

RESEARCH ARTICLE

Benchmarking network algorithms for contextualizing genes of interest

Abby Hill¹ , Scott Gleim¹ , Florian Kiefer², Frederic Sigoillot¹ , Joseph Loureiro¹ , Jeremy Jenkins¹ , Melody K. Morris^{3*} 

1 Chemical Biology and Therapeutics, Novartis Institutes for Biomedical Research, Cambridge, Massachusetts, United States of America, **2** Novartis Informatics, Novartis Institutes for Biomedical Research, Basel, Switzerland, **3** Respiratory Disease Area Department, Novartis Institutes for Biomedical Research, Cambridge, Massachusetts, United States of America

✉ Current address: Rare Disease Research Unit, Pfizer, Cambridge, Massachusetts, United States of America

* melody.morris@novartis.com



 OPEN ACCESS

Citation: Hill A, Gleim S, Kiefer F, Sigoillot F, Loureiro J, Jenkins J, et al. (2019) Benchmarking network algorithms for contextualizing genes of interest. *PLoS Comput Biol* 15(12): e1007403. <https://doi.org/10.1371/journal.pcbi.1007403>

Editor: Luonan Chen, Chinese Academy of Sciences, CHINA

Received: January 16, 2019

Accepted: September 11, 2019

Published: December 20, 2019

Copyright: © 2019 Hill et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The majority of the data used for benchmarking are publically available, and their locations are described within the manuscript. A small subset of the datasets used were results from internal Novartis CRISPR screens that are proprietary to Novartis. Overall conclusions from the proprietary data was similar to the publically available datasets. All algorithms have been previously published and are cited within the manuscript. For this specific work, we used a re-implementation of algorithms into CBDD software package. This software is proprietary to Clarivate. For those interested in accessing the

Abstract

Computational approaches have shown promise in contextualizing genes of interest with known molecular interactions. In this work, we evaluate seventeen previously published algorithms based on characteristics of their output and their performance in three tasks: cross validation, prediction of drug targets, and behavior with random input. Our work highlights strengths and weaknesses of each algorithm and results in a recommendation of algorithms best suited for performing different tasks.

Author summary

In our labs, we aimed to use network algorithms to contextualize hits from functional genomics screens and gene expression studies. In order to understand how to apply these algorithms to our data, we characterized seventeen previously published algorithms based on characteristics of their output and their performance in three tasks: cross validation, prediction of drug targets, and behavior with random input.

This is a *PLOS Computational Biology* Benchmarking paper.

Introduction

In 2000, Schwikowski et al. demonstrated the utility of the guilt by association principle to assign function of yeast genes by examining the function of neighboring genes in a protein-protein interaction [1]. Since then, the scientific community has launched a massive effort to determine protein-protein interaction (PPI) networks for model organisms [2–5] and humans [4, 6]. At the same time, a multitude of computational approaches have been developed for contextualizing genes of interest with known molecular interactions in order to aide

CBDD software, please visit www.clarivate.com for company contact information. Networks used in this work are a combination of a resource proprietary to Clarivate (see www.clarivate.com for company contact information) and a publicly available network (STRING). Generality of the results to other networks was confirmed with a publically available network, HumanNet, as described in the manuscript.

Funding: This research was funded by Novartis Institutes for BioMedical Research. Novartis provided support in the form of salaries for all authors. Army Research Office Institute for Collaborative Biotechnologies (W911NF-09-0001) funded the graduate school tuition of Abby Hill. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: All authors were employed by Novartis when the work was completed, and some have equity interest in Novartis. AH is currently employed by Pfizer.

interpretation of high throughput data. The promise of these algorithms is to connect genes of interest into functional networks and extend the findings with additional genes relevant to the initial list.

In our labs, we aimed to use these algorithms to contextualize hits from functional genomics screens. The hits from a functional genomic screen represent a list of genes that affect a given cellular phenotype (eg. survival [7], autophagy [8], etc.) and that are hypothesized to belong to pathways involved in regulating the phenotype. In these screens, false negatives are also a common concern. In the case of false negatives, genes that affect a given phenotype are missing from the final gene list due to technical factors (eg. editing efficiency) or biological factors (eg. gene redundancy). We aimed to use network algorithms in combination with a protein-protein interaction (PPI) network to both organize hit lists into pathways and extend the hit list through the identification of potential false negatives (i.e. genes that are connected to hits through many PPIs but missing from the hit list).

While many of these network contextualization algorithms have been developed in academia in the context of specific biological questions [9, 10], others are part of commercially available tools (eg. Metacore, Ingenuity Pathway Analysis). However, despite the growing number of available algorithms, to our knowledge there has been no systematic effort to benchmark their ability to return meaningful, actionable hypotheses. In this work, we evaluate network contextualization algorithms available in the Computational Biology for Drug Discovery (CBDD) R package developed by Clarivate, Inc. While we were initially interested in applying these algorithms to hits from functional genomics screens, we appreciated that these algorithms might have utility for other data types with similar interpretation (eg. genes genetically associated to a disease) or for different tasks altogether (eg. target prediction from gene expression signatures). Thus, we assessed the algorithms for three data types: genetic associations; hits from functional genomics CRISPR screens; and gene expression signatures of drug response. We first characterized the algorithms in terms of the novelty and number of connections (i.e. degree) of returned output nodes. We then assessed their performance using cross validation and target prediction, with the ultimate aim of applying appropriate algorithms to contextualize gene lists from gene expression studies or functional genomics screens.

Results

Overview of benchmarking workflow

This work evaluates the ability of seventeen algorithms to use a protein-protein interaction (PPI) network to contextualize and extend a list of genes of interest. [Fig 1](#) exemplifies our workflow with a published pooled CRISPR screen of survival [7]. In this case, the hits from the screen were provided to the network algorithms as the input “start nodes”. The type of output depended on the type of algorithm under investigation. In the case of node prioritization and causal regulator algorithms, the output consisted of a list of ranked network nodes (i.e. the “output nodes”) while subnetwork ID algorithms returned a sub-network consisting of output nodes and the connections between them.

In this work, we considered seventeen algorithms ([Table 1](#)) implemented as part of the Computational Biology for Drug Discovery (CBDD) collaboration between Clarivate Analytics and sixteen pharmaceutical companies. A key deliverable of CBDD is the CBDD R package which implements published algorithms in a consistent interface. Algorithms chosen were available in CBDD version 8.2 and had no major performance considerations that would limit systematic benchmarking efforts. Additionally, the aim of these algorithms was consistent with our aim: to use the network to contextualize and extend genes of interest.

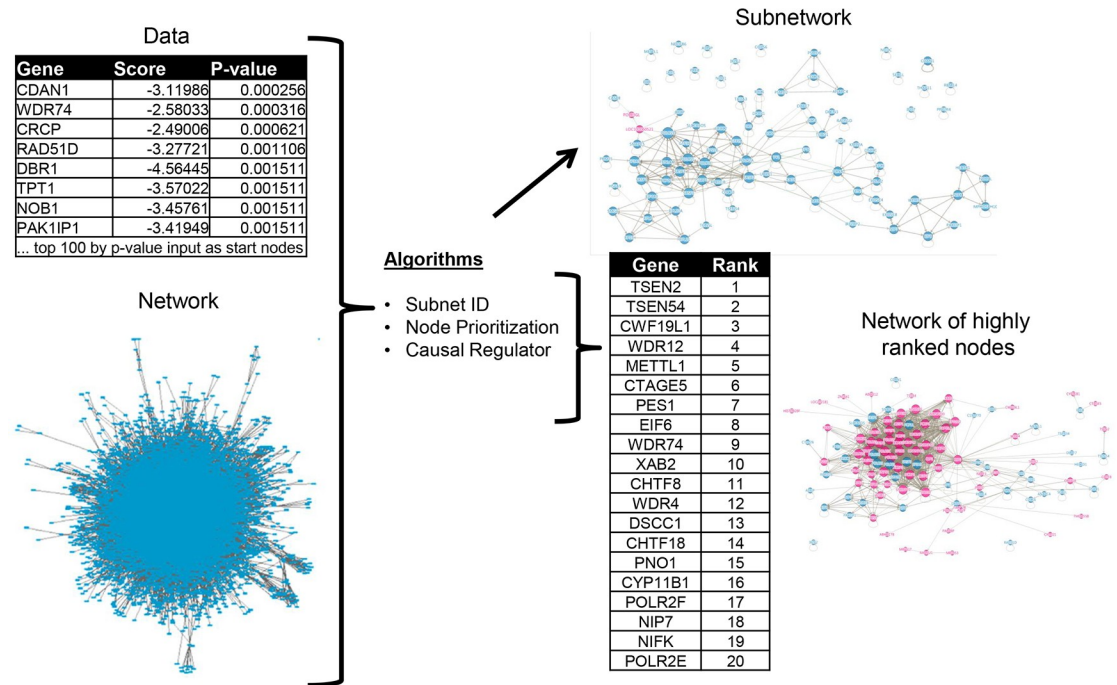


Fig 1. Overview of network algorithm benchmarking workflow: All algorithms considered in this work required a set of identified genes of relevant to a disease, pathway, or treatment (i.e. “start nodes”) as inputs while some also required fold changes and/or p-values. The output of algorithms differed depending on algorithm class, with subnetwork ID algorithms returning highly connected subnetworks; node prioritization algorithms returning ranked lists of genes; and causal regulator algorithms returning ranked lists of hypotheses corresponding to a positive or negative effect of a given gene on the observed data. In the case of node prioritization and causal regulator algorithms, we considered the “output nodes” as the top ranked nodes using a rank cutoff equal to the number of input start nodes for each data set. Also, we note that subnetworks could be constructed from the interactions among the most highly ranked genes in the output lists. For illustration purposes for this figure, we have used the list of top 100 hits (based on p-value) from a CRISPR survival screen in the KBM7 cell line [7]. Each output network contains genes that were included in the input start node list (blue) as well as genes that were identified by the algorithms (pink).

<https://doi.org/10.1371/journal.pcbi.1007403.g001>

When considering these algorithms, we noted they could be divided into three main categories: (1) node prioritization algorithms that prioritize network nodes that are near input nodes, where the definition of “near” varies depending on the specific algorithm, (2) causal regulator algorithms that prioritize network nodes that regulate input start nodes based on their network connectivity, and (3) subnet identification (ID) algorithms that identify regions of the network that connect input nodes and include additional nodes for their connection if warranted. In the case of subnetwork identification algorithms, we wanted to be able to compare to the simplest case of network connections between nodes. Thus, we include output from an algorithm called “Start Node Links”, which connects input start nodes to each out.

We applied the algorithms to hundreds of datasets from four sources, aiming to test the algorithms on a large selection of data sets of different types and confidences. Initial characterization was performed using three types of data meant to capture phenotype- or disease-relevant pathways: (1) KEGG and REACTOME pathway genesets provide high-confidence, well characterized data sets; (2) DisGeNET provides data sets describing curated disease-gene associations [11, 12]; and (3) hits from phenotypic CRISPR screens provide a source of real experimental data most similar to our intended use case. We then turned our attention to Connectivity Map gene expression response signatures, where the aim of applying the algorithms is to predict the target of a perturbation from the response signature. The network used in this work was a protein-protein interaction network derived by combining multiple sources:

Table 1. Algorithms evaluated.

Algorithm	Category	Network Requirement	Brief Description	Reference
<i>Node Prioritization Algorithms: ranks nodes in the network based on connectivity or distance from start nodes</i>				
Random Walk	Node Prioritization		Models path of a random walker starting from nodes of interest and walking to other nodes based on edges in the network	[16]
Network Propagation	Node Prioritization		Random walk based approach controlled for degree of nodes	[17]
ToppNet KM	Node Prioritization	Directed	Random walk-based method with limited number of steps	[18]
ToppNet HITS	Node Prioritization	Directed	Random walk-based method that also takes into account hubness and authority of nodes	[18]
Overconnectivity	Node Prioritization		Enrichment of start nodes and gene sets consisting of each network nodes' neighbors	N/A
Interconnectivity	Node Prioritization		Enrichment based method that identifies nodes between other nodes	[19]
Hidden Nodes	Node Prioritization		Enrichment based method that uses shortest paths to identify nodes between other nodes	[20]
GeneMania	Node Prioritization		Ranks nodes by topological closeness to start nodes in an integrated network	[21]
Guilt By Association	Node Prioritization		Fraction of neighbor nodes that appear in the start node list	[1]
Neighborhood Scoring	Node Prioritization		Guilt-by-association based approach with optional weighting for start nodes	[22]
<i>Causal regulator algorithms: ranks nodes based on evidence that a perturbation to the node would result in observed changes in start nodes</i>				
Causal Reasoning	Causal Regulator	Signed and Directed	Processes network and calculates directional consistency and overconnectivity with start nodes	[23, 24]
SigNet	Causal Regulator	Signed and Directed	Processes network and calculates several metrics to infer relationship with start nodes	[25]
<i>Subnetwork ID algorithms: extract a part of the input network containing many start nodes and additional connecting nodes</i>				
DIAMOnD	Subnetwork ID		Evaluates overconnectivity enrichment iteratively until it reaches a user-defined number of nodes	[26]
Pathway Inference	Subnetwork ID		Heuristic methods that identifies subnetworks enriched in start nodes	[27]
Active Modules	Subnetwork ID		Memetic algorithm with addition of encoding/decoding scheme and local search operator	[28]
CASNet	Subnetwork ID	Signed	Considers edge sign to determine relevance to provided start nodes	[29]
HotNet1	Subnetwork ID		Diffusion based method accounting for FDR	[30]
HotNet2	Subnetwork ID	Directed	Extension of HotNet1 approach than incorporates insulated diffusion and edge direction	[31]
Start Node Links	Subnetwork ID		Directly extracts connections between start nodes	N/A

<https://doi.org/10.1371/journal.pcbi.1007403.t001>

the STRING [13, 14] public database, the Metabase (Clarivate) manually curated database, and interactions from affinity purification mass spectrometry experiments (Bioplex [15]).

Algorithms differ in ranking of start nodes

To determine which algorithms extended the list of interesting genes beyond the input list provided, we first sought to determine the proportion of output nodes that were contained in the input start nodes (Fig 2A). Within the node prioritization algorithms, Random Walk, ToppNet HITS, and GeneMANIA showed clear tendency to include start nodes in their outputs. While Neighborhood Scoring showed an intermediate behavior, all other node prioritization algorithms did not rank start nodes highly and, rather, tended to include a large number of non-start nodes in their output. As causal regulator algorithms are intended to identify nodes that influence the start nodes, possibly from several steps away, they generally did not have a strong preference for including the start nodes themselves in output lists. Most subnetwork ID

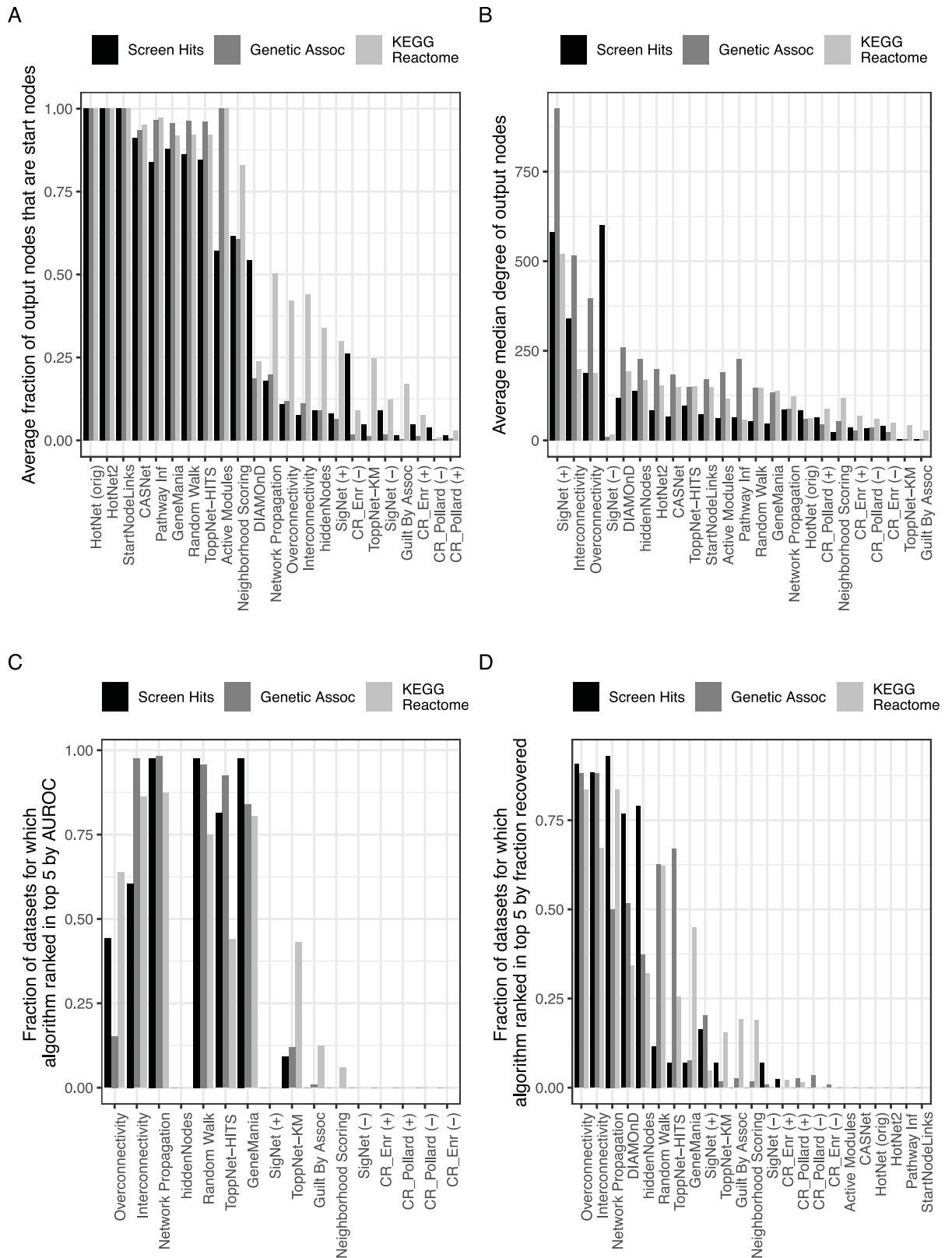


Fig 2. Characterizing algorithms using average fraction of start nodes in the output to indicate tendency to return start nodes in output (A, top left) and degree to indicate tendency to return nodes with many edges (B, top right). Cross-validation performance of algorithms as indicated by the fraction of datasets for which the algorithm appeared in the top five when ranked by AUROC (C, bottom left) or Fraction recovered (D, bottom right). For the fraction recovered analysis, the top nodes were defined as the 200 top-ranked nodes for node prioritization and causal regulator algorithms or any node present in a subnetwork for subnetwork ID algorithms.

<https://doi.org/10.1371/journal.pcbi.1007403.g002>

algorithms showed a strong tendency to include start nodes in their output with the exception of DIAMOnD, which employs the overconnectivity node prioritization algorithm iteratively until it reaches a user-defined number of nodes (in this case 200).

Algorithms differ in preference for node degree

We also sought to understand which algorithms had a tendency to include high degree nodes in the output (i.e. “hub nodes”). Hub nodes are those with many edges (or connections) to other nodes. Across all algorithms, several returned extremely high-degree outputs: DIAMOnD, Interconnectivity, and Overconnectivity (Fig 2B). We noted that these algorithms with high-degree outputs are all enrichment-based methods. Other subnetwork ID and node prioritization algorithms had intermediate but rather variable median degree within the outputs. Several of these algorithms (eg. Pathway Inference, CASNet, HotNet, HotNet2, Active Modules, and GeneMANIA) also ranked start nodes very highly, so the median degree of the output depended heavily on the degree of the start nodes. Of the remaining algorithms that showed intermediate behavior by this metric (ToppNet HITS, Hidden Nodes, Random Walk, and Network Propagation), all are walk-based.

Assessing algorithm performance by cross-validation

In assessment of performance, we performed 10 repeats of 10-fold cross-validation to determine how well the algorithms were able to recover nodes randomly excluded from the input lists. The excluded nodes were true positives in that they were related to the remaining input nodes on the basis of their membership in the original list. Thus, this test determined the ability of the algorithms to identify nodes biologically related to the input list. To summarize the results from cross validation, the area under the receiver operator curve (AUROC) is often evaluated. This metric assumes a perfect gold standard and takes into account both the true positives with the sensitivity metric and false positives with the specificity metric. However, we noted that our input lists were not perfect gold standards in that some nodes returned by the algorithms might appear to be false positives but actually be biologically related to the input list (i.e. nodes designated as false positives by the specificity calculation might actually be false negatives in the original input list). Thus, we also computed the fraction of excluded nodes that were recovered in the top 200 nodes returned by each algorithm (i.e. the fraction recovered). This metric does not take into account false positives and instead asks the question relevant to our intended use of the algorithm: if we were to follow up on the top 200 nodes returned by the algorithms, would nodes known to be biologically relevant to the initial input list be recovered? It is equivalent to the true positive rate (i.e. sensitivity) computed when the top 200 nodes returned by the algorithm are considered the output of the algorithm.

We calculated the AUROC and fraction recovered for each data sets tested. To summarize across individual data sets, we noted that variability in the metrics across datasets made it difficult to determine which were performing better than others (S1 Fig). Thus, we used a ranked-based approach and found the fraction of data sets for which each algorithm appeared in the top five when ranked by AUROC or fraction recovered (Fig 2C and 2D). While performance by AUROC varied across data sources, Random Walk, Network Propagation, GeneMANIA,

Interconnectivity, and ToppNet HITS performed among the top node prioritization algorithms in all datasets tested. Subnetwork ID algorithms could only be quantified by fraction recovered, and for these algorithms, a node was considered ‘recovered’ if it was returned in any subnetwork (in contrast to node prioritization outputs, which were limited to the top 200 nodes). While several different algorithms performed better by the fraction recovered metric than AUC (eg Overconnectivity and Hidden Nodes), the walk-based algorithms Network Propagation and Random Walk performed well by both metrics in all datatypes considered here.

Behavior of algorithms with random input lists

In order to determine whether certain nodes, particularly hub nodes, would be highly ranked by a given algorithm regardless of the input list, we ran the algorithms on 10,000 randomly selected input start node lists. We then compiled the output and calculated the fraction of times that each node appeared in most highly ranked nodes. For most algorithms, a few hundred nodes were ranked in the top 200 nodes in more than 5% of randomly generated list (Table 2). Of greater concern, some algorithms highly ranked a few specific nodes in more than 50% of the output from random input lists (eg. Causal Reasoning, InterConnectivity, SigNet, Random Walk, and ToppNet—HITs), indicating that these nodes were likely to be included in the algorithms’ outputs regardless of their importance for the particular pathway or process of interest. For most algorithms, the tendency of nodes to be highly ranked in the output even with randomly chosen input nodes was related to the degree of the nodes (S2 Fig). However, degree of the nodes did not explain the behavior of all randomly included nodes for all algorithms, and it was clear that other network properties play a role in this finding.

Use of algorithms for target identification using connectivity map

Because causal regulator algorithms were developed to identify upstream regulators of differentially expressed genes, we tested their ability to accomplish this goal using the Connectivity Map [32]. The Connectivity Map dataset captures gene differential expression after treatment

Table 2. Number of nodes ranked in top 200 when algorithms were run with 200 randomly chosen nodes as input start nodes.

Algorithm	Number of nodes highly ranked in 50% of random input tests	Number of nodes highly ranked in 5% of random input tests
Causal Reasoning (Pollard Rank)	64	1129
InterConnectivity	44	1042
Hidden Nodes	0	559
SigNet	200	375
Network Propagation	0	309
ToppNet—HITs	239	289
Random Walk	4	200
Guilt by Association	0	119
ToppNet—KM	0	56
Causal Reasoning (Enrichment Rank)	0	0
Overconnectivity	0	0
Neighborhood Scoring	0	0
GeneMania	0	0

<https://doi.org/10.1371/journal.pcbi.1007403.t002>

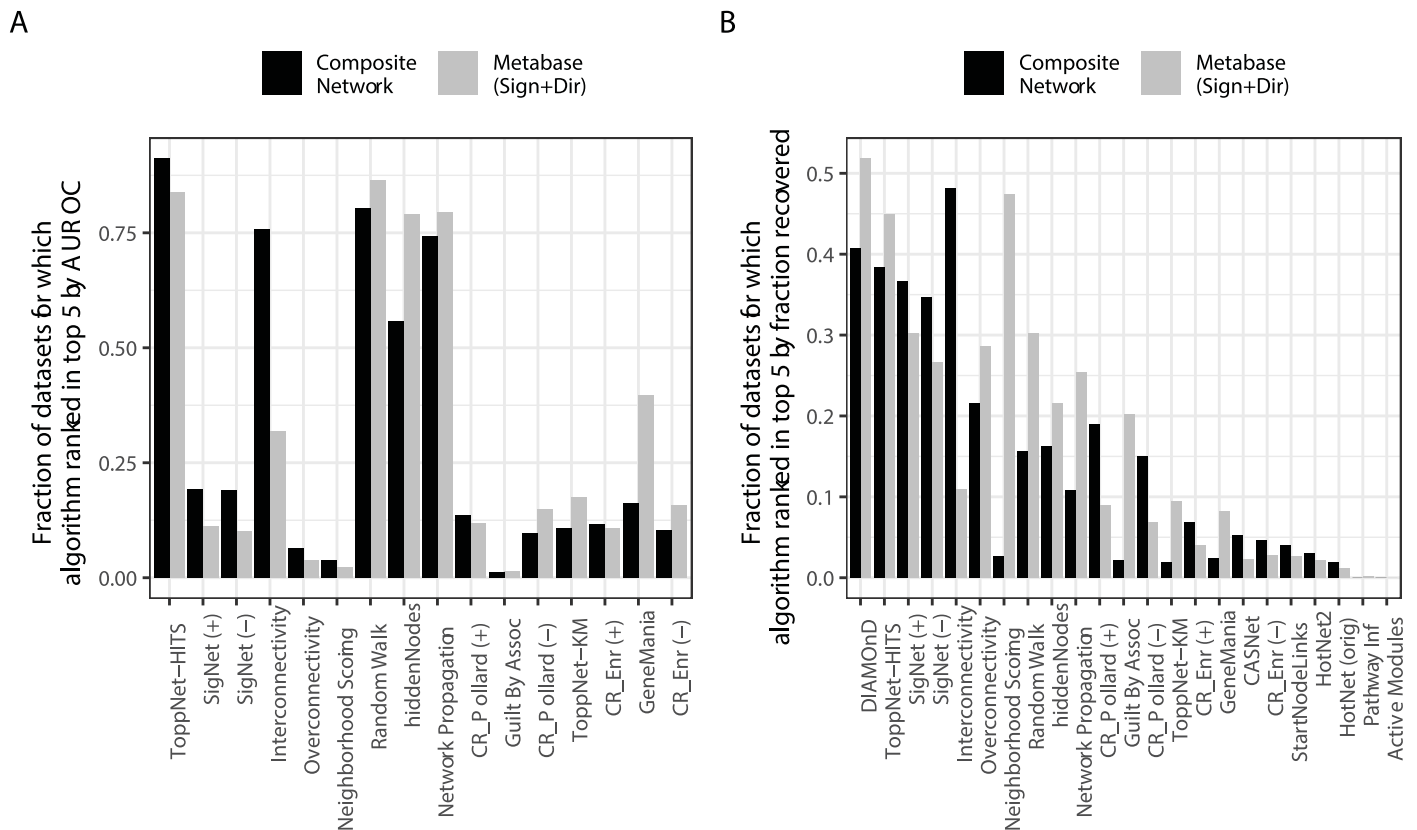


Fig 3. Connectivity Map target prediction in the composite network or metabase signed+directed. Performance was characterized by the ability of the algorithms to highly rank known targets of drugs. (A, top left) Fraction of datasets for which the algorithm appeared in the top five when ranked by fraction of drug targets recovered (B, top right) Fraction of datasets for which the algorithm appeared in the top five when ranked by AUROC.

<https://doi.org/10.1371/journal.pcbi.1007403.g003>

with a drug. Thus, for this analysis, the input start nodes were the differentially expressed genes, and the gold standard we tested was the ability of the algorithms to highly rank the real target(s) of the drugs used for each treatment condition. Our results (Fig 3) indicated that for this type of data, SigNet appeared in the top ranked algorithms. However, it is important to note that, in general, the causal regulator algorithms did not outperform several node prioritization algorithms. We hypothesized that the causal regulator algorithms relied heavily on network information that was not known with sufficient accuracy in the network, which was a composite of signed, unsigned, directed, and undirected edges from multiple sources. Thus, we ran the connectivity map benchmarking workflow with a network that only contained high confidence, signed, and directed edges from the curated Metabase network. With this network, our conclusions were generally consistent (Fig 3, grey bars) although neighborhood scoring performed much better with the Metabase network than composite network.

Discussion

Taken together, our results clearly demonstrate the strengths and weaknesses of several algorithms (Table 3). The benchmarking results shown here suggest that certain categories of algorithms may have different applications, and the choice of algorithm(s) may depend on the specific use case. If the scientist is interested in re-ranking or contextualizing input start nodes,

Table 3. Summary of Algorithm Characteristics and Performance. “Tunable” indicates that the algorithm contains a tunable parameter directly related to the evaluated aspect. Bold italics are used to indicate algorithms that perform well for the indicated metric with flanking asterisks distinguishing the top performers.

Algorithm	Highly ranks start nodes	Output Degree	Highly ranks nodes with random inputs (number of nodes in 50%/5% of test cases)	Number of datatypes for which algorithm is top for gene list extension (AUROC, FR)	Number of networks for which algorithm is top for target prediction task (AUROC, FR)
Network Propagation	tunable		<i>0, 309</i>	<i>* 3, 2 *</i>	<i>* 2, 0 *</i>
Random Walk	Y, tunable		<i>0, 200</i>	<i>* 3, 2 *</i>	<i>* 2, 0 *</i>
GeneMania	Y		<i>* 0, 0 *</i>	<i>3, 1</i>	<i>1, 0</i>
Interconnectivity		High	44, 1042	<i>* 3, 3 *</i>	<i>1, 1</i>
ToppNet-HITS	Y, tunable		239, 289	<i>3, 1</i>	<i>* 2, 2 *</i>
Overconnectivity		High	<i>* 0, 0 *</i>	<i>* 2, 3 *</i>	<i>0, 1</i>
DIAMOnD	tunable		n/a	<i>n/a, 2</i>	<i>* n/a, 2 *</i>
ToppNet-KM	tunable	Low	<i>0, 56</i>	<i>1, 0</i>	0, 0
Hidden Nodes			<i>0, 559</i>	<i>0, 1</i>	<i>* 2, 1 *</i>
Guilt By Association		Low	<i>0, 119</i>	0, 0	n/a, 0
Neighborhood Scoring	Y, tunable	Low	<i>* 0, 0 *</i>	0, 0	<i>0, 1</i>
Pathway Inference	Y, tunable		n/a	n/a, 0	n/a, 0
Active Modules	Y, tunable	tunable	n/a	n/a, 0	n/a, 0
CASNet	Y		n/a	n/a, 0	n/a, 0
HotNet1	Y, tunable		n/a	n/a, 0	n/a, 0
HotNet2	Y, tunable		n/a	n/a, 0	n/a, 0
Start Node Links	Y		n/a	n/a, 0	n/a, 0
Causal Reasoning		Low	64, 1129 (Pollard)	0, 0	n/a, 0
SigNet		High	200, 375	0, 0	<i>* 0, 2 *</i>

<https://doi.org/10.1371/journal.pcbi.1007403.t003>

Random Walk, GeneMANIA, or subnetwork ID methods perform well. Alternatively, if the scientist aims to extend an input list to identify new nodes that may be involved in a disease process or response, Network Propagation or Overconnectivity would be better selections. Of the causal regulator algorithms, SigNet performed well using one metric for tests of target prediction using connectivity map response signatures. However, we note that several node prioritization algorithms also performed well at this task.

In this work, we have characterized the algorithms’ performance using a wide range of data sources in order to understand the broad behavior of the algorithms. However, it is possible that a specific dataset of interest will require a different algorithm than that recommended by these results. For this work, we limited ourselves to algorithms implemented as part of the CBDD collaboration, since the consistent interface resulting from this effort facilitated well our benchmarking study. However, we note that many additional network algorithms are have been developed in the literature (eg. [33–36]), and a comparison of additional algorithms to those studied here in a future benchmarking effort might further refine our understanding in what type of algorithms are appropriate for various tasks.

The majority of these results were obtained using a large network containing PPIs from multiple sources. However, we note that we have run these same characterizations with multiple networks [37] and have included results from a published, undirected network (HumanNet [38]) for the task of extending an initial gene list to include additional biologically relevant nodes (S3 Fig). The results for the HumanNet analysis are consistent overall with our previous results and indicate that network propagation and random walk are top performing algorithms even with an un-directed network. Our goal with this work was to understand which

algorithms performed well for each data type and task. However, another key component to the success of our analysis is the influence of network quality on performance. While we have not undertaken a systematic evaluation of this question with this work, we look forward to future benchmarking efforts to shed further light into this important aspect as well.

Finally, we did not explore individual algorithm parameters, instead relying on author recommendations. However, we note in [Table 3](#) that some algorithms (eg. Network Propagation and Random Walk) contain a parameter meant to alter the number of start nodes included in the output. While a full exploration of parameter landscape for each individual algorithm is out of scope for this work, we have noted key parameters in [S1 Table](#) and would encourage developers of novel algorithms to consider the metrics we have explored here as means to characterize their algorithm across its parameter space and as a starting framework for benchmarking a novel algorithm against existing algorithms.

Materials and methods

Network algorithm parameters

For each algorithm, parameters were chosen to moderate the behavior of the algorithms ([S1 Table](#)). For example, both random walk and network propagation contain a parameter that sets the probability that the random walk will restart at the start nodes at each step; this parameter was set to 0.5 for both to allow for comparison between the two algorithms. If the value of the parameter that would result in moderate behavior was not obvious, it was set based on author recommendations.

Data sets

In the KEGG and Reactome data sets, all sets with 20 or more nodes were included, yielding 165 sets from KEGG and 307 from Reactome. We also used curated gene-disease associations from DisGeNet [[11](#), [12](#)] (accessed 7 June 2016). Nodes were included in a disease set if they had at least 2 Pubmed IDs, and disease sets were kept if the number of associated genes was at least 20, yielding 117 disease sets. For these data sets, where fold changes and p-values are not available, nodes were assigned a \log_2 fold change of 1 and p-value of 0.05 to allow input lists to be run with algorithms that require fold change or p-value.

To test the algorithms using real experimental data, 43 pooled CRISPR screens from Novartis were used as an example set of experimental data with relatively low noise. For CRISPR experiments, cells were transfected with a GFP-tagged target protein of interest and Cas9, then exposed to a pooled library of sgRNA. Cells were FACS-sorted into high- and low-GFP populations, and sgRNA count was used to calculate fold changes and RSA p-values for each targeted gene [[8](#)]. Genes were included in start lists if the RSA p-value $< 1 \times 10^{-4}$ and for each experiment (which may have included multiple comparisons) the start list with length closest to 150 genes was used. Experiments were excluded from the benchmarking data if the longest start list was < 20 genes.

The causal regulator algorithms were originally developed to identify proteins upstream of observed gene expression changes. Since this approach was not specifically relevant to the pathway and screening data described above, we also used data from the Connectivity Map [[32](#)], with more appropriate parameters for the causal regulator algorithms. Data from the connectivity map (v1) was downloaded from <https://portals.broadinstitute.org/cmap/> and genes were included as start nodes if they were differentially expressed more than 2-fold for the indicated treatment. Because connectivity map includes some compounds in multiple settings, we ran the algorithms on each data set independently and then used the average for summarizing algorithm performance.

Networks

Three different network sources were used for this work: (1) The “Composite network” consisting of high-confidence, PPI or transcription factor-gene interactions from the Metabase manually curated network, STRING [13, 14] and BioPlex [15]; (2) “MetabaseSD” consisting of signed and directed high confidence interactions from the Metabase curated network; and (3) HumanNet a previously published undirected network [38]. The composite network was constructed by combining edges from the indicated sources. In the case of the Metabase curated network, nodes are occasionally mapped to multiple genes. In these cases, multiple edges were included in the composite network to capture all genes represented by that network node. In the case of STRING, only the “STRING:actions” network edges were considered high confident, PPI interactions and included in the composite network. The resultant composite network consisted of 597,538 unique edges. Of these edges, 22.6% were signed and 36.8% were directed. For algorithms that required direction, any undirected edge was considered in both directions. For those that required sign, a positive sign was assumed for un-signed edges.

Calculation of start node fraction and median degree

For the purposes of these calculations, “output nodes” were considered to be the top n nodes ranked by the algorithm, where n was the length of the input start list. To quantify preference for start nodes, we calculated the proportion of output nodes that were represented in the input. Thus, an algorithm that ranked all start nodes above all other network nodes would have a start node fraction of 1. To quantify tendency to return hub nodes, we calculated the median degree of output nodes where degree was the total number of edges connected to the node.

Cross-validation and target validation

Ten repeats of 10-fold cross-validation were performed for each data set to calculate the area under the ROC curve (AUC). Each data set was divided into tenths, with one tenth left out each time; then that process was repeated ten times for a total of 100 lists each with 90% of the original input list. Sensitivity and specificity were found using the omitted 10% of nodes as “true” nodes to be found by the algorithms. We also examined Fraction Recovered as the fraction of left out nodes recovered in the top nodes (top 200 nodes for node prioritization or any node present in a subnet for subnet id algorithms). When omitted input nodes were not included in the network, they were excluded from the list of “true” nodes, as the use of that network prevented them from being included in the output regardless of the algorithm used.

For connectivity map data, sensitivity, specificity, and fraction recovered were calculated based on ranking of known drug targets in algorithm outputs where known drug targets were determined as described previously [25].

Empirical null distributions

To determine whether nodes were highly ranked based on the network properties only (irrespective of the input list) we generated lists of randomly selected input nodes. Fold changes were chosen from a random distribution with mean 0 and standard deviation 1, with corresponding p-values. Fold change and p-value pairs were randomly assigned to all possible nodes, and the nodes with highest fold change were used as the input list. We generated 10,000 random gene lists each of length 200 and ran the algorithms on these input lists. We were thus able to determine, for each node and algorithm, the frequency each node was ranked higher than a chosen output rank.

Supporting information

S1 Fig. Performance results using standard summary statistics (mean and standard deviation across datasets) for AUROC (left) and Fraction Recovered (right). Comparison of algorithms was difficult due to variation across datasets. Thus, a rank-based approach was used to establish the fraction of datasets for which the algorithm was performing in top five algorithms for each dataset ([Fig 2C and 2D](#)).

(EPS)

S2 Fig. Fraction of times a node was highly ranked using randomly chosen input start nodes as a function of node degree. Causal regulator algorithms consider each node in two directions—positive (black points) and negative (red points).

(EPS)

S3 Fig. Characterization and performance results generalize across the HumanNet published, undirected network. Average fraction of start nodes in the output (A) and median degree (B) characterization of each algorithm. Cross-validation performance of algorithms as indicated by the fraction of datasets for which the algorithm appeared in the top five when ranked by AUROC (C) or Fraction recovered (D) from the CRISPR screen hits, Genetic Association, and KEGG/REACTOME datasets using HumanNet as the network. Note: Because HumanNet contains no signed or directed edges, the causal regulator algorithms were not examined in this analysis.

(EPS)

S1 Table. Algorithm parameters used (missing algorithms did not have adjustable parameters).

(DOCX)

S1 Data. Summary values to create plots in [Fig 2](#) and [S1 Fig](#).

(CSV)

S2 Data. Summary values to create plots in [Fig 3](#).

(CSV)

Acknowledgments

We wish to thank Alexander Ishkin and the team at Clarivate Analytics for their excellent implementation of the CBDD software. We also thank Douglas Lauffenburger for his guidance and support.

Author Contributions

Conceptualization: Melody K. Morris.

Data curation: Florian Kiefer, Frederic Sigoillot.

Formal analysis: Abby Hill, Melody K. Morris.

Funding acquisition: Joseph Loureiro, Jeremy Jenkins.

Investigation: Scott Gleim, Frederic Sigoillot, Joseph Loureiro, Jeremy Jenkins.

Project administration: Jeremy Jenkins.

Resources: Florian Kiefer, Jeremy Jenkins.

Software: Scott Gleim, Florian Kiefer.

Supervision: Melody K. Morris.

Visualization: Abby Hill, Melody K. Morris.

Writing – original draft: Abby Hill, Melody K. Morris.

Writing – review & editing: Abby Hill, Scott Gleim, Florian Kiefer, Frederic Sigoillot, Joseph Loureiro, Jeremy Jenkins, Melody K. Morris.

References

- Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. *Nat Biotechnol*. 2000; 18(12):1257–61. <https://doi.org/10.1038/82360> PMID: 11101803
- Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, et al. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*. 2000; 403(6770):623–7. <https://doi.org/10.1038/35001009> PMID: 10688190
- Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*. 2001; 98(8):4569–74. <https://doi.org/10.1073/pnas.061034498> PMID: 11283351
- Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*. 2005; 437(7062):1173–8. <https://doi.org/10.1038/nature04209> PMID: 16189514
- Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, et al. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science*. 2000; 287(5450):116–22. <https://doi.org/10.1126/science.287.5450.116> PMID: 10615043
- Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, et al. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*. 2005; 122(6):957–68. <https://doi.org/10.1016/j.cell.2005.08.029> PMID: 16169070
- Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic screens in human cells using the CRISPR-Cas9 system. *Science*. 2014; 343(6166):80–4. <https://doi.org/10.1126/science.1246981> PMID: 24336569
- DeJesus R, Moretti F, McAllister G, Wang Z, Bergman P, Liu S, et al. Functional CRISPR screening identifies the ufm1ylation pathway as a regulator of SQSTM1/p62. *Elife*. 2016; 5.
- Goehler H, Lalowski M, Stelzl U, Waelter S, Stroedicke M, Worm U, et al. A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. *Mol Cell*. 2004; 15(6):853–65. <https://doi.org/10.1016/j.molcel.2004.09.016> PMID: 15383276
- Lim J, Hao T, Shaw C, Patel AJ, Szabo G, Rual JF, et al. A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell*. 2006; 125(4):801–14. <https://doi.org/10.1016/j.cell.2006.03.032> PMID: 16713569
- Pinero J, Bravo A, Queralt-Rosinach N, Gutierrez-Sacristan A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res*. 2017; 45(D1):D833–D9. <https://doi.org/10.1093/nar/gkw943> PMID: 27924018
- Pinero J, Queralt-Rosinach N, Bravo A, Deu-Pons J, Bauer-Mehren A, Baron M, et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database (Oxford)*. 2015; 2015:bav028.
- Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*. 2015; 43(Database issue):D447–52. <https://doi.org/10.1093/nar/gku1003> PMID: 25352553
- von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res*. 2003; 31(1):258–61. <https://doi.org/10.1093/nar/gkg034> PMID: 12519996
- Huttlin EL, Ting L, Bruckner RJ, Gebreab F, Gygi MP, Szpyt J, et al. The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell*. 2015; 162(2):425–40. <https://doi.org/10.1016/j.cell.2015.06.043> PMID: 26186194
- Kohler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet*. 2008; 82(4):949–58. <https://doi.org/10.1016/j.ajhg.2008.02.013> PMID: 18371930
- Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol*. 2010; 6(1):e1000641. <https://doi.org/10.1371/journal.pcbi.1000641> PMID: 20090828

18. Chen J, Aronow BJ, Jegga AG. Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics*. 2009; 10:73. <https://doi.org/10.1186/1471-2105-10-73> PMID: 19245720
19. Hsu CL, Huang YH, Hsu CT, Yang UC. Prioritizing disease candidate genes by a gene interconnectedness-based approach. *BMC Genomics*. 2011; 12 Suppl 3:S25.
20. Dezsó Z, Nikolsky Y, Nikolskaya T, Miller J, Cherba D, Webb C, et al. Identifying disease-specific genes based on their topological significance in protein networks. *BMC Syst Biol*. 2009; 3:36. <https://doi.org/10.1186/1752-0509-3-36> PMID: 19309513
21. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol*. 2008; 9 Suppl 1:S4.
22. Nitsch D, Goncalves JP, Ojeda F, de Moor B, Moreau Y. Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC Bioinformatics*. 2010; 11:460. <https://doi.org/10.1186/1471-2105-11-460> PMID: 20840752
23. Pollard J Jr., Butte AJ, Hoberman S, Joshi M, Levy J, Pappo J. A computational model to define the molecular causes of type 2 diabetes mellitus. *Diabetes Technol Ther*. 2005; 7(2):323–36. <https://doi.org/10.1089/dia.2005.7.323> PMID: 15857235
24. Chindelevitch L, Ziemek D, Enayattallah A, Randhawa R, Sidders B, Brockel C, et al. Causal reasoning on biological networks: interpreting transcriptional changes. *Bioinformatics*. 2012; 28(8):1114–21. <https://doi.org/10.1093/bioinformatics/bts090> PMID: 22355083
25. Jaeger S, Min J, Nigsch F, Camargo M, Hutz J, Cornett A, et al. Causal Network Models for Predicting Compound Targets and Driving Pathways in Cancer. *J Biomol Screen*. 2014; 19(5):791–802. <https://doi.org/10.1177/1087057114522690> PMID: 24518063
26. Ghiassian SD, Menche J, Barabasi AL. A Disease Module Detection (DIAMOND) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput Biol*. 2015; 11(4):e1004120. <https://doi.org/10.1371/journal.pcbi.1004120> PMID: 25853560
27. Rajagopalan D, Agarwal P. Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics*. 2005; 21(6):788–93. <https://doi.org/10.1093/bioinformatics/bti069> PMID: 15509611
28. Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*. 2002; 18 Suppl 1:S233–40.
29. Gaire RK, Smith L, Humbert P, Bailey J, Stuckey PJ, Haviv I. Discovery and analysis of consistent active sub-networks in cancers. *BMC Bioinformatics*. 2013; 14 Suppl 2:S7.
30. Vandin F, Upfal E, Raphael BJ. Algorithms for detecting significantly mutated pathways in cancer. *J Comput Biol*. 2011; 18(3):507–22. <https://doi.org/10.1089/cmb.2010.0265> PMID: 21385051
31. Leiserson MD, Vandin F, Wu HT, Dobson JR, Eldridge JV, Thomas JL, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet*. 2015; 47(2):106–14. <https://doi.org/10.1038/ng.3168> PMID: 25501392
32. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*. 2006; 313(5795):1929–35. <https://doi.org/10.1126/science.1132939> PMID: 17008526
33. Melas IN, Sakellaropoulos T, Iorio F, Alexopoulos LG, Loh WY, Lauffenburger DA, et al. Identification of drug-specific pathways based on gene expression data: application to drug induced lung injury. *Integr Biol (Camb)*. 2015; 7(8):904–20.
34. Pers TH, Karjalainen JM, Chan Y, Westra HJ, Wood AR, Yang J, et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun*. 2015; 6:5890. <https://doi.org/10.1038/ncomms6890> PMID: 25597830
35. Lee JH, Zhao XM, Yoon I, Lee JY, Kwon NH, Wang YY, et al. Integrative analysis of mutational and transcriptional profiles reveals driver mutations of metastatic breast cancers. *Cell Discov*. 2016; 2:16025. <https://doi.org/10.1038/celldisc.2016.25> PMID: 27625789
36. Zhao XM, Li S. HISP: a hybrid intelligent approach for identifying directed signaling pathways. *J Mol Cell Biol*. 2017; 9(6):453–62. <https://doi.org/10.1093/jmcb/mjx054> PMID: 29281081
37. Hill AB. *Integrated Experimental and Computational Analysis of Intercellular Communication with Application to Endometriosis*: Massachusetts Institute of Technology; 2018.
38. Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res*. 2011; 21(7):1109–21. <https://doi.org/10.1101/gr.118992.110> PMID: 21536720