status as one of the

FIGURE 3

The WINNER ranking p-value (p_r) is robust to the addition of noise (STATS?). Genes in all KEGG pathways were ranked *via* WINNER, and WINNER ranking p-values (p_r) were calculated, after varying degrees of noise were added to the network; then, noise robustness was compared for genes with $p_r < 0.05$ and $p_r \geq 0.05$ by determining the likelihood that the gene's ranking changed by 10 or more upon the addition of noise.

FIGURE 4

WINNER gene prioritization more accurately identifies the relationship between breast-cancer genes and patient survival. Genes in the KEGG breast-cancer pathway were ranked *via* WINNER, PageRank, and Dual Rank, and the significance of each gene's relationship to patient survival was determined with an online Kaplan-Meier plotting tool. **(A)** The proportion of genes that were significantly ($p < 0.05$) related to breast-cancer survival was determined for the top 0-50% of ranked genes. **(B)** The precision of the WINNER ranking of genes for breast-cancer survival (Bland and Altman, 1998) was compared for the top 0–30% of ranked genes with $p_r < 0.05$ and $p_r \geq 0.05$.

primary targets for treatment of CML (Valent, 2014), as well as evidence that STAT5 is phosphorylated by IL-2 (Kobayashi et al., 2014; Valent, 2014) and IL-3 (Jiang et al., 1999) signaling, and that GM-CSF is a crucial growth factor for myeloid cells; notably, several of these molecules are currently being investigated as therapeutic targets for CML treatment (Hercus et al., 2012; Broughton et al., 2014; Kobayashi et al., 2014). For the hepatocellular carcinoma pathway (Figure 7), WINNER expansion added KC1G2, a serine-threonine kinase that can activate TGF-β1/Smad signaling (Guo et al., 2008); TMED4, WLS, and PRCN, which mediate Wnt/β-catenin signaling (Guo et al., 2008; Martin-Orozco et al., 2019; Bland et al., 2021); and several genes for proteins in the FGF signaling pathway (FRS2, FRS3, KLB, and PLCG1; Gotoh, 2008; Gyanchandani et al., 2013; Wang et al., 2020), of which KLB is particularly important, because it functions as a co-receptor for the binding of FGF-19/21 to FGFR-1/4 (Yang et al., 2012). Thus, the genes added to the KEGG CML and hepatocellular carcinoma pathways by WINNER expansion have strong, well-established links to multiple binding partners that participate in the mechanisms associated these diseases.

Besides, WINNER ranking correlation with other ranking techniques, including Ingenuity Pathway Analysis (IPA; Kramer et al., 2014), DIAMOnD (Ghiassian et al., 2015), Random Walk (Smedley et al., 2014), Node2Vec (Grover and Leskovec, 2016; Peng et al., 2019), and GenePANDA (Yin et al., 2017), vary from −0.83 (negatively correlated) to −0.05 (insignificant correlation), then to 0.74 (moderate-positively correlated; Figure 6C). This result suggests that the major difference between WINNER and other techniques' ranking appears when
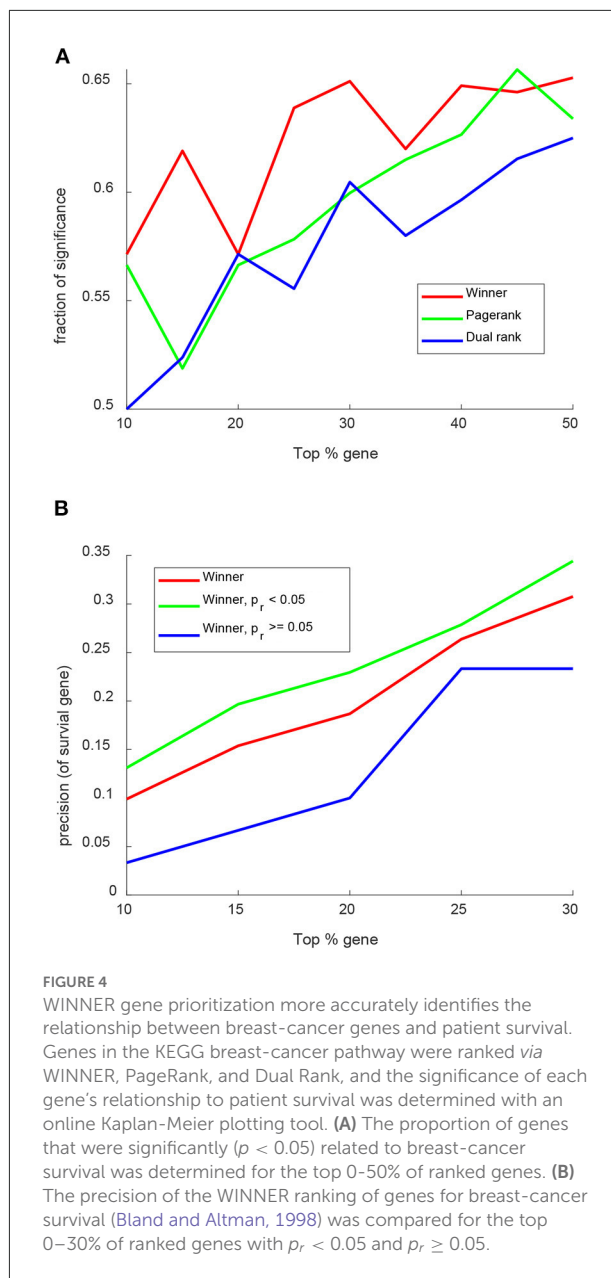
the network expands beyond the seed genes. Thus, a good benchmark among WINNER and other techniques can be performed by a network-expansion scenario.

Benchmarking WINNER ranking by retrieving newly updated genes in KEGG pathways

Gene prioritization algorithms are benchmarked by information retrieval experiments, such as in Guala and Sonnhammer (2017) and Zhang et al. (2021), where some

important regulators are labeled "unknown," and the algorithms are executed to rank these "unknown-labeled" gene such that these regulators are top-ranked. Thus, to benchmark WINNER, we setup the KEGG Pathway retrieval experiment. Here, WINNER took a KEGG pathway release 50 (2009 version; Kanehisa et al., 2010) as the seed genes and gene-gene interactions (expanded network) in HAPPI database (Chen et al., 2017) as the input; the WINNER expansion p-value (p_e) and WINNER score were calculated for candidate genes to include in the KEGG release 50 pathway networks; then, the highly-ranked non-seed (expanded genes) was compared to the same updated pathway network in KEGG release 85 (Ogata et al., 1999; Kanehisa et al., 2017; 2017 version) as the ground-truth. In this experiment, WINNER performance, quantified by precision, recall, and the F1 score, was compared with Ingenuity Pathway Analysis (IPA; Kramer et al., 2014), DIAMOnD (Ghiassian et al., 2015), Random Walk (Smedley et al., 2014), Node2Vec (Grover and Leskovec, 2016; Peng et al., 2019), and GenePANDA (Yin et al., 2017); these techniques were chosen according to Zhang et al. (2021). The same experiment was executed with each KEGG pathway, and the results were aggregated into error bars.

Our results indicated that the WINNER predictions had greater precision but less recall (i.e., the proportion of newly incorporated genes that were retrieved by the prediction) than the predictions generated *via* other comparing methods (Figure 8). The WINNER predictions were also associated with a higher F1 score, which incorporates both precision and recall into a global measure of accuracy, when more than 60% of the extension candidates were examined. Besides, Figure 8 shows that the retrieval recall rate is low (usually <0.2) in all of the algorithms. Precision should be prioritized in comparing the performance among these expansion algorithms.

# WINNER ranking of differentially expressed genes in biological case-studies

WINNER ranking of genes involved in apoptosis and cell-cycle activity

The use of WINNER for prioritizing genes involved in cellular processes was evaluated with the KEGG apoptosis and cell-cycle pathways and node-degree–preserved network randomization. WINNER ranking p-values were highly significant for genes that participate in some of the most essential mechanisms of apoptosis, such as Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit alpha isoform (PIK3CA) ($p_r = 5.01\times10^{-13}$); the Phosphatidylinositol 3-kinase regulatory subunit alpha (P85A; $p_r = 1.34 \times 10^{-12}$) and Cytokine receptor common subunit beta (IL3RB; $p_r = 4.60 \times 10^{-12}$); and genes for several proteins of the cytoskeleton (actin, $p_r = 1.94\times10^{-104}$; Tubulin, $p_r = 1.94\times10^{-104}$; B4DZT3, p_r

$= 8.71\times10^{-87}$; Lamin A/C, $p_r = 8.17\times10^{-87}$; Lamin B1, $p_r = 8.17\times10^{-87}$; actin-G, $p_r = 5.15\times10^{-63}$), which is substantially reorganized to produce the characteristic shrunken morphology of apoptotic cells; notably, actin and actin-binding proteins also initiate and regulate apoptosis (Desouza et al., 2012). However, the KEGG apoptosis pathway also includes genes for a number of proteins that participate IL-3– and NGF-signaling (IL-3, IL-3R, and NGF), which are nonessential (or even irrelevant) for apoptosis, and the ranking p-values calculated for these genes were not significant ($p_r = 0.18$). Similarly, genes in the KEGG cell-cycle pathway that encode proteins directly involved in DNA replication and cell division had highly significant ranking p-values (Cell Division Cycle 14B, $p_r = 9.5\times10^{-297}$ and 14A, $p_r = 2.28\times10^{-22}$) whereas the ranking p-values for genes that participate in TGF-β signaling were nonsignificant (TGF-β, $p_r = 0.29$; SMAD2, $p_r = 0.29$; SMAD3, $p_r = 0.29$; SMAD4, $p_r = 0.29$), which is consistent with the role of TGF-β in cell-proliferation: it interacts with many components of the cell cycle pathway but generally inhibits proliferation in non-mesenchymal cells. Collectively, these observations demonstrate that the WINNER ranking p-value can be a useful guide for distinguishing between genes that are essential or nonessential participants in a particular cellular process.

WINNER ranks important signaling pathway markers in mammalian pig heart regeneration

The hearts of adult mammals cannot regenerate myocardial tissues that are lost to injury; however, when myocardial infarction (MI) was induced in the hearts of one-day-old piglets, the animals recovered with no significant loss of cardiac function and little evidence of myocardial scarring (Zhu et al., 2018). Thus, to identify genes that may contribute to mammalian cardiac regeneration, we used WINNER to rank the list of differentially expressed genes from piglets that had or had not undergone surgically induced MI on postnatal day 1 for a previous report (Zhang et al., 2020; Figure 9, Supplementary Table 1). Here, we used HAPPI version 2 database (Chen et al., 2017) to build the network connecting these genes. The two top-ranked genes (FN1 and JAK3) encoded fibronectin, which is required for cardiac regeneration in zebrafish (Wang et al., 2013), and Janus kinase 3 (JAK3), which has been shown to protect against ischemia-reperfusion injury (Kubin et al., 2011); notably, JAK3 also interacts with oncostatin-M, which is encoded by the tenth-highest WINNER-ranked gene (OSM) and is a primary factor in cardiomyocyte dedifferentiation and remodeling (Singh et al., 2016; Doll et al., 2017). Also among the top 10 were genes encoding subunits of the essential matrix proteins integrin alpha (ITGA8) and beta (ITGB4), which are differentially expressed in adult and fetal cardiac fibroblasts and involved in chamber specification of zebrafish hearts (Singh et al., 2016; Doll et al., 2017), while the 11th-ranked gene, THBS3, encodes another extracellular matrix protein, thrombospontin
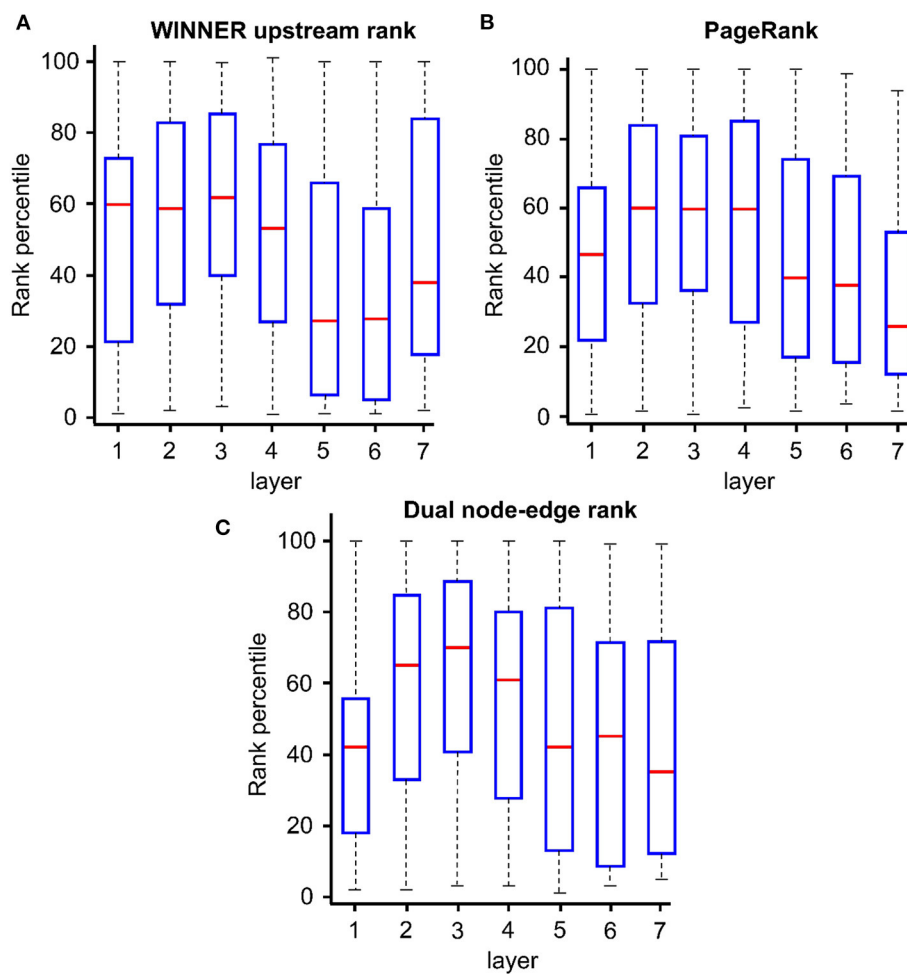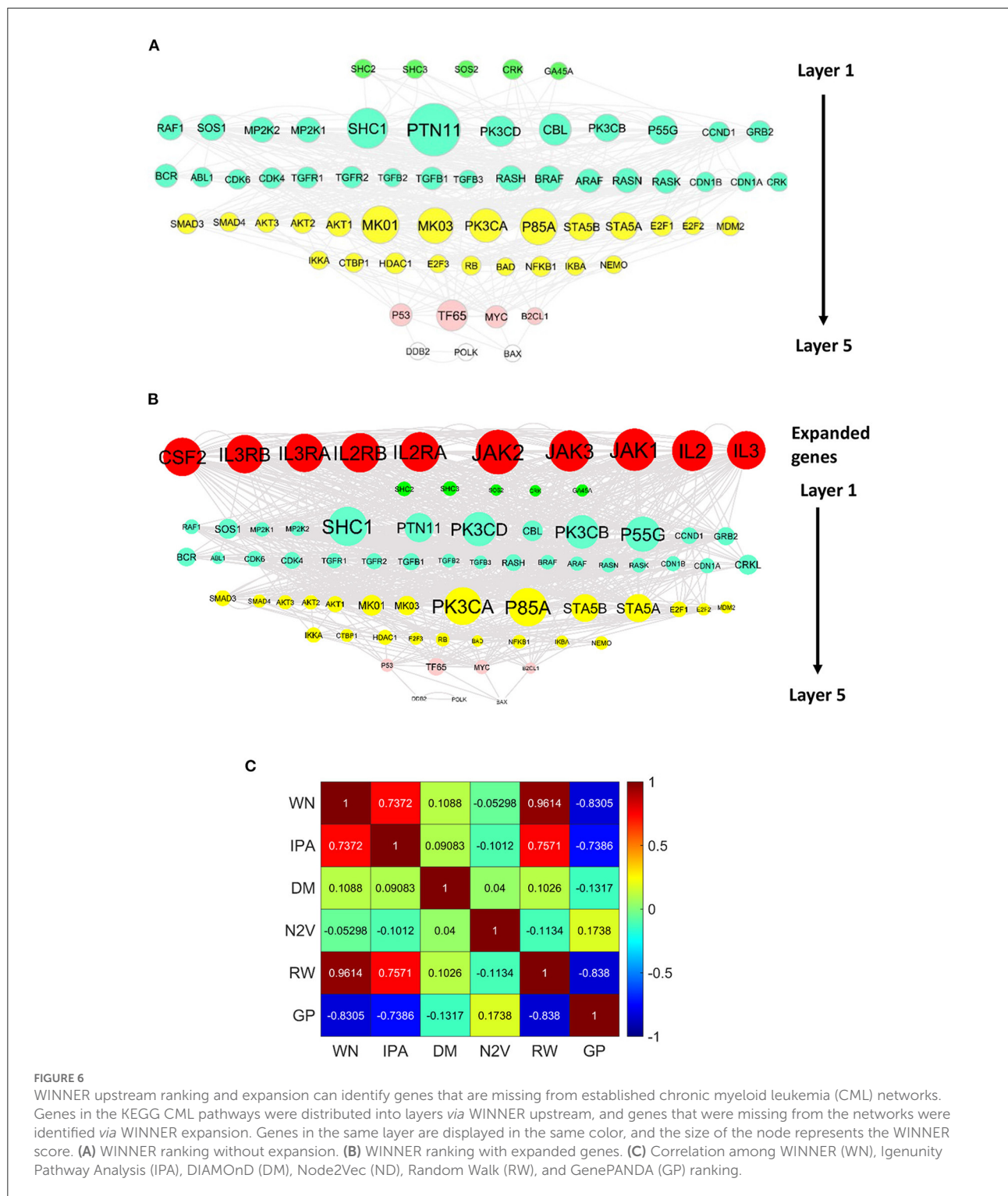
FIGURE 5
WINNER upstream prioritization more accurately identifies the relative position of genes in a pathway. Gene-gene regulatory relationships from STRING v.10.5 were used to distribute genes from all KEGG cancer pathways into 7 layers *via* WINNER (customized for upstream ranking), PageRank, and Dual Rank; genes coding for proteins that function further upstream in the pathways were assigned to the lower-numbered layers. Layers 1−3 are the most upstream layers, usually correspond to the kineases, grow factors, and receptors. Layers 4−7 are downstream, usually correspond to signaling hubs, phospholization, transcription factors, and inside-nucleus genes. The y axis indicates the ranking scores, which were converted into percentile so that the rankings across different pathways could be combined into one boxplot. The red cross implies boxplot outliners (beyond 2 and 98% percentiles). **(A)** WINNER upstream rank. **(B)** PageRank. **(C)** Dual node-edge rank.

3, which is a critical [and clinically relevant (Mustonen et al., 2013)] regulator of cell-cell and cell-matrix signaling that appears to impede integrin function and contribute to injury-induced cardiomyopathy in mice (Costa et al., 2014; Porrello and Olson, 2014; Puente et al., 2014). Other genes ranked among the top 20 by WINNER included the nitrous-oxide–related genes NCF2 and NCF4, and the gene for vasopressin 2 (AVPR2), which collectively modulate the cellular environment to promote cardiac regeneration (Costa et al., 2014; Porrello and Olson, 2014; Puente et al., 2014); ERBB3, which encodes a tyrosine kinase that appears to be crucial for embryonic development (Erickson et al., 1997); and genes for a dynamin protein (DNM1) and a Rho GTPase (RND2), which suggests that at least some of the mechanisms of

mammalian myocardial regeneration are mediated by vesicle-based signaling.

WINNER ranking reflects the important genes supported by co-citations and reveals the upstream events in the r-type pathway-to-pathway network in triple negative breast cancer (TNBC) study
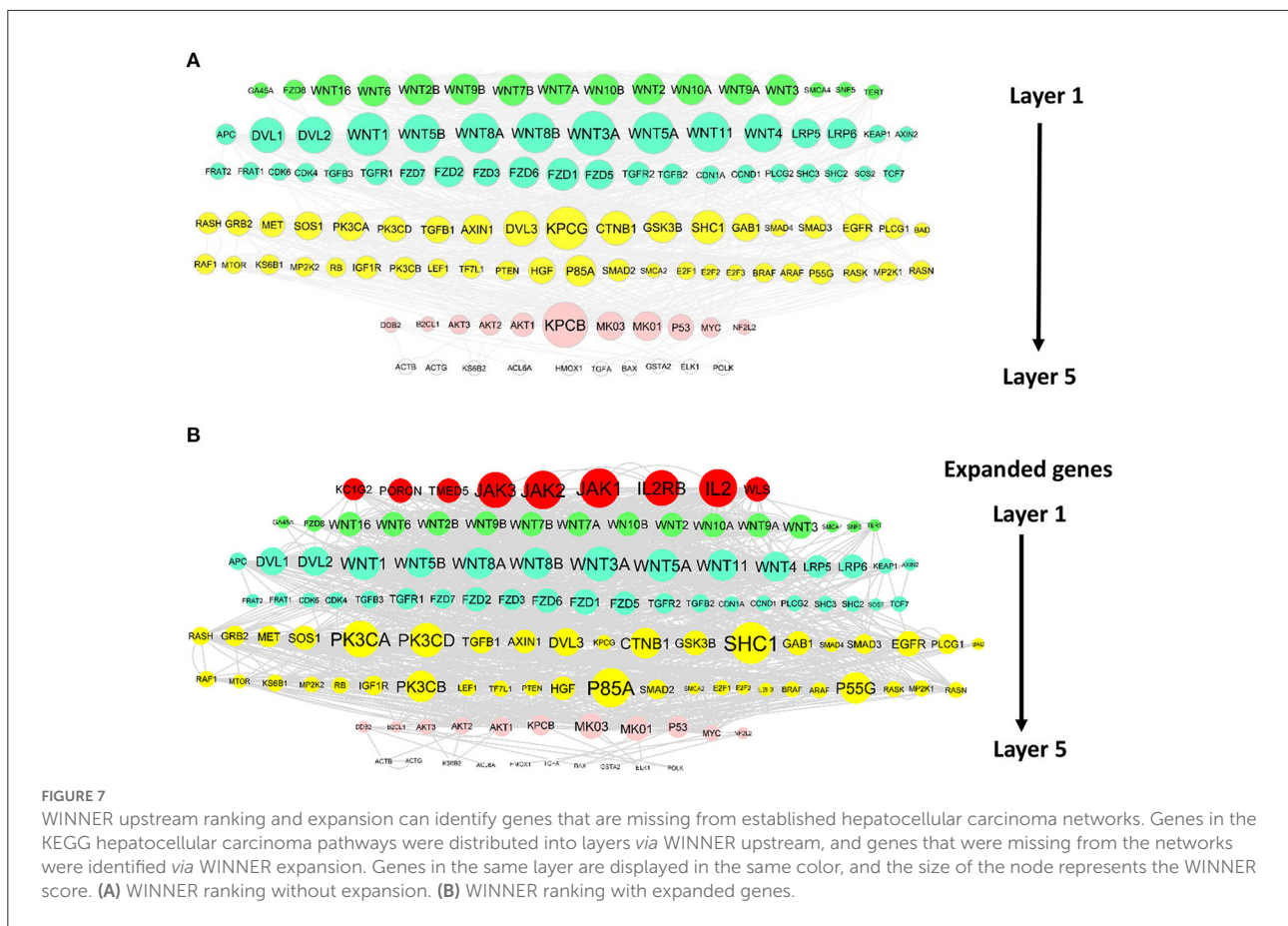
We found 72 significant genes ranked by WINNER using p-value ≤ 0.05 with the WINNER score ranging from 7.4 to 92.5, and the left nonsignificant genes' WINER score ranges from 0 to 68.7. The co-citations analysis shows that the "triple negative breast cancer" co-citations between the significant

FIGURE 6
WINNER upstream ranking and expansion can identify genes that are missing from established chronic myeloid leukemia (CML) networks. Genes in the KEGG CML pathways were distributed into layers *via* WINNER upstream, and genes that were missing from the networks were identified *via* WINNER expansion. Genes in the same layer are displayed in the same color, and the size of the node represents the WINNER score. **(A)** WINNER ranking without expansion. **(B)** WINNER ranking with expanded genes. **(C)** Correlation among WINNER (WN), Igenunity Pathway Analysis (IPA), DIAMOnD (DM), Node2Vec (ND), Random Walk (RW), and GenePANDA (GP) ranking.

ranked genes and the nonsignificant ranked genes have significant difference with Kruskal Wallis test's p-value = 0.027 (Figure 10). The result suggests that WINNER's high-rank genes are more likely lead to biological insights than the WINNER's low-rank genes.

To explore new insights among the high-ranking genes, we performed pathway analysis and built the pathway-to-pathway regulatory networks from these genes using PAGER tool (Yue et al., 2018). The WINNER significantly ranked genes regulated many implicated pathways and processes for TNBC. Thus, we

FIGURE 7
WINNER upstream ranking and expansion can identify genes that are missing from established hepatocellular carcinoma networks. Genes in the KEGG hepatocellular carcinoma pathways were distributed into layers *via* WINNER upstream, and genes that were missing from the networks were identified *via* WINNER expansion. Genes in the same layer are displayed in the same color, and the size of the node represents the WINNER score. **(A)** WINNER ranking without expansion. **(B)** WINNER ranking with expanded genes.

observed the higher ranked gene enriched pathways are more likely to be at upstream side of the regulatory (r-type) enriched pathway-to-pathway network. In general, the add-on pathway levels were positive correlated to the ranked gene bins with Pearson correlation coefficient equal to 0.74 (Figure 11).

We found that the top ranked genes, TOP2A, CDK1, PLK1, and UBE2C, were enriched in the cell cycle related pathways, such as "Phosphorylation of Cyclin B1 in the CRS domain," "Regulation of mitotic cell cycle," "Mitotic Metaphase and Anaphase," and "Free APC/C phosphorylated by Plk1."

Topoisomerase II a (TOP2A) can be a useful gene in determining whether TNBC patients would have a good response to anthracycline therapy, which is the mainstay treatment in TNBC cancer (Brase et al., 2010; Di Leo et al., 2011; Eltohamy et al., 2018). Both Eltohamy et al. and Di Leo et al. found that patients with aberrant expression of TOP2A have better response to anthracycline treatment (Di Leo et al., 2011; Eltohamy et al., 2018).

Cyclin dependent kinase 1 (CDK1) play a critical role how the cell cycle is regulated, specifically during mitosis. Liu et al. used nanoparticles with siRNA to target CDK1, and it has been found to successfully inhibit the TNBC cell line that has been injected in mice (Liu et al., 2014). Xia et al. has found that the

CDK1 inhibitor can inhibit the growth of the TNBC cells by arresting them in the G2/M cell phase (Xia et al., 2014).

Polo like kinase-1 (PLK1) has been found to be one of the key regulators in the cell cycle. Targeting and knocking out of PLK1 has been found to cause the TNBC tumor cells to be arrested in the G2-M cell cycle (Ueda et al., 2019; Zhao et al., 2021; Patel et al., 2022). Morray et al. found that a nanoparticle with siRNA targeting PLK1 can inhibit growth in the TNBC tumor cell line (Morry et al., 2017). Patel et al. used the allosteric inhibitor RK-10 to target the PLK1 in TNBC cell lines, and it has inhibited growth through the S phase and G2/M (Patel et al., 2022).

Overexpression of Ubiquitin-conjugated enzyme (UBE2C) can play a role in the pathogenesis of TNBC (Chou et al., 2014; Kim et al., 2019). Chou et al had found that UBE2C has been highly expressed in cancer tissue cells, and that when UBE2C has been targeted with siRNA, the tumor cells have stopped proliferating (Chou et al., 2014).

Discussion and conclusion

In this paper, we introduce WINNER, a new network-based ranking tool that addresses several of the limitations associated with other gene prioritization techniques. Our
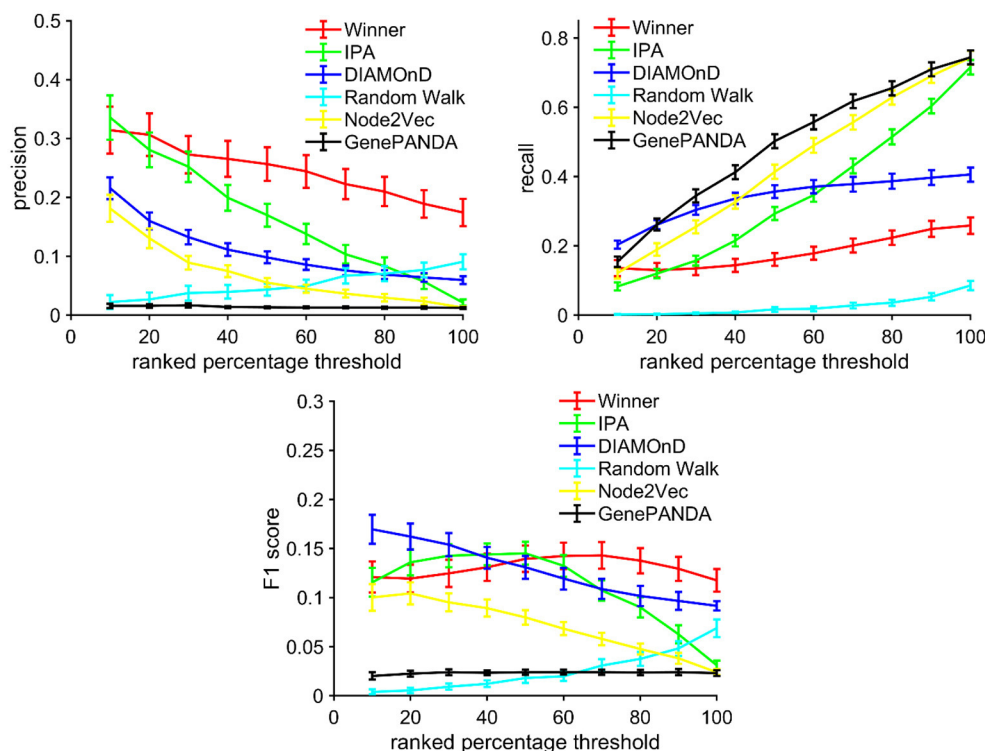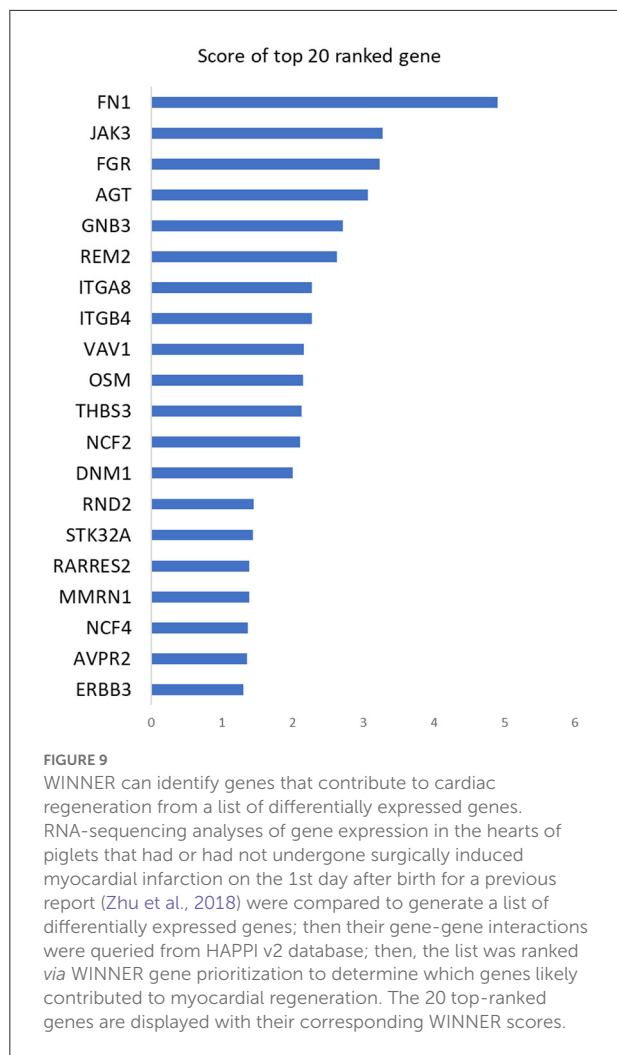
FIGURE 8
Benchmark: WINNER expansion more accurately identifies the addition of new genes to established networks. The pathway networks in KEGG (https://www.genome.jp/kegg/network.html) release 50 was expanded *via* WINNER (i.e., calculation of the WINNER expansion *p*-value), Ingenuity Pathway Analysis (IPA), DIAMOND, Random Walk, Node2Vec, and GenePANDA. Then, the expanded networks were compared to the updated network in KEGG release 85 to determine the precision, recall, and F1 scores for each expansion technique.

novel use of node-degree–preserved and modularity-preserved randomization produced randomized networks that retained some of the original network topology and were more normally distributed, which increased the precision and robustness of our ranking *p*-value (p_r) calculations, while the expansion *p*-value (p_e) better accommodated the incomprehensiveness and redundancy of the input gene list. However, WINNER rankings were not well-correlated with the clustering coefficient, which represents the presence of network cliques (Newman, 2008; i.e., semi-isolated groups of genes that collectively function like a single node), which suggests that WINNER ranking may be somewhat compromised in dense networks, such as those containing families of proteins, where the scale-free property (Timar et al., 2016) does not apply. Nevertheless, many biological networks are scale-free (Khanin and Wit, 2006), and since degree-preserved randomization tends to produce near-normal ranking distributions, the WINNER p_r value is likely more accurate than the empirical *p*-value, even for networks that are not perfectly scale-free.

WINNER network ranking belongs to the "eigenvector ranking" (Newman, 2008) class of algorithm. Therefore, it has the same "big-O" computational cost to PageRank [$O(N^3)$,

where N is the number of network genes] if implemented using iterative matrix multiplication. However, this class of algorithm can be implemented in parallel, which significantly reduced the computational time in practice.

The performance of gene network prioritization significantly depends on the disease (Zhang et al., 2021), or the biological case-study. Therefore, we demonstrate WINNER's performance in various disease and biological study scenarios. The comprehensive KEGG pathway results reflect the case when lacking biological samples and expression data. Then, prioritization needs to be performed only using the domain-knowledge available network to generate hypotheses. Cardiac regeneration, which focuses on cardiomyocyte proliferation, case-study is an example when a significant biological process, not a disease, that does not naturally happen in matured mammals (Porrello et al., 2011; Lam and Sadek, 2018; Ye et al., 2018; Zhu et al., 2018; Zhao et al., 2020; Nakada et al., 2021; Nguyen et al., 2022). In this case, the focus is finding the regulating mechanism to create new cells and to apply this knowledge in biomedical engineering research. Cancer and other disease case studies (leukemia, TNBC, and Vitamin D) are directly related to the disease, and targeted therapies to kill cells are available or proposed. In this case, the focus is to find

FIGURE 9
WINNER can identify genes that contribute to cardiac regeneration from a list of differentially expressed genes. RNA-sequencing analyses of gene expression in the hearts of piglets that had or had not undergone surgically induced myocardial infarction on the 1st day after birth for a previous report (Zhu et al., 2018) were compared to generate a list of differentially expressed genes; then their gene-gene interactions were queried from HAPPI v2 database; then, the list was ranked *via* WINNER gene prioritization to determine which genes likely contributed to myocardial regeneration. The 20 top-ranked genes are displayed with their corresponding WINNER scores.

FIGURE 10
The literature validation of triple negative breast cancer genes using co-citations from PubMed. The co-citations of gene and TNBC are grouped by the WINNER reported *p*-values. The non-significant gene *p*-values are larger than 0.05 in WINNER, and the significant gene *p*-values are $\leq$0.05 in WINNER. The Kruskal Wallis test *p*-value is 0.027.

FIGURE 11
The correlation between the add-on pathways enriched in the top 2^x bins and the bin size. The violin plot shows the pathway level distribution. The red points connected by solid red lines are the means of pathway levels.

markers, especially the "cell-killer ones" associated with the disease outcomes, and there is less emphasis rather than the regulating growing mechanism. WINNER results are insightful in all of these cases, whereas whether other techniques have insightful results is yet to be examined in multiple studies.

In conclusion, WINNER gene prioritization is generally more accurate and robust than other network-based prioritization techniques, such as PageRank and node-degree ranking, and can be effective for identifying genes that may be missing from established gene networks, for determining the relative position (i.e., upstream or downstream) of genes within a pathway, and for ranking a list of differentially expressed genes. The superior performance is linked to better retrieval precision when expanding the network among the seed genes. The important case studies presented in this work are in a scenario where new disease-specific gene-expression data were generated, and novel genes associated with the disease and phenotype are expected. Then, network expansion is required. In this expansion, WINNER emphasizes precision, where only
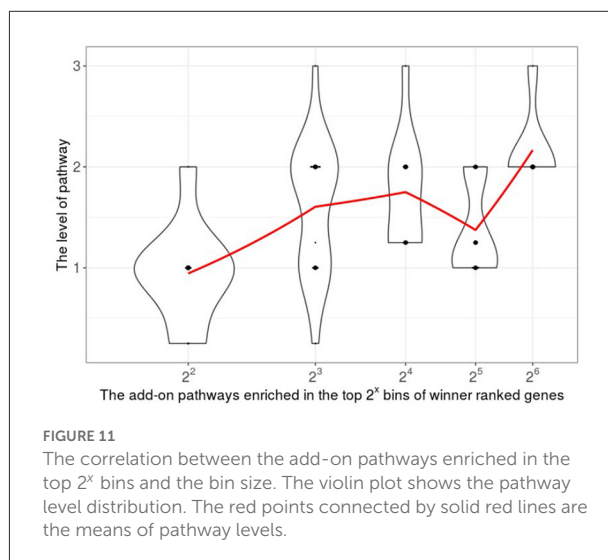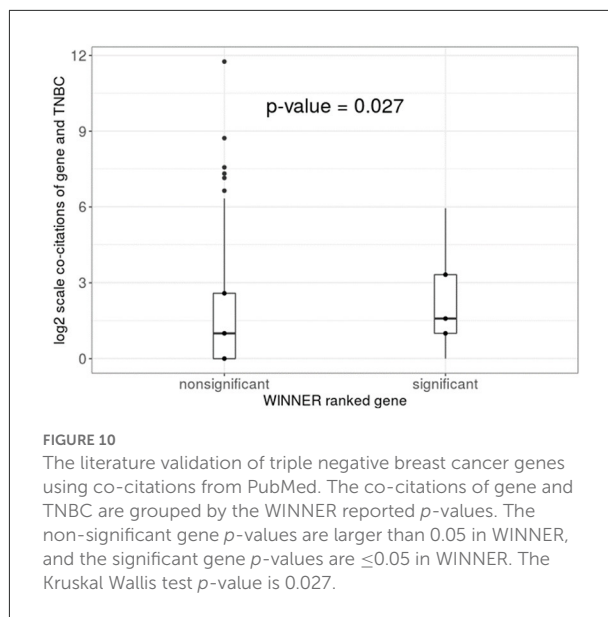
a small expanded but highly relevant candidates are explored, over recall, where more comprehensive candidate genes were explored but may involve many irrelevant ones. Other methods tend to emphasize recall; therefore, they may computationally retrieve more candidates; however, at the same time, make it much more difficult for the user to choose the rightly relevant ones. Also, having too many irrelevant genes in the network significantly affects the ranks of the well-known disease-specific genes. This scenario explains the advantage of WINNER over other methods. Future investigations are warranted to determine what additional biological insights can be obtained by using WINNER to rank genes that participate in other cellular processes, in metabolic regulatory pathways (Berkhout et al., 2013), and in co-expression networks (Radulescu et al., 2018).

Data availability statement

The gene expression data used in this work are publicly available at the Gene Expression Omnibus database, accession number GSE144883, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE144883.

Author contributions

TN developed the algorithm, performed case studies, and wrote the manuscript. ZY and RS performed case studies and performed the literature validation of the results. ZY built the website. RW and JZ provided data and participated in the case studies. JC conceptualized the ideas, helped design the analytical experiments, and revised the final manuscript. All authors read, edited, and approved the manuscript.

Funding

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdata.2022.1016606/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Schematic diagrams of WINNER gene prioritization and network expansion. **(a)** Seeded genes (green) and candidate expansion genes (yellow) are assembled into a network as indicated by their pairwise interactions. **(b)** The expansion p-value (p_e) are calculated among the expansion-candidate genes, then genes with $p_e < 0.05$ will be further evaluate and added into and expand the network, one gene at a time. Then **(c)** the expansion score **(e)** are calculated for the candidate expansion genes; then, the highest-scored gene is added to the network; this process is repeated until all candidates are added or being halted (not adding all candidates). And **(d)**, after completing the expansion, the statistical significance of the rankings are recalculated for the expanded network.

SUPPLEMENTARY FIGURE 2
WINNER filtering of candidate genes for network expansion. Red nodes represent seeded genes, open nodes represent candidate expansion genes, black lines represent interactions between two seeded genes, and gray lines represent interactions between one seeded gene and one expansion gene or between two expansion genes. Candidate genes for network expansion were filtered *via* two tests: (1) the likelihood of the candidate expansion gene (E.Gene) having a seeded interaction relative to its total number of interactions (bottom left table), and (2) the likelihood of the candidate expansion gene having seeded interactions relative to the seeded interactions of its most similar seeded gene (S.Gene), with similarity determined by node degree (bottom right table).

SUPPLEMENTARY FIGURE 3
WINNER ranking of the network of Alzheimer's disease pathways in KEGG release 50. The network graph was constructed with Cytoscape (Shannon et al., 2003) version 3.6.0 and the force-directed layout; the size of the node represents the WINNER score.

SUPPLEMENTARY TABLE 1
WINNER ranking for genes in cardiac regeneration dataset. The table includes gene symbol, the indication of whether a gene is a seeded (S) or expanded (E) gene, and WINNER score.

SUPPLEMENTARY TABLE 2
Gene-gene interaction network in the cardiac regeneration dataset.

SUPPLEMENTARY TABLE 3
WINNER ranking for genes in triple negative breast cancer (TNBC) dataset. The table includes gene symbol, the indication of whether a gene is a seeded (S) or expanded (E) gene, WINNER score, and p-value.

SUPPLEMENTARY TABLE 4
Gene-gene interaction network in triple negative breast cancer (TNBC) dataset.

SUPPLEMENTARY VIDEO 1
The .cys (cytoscape) file of the regulatory (r-type) pathway-to-pathway network in the triple negative breast cancer study.

References

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., et al. (2006). Gene prioritization through genomic data fusion. *Nat. Biotechnol.* 24, 537–544. doi: 10.1038/nbt 1203

Alvarez-Ponce, D., Lopez, P., Bapteste, E., and McInerney, J. O. (2013). Gene similarity networks provide tools for understanding eukaryote origins and evolution. *Proc. Natl. Acad. Sci. U. S. A.* 110, E1594–1603. doi: 10.1073/pnas.1211371110

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169. doi: 10.1093/bioinformatics/btu638

Antanaviciute, A., Daly, C., Crinnion, L. A., Markham, A. F., Watson, C. M., Bonthron, D. T., et al. (2015). GeneTIER: prioritization of candidate disease genes using tissue-specific gene expression profiles. *Bioinformatics* 31, 2728–2735. doi: 10.1093/bioinformatics/btv196

Beissbarth, T., and Speed, T. P. (2004). GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* 20, 1464–1465. doi: 10.1093/bioinformatics/bth088

Berkhout, J., Teusink, B., and Bruggeman, F. J. (2013). Gene network requirements for regulation of metabolic gene expression to a desired state. *Sci. Rep.* 3, 1417. doi: 10.1038/srep01417

Bian, W., Chen, W., Nguyen, T., Zhou, Y., and Zhang, J. (2021). miR-199a overexpression enhances the potency of human induced-pluripotent stem-cell-derived cardiomyocytes for myocardial repair. *Front. Pharmacol.* 12, 673621. doi: 10.3389/fphar.2021.673621

Bland, J. M., and Altman, D. G. (1998). Survival probabilities (the Kaplan-Meier method). *Br. Med. J.* 317, 1572. doi: 10.1136/bmj.317.7172.1572

Bland, T., Wang, J., Yin, L., Pu, T., Li, J., Gao, J., et al. (2021). WLS-Wnt signaling promotes neuroendocrine prostate cancer. *iScience* 24, 101970. doi: 10.1016/j.isci.2020.101970

Bowman, A. W., and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations, vol. 18.* Oxford: Oxford University Press.

Brase, J. C., Schmidt, M., Fischbach, T., Sultmann, H., Bojar, H., Koelbl, H., et al. (2010). ERBB2 and TOP2A in breast cancer: a comprehensive analysis of gene amplification, RNA levels, and protein expression and their influence on prognosis and prediction. *Clin. Cancer Res.* 16, 2391–2401. doi: 10.1158/1078-0432.CCR-09-2471

Bromberg, Y. (2013). Chapter 15: disease gene prioritization. *PLoS Comput. Biol.* 9, e1002902. doi: 10.1371/journal.pcbi.1002902

Broughton, S. E., Hercus, T. R., Hardy, M. P., McClure, B. J., Nero, T. L., Dottore, M., et al. (2014). Dual mechanism of interleukin-3 receptor blockade by an anti-cancer antibody. *Cell Rep.* 8, 410–419. doi: 10.1016/j.celrep.2014.06.038

Cantor, R. M., Lange, K., and Sinsheimer, J. S. (2010). Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *Am. J. Hum. Genet.* 86, 6–22. doi: 10.1016/j.ajhg.2009.11.017

Chandrashekar, D. S., Karthikeyan, S. K., Korla, P. K., Patel, H., Shovon, A. R., Athar, M., et al. (2022). UALCAN: an update to the integrated cancer data analysis platform. *Neoplasia* 25, 18–27. doi: 10.1016/j.neo.2022.01.001

Chatr-Aryamontri, A., Breitkreutz, B. J., Heinicke, S., Boucher, L., Winter, A., Stark, C., et al. (2013). The BioGRID interaction database: 2013 update. *Nucleic Acids Res.* 41, D816–823. doi: 10.1093/nar/gks1158

Chen, J., Bardes, E. E., Aronow, B. J., and Jegga, A. G. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* 37, W305–311. doi: 10.1093/nar/gkp427

Chen, J. Y., Mamidipalli, S., and Huan, T. (2009). HAPPI" an online database of comprehensive human annotated and predicted protein interactions. *BMC Genomics* 10, S16. doi: 10.1186/1471-2164-10-S1-S16

Chen, J. Y., Pandey, R., and Nguyen, T. M. (2017). HAPPI-2: a comprehensive and high-quality map of human annotated and predicted protein interactions. *BMC Genomics* 18, 182. doi: 10.1186/s12864-017-3512-1

Chen, J. Y., Pinkerton, S. L., Shen, C., and Wang, M. (2006b). "An integrated computational proteomics method to extract protein targets for fanconi anemia studies," in *21st Annual ACM Symposium on Applied Computing.* Dijon, 173–179. doi: 10.1145/1141277.1141316

Chen, J. Y., Piquette-Miller, M., and Smith, B. P. (2013). Network medicine: finding the links to personalized therapy. *Clin. Pharmacol. Therapeut.* 94, 613–616. doi: 10.1038/clpt.2013.195

Chen, J. Y., Shen, C., and Sivachenko, A. Y. (2006a). Mining Alzheimer disease relevant proteins from integrated protein interactome data. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing* 2006, 367–378. doi: 10.1142/9789812701626_0034

Chou, C. P., Huang, N. C., Jhuang, S. J., Pan, H. B., Peng, N. J., Cheng, J. T., et al. (2014). Ubiquitin-conjugating enzyme UBE2C is highly expressed in breast microcalcification lesions. *PLoS ONE* 9, e93934. doi: 10.1371/journal.pone.0093934

Cornish, A. J., David, A., and Sternberg, M. J. E. (2018). PhenoRank: reducing study bias in gene prioritization through simulation. *Bioinformatics* 34, 2087–2095. doi: 10.1093/bioinformatics/bty028

Costa, A., Rossi, E., Scicchitano, B. M., Coletti, D., Moresi, V., Adamo, S., et al. (2014). Neurohypophyseal hormones: novel actors of striated muscle development and homeostasis. *Eur. J. Transl. Myol.* 24, 3790. doi: 10.4081/bam.2014.3.217

Cowen, L., Ideker, T., Raphael, B. J., and Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.* 18, 551–562. doi: 10.1038/nrg.2017.38

D'Amelio, M., Cavallucci, V., Middei, S., Marchetti, C., Pacioni, S., Ferri, A., et al. (2011). Caspase-3 triggers early synaptic dysfunction in a mouse model of Alzheimer's disease. *Nat. Neurosci.* 14, 69–76. doi: 10.1038/nn.2709

Desouza, M., Gunning, P. W., and Stehn, J. R. (2012). The actin cytoskeleton as a sensor and mediator of apoptosis. *Bioarchitecture* 2, 75–87. doi: 10.4161/bioa.20975

Di Leo, A., Desmedt, C., Bartlett, J. M., Piette, F., Ejlertsen, B., Pritchard, K. I., et al. (2011). HER2 and TOP2A as predictive markers for anthracycline-containing chemotherapy regimens as adjuvant treatment of breast cancer: a meta-analysis of individual patient data. *Lancet Oncol.* 12, 1134–1142. doi: 10.1016/S1470-2045(11)70231-5

do Valle, I. F., Menichetti, G., Simonetti, G., Bruno, S., Zironi, I., Durso, D. F., et al. (2018). Network integration of multi-tumour omics data suggests novel targeting strategies. *Nat. Commun.* 9, 4514. doi: 10.1038/s41467-018-06992-7

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Doll, S., Dressen, M., Geyer, P. E., Itzhak, D. N., Braun, C., Doppler, S. A., et al. (2017). Region and cell-type resolved quantitative proteomic map of the human heart. *Nat. Commun.* 8, 1469. doi: 10.1038/s41467-017-01747-2

ElShal, S., Tranchevent, L. C., Sifrim, A., Ardeshirdavani, A., Davis, J., and Moreau, Y. (2016). Beegle: from literature mining to disease-gene discovery. *Nucleic Acids Res.* 44, e18. doi: 10.1093/nar/gkv905

Eltohamy, M. I., Badawy, O. M., El kinaai, N., Loay, I., Nassar, H. R., Allam, R. M., et al. (2018). Topoisomerase II alpha gene alteration in triple negative breast cancer and its predictive role for anthracycline-based chemotherapy (Egyptian NCI Patients). *Asian Pac. J. Cancer Prev.* 19, 3581–3589. doi: 10.31557/APJCP.2018.19.12.3581

Erickson, S. L., O'Shea, K. S., Ghaboosi, N., Loverro, L., Frantz, G., Bauer, M., et al. (1997). ErbB3 is required for normal cerebellar and cardiac development: a comparison with ErbB2-and heregulin-deficient mice. *Development* 124, 4999–5011. doi: 10.1242/dev.124.24.4999

Erten, S., Bebek, G., Ewing, R. M., and Koyuturk, M. (2011). DADA: degree-aware algorithms for network-based disease gene prioritization. *BioData Min.* 4, 19. doi: 10.1186/1756-0381-4-19

Eschenhagen, T., Bolli, R., Braun, T., Field, L. J., Fleischmann, B. K., Frisen, J., et al. (2017). Cardiomyocyte regeneration: a consensus statement. *Circulation* 136, 680–686. doi: 10.1161/CIRCULATIONAHA.117.029343

Espinoza, M. (2012). *On Network Randomization Methods: A Negative Control Study.* Fairfield, CT: Fairfield University.

Gene Ontology, C., Blake, J. A., Dolan, M., Drabkin, H., Hill, D. P., Li, N., et al. (2013). Gene Ontology annotations and resources. *Nucleic Acids Res.* 41, D530–535. doi: 10.1093/nar/gks1050

Ghiassian, S. D., Menche, J., and Barabasi, A. L. A. (2015). DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput. Biol.* 11, e1004120. doi: 10.1371/journal.pcbi.1004120

Gotoh, N. (2008). Regulation of growth factor signaling by FRS2 family docking/scaffold adaptor proteins. *Cancer Sci.* 99, 1319–1325. doi: 10.1111/j.1349-7006.2008.00840.x

Gottlieb, A., Magger, O., Berman, I., Ruppin, E., and Sharan, R. (2011). PRINCIPLE: a tool for associating genes with diseases *via* network propagation. *Bioinformatics* 27, 3325–3326. doi: 10.1093/bioinformatics/btr584

Grover, A., and Leskovec, J. (2016). "node2vec: scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, CA), 855–864. doi: 10.1145/2939672.2939754

Guala, D., Sjolund, E., and Sonnhammer, E. L. (2014). MaxLink: network-based prioritization of genes tightly linked to a disease seed set. *Bioinformatics* 30, 2689–2690. doi: 10.1093/bioinformatics/btu344

Guala, D., and Sonnhammer, E. L. L. (2017). A large-scale benchmark of gene prioritization methods. *Sci. Rep.* 7, 46598. doi: 10.1038/srep46598

Guney, E., and Oliva, B. (2012). Exploiting protein-protein interaction networks for genome-wide disease-gene prioritization. *PLoS ONE* 7, e43557. doi: 10.1371/journal.pone.0043557

Guo, X., Waddell, D. S., Wang, W., Wang, Z., Liberati, N. T., Yong, S., et al. (2008). Ligand-dependent ubiquitination of Smad3 is regulated by casein kinase 1 gamma 2, an inhibitor of TGF-beta signaling. *Oncogene* 27, 7235–7247. doi: 10.1038/onc.2008.337

Gyanchandani, R., Ortega Alves, M. V., Myers, J. N., and Kim, S. (2013). A proangiogenic signature is revealed in FGF-mediated bevacizumab-resistant head and neck squamous cell carcinoma. *Mol. Cancer Res.* 11, 1585–1596. doi: 10.1158/1541-7786.MCR-13-0358

Gyorffy, B., Lanczky, A., Eklund, A. C., Denkert, C., Budczies, J., Li, Q., et al. (2010). An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Res. Treat.* 123, 725–731. doi: 10.1007/s10549-009-0674-9

Hale, P. J., Lopez-Yunez, A. M., and Chen, J. Y. (2012). Genome-wide meta-analysis of genetic susceptible genes for Type 2 Diabetes. *BMC Syst. Biol.* 6(Suppl.3), S16. doi: 10.1186/1752-0509-6-S3-S16

Hercus, T. R., Broughton, S. E., Ekert, P. G., Ramshaw, H. S., Perugini, M., Grimbaldeston, M., et al. (2012). The GM-CSF receptor family: mechanism of activation and implications for disease. *Growth Fact.* 30, 63–75. doi: 10.3109/08977194.2011.649919

Huang, de. W., Sherman, B. T., and Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37, 1–13. doi: 10.1093/nar/gkn923

Huang, H., Li, J., and Chen, J. Y. (2009). "Disease gene-fishing in molecular interaction networks: a case study in colorectal cancer," in *Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Conference* (Minneapolis, MN), 6416–6419.

Jiang, X., Lopez, A., Holyoake, T., Eaves, A., and Eaves, C. (1999). Autocrine production and action of IL-3 and granulocyte colony-stimulating factor in chronic myeloid leukemia. *Proc. Natl. Acad. Sci. U. S. A.* 96, 12804–12809. doi: 10.1073/pnas.96.22.12804

Johnston, S. R. (2006). Targeting downstream effectors of epidermal growth factor receptor/HER2 in breast cancer with either farnesyltransferase inhibitors or mTOR antagonists. *Int. J. Gynecol. Cancer* 16(Suppl.2), 543–548. doi: 10.1111/j.1525-1438.2006.00692.x

Jonsson, T., Atwal, J. K., Steinberg, S., Snaedal, J., Jonsson, P. V., Bjornsson, S., et al. (2012). A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature* 488, 96–99. doi: 10.1038/nature11283

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361. doi: 10.1093/nar/gkw1092

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* 38, D355–360. doi: 10.1093/nar/gkp896

KEGG (2022). *Chronic Myeloid Leukemia - Homo Sapiens (Human)*. Kyoto: Human Genome Center, Institute of Medical Science; University of Tokyo, Bioinformatics Center; Institute for Chemical Research, Kyoto University.

Khanin, R., and Wit, E. (2006). How scale-free are biological networks. *J. Comput. Biol.* 13, 810–818. doi: 10.1089/cmb.2006.13.810

Kim, J., and Bang, H. (2016). Three common misuses of *P*-values. *Dent. Hypotheses* 7, 73–80. doi: 10.4103/2155-8213.190481

Kim, Y. J., Lee, G., Han, J., Song, K., Choi, J. S., Choi, Y. L., et al. (2019). UBE2C overexpression aggravates patient outcome by promoting estrogen-dependent/independent cell proliferation in early hormone receptor-positive and HER2-negative breast cancer. *Front. Oncol.* 9, 1574. doi: 10.3389/fonc.2019.01574

Kobayashi, C. I., Takubo, K., Kobayashi, H., Nakamura-Ishizu, A., Honda, H., Kataoka, K., et al. (2014). The IL-2/CD25 axis maintains distinct subsets of chronic myeloid leukemia-initiating cells. *Blood* 123, 2540–2549. doi: 10.1182/blood-2013-07-517847

Koschmann, J., Bhar, A., Stegmaier, P., Kel, A. E., and Wingender, E. (2015). "Upstream analysis": an integrated promoter-pathway analysis approach to causal interpretation of microarray data. *Microarrays* 4, 270–286. doi: 10.3390/microarrays4020270

Krallinger, M., Valencia, A., and Hirschman, L. (2008). Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol.* 9(Suppl.2), S8. doi: 10.1186/gb-2008-9-s2-s8

Kramer, A., Green, J., and Pollard, J. Jr., and Tugendreich, S. (2014). Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* 30, 523–530. doi: 10.1093/bioinformatics/btt703

Krueger, F. (2015). *Trim galore. A wrapper tool around Cutadapt and FastQC to Consistently Apply Quality and Adapter Trimming to FastQ Files*, 516.

Kubin, T., Poling, J., Kostin, S., Gajawada, P., Hein, S., Rees, W., et al. (2011). Oncostatin M is a major mediator of cardiomyocyte dedifferentiation and remodeling. *Cell Stem Cell.* 9, 420–432. doi: 10.1016/j.stem.2011.08.013

La Bella, V., Liguori, M., Cittadella, R., Settipani, N., Piccoli, T., Manna, I., et al. (2004). A novel mutation (Thr116Ile) in the presenilin 1 gene in a patient with early-onset Alzheimer's disease. *Eur. J. Neurol.* 11, 521–524. doi: 10.1111/j.1468-1331.2004.00828.x

Lam, N. T., and Sadek, H. A. (2018). Neonatal heart regeneration: comprehensive literature review. *Circulation* 138, 412–423. doi: 10.1161/CIRCULATIONAHA.118.033648

Lanczky, A., Nagy, A., Bottai, G., Munkacsy, G., Szabo, A., Santarpia, L., et al. (2016). miRpower: a web-tool to validate survival-associated miRNAs utilizing expression data from 2178 breast cancer patients. *Breast Cancer Res. Treat.* 160, 439–446. doi: 10.1007/s10549-016-4013-7

Li, J., Zhu, X., and Chen, J. Y. (2009). Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts. *PLoS Comput. Biol.* 5, e1000450. doi: 10.1371/journal.pcbi.1000450

Li, R., and Campos, J. (2015). Iida J: a gene regulatory program in human breast cancer. *Genetics* 201, 1341–1348. doi: 10.1534/genetics.115.180125

Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J. P., Tamayo, P., et al. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* 1, 417–425. doi: 10.1016/j.cels.2015.12.004

Liu, Y., Liang, Y., and Wishart, D. (2015). PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. *Nucleic Acids Res.* 43, W535–542. doi: 10.1093/nar/gkv383

Liu, Y., Zhu, Y. H., Mao, C. Q., Dou, S., Shen, S., Tan, Z. B., et al. (2014). Triple negative breast cancer therapy with CDK1 siRNA delivered by cationic lipid assisted PEG-PLA nanoparticles. *J. Control Release.* 192, 114–121. doi: 10.1016/j.jconrel.2014.07.001

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550. doi: 10.1186/s13059-014-0550-8

Lu, Y., Yang, G., Xiao, Y., Zhang, T., Su, F., Chang, R., et al. (2020). Upregulated cyclins may be novel genes for triple-negative breast cancer based on bioinformatic analysis. *Breast Cancer.* 27, 903–911. doi: 10.1007/s12282-020-01086-z

Martin-Orozco, E., Sanchez-Fernandez, A., Ortiz-Parra, I., and Ayala-San Nicolas, M. (2019). WNT: signaling in tumors: the way to evade drugs and immunity. *Front. Immunol.* 10, 2854. doi: 10.3389/fimmu.2019.02854

Moreau, Y., and Tranchevent, L. C. (2012). Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat. Rev. Genet.* 13, 523–536. doi: 10.1038/nrg3253

Morry, J., Ngamcherdtrakul, W., Gu, S., Reda, M., Castro, D. J., Sangvanich, T., et al. (2017). Targeted treatment of metastatic breast cancer by PLK1 siRNA delivered by an antioxidant nanoparticle platform. *Mol. Cancer Ther.* 16, 763–772. doi: 10.1158/1535-7163.MCT-16-0644

Muhammad, S. A., Raza, W., Nguyen, T., Bai, B., Wu, X., Chen, J., et al. (2017). Cellular signaling pathways in insulin resistance-systems biology analyses of microarray dataset reveals new drug target gene signatures of type 2 diabetes mellitus. *Front. Physiol.* 8, 13. doi: 10.3389/fphys.2017.00013

Mustonen, E., Ruskoaho, H., and Rysa, J. (2013). Thrombospondins, potential drug targets for cardiovascular diseases. *Basic Clin. Pharmacol. Toxicol.* 112, 4–12. doi: 10.1111/bcpt.12026

Nakada, Y., Zhou, Y., Gong, W., Zhang, E., Skie, E., Nguyen, T., et al. (2021). Single nucleus transcriptomics: apical resection in newborn pigs extends the time-window of cardiomyocyte proliferation and myocardial regeneration. *Circulation* 121, 56995. doi: 10.1161/CIRCULATIONAHA.121.056995

Nedeljkovic, M., and Damjanovic, A. (2019). Mechanisms of chemotherapy resistance in triple-negative breast cancer-how we can rise to the challenge. *Cells* 8, 90957. doi: 10.3390/cells8090957

Neupane, M., and Kiser, J. N. (2018). Bovine respiratory disease complex coordinated agricultural project research T, Neibergs HL: gene set enrichment analysis of SNP data in dairy and beef cattle with bovine respiratory disease. *Anim. Genet.* 49, 527–538. doi: 10.1111/age.12718

Newman, M. E. (2005). A measure of betweenness centrality based on random walks. *Social Netw.* 27, 39–54. doi: 10.1016/j.socnet.2004.11.009

Newman, M. E. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U. S. A.* 103, 8577–8582. doi: 10.1073/pnas.0601602103

Newman, M. E. J. (2008). "Mathematics of networks," in *The New Palgrave Encyclopedia of Economics, 2 Edn*, eds L. E. Blume, S. N. Durlauf. London: Palgrave Macmillan UK.

Nguyen, T., Wei, Y., Nakada, Y., Zhou, Y., and Zhang, J. (2022). Cardiomyocyte cell-cycle regulation in neonatal large mammals: single nucleus RNA-sequencing

data analysis *via* an artificial-intelligence-based pipeline. *Front. Bioeng. Biotechnol.* 10, 914450. doi: 10.3389/fbioe.2022.914450

Nitsch, D., Tranchevent, L. C., Goncalves, J. P., Vogt, J. K., Madeira, S. C., Moreau, Y., et al. (2011). PINTA: a web server for network-based gene prioritization from expression data. *Nucleic Acids Res.* 39, W334–338. doi: 10.1093/nar/gkr289

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., Kanehisa, M., et al. (1999). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 27, 29–34. doi: 10.1093/nar/27.1.29

Olsen, C., Fleming, K., Prendergast, N., Rubio, R., Emmert-Streib, F., Bontempi, G., et al. (2014). Inference and validation of predictive gene networks from biomedical literature and gene expression data. *Genomics* 103, 329–336. doi: 10.1016/j.ygeno.2014.03.004

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). *The PageRank Citation Ranking: Bringing Order to the Web.* Stanford, CA: Stanford InfoLab.

Patel, J. R., Thangavelu, P., Terrell, R. M., Israel, B., Sarkar, A. B., Davidson, A. M., et al. (2022). Novel allosteric inhibitor targets PLK1 in triple negative breast cancer cells. *Biomolecules* 12, 40531. doi: 10.3390/biom12040531

Peng, J., Guan, J., and Shang, X. (2019). Predicting Parkinson's disease genes based on node2vec and autoencoder. *Front. Genet.* 10, 226. doi: 10.3389/fgene.2019.00226

Perez-Iratxeta, C., Bork, P., and Andrade, M. A. (2002). Association of genes to genetically inherited diseases using data mining. *Nat. Genet.* 31, 316–319. doi: 10.1038/ng895

Peters, L. A., Perrigoue, J., Mortha, A., Iuga, A., Song, W. M., Neiman, E. M., et al. (2017). A functional genomics predictive network model identifies regulators of inflammatory bowel disease. *Nat. Genet.* 49, 1437–1449. doi: 10.1038/ng.3947

Porrello, E. R., Mahmoud, A. I., Simpson, E., Hill, J. A., Richardson, J. A., Olson, E. N., et al. (2011). Transient regenerative potential of the neonatal mouse heart. *Science* 331, 1078–1080. doi: 10.1126/science.1200708

Porrello, E. R., and Olson, E. N. (2014). A neonatal blueprint for cardiac regeneration. *Stem Cell Res.* 13, 556–570. doi: 10.1016/j.scr.2014.06.003

Puente, B. N., Kimura, W., Muralidhar, S. A., Moon, J., Amatruda, J. F., Phelps, K. L., et al. (2014). The oxygen-rich postnatal environment induces cardiomyocyte cell-cycle arrest through DNA damage response. *Cell* 157, 565–579. doi: 10.1016/j.cell.2014.03.032

Radulescu, E., Jaffe, A. E., Straub, R. E., Chen, Q., Shin, J. H., Hyde, T. M., et al. (2018). Identification and prioritization of gene sets associated with schizophrenia risk by co-expression network analysis in human brain. *Mol. Psychiatry* 2018, 286559. doi: 10.1101/286559

Rajab, A., Straub, V., McCann, L. J., Seelow, D., Varon, R., Barresi, R., et al. (2010). Fatal cardiac arrhythmia and long-QT syndrome in a new form of congenital generalized lipodystrophy with muscle rippling (CGL4) due to PTRF-CAVIN mutations. *PLoS Genet.* 6, e1000874. doi: 10.1371/journal.pgen.1000874

Rao, A. R., Jana, R., and Bandyopadhyay, S. A. (1996). Markov chain Monte Carlo method for generating random (0, 1)-matrices with given marginals. *Sankhyā* 1996, 225–242.

Rolland, T., Tasan, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., et al. (2014). A proteome-scale map of the human interactome network. *Cell* 159, 1212–1226. doi: 10.1016/j.cell.2014.10.050

Saha, S., Harrison, S. H., and Chen, J. Y. (2008). Dissecting the human plasma proteome and inflammatory response biomarkers. *Proteomics* 2008, 507. doi: 10.1002/pmic.200800507

Sayers, E. (2008). *E-utilities Quick Start.* Bethesda, MD: Entrez Programming Utilities Help.

Schlotterer, C., Tobler, R., Kofler, R., and Nolte, V. (2014). Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. *Nat. Rev. Genet.* 15, 749–763. doi: 10.1038/nrg3803

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Singh, A. R., Sivadas, A., Sabharwal, A., Vellarikal, S. K., Jayarajan, R., Verma, A., et al. (2016). Chamber specific gene expression landscape of the zebrafish heart. *PLoS ONE* 11, e0147823. doi: 10.1371/journal.pone.0147823

Singh-Blom, U. M., Natarajan, N., Tewari, A., Woods, J. O., Dhillon, I. S., Marcotte, E. M., et al. (2013). Prediction and validation of gene-disease associations using methods inspired by social network analyses. *PLoS ONE* 8, e58977. doi: 10.1371/journal.pone.0058977

Smedley, D., Kohler, S., Czeschik, J. C., Amberger, J., Bocchini, C., Hamosh, A., et al. (2014). Walking the interactome for candidate prioritization in

exome sequencing studies of Mendelian diseases. *Bioinformatics* 30, 3215–3222. doi: 10.1093/bioinformatics/btu508

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Sun, J., and Zhao, Z. (2010). A comparative study of cancer proteins in the human protein-protein interaction network. *BMC Genomics* 11(Suppl.3), S5. doi: 10.1186/1471-2164-11-S3-S5

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–452. doi: 10.1093/nar/gku1003

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 45, D362–D368. doi: 10.1093/nar/gkw937

Talesa, V. N. (2001). Acetylcholinesterase in Alzheimer's disease. *Mech. Ageing Dev.* 122, 1961–1969. doi: 10.1016/S0047-6374(01)00309-8

Tessier, L., Cote, O., Clark, M. E., Viel, L., Diaz-Mendez, A., Anders, S., et al. (2018). Gene set enrichment analysis of the bronchial epithelium implicates contribution of cell cycle and tissue repair processes in equine asthma. *Sci. Rep.* 8, 16408. doi: 10.1038/s41598-018-34636-9

Timar, G., Dorogovtsev, S. N., and Mendes, J. F. (2016). Scale-free networks with exponent one. *Phys. Rev. E.* 94, 022302. doi: 10.1103/PhysRevE.94.022302

Tiong, K. L., and Yeang, C. H. (2019). MGSEA - a multivariate Gene set enrichment analysis. *BMC Bioinformatics* 20, 145. doi: 10.1186/s12859-019-2716-6

Tyner, C., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Eisenhart, C., et al. (2017). The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res.* 45, D626–D634. doi: 10.1093/nar/gkw1134

Ueda, A., Oikawa, K., Fujita, K., Ishikawa, A., Sato, E., Ishikawa, T., et al. (2019). Therapeutic potential of PLK1 inhibition in triple-negative breast cancer. *Lab. Invest.* 99, 1275–1286. doi: 10.1038/s41374-019-0247-4

UniProt Consortium, T. (2018). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 46, 2699. doi: 10.1093/nar/gky092

Valent, P. (2014). Targeting the JAK2-STAT5 pathway in CML. *Blood* 124, 1386–1388. doi: 10.1182/blood-2014-07-585943

van Dam, S., Cordeiro, R., Craig, T., van Dam, J., Wood, S. H., de Magalhaes, J. P., et al. (2012). GeneFriends: an online co-expression analysis tool to identify novel gene targets for aging and complex diseases. *BMC Genomics* 13, 535. doi: 10.1186/1471-2164-13-535

Van Vooren, S., Thienpont, B., Menten, B., Speleman, F., De Moor, B., Vermeesch, J., et al. (2007). Mapping biomedical concepts onto the human genome by mining literature on chromosomal aberrations. *Nucleic Acids Res.* 35, 2533–2543. doi: 10.1093/nar/gkm054

Waksman, A. (1968). A permutation network. *J. ACM* 15, 159–163. doi: 10.1145/321439.321449

Wang, C., Li, Y., Li, H., Zhang, Y., Ying, Z., Wang, X., et al. (2020). Disruption of FGF signaling ameliorates inflammatory response in hepatic stellate cells. *Front. Cell Dev. Biol.* 8, 601. doi: 10.3389/fcell.2020.00601

Wang, J., Karra, R., Dickson, A. L., and Poss, K. D. (2013). Fibronectin is deposited by injury-activated epicardial cells and is necessary for zebrafish heart regeneration. *Dev. Biol.* 382, 427–435. doi: 10.1016/j.ydbio.2013.08.012

Wang, S. L., Li, X. L., and Fang, J. (2012). Finding minimum gene subsets with heuristic breadth-first search algorithm for robust tumor classification. *BMC Bioinformatics* 13, 178. doi: 10.1186/1471-2105-13-178

Wang, Z., Duenas-Osorio, L., and Padgett, J. E. (2015). A new mutually reinforcing network node and link ranking algorithm. *Sci. Rep.* 5, 15141. doi: 10.1038/srep15141

Wei, W., Norton, D. D., Wang, X., and Kusiak, J. W. (2002). Abeta 17-42 in Alzheimer's disease activates JNK and caspase-8 leading to neuronal apoptosis. *Brain* 125, 2036–2043. doi: 10.1093/brain/awf205

Winter, C., Kristiansen, G., Kersting, S., Roy, J., Aust, D., Knosel, T., et al. (2012). Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Comput. Biol.* 8, e1002511. doi: 10.1371/journal.pcbi.1002511

Wu, X., Chen, J. Y., Alterovitz, G., Benson, R., and Ramoni, M. (2009). Molecular interaction networks: topological and functional characterizations. *Automat. Proteom. Genom.* 145, 6. doi: 10.1002/9780470741191.ch6

Xia, Q., Cai, Y., Peng, R., Wu, G., Shi, Y., Jiang, W., et al. (2014). The CDK1 inhibitor RO3306 improves the response of BRCA-proficient breast cancer cells to PARP inhibition. *Int. J. Oncol.* 44, 735–744. doi: 10.3892/ijo.2013.2240

Xie, B., Agam, G., Balasubramanian, S., Xu, J., Gilliam, T. C., Maltsev, N., et al. (2015). Disease gene prioritization using network and feature. *J. Comput. Biol.* 22, 313–323. doi: 10.1089/cmb.2015.0001

Yang, C., Jin, C., Li, X., Wang, F., McKeehan, W. L., Luo, Y., et al. (2012). Differential specificity of endocrine FGF19 and FGF21 to FGFR1 and FGFR4 in complex with KLB. *PLoS ONE* 7, e33870. doi: 10.1371/journal.pone.0033870

Ye, L., D'Agostino, G., Loo, S. J., Wang, C. X., Su, L. P., Tan, S. H., et al. (2018). Early regenerative capacity in the porcine heart. *Circulation* 138, 2798–2808. doi: 10.1161/CIRCULATIONAHA.117.031542

Yin, T., Chen, S., Wu, X., and Tian, W. (2017). GenePANDA-a novel network-based gene prioritizing tool for complex diseases. *Sci. Rep.* 7, 43258. doi: 10.1038/srep43258

Yu, L., Fernandez, S., and Brock, G. (2017). Power analysis for RNA-Seq differential expression studies. *BMC Bioinformatics* 18, 234. doi: 10.1186/s12859-017-1648-2

Yu, W., Wulf, A., Liu, T., Khoury, M. J., and Gwinn, M. (2008). Gene Prospector: an evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases. *BMC Bioinformatics* 9, 528. doi: 10.1186/1471-2105-9-528

Yue, Z., Arora, I., Zhang, E. Y., Laufer, V., Bridges, S. L., Chen, J. Y., et al. (2017). Repositioning drugs by targeting network modules: a Parkinson's disease case study. *BMC Bioinformatics* 18, 532. doi: 10.1186/s12859-017-1889-0

Yue, Z., Kshirsagar, M. M., Nguyen, T., Suphavilai, C., Neylon, M. T., Zhu, L., et al. (2015). PAGER: constructing PAGs and new PAG-PAG relationships for network biology. *Bioinformatics* 31, i250–i257. doi: 10.1093/bioinformatics/btv265

Yue, Z., Zheng, Q., Neylon, M. T., Yoo, M., Shin, J., Zhao, Z., et al. (2018). 2.0: an update to the pathway, annotated-list and gene-signature electronic repository for Human Network Biology. *Nucleic Acids Res.* 46, D668–D676. doi: 10.1093/nar/gkx1040

Zhang, E., Nguyen, T., Zhao, M., Dang, S. D. H., Chen, J. Y., Bian, W., et al. (2020). Identifying the key regulators that promote cell-cycle activity in the hearts of early neonatal pigs after myocardial injury. *PLoS ONE* 15, e0232963. doi: 10.1371/journal.pone.0232963

Zhang, F., and Chen, J. Y. (2010). Discovery of pathway biomarkers from coupled proteomics and systems biology methods. *BMC Genomics* 11(Suppl.2), S12. doi: 10.1186/1471-2164-11-S2-S12

Zhang, F., and Chen, J. Y. (2013). Breast cancer subtyping from plasma proteins. *BMC Medical Genom.* 6(Suppl.1), S6. doi: 10.1186/1755-8794-6-S1-S6

Zhang, H., Ferguson, A., Robertson, G., Jiang, M., Zhang, T., Sudlow, C., et al. (2021). Benchmarking network-based gene prioritization methods for cerebral small vessel disease. *Brief Bioinform.* 22, bbab006. doi: 10.1093/bib/bbab006

Zhao, M., Zhang, E., Wei, Y., Zhou, Y., Walcott, G. P., Zhang, J., et al. (2020). Apical resection prolongs the cell cycle activity and promotes myocardial regeneration after left ventricular injury in neonatal pig. *Circulation* 142, 913–916. doi: 10.1161/CIRCULATIONAHA.119.044619

Zhao, S., Geng, Y., Cao, L., Yang, Q., Pan, T., Zhou, D., et al. (2021). Deciphering the performance of polo-like kinase 1 in triple-negative breast cancer progression according to the centromere protein U-phosphorylation pathway. *Am. J. Cancer Res.* 11, 2142–2158.

Zhao, Z. Q., Han, G. S., Yu, Z. G., and Li, J. (2015). Laplacian: normalization and random walk on heterogeneous networks for disease-gene prioritization. *Comput. Biol. Chem.* 57, 21–28. doi: 10.1016/j.compbiolchem.2015.02.008

Zhu, W., Zhang, E., Zhao, M., Chong, Z., Fan, C., Tang, Y., et al. (2018). Regenerative potential of neonatal porcine hearts. *Circulation* 138, 2809–2816. doi: 10.1161/CIRCULATIONAHA.118.034886