Network-Based Disease Module Discovery by a Novel Seed Connector Algorithm with Pathobiological Implications

Rui-Sheng Wang, Joseph Loscalzo

PII: S0022-2836(18)30427-3
DOI: doi:10.1016/j.jmb.2018.05.016
Reference: YJMBI 65724

To appear in:

Received date: 7 March 2018
Revised date: 30 April 2018
Accepted date: 10 May 2018

Please cite this article as: Rui-Sheng Wang, Joseph Loscalzo , Network-Based Disease Module Discovery by a Novel Seed Connector Algorithm with Pathobiological Implications. The address for the corresponding author was captured as affiliation for all authors. Please check if appropriate. Yjmbi(2018), doi:10.1016/j.jmb.2018.05.016

May 16, 2018

Network-based Disease Module Discovery by a Novel Seed Connector Algorithm with Pathobiological Implications

Rui-Sheng Wang and Joseph Loscalzo*

Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA

*To whom correspondence should be addressed:

Dr. Joseph Loscalzo

Department of Medicine

Brigham and Women's Hospital

75 Francis Street

Boston, MA 02115

617-732-6340; 617-732-6439 (fax)

E-mail: jloscalzo@rics.bwh.harvard.edu

Conflicts of Interest:

The authors declare that no conflict of interest exists**.**

Abstract

Understanding the genetic basis of complex diseases is challenging. Prior work shows that disease-related proteins do not typically function in isolation. Rather, they often interact with each other to form a network module that underlies dysfunctional mechanistic pathways. Identifying such disease modules will provide insights into a systems-level understanding of molecular mechanisms of diseases. Owing to the incompleteness of our knowledge of disease proteins and limited information on the biological mediators of pathobiological processes, the key proteins (seed proteins) for many diseases appear scattered over the human protein-protein interactome and form a few small branches, rather than coherent network modules. In this paper, we develop a network-based algorithm, called Seed Connector algorithm (SCA), to pinpoint disease modules by adding as few additional linking proteins (seed connectors) to the seed protein pool as possible. Such seed connectors are hidden disease module elements that are critical for interpreting the functional context of disease proteins. The SCA aims to connect seed disease proteins so that disease mechanisms and pathways can be decoded based on predicted coherent network modules. We validate the algorithm using a large corpus of 70 complex diseases and binding targets of over 200 drugs, and demonstrate the biological relevance of the seed connectors. Lastly, as a specific proof-of-concept, we apply SCA to a set of seed proteins for coronary artery disease (CAD) derived from a meta-analysis of large-scale genome-wide association studies (GWAS) and obtain a CAD module enriched with important disease-related signaling pathways and drug targets not previously recognized.

Introduction

After decades of research, many susceptibility alleles and genes associated with complex diseases have been identified by genome-wide association studies (GWAS) and other ‘omic’ or bioinformatic approaches [1]. Nevertheless, our knowledge of the mechanisms underlying these associations that are responsible for the diseases remains largely undefined. There is increasing evidence that a set of proteins associated with a given disease do not function in an isolated way. Rather, these causal proteins interact with each other to form a distinct network module within the universe of (physical) protein-protein interactions (the human protein-protein interactome) representing perturbed, dysfunctional pathways [2-4]. For this reason, traditional single protein or single pathway-based approaches for studying complex diseases have limited utility. Network-based approaches can aid in identifying such disease modules in the human interactome; provide insights into systems-level understanding of disease mechanisms and pathophenotypes; and guide the search for therapeutic targets.

Many heuristic methods have been proposed to integrate protein interaction networks with different types of omics data to prioritize or predict disease-associated genes [5-7]. These prioritization methods, however, generally do not yield a coherent connected module in the human interactome nor offer insights into causal mechanisms. Most of the existing algorithms for identifying disease modules aim to expand the set of seed proteins (nodes in the interactome network) by adding many predicted associated proteins (linked nodes) in the interactome while ignoring the small branches of loosely connected seed proteins, leading many proteins associated with diseases to be ‘orphaned’ from the disease module in the network and to be of unclear mechanistic significance [8-10]. In fact, these isolated seed proteins are often more reliable than the predicted associated disease proteins as they have documented evidence of association with the disease. Their isolation in the network may simply be a consequence of the incompleteness of the pool of seed proteins or the scarcity of information on biological mediators linked thereto.

Three examples of existing algorithms serve to illustrate these shortcomings. DIAMOnD, a disease module detection algorithm based on connectivity patterns of disease proteins [8], it ranks candidate proteins according to their numbers of connections to seed proteins. While it can identify a connected disease module, the coverage of seed proteins in the module may be very low, with many isolated seed proteins unexplored. Importantly, these isolated proteins may be only one or two links away from other seed proteins. The prize-collecting Steiner tree (PCST) algorithm searches for a subtree minimizing the sum of the total ‘cost’ of all edges in the subtree plus the total ‘profit’ of all nodes, and has been used to identify optimal subnetworks for a given set of seed genes or proteins [11-14]. This algorithm requires additional weights for nodes (denoted as profits) and edges (denoted as costs), which are usually not available for disease proteins. Moreover, PCST gives a tree-like or forest-like network without cycles, not a subnetwork with general topological modular structure. The recently developed method, GLADIATOR, ascertains disease modules based on disease-disease phenotype similarity, which indicates that it requires knowledge of multiple linked diseases and, therefore, cannot predict disease modules individually [9].

In this work, we develop a novel network-based Seed Connector Algorithm (SCA) to discover disease modules in the human interactome by adding as few extra hidden nodes as possible in order to link seed disease proteins. This method facilitates decoding of disease mechanisms and pathways based on predicted coherent network modules. These seed connectors, serving as bridges of different network branches induced by seed proteins, are otherwise hidden components (i.e., not previously recognized by genetic linkage or reductionist studies that define the seed gene/protein pool) that are critical for interpreting the functional context of disease proteins and for understanding the dysfunctional pathways of diseases. We validate our algorithm using a list of 70 complex diseases and a large corpus of binding targets of over 200 drugs, and demonstrate the biological relevance of these seed connectors. Lastly, we apply this algorithm to a set of seed genes for coronary artery disease (CAD) derived from a meta-analysis of large-scale genome-wide association studies (GWAS) and obtain a CAD module enriched with important disease-related signaling pathways and drug targets not previously recognized.

Materials and Methods

Sources of the Human Interactome

The molecular mechanisms underlying disease modules involve multiple types of molecular interactions. We used a consolidated human protein-protein interactome, which combines physical macromolecular interactions including protein-protein interactions, protein complexes, protein-DNA interactions, kinase-substrate interactions, metabolic interactions, and signaling pathways from different sources, to ascertain disease modules. Protein-protein interactions are derived from several high-throughput yeast two-hybrid studies [15-18] and also include binary interactions from IntAct and MINT databases [19, 20], as well as literature-curated interactions obtained from low-throughput experiments reported in the IntAct, MINT, HPRD, and BioGRID databases [21, 22]. The manually curated dataset of mammalian protein complexes (CORUM) and experimentally determined human protein complexes are also incorporated in the comprehensive set of protein-protein interactions [23, 24] Protein-DNA regulatory interactions are taken from the TRANSFAC database [25], and kinase-substrate interactions are obtained from the PhosphositePlus database [26]. Metabolic enzyme-coupled interactions are derived from the KEGG and BiGG databases [27]. In addition, protein interactions from 3D structural prediction and signaling interactions are also included in the construction of the interactome [28, 29] . This consolidated human interactome, after removing duplicate interactions and self-loops, has 14,174 proteins with 170,303 interactions.

Collecting Disease Proteins

As described in [8], a list of 70 diseases and their associated proteins (seed proteins) were compiled from the Online Mendelian Inheritance in Man (OMIM) [30], UniProtKB/SwissProt [31], and GWAS data from the Phenotype-Genotype Integrator database (PheGenI) [32]. The list was manually chosen according to unique (i.e., non-redundant) nosology, omitting those phenotypes that reflect symptoms or common disease mechanisms (i.e., endopathophenotypes such as fibrosis or inflammation). The diseases chosen also had to have at least 20 associated proteins to be mapped to the human interactome. The genome-wide significance cutoff we used is P <5.0E-8. The Medical Subject Headings ontology (MeSH) was used to combine the different disease nomenclatures from the two sources into a single standard vocabulary.

Compiling Drug Targets

The draft set of drug-target pairs were collected from DrugBank [33], TTD (Therapeutic Target Database) [34], and PharmGKB [35]. We then used the bioactivity data of drugs to filter out some drug targets. The bioactivity data were collected from three commonly used databases: ChEMBL [36], BindingDB [37], and IUPHAR/BPS Guide to PHARMACOLOGY (GtoPdb) [38]. Only those drug targets meeting the following three criteria were retained: (1) binding affinities, including K_i, K_d, IC_{50}, or EC_{50}, each ≤ 10 μM; (2) the target protein can be represented by a unique UniProt accession number and also marked as 'reviewed' in the UniProt database [39]; and (3) the target protein is a human protein. The draft set of drug targets was refined with the bioactivity data and then mapped to their Entrez ID from the NCBI database [40], as well as their official gene symbols based on GeneCards [41]. We mapped the drug targets to the human interactome and only considered those drugs that have at least 10 targets in the interactome, with 220 drugs included for further analysis.

The Seed Connector Algorithm (SCA)

Owing to the incompleteness of the set of disease seed proteins, these proteins may not be closely connected (immediately proximate) to each other to form a network module in the human interactome. We, therefore, developed a network-based algorithm called the Seed Connector Algorithm (SCA) which aims to connect the disease seed proteins by introducing as few extra connector nodes as possible. This algorithm is sort of similar to a Steiner tree problem [42] which aims to create a tree of minimum weight that contains all of the given nodes (but may include additional nodes) on an undirected graph with non-negative edge weights; however, instead of a tree structure, the SCA can generate a subnetwork with a general topological modular structure. The principle underlying this algorithm is that seed proteins associated with the same disease should not be very far from each other and, thus, should link to each other through very short paths (e.g., not more than one intermediary node). SCA is an iterative algorithm as follows:

I. Assume that the seed protein pool P={P_1, P_2, …, P_s} induces a subnetwork G_t. Calculate the size of the largest connected component (LCC) of the subnetwork G_t.

II. Consider all of the first-order interactors of the seed proteins as identified in the human interactome. Add each interactor temporarily to the seed protein pool one-by-one: P_t=P∪{P_I}. Obtain the subnetwork induced by this temporary seed protein pool and determine the size of its LCC.

III. Select those interactors that can increase the coverage of seed proteins in the LCC of the subnetwork maximally. If there are multiple candidates, select those whose neighbors contain the greatest number of seed proteins compared to its total number of neighbors and add it to the seed protein pool *P*.

Steps I-III repeat until none of the first-order interactors, when added, increases the coverage of seed proteins in the LCC of the induced subnetwork G_t. The final subnetwork is the predicted coherent disease module, which is obtained by introducing as few additional nodes as possible. The disease module obtained by this algorithm has a very high ratio of seed proteins to connector proteins. A preliminary version of the SCA has been used in constructing the placebome module in one of our previous studies [43]. As we realized that for some isolated seed proteins, there can be too many candidate connectors with equal roles. Therefore, in this current study, we add one more step to the algorithm, retaining only those candidates whose neighbors are enriched with the greatest number of seed proteins. This step ensures that a minimum number of seed connectors are added to network modules.

Figure 1 gives an overview of SCA. First, the algorithm starts with known seed proteins and induces a loosely connected subnetwork consisting of only seed proteins. Next, one first-order interactor (in grey) of seed proteins that increases the size of LCC of the subnetwork maximally is selected as a seed connector. More proteins (in grey) that maximally increase the size of LCC of the subnetwork are selected sequentially as seed connectors until there is no additional first-order interactor that can be selected as a seed connector. Note that after all the iterations are completed, there may still be isolated seed proteins not connected to the subnetwork. This observation likely means that these isolated proteins are, indeed, far from the disease module, and simply cannot be connected through short paths given our current (incomplete) knowledge of the human interactome (~25% of all likely interactions). SCA was implemented by Python with the assistance of the NetworkX package.

Statistical Analysis and Tools

All network visualization was performed with the open source platform Cytoscape 3.30 [44]. When we assessed the topological properties of a disease module or a drug target module, we created a random control for determining its statistical significance, i.e., we randomly selected a protein set of the same size from the human interactome and calculated the topological properties of the randomly defined modules. *P* values were obtained by fitting the histograms to normal distributions using the 'normfit' command in Matlab (Mathworks, Inc). *P* values (adjusted by the Benjamini-Hochberg procedure if applicable) less than 0.05 were considered significant. Gene Ontology (GO)-based functional similarity of pairs of proteins was quantified by GS2 (GO-based similarity of gene sets) developed in a previous study [45]. The daily snapshots of the GO tree and human gene annotations were downloaded from the GO web site (http://www.geneontology.org)

Results

Discovery of Disease Modules

We first examined the performance of SCA using a corpus of 70 diseases. Importantly, the full set (i.e., ground truth) of disease proteins is unknown, so we cannot evaluate the performance of the algorithm in terms of sensitivity and specificity. We first checked the coverage rate of seed proteins in the disease modules predicted by SCA and compared it with DIAMOnD [9]. Since DIAMOnD needs manual input (the number of candidate proteins) to stop the algorithm, we set this number equal to the number of seed connectors, with the result summarized in **Table 1**. Among 70 diseases, the SCA is able to ascertain network modules for 67 diseases. Three diseases -- exophthalmos, glomerulonephritis, and Graves disease -- have too few seed proteins which are sufficiently remote from each other that first-order interactors are unable to link them. Of these 67 diseases, 45 disease modules constructed with the SCA have significantly higher coverage rates of seed proteins in the LCC than the disease modules predicted by DIAMOnD. For the remainder of the diseases, the modules generated by DIAMOnD were not significantly different in coverage of the seed proteins compared to the SCA; yet the total number of seed proteins covered by the LCC of these DIAMOnD-defined modules were much less than the seed protein coverage in disease modules defined by the SCA in all cases.

Figure 2 provides an example showing the typical characteristics of a disease module constructed using the SCA and DIAMOnD. The SCA links the seed proteins together by introducing additional hidden components (biological intermediaries) such that meaningful signaling pathways relevant to diseases become explicit. Disease modules constructed by SCA usually cover the majority of disease seed proteins, with only a few proteins isolated from the disease modules. In contrast, DIAMOnD continually expands the LCC of the module, but fails to incorporate more seed proteins with iterative expansion for many of the disease modules.

We next used publicly available gene function annotation data from Gene Ontology (GO) to validate the biological relevance of seed connectors and the functional rationale of the disease modules. The hypothesis here is that seed connectors act as intermediaries of seed proteins and should have functions similar to the seed proteins. We, therefore, examined the GO functional similarity between seed connectors and seed proteins. As a background control, we randomly selected the same number of proteins in the human interactome 1,000 times and calculated their GO functional similarity to seed proteins. The statistical significance of this analysis is shown in **Figure 3A**. Of 67 disease modules, the seed connectors of 65 disease modules are statistically significantly functionally similar to the seed proteins. This finding demonstrates the functional relevance of the seed connectors and the reliability of the constructed disease modules as representations of a functionally integrated pathobiology. We also used another background control to examine the significance of the functional similarity of seed connectors to seed proteins: the random proteins were selected from other first-order interactors of the seed proteins that were not seed connectors. This approach was used to determine whether seed connectors are functionally more similar to seed proteins than other first-order interactors. Even with this strict background control, the seed connectors of 60 out of 67 diseases are functionally more similar to seed proteins than other first-order interactors. The result, shown in **Supplementary Figure 1**, further substantiates the biological relevance of the seed connectors as functional intermediaries and their potential mechanistic role in disease pathogenesis.

We also compared the seed connectors from the SCA and DIAMOnD proteins in terms of their functional similarity to seed proteins. As DIAMOnD proteins are candidate disease proteins in the neighborhood of seed proteins, their functional similarities to seed proteins are comparable as seed connectors (**Figure 3B**); however, seed connectors are somewhat more robust in this regard (**Figure 3C**): for 34 disease modules (51%), seed connectors identified by the SCA are more functionally similar to seed proteins than candidate disease proteins identified by DIAMOnD; for 26 disease modules (39%), DIAMOnD candidate disease proteins are more functionally similar to seed proteins than SCA seed connectors; and for 7 disease modules (10%), SCA seed connectors and DIAMOnD candidate disease proteins are equally functionally similar to disease modules. Note that owing to the lack of ground truth, functional similarity is simply an indirect way of comparing the two algorithms without statistical validation.

Coherent Network Modules for Drug Targets

The SCA does not only facilitate identification of disease modules, but can also be used to determine drug target modules. Identification of drug target modules is helpful for understanding the molecular mechanisms of action of drugs. Owing to the incompleteness of target information, some targets may not be connected to principal mechanistic pathways. We used a total of 220 drugs with at least 10 targets in the human interactome in this analysis. Interestingly, of the 220 drugs, the targets of over 200 drugs form a connected module with significantly more interactions and larger sized LCC compared to the same number of random proteins chosen from the human interactome (**Supplementary Table 1** and **Supplementary Table 2**). This observation can be interpreted to mean that the majority of drugs act on a local neighborhood of the human interactome.

Some drug targets are loosely connected to and isolated from the main modules. We applied our SCA to the binding targets of all of the 220 drugs, and obtained larger drug target modules by adding connectors. Of 220 drugs, the SCA enables the expansion of the target modules of 199 drugs. **Figure 4** provides an example of a drug target module identified by the SCA. Palbociclib (DB09073) is a drug used for the treatment of ER-positive and HER2-negative breast cancer and has 30 drug targets in the interactome that induce a subnetwork of 16 proteins (nodes) and 13 interactions (edges) (**Figure 4A**). After expansion using the SCA, the target module of this drug contains 39 proteins and 50 interactions and covers all (known) targets (**Figure 4B**). The functional context of the CDK pathway and the MAPK pathway is explicit in the expanded drug target module.

To evaluate the biological relevance of the seed connectors for ascertaining drug target modules, we collected an expanded drug target set from Drugbank, including drug targets, drug carriers, drug transporters, and drug enzymes. We examined whether or not the seed connectors are significantly enriched with drug targets. We found that of 199 drugs whose target modules are expandable, the seed connectors for 74 drugs are significantly enriched with drug targets compared to random expectation, indicating the pharmaceutical relevance of the seed connectors. For example, the SCA added 8 extra intermediaries to the target set of palbociclib**,** 5 of which are known drug targets: MAP3K1, PTPN1, LIMK1, SNCA, and PRKAR2A (**Figure 4B**). MAP3K1 is the target of DB06061 --- an MEK inhibitor that blocks signal transduction pathways implicated in cancer cell proliferation and survival. LIMK1 is a target of DB08912 (dabrafenib)--- a reversible ATP-competitive kinase inhibitor and targets the MAPK pathway. PRKAR2A is the target of DB05798, an antisense oligonucleotide being investigated for treatment of many solid tumors. This target module is enriched with target proteins of drugs used for cancer treatment.

Application to Coronary Artery Disease

We finally applied the SCA to a set of seed proteins associated with coronary artery disease (CAD) derived from a large-scale meta-analysis of 48 genome-wide association studies assembling 60,801 cases and 123,504 controls [46]. This meta-analysis confirmed most of the known CAD-associated loci and also identified 10 new loci. The seed protein pool has 81 proteins, 65 of which can be found in the human interactome. The subnetwork induced by the seed protein pool has 18 proteins and 15 interactions (**Figure 5A**). Compared to a random protein set of the same size, this subnetwork has significantly more interactions and larger sized LCC (*P*= 1.9E-06 and 4.0E-07, respectively) (**Supplementary Figure 2**), confirming that disease proteins tend to interact with each other and function together to form a network module representing dysfunctional pathways of diseases.

After adding seed connectors as pathway mediators, the CAD network module has 88 proteins and 111 interactions, as shown in **Figure 5B**. This module integrates the physiological pathways related to genetic loci associated with coronary artery disease, e.g. LDL cholesterol and lipoproteins (*SORT1*, *APOB, APOE, LDLR*, *PCSK9, LPA*), triglyceride-rich lipoproteins (*LPL*, *APOA1*, *APOC1*), inflammation (*IL6R*, *CXCL12*), cellular proliferation and vascular remodeling (*MIA3*, *COL4A1*, *COL4A2*, *REST*, *NOA1*, *SMAD3*, *SWAP70*, *BCAS3*,*FLT1*, *PDGFD*) and vascular tone and nitric oxide signaling (*GUCY1A3*, *NOS3*, *EDNRA*) [47], providing an overall picture of pathobiological implications of these pathways. The Cardiovascular GO Annotation Initiative aims to manually annotate cardiovascular-associated genes or proteins by curating scientific literature and integrating results from high-quality high-throughput experiments [48]. So far over 4000 cardiovascular-associated genes have been prioritized as targets for annotation with GO terms. We assess whether the seed connectors are cardiovascular-associated proteins by using this resource of manual GO annotation and found that of 28 seed connectors, 18 are cardiovascular-associated proteins. The enrichment is significant compared to a random protein set of the same size from the human interactome (*P*=2.86E-06, **Figure 5C**), confirming the likely functional role of the seed connectors.

In addition, of 28 seed connectors, 14 are drug targets from the expanded set of drug targets in Drugbank. This enrichment is significant compared to random expectation (*P*= 3.9E-06, **Figure 5D**), demonstrating the likely biological relevance of the seed connectors. One seed connector

NRP1 (Neuropilin-1) is the target protein of pegaptanib (DB04895) -- an anti-angiogenic medicine for the treatment of neovascular age-related macular degeneration. NRP1 is a receptor regulating developmental and pathological angiogenesis and arteriogenesis, and mediating vascular permeability independently of vascular endothelial growth factor receptor-2 (VEGFR-2) activation [49]. In the CAD module, NRP1 connects REST, FLT1, and PDGFD (through PGF) which involve in cellular proliferation and vascular remodeling. *MFGE8* contains genetic loci associated with CAD, but its physiology pathway is uncertain yet [47]. We predict that it also functions in cellular proliferation and vascular remodeling since it connects to REST and FLT1 through NRP1.

Conclusion and Discussion

Identification of coherent disease modules -- a fundamental tenet of network medicine -- is important for deciphering the molecular mechanisms of diseases. Many algorithms have been developed to address this challenge. In this study, we provided an alternative method to ascertain disease modules in the neighborhood of the human interactome. Unlike many existing algorithms which ignore many loosely connected disease proteins, our algorithm, the Seed Connector Algorithm (SCA), aims to connect seed disease proteins maximally and efficiently by introducing as few extra intermediaries as possible. We demonstrate the biological relevance of the extra mediators (seed connectors) identified in terms of their functional similarity to seed proteins and their enrichment of drug targets. The seed connectors in disease modules provide the functional context of disease proteins and serve as guide for experimental validation of dysfunctional pathways.

This algorithm has a variety of applications, including construction of disease modules based on seed proteins; identification of drug target modules; and determination of network modules based on differentially expressed genes. For example, RNA-Seq is now frequently used to obtain genome-wide gene expression data under different conditions. For some studies, only a few differential expressed genes are identified, and these are scattered over the human interactome. Our algorithm can help to obtain an overall picture as to how these differentially expressed genes (gene products) may localize or cluster at the network level.

The human organism is an integrated network in which different layers of complex physiological systems interact to execute physiological functions [50, 51]. Prior work has shown that network topology determines physiological states and functions, which can affect human health and diseases [50]. In this construct, human diseases are closely related to perturbations in the physiological states of cells, tissues, and organ systems. Given a disease that involves specific tissues and organs, our algorithm could potentially be adapted to assist in identifying otherwise cryptic dynamical interactions among tissues and organ systems that are relevant to the disease phenotype by applying it to the global physiological network of the body.

We note that network modules are always constructed based on our current knowledge of seed proteins. We would expect that network modules induced by seed proteins will topologically change if we remove a few seed proteins with weak evidence or add more seed proteins once experimental evidence is available. However, we must emphasize that the location and neighborhood of the network modules in the human interactome will not change very much, so long as seed proteins with strong evidence have been included in the module ascertainment process.

Acknowledgements:

This work was supported by NIH grants GM107618, HL061795, and HG007690 (to J.L.). We wish to thank Stephanie Tribuna for her excellent technical assistance in preparing this article, and A.-L. Barabasi's laboratory at Northeastern University for providing access to the consolidated human interactome data and drug target data (to Dr. Feixiong Cheng).

References

[1] Bush WS, Moore JH. Chapter 11: Genome-wide association studies. PLoS Comput Biol. 2012;8:e1002822.

[2] Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12:56-68.

[3] Gustafsson M, Nestor CE, Zhang H, Barabasi AL, Baranzini S, Brunak S, et al. Modules, networks and systems medicine for understanding disease and aiding diagnosis. Genome Med. 2014;6:82.

[4] Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, et al. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science. 2015;347:1257601.

[5] 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;21:1109-21.

[6] Gill N, Singh S, Aseri TC. Computational disease gene prioritization: an appraisal. J Comput Biol. 2014;21:456-65.

[7] Piro RM, Di Cunto F. Computational approaches to disease-gene prediction: rationale, classification and successes. FEBS J. 2012;279:678-96.

[8] 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:e1004120.

[9] Silberberg Y, Kupiec M, Sharan R. GLADIATOR: a global approach for elucidating disease modules. Genome Med. 2017;9:48.

[10] Kohler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet. 2008;82:949-58.

[11] Scott MS, Perkins T, Bunnell S, Pepin F, Thomas DY, Hallett M. Identifying regulatory subnetworks for a set of genes. Mol Cell Proteomics. 2005;4:683-92.

[12] Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Muller T. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. Bioinformatics. 2008;24:i223-31.

[13] Huang SS, Fraenkel E. Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. Sci Signal. 2009;2:ra40.

[14] Tuncbag N, Braunstein A, Pagnani A, Huang SS, Chayes J, Borgs C, et al. Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem. J Comput Biol. 2013;20:124-36.

[15] 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;122:957-68.

[16] 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;437:1173-8.

[17] Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, et al. An empirical framework for binary interactome mapping. Nat Methods. 2009;6:83-90.

[18] Rolland T, Tasan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, et al. A proteome-scale map of the human interactome network. Cell. 2014;159:1212-26.

[19] Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, et al. The IntAct molecular interaction database in 2012. Nucleic Acids Res. 2012;40:D841-6.

[20] Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, et al. MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. 2012;40:D857-61.

[21] Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, et al. The BioGRID interaction database: 2015 update. Nucleic Acids Res. 2015;43:D470-8.

[22] Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database--2009 update. Nucleic Acids Res. 2009;37:D767-72.

[23] Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, et al. CORUM: the comprehensive resource of mammalian protein complexes--2009. Nucleic Acids Res. 2010;38:D497-501.

[24] Havugimana PC, Hart GT, Nepusz T, Yang H, Turinsky AL, Li Z, et al. A census of human soluble protein complexes. Cell. 2012;150:1068-81.

[25] Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, et al. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. 2003;31:374-8.

[26] Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, et al. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res. 2012;40:D261-70.

[27] Lee DS, Park J, Kay KA, Christakis NA, Oltvai ZN, Barabasi AL. The implications of human metabolic network topology for disease comorbidity. Proc Natl Acad Sci U S A. 2008;105:9880-5.

[28] Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, et al. Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature. 2012;490:556-60.

[29] Vinayagam A, Stelzl U, Foulle R, Plassmann S, Zenkner M, Timm J, et al. A directed protein interaction network for investigating intracellular signal transduction. Sci Signal. 2011;4:rs8.

[30] Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33:D514-7.

[31] Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, et al. UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. Methods Mol Biol. 2016;1374:23-54.

[32] Ramos EM, Hoffman D, Junkins HA, Maglott D, Phan L, Sherry ST, et al. Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. Eur J Hum Genet. 2014;22:144-7.

[33] Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 2008;36:D901-6.

[34] Zhu F, Han B, Kumar P, Liu X, Ma X, Wei X, et al. Update of TTD: Therapeutic Target Database. Nucleic Acids Res. 2010;38:D787-91.

[35] Thorn CF, Klein TE, Altman RB. PharmGKB: the Pharmacogenomics Knowledge Base. Methods Mol Biol. 2013;1015:311-20.

[36] Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 2012;40:D1100-7.

[37] Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res. 2007;35:D198-201.

[38] Pawson AJ, Sharman JL, Benson HE, Faccenda E, Alexander SP, Buneman OP, et al. The IUPHAR/BPS Guide to PHARMACOLOGY: an expert-driven knowledgebase of drug targets and their ligands. Nucleic Acids Res. 2014;42:D1098-106.

[39] Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. 2004;32:D115-9.

[40] Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. 2011;39:D52-7.

[41] Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, et al. GeneCards Version 3: the human gene integrator. Database (Oxford). 2010;2010:baq020.

[42] Chlebik M, Chlebikova J. The Steiner tree problem on graphs: Inapproximability results. Theoretical Computer Science. 2008;406:207-14.

[43] Wang RS, Hall KT, Giulianini F, Passow D, Kaptchuk TJ, Loscalzo J. Network analysis of the genomic basis of the placebo effect. JCI Insight. 2017;2.

[44] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498-504.

[45] Ruths T, Ruths D, Nakhleh L. GS2: an efficiently computable measure of GO-based similarity of gene sets. Bioinformatics. 2009;25:1178-84.

[46] Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat Genet. 2015;47:1121-30.

[47] Khera AV, Kathiresan S. Genetics of coronary artery disease: discovery, biology and clinical translation. Nat Rev Genet. 2017;18:331-44.

[48] Lovering RC, Roncaglia P, Howe DG, Laulederkind SJF, Khodiyar VK, Berardini TZ, et al. Improving Interpretation of Cardiac Phenotypes and Enhancing Discovery With Expanded Knowledge in the Gene Ontology. Circ Genom Precis Med. 2018;11:e001813.

[49] Roth L, Prahst C, Ruckdeschel T, Savant S, Westrom S, Fantin A, et al. Neuropilin-1 mediates vascular permeability independently of vascular endothelial growth factor receptor-2 activation. Sci Signal. 2016;9:ra42.

[50] Bashan A, Bartsch RP, Kantelhardt JW, Havlin S, Ivanov P. Network physiology reveals relations between network topology and physiological function. Nat Commun. 2012;3:702.

[51] Bartsch RP, Liu KK, Bashan A, Ivanov P. Network Physiology: How Organ Systems Dynamically Interact. PLoS One. 2015;10:e0142143.

Tables

Table 1. Coverage Rate of Seed Proteins in the LCC of the Disease Modules Constructed using the SCA and DIAMOnD. The *P* values were obtained by proportion test with module size as the sample size.

Disease	**# Seeds (LCC)**	**LCC of the predicted module (seed proteins)**		**Ratio of seed proteins in the LCC**		
		SCA	**DIAMOnD**	**SCA**	**DIAMOnD**	***P***
Adrenal gland diseases	18 (7)	23 (16)	15 (8)	69.6%	53.3%	0.24
Alzheimer disease	29 (7)	40 (28)	21 (9)	70%	42.3%	0.012
Amino acid metabolism, inborn errors of	52 (16)	58 (48)	17 (16)	82.3%	94.1%	0.44
Amyotrophic lateral sclerosis	21 (2)	31 (20)	16 (5)	64.5%	31.2%	0.008
Anemia, aplastic	21 (9)	24 (21)	12 (9)	87.5%	75.0%	0.27
Anemia, hemolytic	29 (7)	35 (27)	15 (7)	77.1%	46.7%	0.0075
Aneurysm	15 (4)	23 (15)	17 (9)	65.2%	52.9%	0.40
Arrhythmias cardiac	30 (6)	43 (29)	28 (15)	67.4%	53.6%	0.188
Arterial occlusive diseases	44 (4)	61 (40)	25 (5)	65.6%	20.0%	< 0.0001
Arteriosclerosis	38 (4)	54 (34)	24 (5)	63.0%	20.8%	< 0.0001
Arthritis, rheumatoid	42 (9)	57 (39)	38 (17)	68.4%	48.6%	0.028
Asthma	37 (7)	55 (35)	27 (7)	63.6%	25.9%	< 0.0001
Basal ganglia diseases	45 (9)	62 (41)	37 (16)	66.1%	43.2%	0.0085
Behçet syndrome	12 (3)	17 (10)	10 (3)	58.8%	30.0%	0.078
Bile duct diseases	31 (3)	46 (28)	25 (7)	60.9%	28.0%	0.001
Blood coagulation disorders	40 (28)	43 (40)	31 (28)	93.0%	90.3%	0.65
Blood platelet disorders	26 (7)	32 (26)	13 (7)	81.3%	53.5%	0.019
Breast neoplasms	40 (19)	43 (34)	28 (19)	79.1%	67.9%	0.21
Carbohydrate metabolism, inborn errors of	77 (11)	95 (69)	26 (14)	72.6%	53.8%	0.005
Carcinoma, renal cell	18 (3)	28 (17)	13 (3)	60.7%	23.1%	0.004
Cardiomyopathies	49 (27)	63 (48）	48 (33)	76.2%	68.8%	0.35
Cardiomyopathy, hypertrophic	22 (6)	26 (20)	21 (15)	76.9%	71.4%	0.65
Celiac disease	36 (3)	56 (34)	34 (12)	60.7%	35.3%	0.006
Cerebellar ataxia	30 (2)	42 (27)	18 (4)	64.3%	22.2%	< 0.0001
Cerebrovascular disorders	47 (7)	70 (45)	29 (4)	64.3%	13.8%	< 0.0001
Charcot Marie Tooth disease	24 (5)	36 (24)	17 (7)	66.7%	41.2%	0.031
Colitis, ulcerative	56 (4)	82 (53)	39 (11)	64.6%	28.2%	< 0.0001
Colorectal neoplasms	42 (18)	57 (42)	36 (21)	73.7%	58.3%	0.084
Coronary artery disease	31 (2)	46 (28)	22 (5)	60.9%	22.7%	0.0001
Crohn disease	72 (10)	99 (66)	58 (25)	66.7%	43.1%	0.0006
Death, sudden	19 (1)	31 (17)	20 (3)	54.8%	15.0%	0.0008
Diabetes mellitus, type 2	73 (12)	98 (65)	53 (20)	66.3%	37.7%	< 0.0001
Dwarfism	20 (4)	30 (17)	20 (5)	56.7%	25.0%	0.009
Esophageal diseases	24 (3)	39 (24)	22 (7)	61.5%	31.8%	0.009
Exophthalmos	13 (2)	N/A	N/A	N/A	N/A	N/A
Glomerulonephritis	18 (3)	N/A	N/A	N/A	N/A	N/A
Gout	13 (1)	15 (8)	9 (2)	53.3%	22.2%	0.045
Graves disease	13 (2)	N/A	N/A	N/A	N/A	N/A
Head and neck neoplasms	35 (4)	51 (35)	26 (10)	68.6%	38.5%	0.0024
Hypothalamic diseases	23 (2)	31 (20)	15 (7)	64.5%	46.7%	0.14
Leukemia B-cell	16(2)	22(14)	14(6)	63.6%	42.9%	0.16
Leukemia, myeloid	43 (17)	54 (41)	38 (25)	75.9%	65.8%	0.24

Lipid metabolism disorders	50 (14)	61 (47)	32 (14)	77.1%	43.8%	0.0001
Liver cirrhosis_	24 (3)	33 (20)	15 (3)	60.6%	20.0%	0.0004
Liver cirrhosis, biliary	23 (3)	33 (20)	15 (3)	60.6%	20.0%	0.0005
Lung diseases, obstructive	40 (7)	61 (38)	30 (7)	62.3%	23.3%	< 0.0001
Lupus erythematosus	74 (8)	99 (68)	52 (22)	68.7%	42.3%	0.0001
Lymphoma	24 (2)	35 (23)	15 (3)	65.7%	20.0%	0.0001
Lysosomal storage diseases	45 (24)	58 (44)	33 (24)	75.9%	72.7%	0.69
Macular degeneration	44 (8)	69 (40)	39 (10)	58.0%	25.6%	0.0001
Metabolic syndrome	14 (3)	18 (11)	10 (3)	61.1%	30.0%	0.046
Motor neuron disease	31 (2)	44 (29)	30 (15)	65.9%	50.0%	0.12
Multiple sclerosis	68(21)	94 (65)	57 (29)	69.1%	50.9%	0.0099
Muscular dystrophies	36 (17)	42 (33)	32 (23)	78.6%	71.9%	0.46
Mycobacterium infections	22 (6)	30 (22)	14 (6)	73.3%	42.9%	0.018
Myeloproliferative disorders	19 (6)	29 (19)	17 (7)	65.5%	41.2%	0.066
Nutritional and metabolic diseases	598 (327)	721 (597)	482 (358)	82.8%	74.3%	0.0001
Peroxisomal disorders	20 (19)	21 (20)	20 (19)	95.2%	95.0%	0.98
Psoriasis	53 (10)	72 (49)	37 (15)	68.1%	40.5%	0.0007
Purine-pyrimidine metabolism, inborn errors of	16 (2)	20 (11)	15 (6)	55.0%	40.0%	0.29
Renal tubular transport, inborn errors of	34 (3)	44 (29)	24 (10)	65.9%	41.7%	0.017
Sarcoma	25 (7)	32 (24)	21 (13)	75%	61.9%	0.26
Spastic paraplegia, hereditary	20 (2)	27 (16)	14 (3)	59.3%	21.4%	0.0026
Spinocerebellar ataxias	28 (2)	38 (25)	16 (4)	65.8%	25.0%	0.0002
Spinocerebellar degeneration	30 (2)	41 (27)	25 (11)	65.9%	44.0%	0.04
Spondyloarthropathies	17 (4)	24 (14)	13 (4)	58.3%	30.8%	0.044
Tauopathies	35 (10)	49 (34)	26 (14)	69.4%	53.8%	0.11
Uveal diseases	16 (4)	22 (13)	13 (4)	59.1%	30.8%	0.047
Varicose veins	20 (1)	26 (15)	15 (4)	57.7%	26.7%	0.014
Vasculitis	14 (3)	21 (12)	12 (3)	57.1%	25.0%	0.029

Figure Legends

Figure 1. Illustration of the SCA. In Iteration 0, the algorithm starts with seed proteins and induces a subnetwork that consists of only seed proteins. In Iteration 1, a protein (in grey) that maximally increases the size of LCC of the subnetwork is selected as a seed connector. In iterations 2-4, more proteins (in grey) that maximally increase the size of LCC of the subnetwork are selected as seed connectors. In iteration 5, 4 proteins are selected as seed connectors simultaneously as this is the only way to connect the remaining seed proteins to the LCC of the subnetwork.

Figure 2. The disease modules for cerebrovascular disorders constructed using the Seed Connector algorithm (A) and DIAMOnD (B). The blue nodes are disease seed proteins and the gray nodes are the seed connectors (A) or DIAMOnD proteins (B).

Figure 3. Functional similarity of seed connectors to seed proteins. (A) The significance of the functional similarity of seed connectors to seed proteins. (B) Functional similarity of seed connectors and DIAMOnD proteins to seed proteins. (C) Percentage of disease modules where predicted candidate disease proteins are functionally similar to seed proteins.

Figure 4. Drug target modules of DB09073 (palbociclib). (A). The target module of palbociclib induced by seed targets. (B) The target module of palbociclib identified by the SCA.

Figure 5. The CAD disease module discovered by the Seed Connector algorithm based on the seed proteins derived from a large-scale meta-analysis of GWAS. (A). The network module induced by the seed proteins. (B). The network module constructed using the Seed Connector algorithm. (C). The seed connectors are significantly enriched with cardiovascular-associated proteins. (D). The seed connectors are significantly enriched with drug targets.

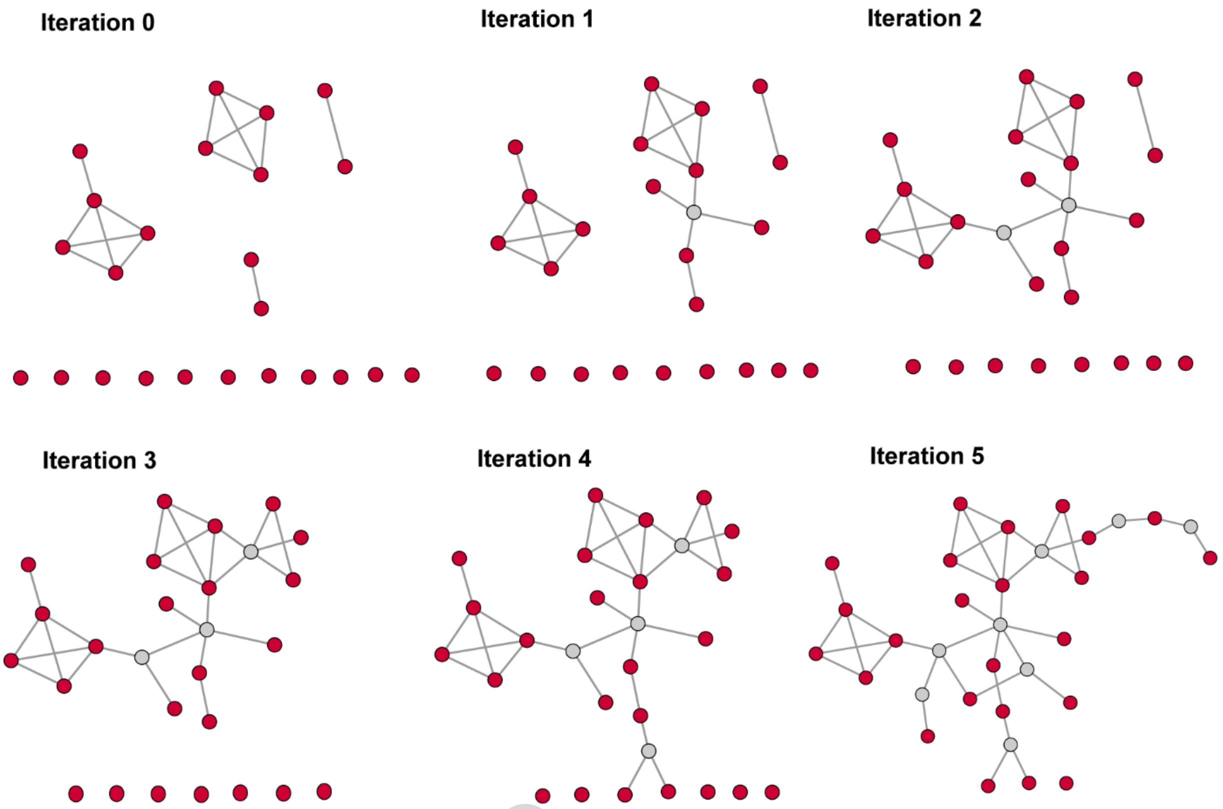

Iteration 0
Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 5

Figure 1

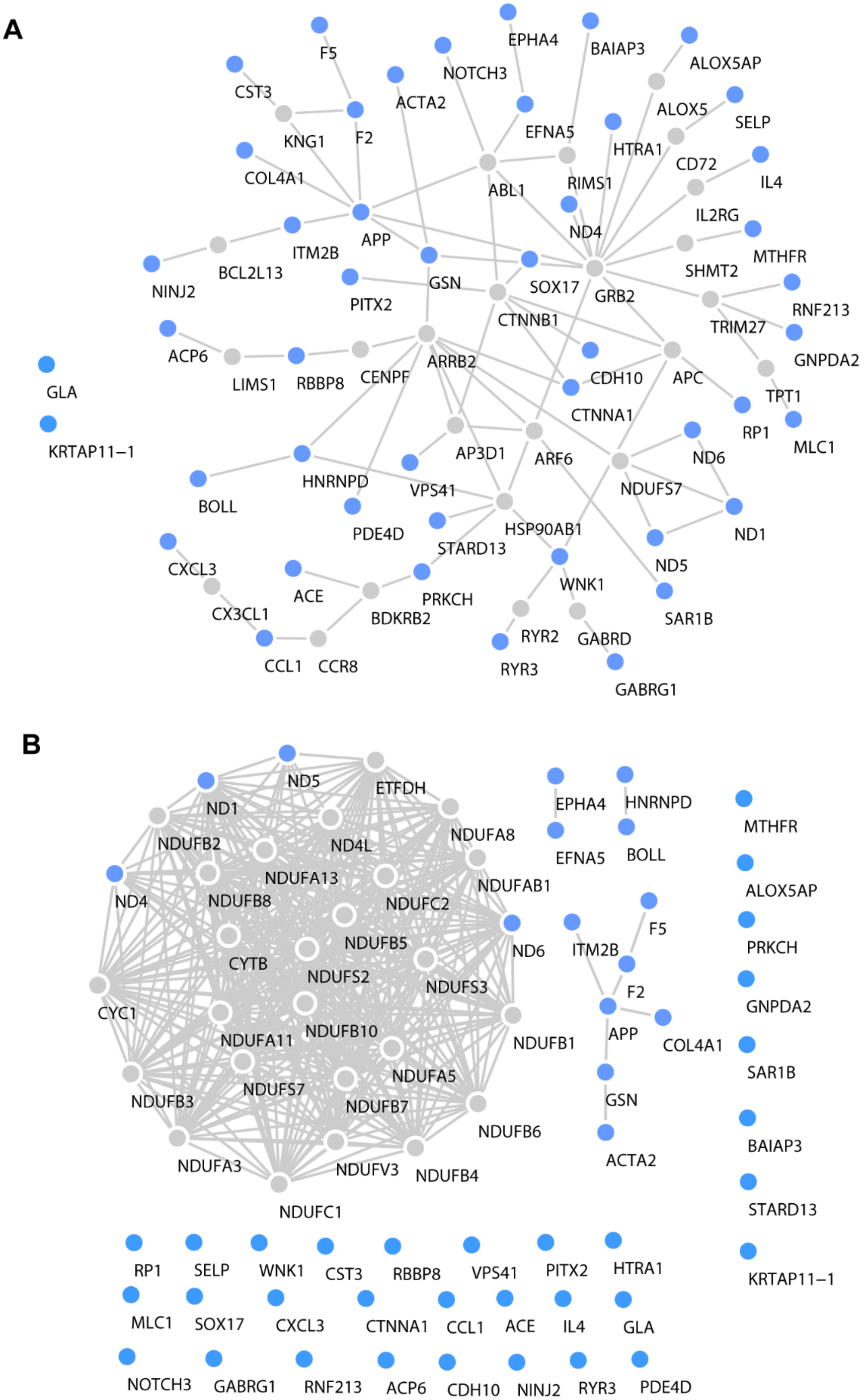

A
F5
EPHA4
BAIAP3
ALOX5AP
CST3
NOTCH3
ACTA2
ALOX5
SELP
KNG1
F2
EFNA5
HTRA1
COL4A1
CD72
IL4
ABL1
RIMS1
IL2RG
ND4
APP
MTHFR
ITM2B
SHMT2
BCL2L13
GSN
SOX17
GRB2
NINJ2
PITX2
RNF213
CTNNB1
TRIM27
GNPDA2
ACP6
ARRB2
CENPF
CDH10
APC
TPT1
LIMS1
RBBP8
GLA
CTNNA1
RP1
MLC1
KRTAP11−1
AP3D1
ARF6
ND6
HNRNPD
NDUFS7
BOLL
VPS41
PDE4D
HSP90AB1
ND1
STARD13
CXCL3
ND5
WNK1
ACE
PRKCH
CX3CL1
BDKRB2
SAR1B
RYR2
GABRD
CCL1
CCR8
RYR3
GABRG1
B
ND5
ETFDH
ND1
NDUFA8
EPHA4
HNRNPD
MTHFR
NDUFB2
ND4L
EFNA5
BOLL
NDUFA13
NDUFAB1
ALOX5AP
ND4
NDUFB8
NDUFC2
F5
NDUFB5
ND6
ITM2B
PRKCH
CYTB
NDUFS2
NDUFS3
F2
GNPDA2
CYC1
NDUFB10
APP
COL4A1
NDUFA11
NDUFB1
SAR1B
NDUFA5
NDUFS7
GSN
NDUFB3
NDUFB7
NDUFB6
BAIAP3
NDUFA3
ACTA2
NDUFV3
NDUFB4
STARD13
NDUFC1
RP1
SELP
WNK1
CST3
RBBP8
VPS41
PITX2
HTRA1
KRTAP11−1
MLC1
SOX17
CXCL3
CTNNA1
CCL1
ACE
IL4
GLA
NOTCH3
GABRG1
RNF213
ACP6
CDH10
NINJ2
RYR3
PDE4D

Figure 2

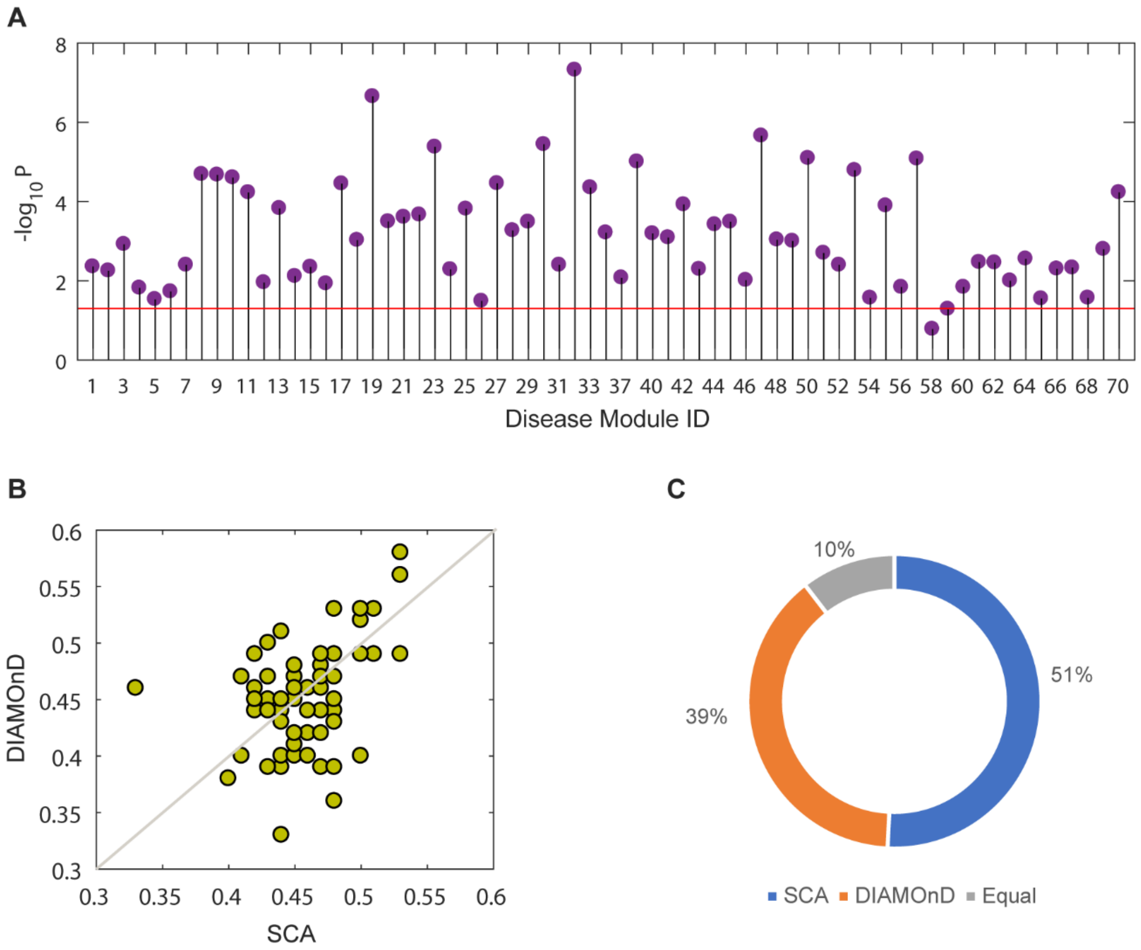

A
-log$_{10}$ P
0
2
4
6
8
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 37 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70
Disease Module ID
B
DIAMOnD
0.3
0.35
0.4
0.45
0.5
0.55
0.6
SCA
C
10%
51%
39%
■ SCA ■ DIAMOnD ■ Equal

Figure 3

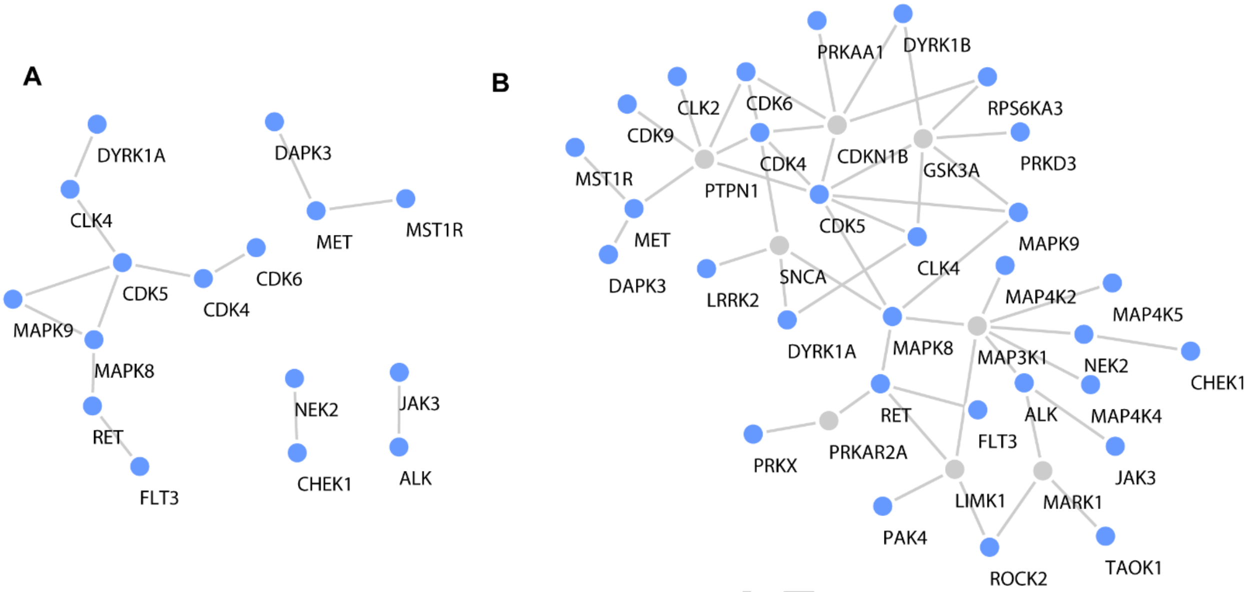

A
DYRK1A
DAPK3
CLK4
MET
MST1R
CDK6
CDK5
CDK4
MAPK9
MAPK8
RET
FLT3
NEK2
CHEK1
JAK3
ALK
B
PRKAA1
DYRK1B
CLK2
CDK6
RPS6KA3
CDK9
CDK4
CDKN1B
GSK3A
PRKD3
MST1R
PTPN1
CDK5
MAPK9
MET
DAPK3
LRRK2
SNCA
CLK4
MAP4K2
MAP4K5
DYRK1A
MAPK8
MAP3K1
NEK2
CHEK1
RET
ALK
MAP4K4
PRKX
PRKAR2A
FLT3
JAK3
LIMK1
MARK1
PAK4
ROCK2
TAOK1

Figure 4

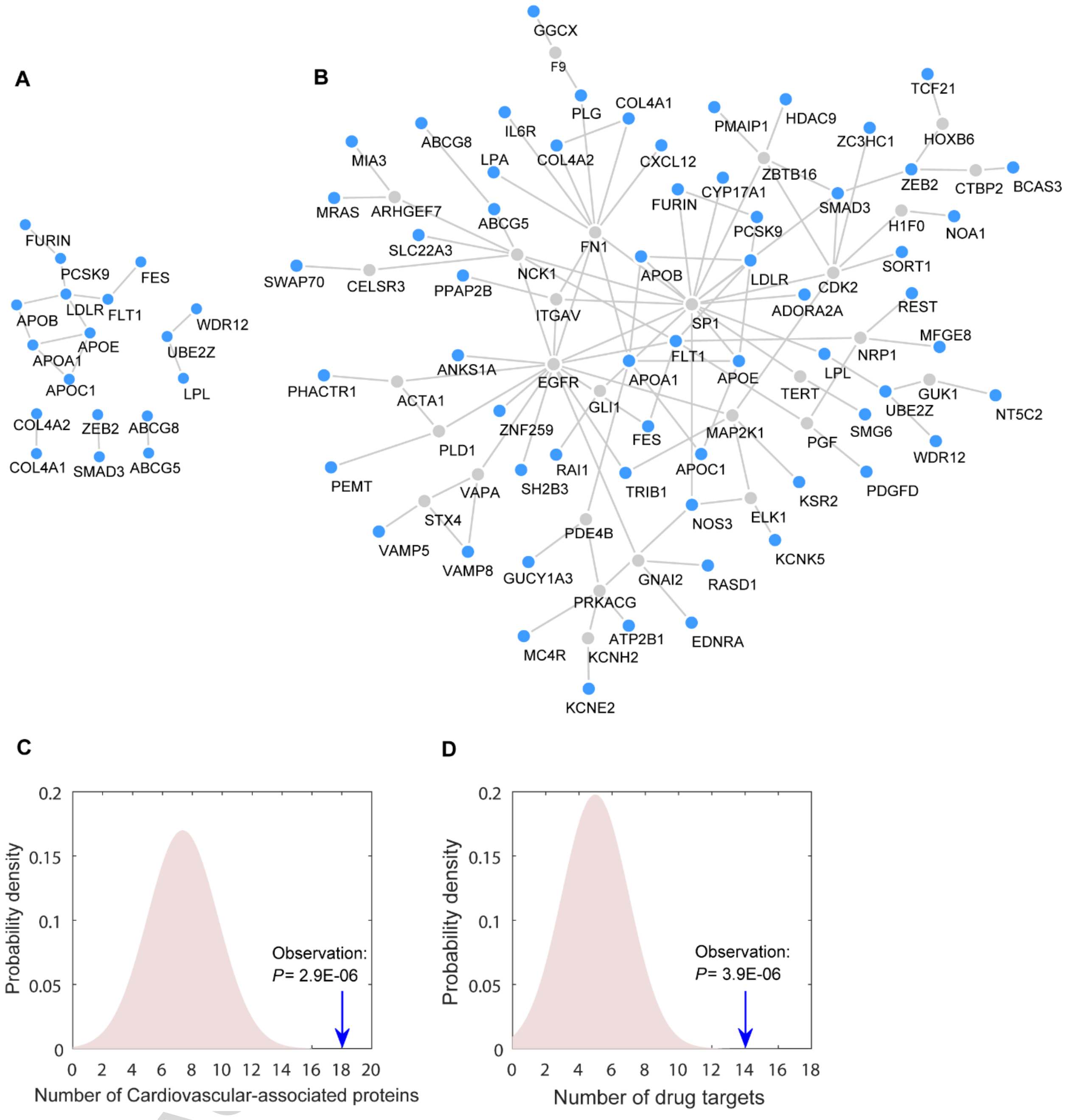
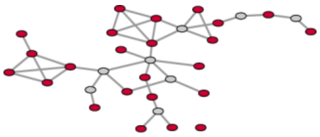

A
FURIN
PCSK9
FES
APOB
LDLR
FLT1
WDR12
APOE
UBE2Z
APOA1
APOC1
LPL
COL4A2
ZEB2
ABCG8
COL4A1
SMAD3
ABCG5
B
GGCX
F9
TCF21
PLG
COL4A1
PMAIP1
HDAC9
IL6R
ABCG8
ZC3HC1
HOXB6
MIA3
LPA
COL4A2
CXCL12
ZBTB16
ZEB2
CTBP2
BCAS3
CYP17A1
FURIN
MRAS
ARHGEF7
ABCG5
SMAD3
H1F0
NOA1
PCSK9
FN1
SLC22A3
SORT1
SWAP70
CELSR3
NCK1
APOB
LDLR
CDK2
PPAP2B
ADORA2A
ITGAV
REST
SP1
MFGE8
FLT1
NRP1
ANKS1A
PHACTR1
EGFR
APOA1
APOE
LPL
GUK1
NT5C2
ACTA1
GLI1
TERT
UBE2Z
SMG6
ZNF259
FES
MAP2K1
PGF
PLD1
WDR12
RAI1
APOC1
PEMT
VAPA
SH2B3
TRIB1
KSR2
PDGFD
STX4
NOS3
ELK1
PDE4B
VAMP5
KCNK5
VAMP8
GUCY1A3
GNAI2
RASD1
PRKACG
ATP2B1
EDNRA
MC4R
KCNH2
KCNE2
C
0.2
0.15
0.1
0.05
0
Probability density
Observation:
P= 2.9E-06
0 2 4 6 8 10 12 14 16 18 20
Number of Cardiovascular-associated proteins
D
0.2
0.15
0.1
0.05
0
Probability density
Observation:
P= 3.9E-06
0 2 4 6 8 10 12 14 16 18
Number of drug targets

Figure 5



Graphical abstract

Research highlights

- Identifying disease modules help to understand the molecular mechanisms of diseases
- We develop a network-based seed connector algorithm to pinpoint disease modules
- Seed connectors are critical for explaining the functional context of disease genes
- We validate the algorithm using 70 diseases and the binding targets of 220 drugs
- A coronary artery disease module is derived based on genetic loci from GWAS data