A statistical physics approach for disease module detection

Supplemental Material

Xu-Wen Wang[1], Dandi Qiao[1], Michael H. Cho[1], Dawn L. DeMeo[1], Edwin K. Silverman[1], Yang-Yu Liu[1]

[1]*Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA*

1. EXISTING DISEASE MODULE DETECTION METHOD USED FOR COMPARISON 2

1.1 DIAMOnD. 2

1.2 ModuleDiscoverer. 2

1.3 DOMINO. 2

1.4 ROBUST. 3

1.5 Hierarchical HotNet. 3

2. EXISTING OPTIMAL SUBGRAPHS IDENTIFICATION METHODS. 3

2.1 Maximum-Weight Connected Subgraph (MWCS). 3

2.2 Prize-Collecting Steiner Tree (PCST). 4

REFERENCES 5

SUPPLEMENTAL FIGURES 7

1. Existing disease module detection method used for comparison

Numerous computational methods have been developed to identify disease modules (de Weerd et al. 2020; Ulitsky et al. 2010; Bersanelli et al. 2016) by projecting the molecular profiles into the network and assigning scores/weights to nodes/edges. Those methods can be briefly classified into three categories: (1) Permutation-based methods (e.g., Hierarchical HotNet (Reyna et al. 2018) and DOMINO (Levi et al. 2021)), which involve a network randomization process to correct for typical PPI network biases. (2) Seed-genes-based methods (e.g., DIAMOnD (Ghiassian et al. 2015)), which "grow" the disease module from the seed genes in a local agglomerative process. (3) Heuristic-based methods, e.g., ModuleDiscoverer (Vlaic et al. 2018), ROBUST (Bernett et al. 2022), attempt to find optimal subgraphs within the molecular interaction network. We compared our RFIM method with the following five existing methods.

1.1 DIAMOnD.

DIseAse Module Detection (DIAMOnD) algorithm (Ghiassian et al. 2015) starts from a set of seed genes and prioritizes the other proteins of the interactome for their putative disease relevance. Therefore, the identified disease module might significantly rely on the initial seed genes. In this work, we used the implementation of DIAMOnD incorporated in R (Team 2013) package MODifieR (de Weerd et al. 2020) which can simultaneously use the PPI network and gene-wise *p*-values. The number of genes in the disease module in DIAMOnD set to be 5% of the total genes, just the same as the RFIM. We used the default values for other parameters: seed weight (10); The *p*-value cutoff for differentially expressed genes is 0.05 and these significantly expressed genes were selected as seed genes and the seed genes were excluded from the disease module in MODifieR (de Weerd et al. 2020).

1.2 ModuleDiscoverer.

ModuleDiscoverer (Vlaic et al. 2018) uses a randomization heuristic-based approximation of the community structure based on maximal clique enumeration problem. ModuleDiscoverer first approximates the underlying community structure by iterative enumeration of gene cliques from random seed genes. Then, the union of all significantly enriched cliques are ensembled into a large module (Vlaic et al. 2018). Number of permutations is set to be 1,000. Number of times the algorithm is repeated is set to be 3.

1.3 DOMINO.

DIMINO (Levi et al. 2021) (Discovery of active Modules In Networks using Omics) finds disjoint connected subnetworks that the active genes are over-represented based on a permutation-based method that empirically evaluate GO terms. The main steps of DOMINO are: (1) Partition the network into disjoint and highly connected slices. (2) Detect relevant slices including over-represented active genes. (3) For each slice, refine it to sub-slice and repartition slice into putative modules. (4) Final modules are those with over-represented by active genes. The active genes were selected as genes with *p*-value lower than 10^{-4}.

1.4 ROBUST.

ROBUST (Bernett et al. 2022) is a disease module mining method via enumeration of diverse prize-collecting Steiner trees. In naïve approach, disease module is identified by running mining method many times on shuffled input and return the subgraph induced by nodes contained in many of the return modules. Robust overcome the runtime limitation of naïve approach by enumerating pairwise diverse rather than merely pairwise non-identical disease modules (Bernett et al. 2022). We used the default parameters for ROBUST: initial fraction 0.25; reduction factor 0.9; number of Steiner trees to be computed 30 and threshold 0.1.

1.5 Hierarchical HotNet.

Hierarchical HotNet (Reyna et al. 2018) is an algorithm that finds a hierarchy of altered subnetworks and identify statistically significant subnetworks in the hierarchy. Hierarchical HotNet (1) combines network topology and node scores, (2) defines a similarity matrix from network using a random walk, (3) construct a hierarchy of clusters consisting of strongly connected components and (4) assesses the statistical significance of clusters in the hierarchy. As Hierarchical HotNet requires much longer computation time, we reduced the number of permutations into 20.

2. Existing optimal subgraphs identification methods.

This type of optimization procedure, i.e., finding optimal subgraphs within the molecular interaction network, turns out to be computationally intractable in general due to their NP-hard nature. For example, MWCS and PCST are two classical formalisms of this optimization procedure.

2.1 Maximum-Weight Connected Subgraph (MWCS).

In cases where we have no prior-knowledge of the edge weights that capture the confidence level of the protein-protein interactions, we focus on the node weights (i.e., the gene-wise p-values). The optimization problem is referred to as the MWCS problem, where given a graph $G(V, E)$ with node weights z_i, we try to find the connected subgraph with maximum weight:

$$\max_{A \subseteq G} S_A, \text{ with } S_A \equiv \frac{Z_A - \mu_k}{\sigma_k} \text{ and } Z_A \equiv \frac{1}{\sqrt{k}} \sum_{i \in A} z_i \qquad [1]$$

Here $z_i \equiv \Phi^{-1}(1 - p_i)$, p_i is the p-value of node (gene) i obtained from a genome-wide association study (GWAS) and Φ^{-1} denotes the inverse normal cumulated distribution function (CDF) that converts the p-values to Z-scores (node weights). Z_A is the aggregate score of subgraph A with size $k = |A|$, and S_A is the normalized score of subgraph A, calculated by comparing Z_A to the distribution of aggregate scores for node sets of size k randomly chosen from the graph. Two heuristic algorithms have been developed to solve the MWCS problem approximately: (i) the simulated annealing algorithm (Van Laarhoven and Aarts 1987) and (ii) the greedy search algorithm (Nacu et al. 2007). The exact solution of the MWCS problem can only be obtained by solving the corresponding integer linear programming (ILP) problem, which is NP-hard in general.

Dittrich *et al.* introduced a scalable scoring function with false discovery rate (FDR) as a meaningful adjustment parameter (Dittrich et al. 2008). The additivity of this logarithmic score enables them to exactly solve the MWCS problem using an ILP approach (Dittrich et al. 2008; El-Kebir and Klau 2014). A fundamental limitation of the MWCS formalism is that edge weights are completely ignored, which might cause serious problems in the presence of many false positive interactions. Moreover, the MWCS formalism could be confounded by nodes with high degrees, such as *TP53* and *UBC*, which interact with hundreds of other proteins in the cellular molecular interactome and hence may make them more likely to be selected in the optimal subgraphs.

2.2 Prize-Collecting Steiner Tree (PCST).

Molecular interaction data (especially PPIs) typically contain large numbers of false positives that can diminish the predictive power of integrative approaches. To optimally exploit the large amount of human interaction data that have been generated in the past decade, the detection of optimal subgraphs of the functional network has been formalized as a PCST problem (Bailly-Bechet et al. 2011): Given a graph $G(V, E)$ with positive node weights (also called node prizes) b_i and positive edge weights (also called edge costs) c_e, find the connected subgraph A that maximizes the following objective function:

$$\max_{A \subseteq G} S_A, \text{ with } S_A \equiv \lambda \sum_{i \in V(A)} b_i - \sum_{e \in E(A)} c_e \qquad [2]$$

Here, the edge costs c_e's are chosen such that edges with high confidence interactions (e.g., protein interactions verified in small-scale experiments or found in many large-scale datasets) have lower values with respect to low confidence ones (Bailly-Bechet et al. 2011; Biazzo et al. 2012). The node prizes are computed as $b_i = -\log p_i$ with p_i the p-value of node (gene) i. (Note that b_i is, by definition, positive.) The control parameter λ regulates the trade-off between the edge costs and node prizes, and its value indirectly controls the size of the optimal subgraph A. In spite of its apparent simplicity, the PCST problem is computationally intractable (NP-hard) (Bailly-Bechet et al. 2011; Biazzo et al. 2012). A fundamental limitation of the PCST formalism is that there is no clear theoretical criterion to choose the control parameter λ. In practice, one has to solve the PCST problem for a wide range of λ values and then use some heuristics to select the best λ value (Bailly-Bechet et al. 2011; Gitter et al. 2014; Balbin et al. 2013). Moreover, by definition, the output of the PCST formalism is always a tree, i.e., an acyclic connected subgraph. Recently, the PCST problem has been extended to the prize-collecting Steiner forest (PCSF) problem by introducing another control parameter, where the optimal subgraph is a set of disjoint tree graphs (Gitter et al. 2014). However, the disease module is not necessarily a tree or a forest graph.

References

Bailly-Bechet M, Borgs C, Braunstein A, Chayes J, Dagkessamanskaia A, François J-M, Zecchina R. 2011. Finding undetected protein associations in cell signaling by belief propagation. *Proc Natl Acad Sci* **108**: 882–887.

Balbin OA, Prensner JR, Sahu A, Yocum A, Shankar S, Malik R, Fermin D, Dhanasekaran SM, Chandler B, Thomas D. 2013. Reconstructing targetable pathways in lung cancer by integrating diverse omics data. *Nat Commun* **4**: 1–13.

Bernett J, Krupke D, Sadegh S, Baumbach J, Fekete SP, Kacprowski T, List M, Blumenthal DB. 2022. Robust disease module mining via enumeration of diverse prize-collecting Steiner trees ed. L. Cowen. *Bioinformatics* **38**: 1600–1606.

Bersanelli M, Mosca E, Remondini D, Castellani G, Milanesi L. 2016. Network diffusion-based analysis of high-throughput data for the detection of differentially enriched modules. *Sci Rep* **6**: 34841.

Biazzo I, Braunstein A, Zecchina R. 2012. Performance of a cavity-method-based algorithm for the prize-collecting Steiner tree problem on graphs. *Phys Rev E* **86**: 026706.

de Weerd HA, Badam TVS, Martínez-Enguita D, Åkesson J, Muthas D, Gustafsson M, Lubovac-Pilav Z. 2020. MODifieR: an Ensemble R Package for Inference of Disease Modules from Transcriptomics Networks ed. L. Cowen. *Bioinformatics* **36**: 3918–3919.

Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Müller T. 2008. Identifying functional modules in protein–protein interaction networks: an integrated exact approach. *Bioinformatics* **24**: i223–i231.

El-Kebir M, Klau GW. 2014. Solving the maximum-weight connected subgraph problem to optimality. *ArXiv Prepr ArXiv14095308*.

Ghiassian SD, Menche J, Barabási A-L. 2015. A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome ed. A. Rzhetsky. *PLOS Comput Biol* **11**: e1004120.

Gitter A, Braunstein A, Pagnani A, Baldassi C, Borgs C, Chayes J, Zecchina R, Fraenkel E. 2014. Sharing information to reconstruct patient-specific pathways in heterogeneous diseases. In *Biocomputing 2014*, pp. 39–50, World Scientific.

Kolberg L, Raudvere U, Kuzmin I, Vilo J, Peterson H. 2020. gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. *F1000Research* **9**: ELIXIR-709.

Levi H, Elkon R, Shamir R. 2021. DOMINO: a network-based active module identification algorithm with reduced rate of false calls. *Mol Syst Biol* **17**: e9593.

Nacu Ş, Critchley-Thorne R, Lee P, Holmes S. 2007. Gene expression network analysis and applications to immunology. *Bioinformatics* **23**: 850–858.

Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, García-García J, Sanz F, Furlong LI. 2016. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res* gkw943.

Reyna MA, Leiserson MD, Raphael BJ. 2018. Hierarchical HotNet: identifying hierarchies of altered subnetworks. *Bioinformatics* **34**: i972–i980.

Team RC. 2013. R: A language and environment for statistical computing.

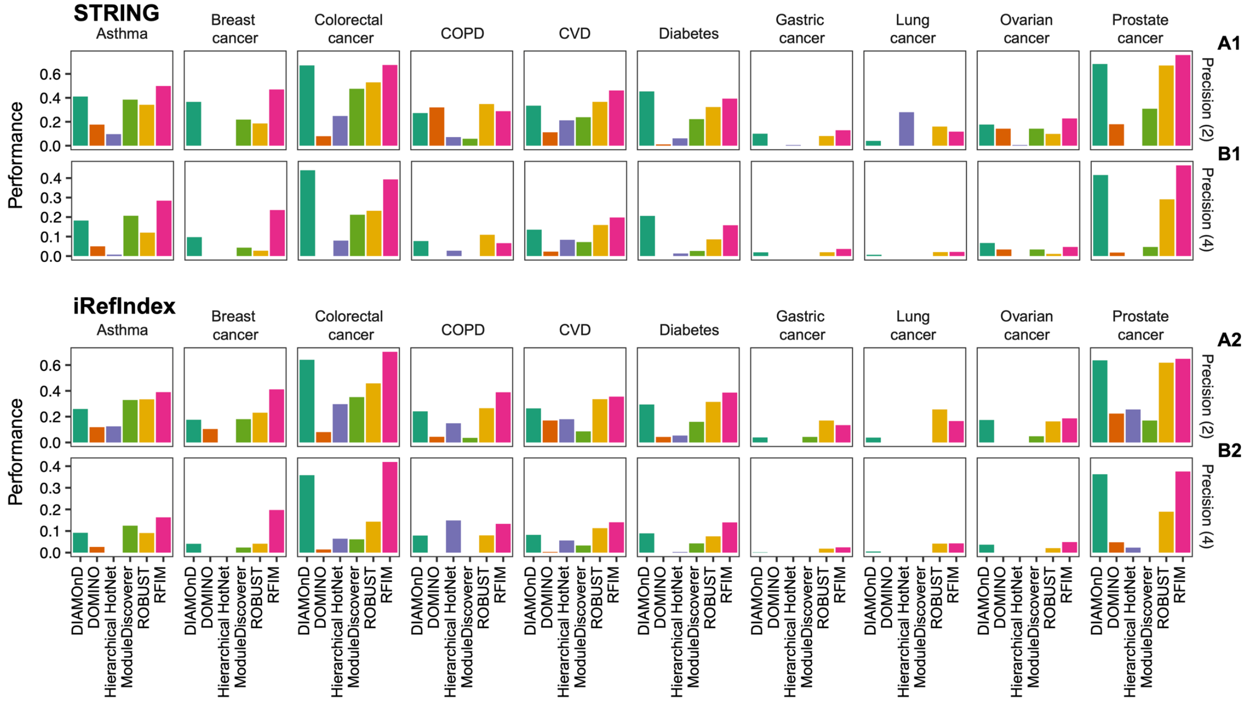
Ulitsky I, Krishnamurthy A, Karp RM, Shamir R. 2010. DEGAS: De Novo Discovery of Dysregulated Pathways in Human Diseases ed. T. Ravasi. *PLoS ONE* **5**: e13367.

Van Laarhoven PJ, Aarts EH. 1987. Simulated annealing. In *Simulated annealing: Theory and applications*, pp. 7–15, Springer.

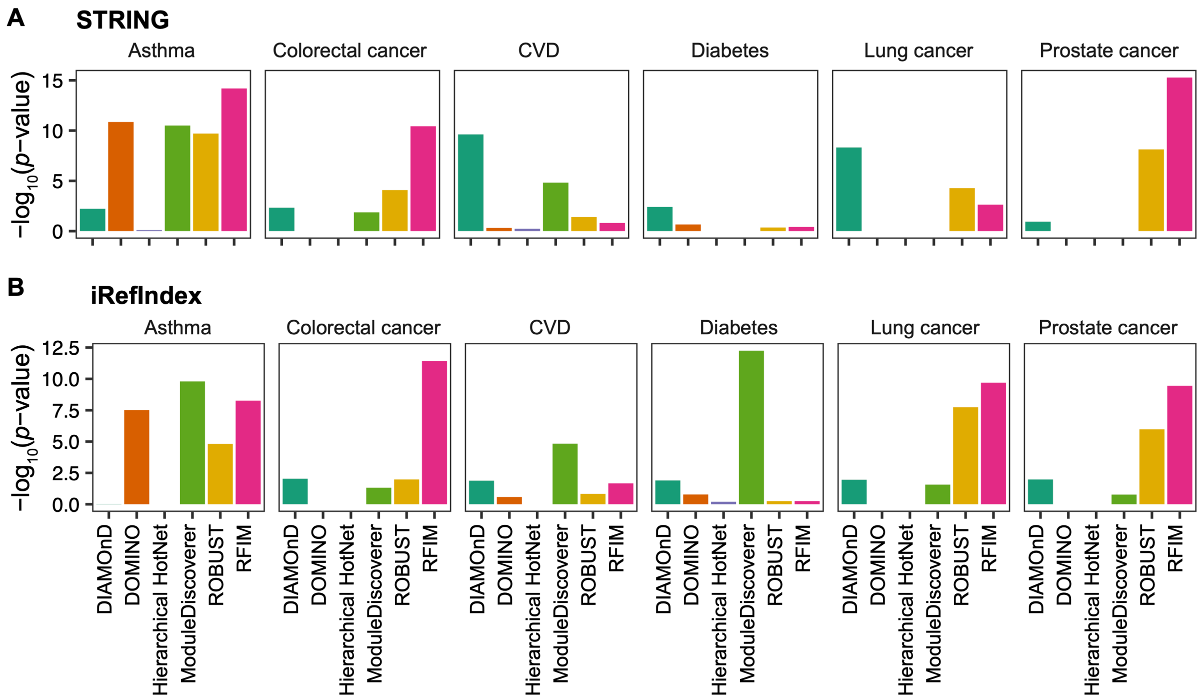
Vlaic S, Conrad T, Tokarski-Schnelle C, Gustafsson M, Dahmen U, Guthke R, Schuster S. 2018. ModuleDiscoverer: Identification of regulatory modules in protein-protein interaction networks. *Sci Rep* **8**: 433.

Yu G, He Q-Y. 2016. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol Biosyst* **12**: 477–479.

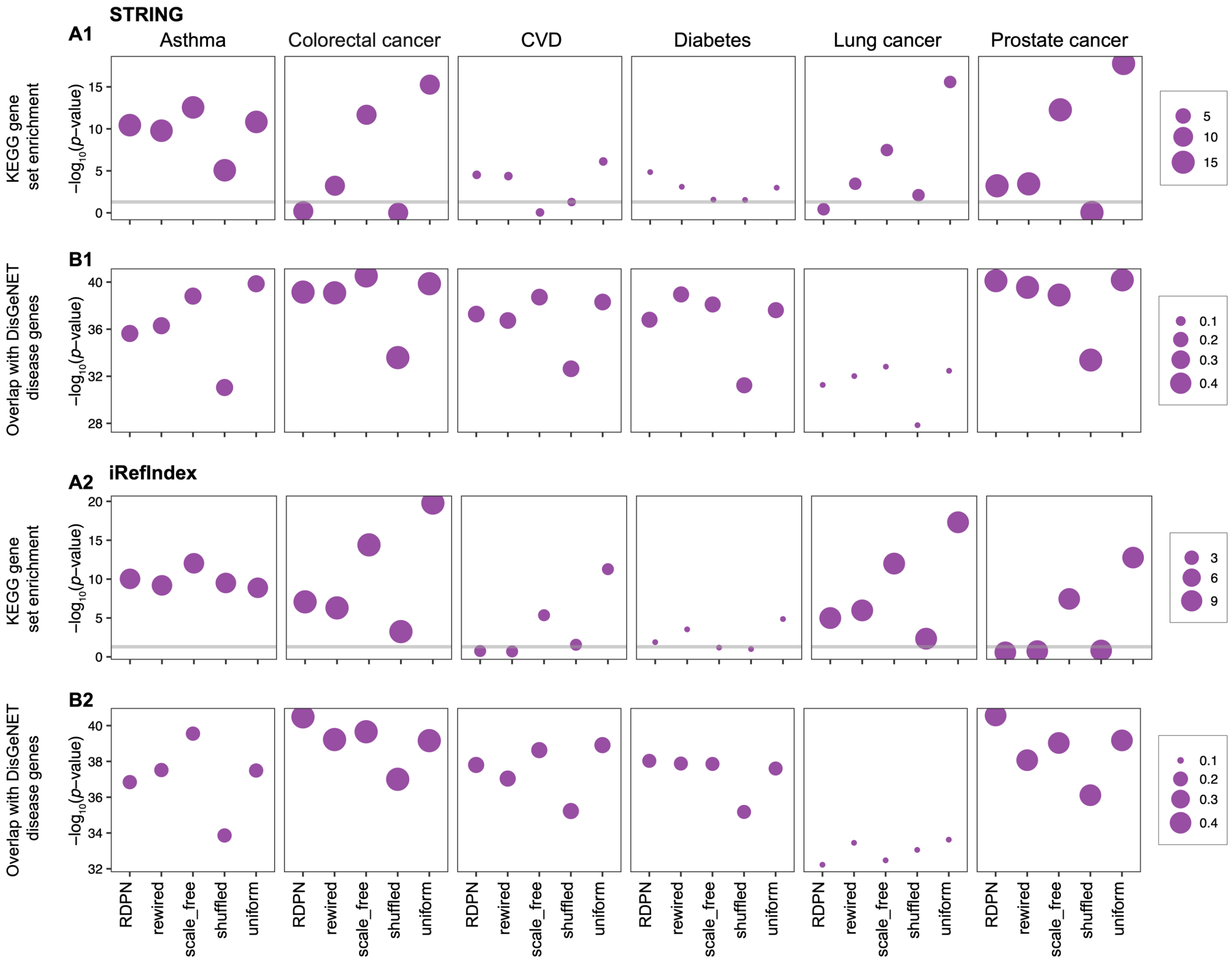
Supplemental Figures



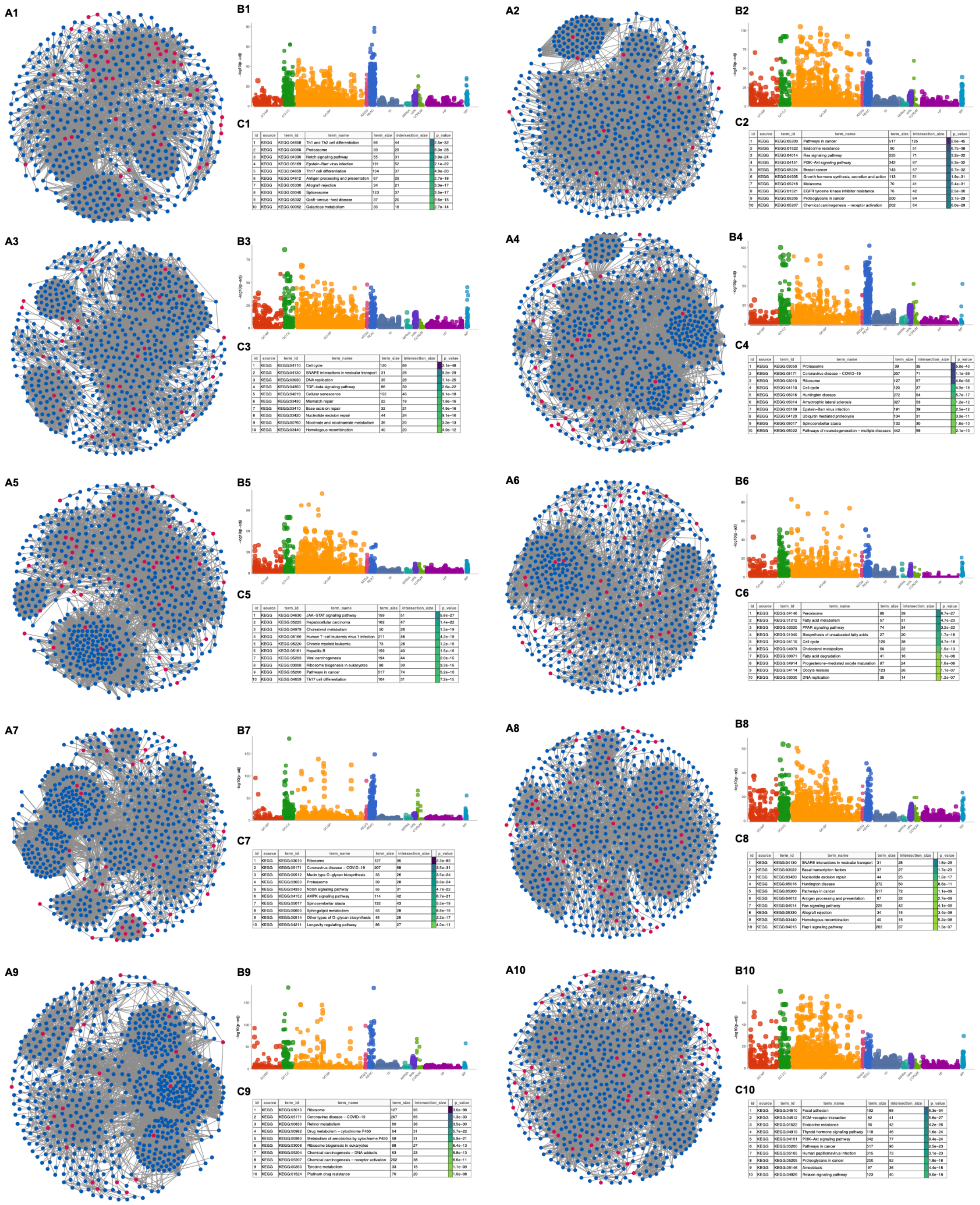
Supplemental Fig S1: Enrichment analysis for genes in disease module detected from integrating the GWAS with two interactomes for each phenotype and method. We obtained the enriched pathways for disease genes using the ReactomePA package (Yu and He 2016) whose *p*-values are lower than 0.05 cutoff adjust by False Discovery Rate (FDR). Then, we extracted the disease-associated genes of each phenotype using the DisGeNET (Piñero et al. 2016) database and calculated the precision defined as fraction of enriched pathways with at least two (A) or four (B) disease-associated genes over all enriched pathways.



Supplemental Fig S2: KEGG enrichment analysis for genes in disease module detected from integrating the GWAS with two interactomes for each phenotype and method. We first extracted the enriched pathways of genes in disease module with *p*-value cutoff 0.05. Then, we computed the average *p*-values of pathways related to the phenotype in KEGG database. KEGG pathways: asthma: hsa05310; breast cancer: hsa05224; lung cancer: hsa05222, hsa05223; colorectal cancer: hsa05210; gastric cancer: hsa05226; prostate cancer: hsa05215; CVD: hsa05410, hsa05412, hsa05414, hsa05416; diabetes: hsa04940 and hsa04930. Note that there are no KEGG pathways associated with COPD and ovarian cancer; and there are no significant KEGG pathways (with *p*<0.05, based on GSEA) associated with breast cancer and gastric cancer. Hence, those four diseases were not considered here.



Supplemental Fig S3: Meaningfulness test of disease module identified by RFIM. Two meaningfulness tests were performed: (A1-A2) KEGG gene set enrichment analysis (GSEA). We compared the *p*-value of gene set in the disease module identified from the original PPI network with that identified from 10 randomized PPI networks generated by each of the following five random network generators: degree preserving (RDPN), expected degree preserving (rewired), topology preserving (shuffled), scale-free (scale_free) and uniform (uniform). (B1-B2) Overlap with DisGeNET disease gene. We compared the overlap genes with DisGeNET genes in the original PPI network with that in 10 randomized PPI networks generated by each of the above five random network generators. Then we showed the *p*-values using one-sided one-sample t-test. (A): STRING and (B) iRefIndex. Node size represents the *p*-value. KEGG pathways: asthma: hsa05310; breast cancer: hsa05224; lung cancer: hsa05222, hsa05223; colorectal cancer: hsa05210; gastric cancer: hsa05226; prostate cancer: hsa05215; CVD: hsa05410, hsa05412, hsa05414, hsa05416; diabetes: hsa04940 and hsa04930. Note that there are no KEGG pathways associated with COPD and ovarian cancer; and there are no significant KEGG pathways (with p<0.05, based on GSEA) associated with breast cancer and gastric cancer. Hence, those four diseases were not considered here.



C1

id	source	term_id	term_name	term_size	intersection_size		p_value
1	KEGG	KEGG:04658	Th1 and Th2 cell differentiation	86	44	■	2.5e−32
2	KEGG	KEGG:03050	Proteasome	39	29	■	8.3e−28
3	KEGG	KEGG:04330	Notch signaling pathway	55	31	■	3.9e−24
4	KEGG	KEGG:05169	Epstein−Barr virus infection	191	52	■	2.1e−22
5	KEGG	KEGG:04659	Th17 cell differentiation	104	37	■	4.9e−20
6	KEGG	KEGG:04612	Antigen processing and presentation	67	29	■	2.7e−18
7	KEGG	KEGG:05330	Allograft rejection	34	21	■	3.3e−17
8	KEGG	KEGG:03040	Spliceosome	123	37	■	3.5e−17
9	KEGG	KEGG:05332	Graft−versus−host disease	37	20	■	9.6e−15
10	KEGG	KEGG:00052	Galactose metabolism	30	18	■	2.7e−14

C2

id	source	term_id	term_name	term_size	intersection_size		p_value
1	KEGG	KEGG:05200	Pathways in cancer	517	125	■	2.6e−45
2	KEGG	KEGG:01522	Endocrine resistance	90	51	■	6.7e−38
3	KEGG	KEGG:04014	Ras signaling pathway	225	71	■	3.2e−32
4	KEGG	KEGG:04151	PI3K−Akt signaling pathway	342	87	■	5.3e−32
5	KEGG	KEGG:05224	Breast cancer	143	57	■	9.7e−32
6	KEGG	KEGG:04935	Growth hormone synthesis, secretion and action	113	51	■	1.9e−31
7	KEGG	KEGG:05218	Melanoma	70	41	■	5.4e−31
8	KEGG	KEGG:01521	EGFR tyrosine kinase inhibitor resistance	76	42	■	2.5e−30
9	KEGG	KEGG:05205	Proteoglycans in cancer	200	64	■	3.1e−29
10	KEGG	KEGG:05207	Chemical carcinogenesis − receptor activation	202	64	■	6.0e−29

C3

id	source	term_id	term_name	term_size	intersection_size		p_value
1	KEGG	KEGG:04110	Cell cycle	120	68	■	2.1e−48
2	KEGG	KEGG:04130	SNARE interactions in vesicular transport	31	28	■	9.2e−29
3	KEGG	KEGG:03030	DNA replication	35	28	■	1.1e−25
4	KEGG	KEGG:04350	TGF−beta signaling pathway	86	39	■	2.6e−22
5	KEGG	KEGG:04218	Cellular senescence	152	46	■	9.1e−18
6	KEGG	KEGG:03430	Mismatch repair	22	18	■	1.8e−16
7	KEGG	KEGG:03410	Base excision repair	32	21	■	4.9e−16
8	KEGG	KEGG:03420	Nucleotide excision repair	44	24	■	9.1e−16
9	KEGG	KEGG:00760	Nicotinate and nicotinamide metabolism	36	20	■	3.3e−13
10	KEGG	KEGG:03440	Homologous recombination	40	20	■	4.9e−12

C4

id	source	term_id	term_name	term_size	intersection_size		p_value
1	KEGG	KEGG:03050	Proteasome	39	35	■	5.8e−40
2	KEGG	KEGG:05171	Coronavirus disease − COVID−19	207	71	■	1.1e−39
3	KEGG	KEGG:03010	Ribosome	127	57	■	4.6e−39
4	KEGG	KEGG:04110	Cell cycle	120	37	■	4.9e−18
5	KEGG	KEGG:05016	Huntington disease	272	54	■	5.7e−17
6	KEGG	KEGG:05014	Amyotrophic lateral sclerosis	327	53	■	1.2e−12
7	KEGG	KEGG:05169	Epstein−Barr virus infection	191	39	■	2.5e−12
8	KEGG	KEGG:04120	Ubiquitin mediated proteolysis	134	31	■	3.9e−11
9	KEGG	KEGG:05017	Spinocerebellar ataxia	132	30	■	1.6e−10
10	KEGG	KEGG:05022	Pathways of neurodegeneration − multiple diseases	442	59	■	2.1e−10

C5

id	source	term_id	term_name	term_size	intersection_size		p_value
1	KEGG	KEGG:04630	JAK−STAT signaling pathway	159	51	■	5.8e−27
2	KEGG	KEGG:05225	Hepatocellular carcinoma	162	47	■	1.4e−22
3	KEGG	KEGG:04979	Cholesterol metabolism	50	26	■	1.5e−19
4	KEGG	KEGG:05166	Human T−cell leukemia virus 1 infection	211	48	■	4.2e−18
5	KEGG	KEGG:05220	Chronic myeloid leukemia	73	28	■	1.2e−16
6	KEGG	KEGG:05161	Hepatitis B	159	40	■	1.5e−16
7	KEGG	KEGG:05203	Viral carcinogenesis	194	44	■	2.0e−16
8	KEGG	KEGG:03008	Ribosome biogenesis in eukaryotes	88	30	■	3.3e−16
9	KEGG	KEGG:05200	Pathways in cancer	517	74	■	5.2e−16
10	KEGG	KEGG:04659	Th17 cell differentiation	104	31	■	7.2e−15

C6

id	source	term_id	term_name	term_size	intersection_size		p_value
1	KEGG	KEGG:04146	Peroxisome	80	39	■	6.7e−27
2	KEGG	KEGG:01212	Fatty acid metabolism	57	31	■	4.7e−23
3	KEGG	KEGG:03320	PPAR signaling pathway	74	34	■	3.2e−22
4	KEGG	KEGG:01040	Biosynthesis of unsaturated fatty acids	27	20	■	1.7e−18
5	KEGG	KEGG:04110	Cell cycle	120	38	■	4.7e−18
6	KEGG	KEGG:04979	Cholesterol metabolism	50	22	■	1.5e−13
7	KEGG	KEGG:00071	Fatty acid degradation	41	16	■	1.1e−08
8	KEGG	KEGG:04914	Progesterone−mediated oocyte maturation	97	24	■	1.6e−08
9	KEGG	KEGG:04114	Oocyte meiosis	123	26	■	1.1e−07
10	KEGG	KEGG:03030	DNA replication	35	14	■	1.2e−07

C7

id	source	term_id	term_name	term_size	intersection_size		p_value
1	KEGG	KEGG:03010	Ribosome	127	95	■	2.3e−89
2	KEGG	KEGG:05171	Coronavirus disease − COVID−19	207	68	■	3.5e−31
3	KEGG	KEGG:00512	Mucin type O−glycan biosynthesis	33	26	■	3.5e−24
4	KEGG	KEGG:03050	Proteasome	39	28	■	3.6e−24
5	KEGG	KEGG:04330	Notch signaling pathway	55	31	■	4.7e−22
6	KEGG	KEGG:04152	AMPK signaling pathway	114	42	■	6.7e−21
7	KEGG	KEGG:05017	Spinocerebellar ataxia	132	43	■	5.5e−19
8	KEGG	KEGG:00600	Sphingolipid metabolism	53	28	■	8.8e−19
9	KEGG	KEGG:00514	Other types of O−glycan biosynthesis	45	25	■	2.2e−17
10	KEGG	KEGG:04211	Longevity regulating pathway	86	27	■	4.0e−11

C8

id	source	term_id	term_name	term_size	intersection_size		p_value
1	KEGG	KEGG:04130	SNARE interactions in vesicular transport	31	28	■	1.8e−29
2	KEGG	KEGG:03022	Basal transcription factors	37	27	■	1.7e−23
3	KEGG	KEGG:03420	Nucleotide excision repair	44	25	■	1.2e−17
4	KEGG	KEGG:05016	Huntington disease	272	50	■	8.8e−11
5	KEGG	KEGG:05200	Pathways in cancer	517	72	■	1.1e−09
6	KEGG	KEGG:04612	Antigen processing and presentation	67	22	■	2.7e−09
7	KEGG	KEGG:04014	Ras signaling pathway	225	42	■	4.1e−09
8	KEGG	KEGG:05330	Allograft rejection	34	15	■	3.4e−08
9	KEGG	KEGG:03440	Homologous recombination	40	16	■	5.2e−08
10	KEGG	KEGG:04015	Rap1 signaling pathway	203	37	■	1.3e−07

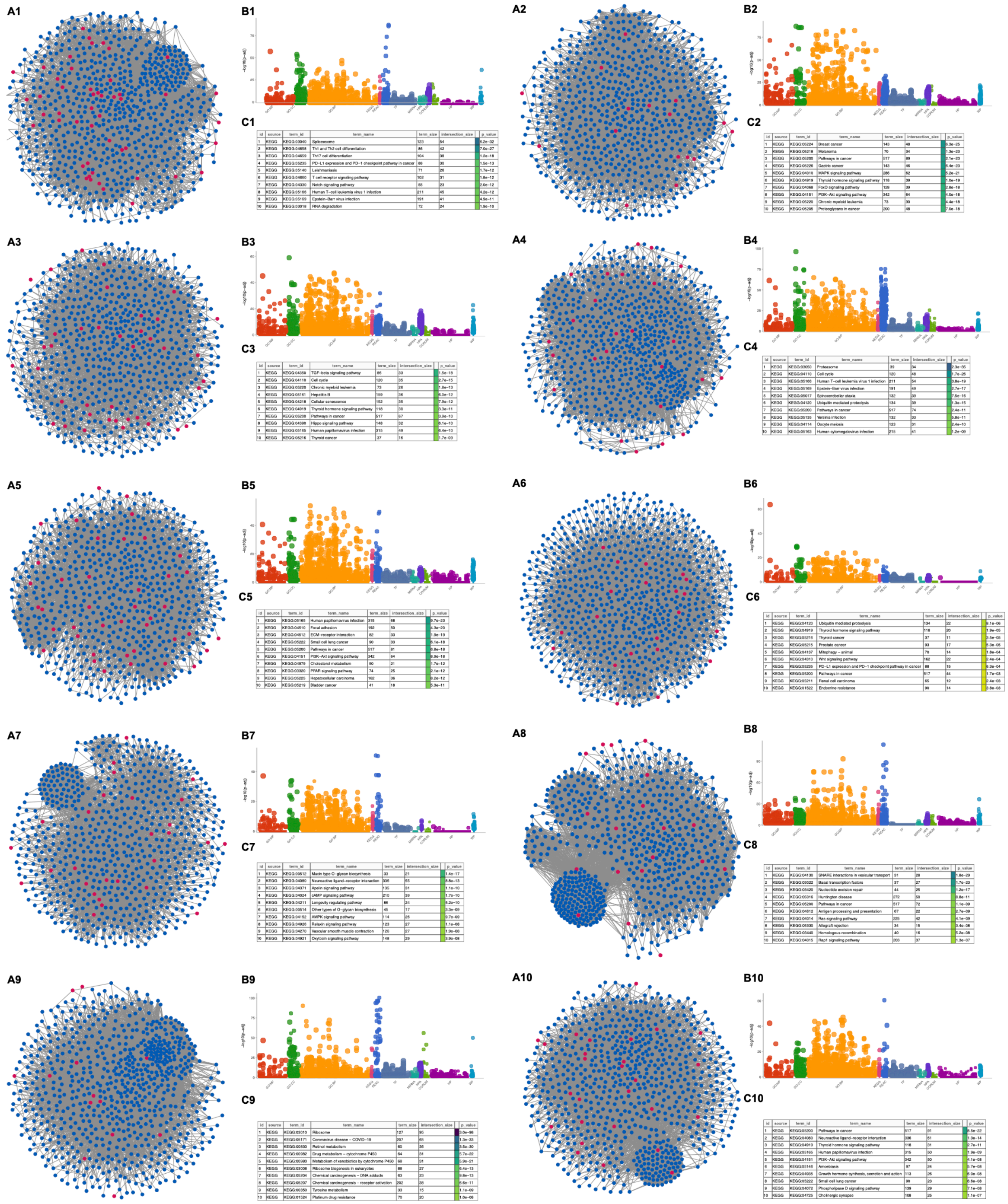
C9

id	source	term_id	term_name	term_size	intersection_size		p_value
1	KEGG	KEGG:03010	Ribosome	127	95	■	3.0e−98
2	KEGG	KEGG:05171	Coronavirus disease − COVID−19	207	65	■	1.3e−33
3	KEGG	KEGG:00830	Retinol metabolism	60	36	■	3.5e−30
4	KEGG	KEGG:00982	Drug metabolism − cytochrome P450	64	31	■	5.7e−22
5	KEGG	KEGG:00980	Metabolism of xenobiotics by cytochrome P450	68	31	■	5.9e−21
6	KEGG	KEGG:03008	Ribosome biogenesis in eukaryotes	88	27	■	6.4e−13
7	KEGG	KEGG:05204	Chemical carcinogenesis − DNA adducts	63	23	■	9.8e−13
8	KEGG	KEGG:05207	Chemical carcinogenesis − receptor activation	202	38	■	6.6e−11
9	KEGG	KEGG:00350	Tyrosine metabolism	33	15	■	1.1e−09
10	KEGG	KEGG:01524	Platinum drug resistance	70	20	■	1.0e−08

C10

id	source	term_id	term_name	term_size	intersection_size		p_value
1	KEGG	KEGG:04510	Focal adhesion	192	68	■	6.3e−34
2	KEGG	KEGG:04512	ECM−receptor interaction	82	41	■	5.6e−27
3	KEGG	KEGG:01522	Endocrine resistance	90	42	■	4.2e−26
4	KEGG	KEGG:04919	Thyroid hormone signaling pathway	118	46	■	1.6e−24
5	KEGG	KEGG:04151	PI3K−Akt signaling pathway	342	77	■	9.4e−24
6	KEGG	KEGG:05200	Pathways in cancer	517	96	■	2.5e−23
7	KEGG	KEGG:05165	Human papillomavirus infection	315	73	■	3.1e−23
8	KEGG	KEGG:05205	Proteoglycans in cancer	200	52	■	1.8e−18
9	KEGG	KEGG:05146	Amoebiasis	97	36	■	4.4e−18
10	KEGG	KEGG:04926	Relaxin signaling pathway	123	40	■	9.0e−18

Supplemental Fig S4: Subnetworks of genes in the disease module and enriched functional terms using STRING interactome. (A1-A10): Subnetwork of genes in disease modules of asthma, breast cancer, colorectal cancer, COPD, CVD, diabetes, gastric cancer, lung cancer, ovarian cancer and prostate cancer. Genes with *p*-values lower (higher) than 0.001 are colored with red (blue). **(B1-B10):** Enriched functional terms of genes in disease modules of asthma, breast cancer, COPD, CVD, diabetes and lung cancer using gprofier2 (Kolberg et al. 2020, 2). **(C1-C10):** Top-10 enriched KEGG pathways of genes in the disease modules of asthma, breast cancer, colorectal cancer, COPD, CVD, diabetes, gastric cancer, lung cancer, ovarian cancer and prostate cancer.



C1

id	source	term_id	term_name	term_size	intersection_size		p_value
1	KEGG	KEGG:03040	Spliceosome	123	54	■	6.2e−32
2	KEGG	KEGG:04658	Th1 and Th2 cell differentiation	86	42	■	7.0e−27
3	KEGG	KEGG:04659	Th17 cell differentiation	104	38	■	1.2e−18
4	KEGG	KEGG:05235	PD−L1 expression and PD−1 checkpoint pathway in cancer	88	30	■	1.5e−13
5	KEGG	KEGG:05140	Leishmaniasis	71	26	■	1.7e−12
6	KEGG	KEGG:04660	T cell receptor signaling pathway	102	31	■	1.8e−12
7	KEGG	KEGG:04330	Notch signaling pathway	55	23	■	2.0e−12
8	KEGG	KEGG:05166	Human T−cell leukemia virus 1 infection	211	45	■	4.2e−12
9	KEGG	KEGG:05169	Epstein−Barr virus infection	191	41	■	4.9e−11
10	KEGG	KEGG:03018	RNA degradation	72	24	■	1.9e−10

C2

id	source	term_id	term_name	term_size	intersection_size		p_value
1	KEGG	KEGG:05224	Breast cancer	143	48	■	6.3e−25
2	KEGG	KEGG:05218	Melanoma	70	34	■	1.3e−23
3	KEGG	KEGG:05200	Pathways in cancer	517	89	■	2.1e−23
4	KEGG	KEGG:05226	Gastric cancer	143	46	■	6.4e−23
5	KEGG	KEGG:04010	MAPK signaling pathway	286	62	■	5.2e−21
6	KEGG	KEGG:04919	Thyroid hormone signaling pathway	118	39	■	1.0e−19
7	KEGG	KEGG:04068	FoxO signaling pathway	128	39	■	2.8e−18
8	KEGG	KEGG:04151	PI3K−Akt signaling pathway	342	64	■	4.0e−18
9	KEGG	KEGG:05220	Chronic myeloid leukemia	73	30	■	4.4e−18
10	KEGG	KEGG:05205	Proteoglycans in cancer	200	48	■	7.0e−18

C3

id	source	term_id	term_name	term_size	intersection_size		p_value
1	KEGG	KEGG:04350	TGF−beta signaling pathway	86	33	■	1.5e−18
2	KEGG	KEGG:04110	Cell cycle	120	35	■	2.7e−15
3	KEGG	KEGG:05220	Chronic myeloid leukemia	73	26	■	1.8e−13
4	KEGG	KEGG:05161	Hepatitis B	159	36	■	6.0e−12
5	KEGG	KEGG:04218	Cellular senescence	152	35	■	7.9e−12
6	KEGG	KEGG:04919	Thyroid hormone signaling pathway	118	30	■	3.3e−11
7	KEGG	KEGG:05200	Pathways in cancer	517	67	■	3.9e−10
8	KEGG	KEGG:04390	Hippo signaling pathway	148	32	■	6.1e−10
9	KEGG	KEGG:05165	Human papillomavirus infection	315	49	■	6.4e−10
10	KEGG	KEGG:05216	Thyroid cancer	37	16	■	1.7e−09

C4

id	source	term_id	term_name	term_size	intersection_size		p_value
1	KEGG	KEGG:03050	Proteasome	39	34	■	2.3e−35
2	KEGG	KEGG:04110	Cell cycle	120	48	■	1.7e−26
3	KEGG	KEGG:05166	Human T−cell leukemia virus 1 infection	211	54	■	3.8e−19
4	KEGG	KEGG:05169	Epstein−Barr virus infection	191	49	■	2.7e−17
5	KEGG	KEGG:05017	Spinocerebellar ataxia	132	39	■	7.5e−16
6	KEGG	KEGG:04120	Ubiquitin mediated proteolysis	134	39	■	1.3e−15
7	KEGG	KEGG:05200	Pathways in cancer	517	74	■	2.4e−11
8	KEGG	KEGG:05135	Yersinia infection	132	33	■	5.8e−11
9	KEGG	KEGG:04114	Oocyte meiosis	123	31	■	2.4e−10
10	KEGG	KEGG:05163	Human cytomegalovirus infection	215	41	■	1.2e−09

C5

id	source	term_id	term_name	term_size	intersection_size		p_value
1	KEGG	KEGG:05165	Human papillomavirus infection	315	68	■	9.7e−23
2	KEGG	KEGG:04510	Focal adhesion	192	50	■	4.3e−20
3	KEGG	KEGG:04512	ECM−receptor interaction	82	33	■	1.9e−19
4	KEGG	KEGG:05222	Small cell lung cancer	90	33	■	6.1e−18
5	KEGG	KEGG:05200	Pathways in cancer	517	81	■	6.8e−18
6	KEGG	KEGG:04151	PI3K−Akt signaling pathway	342	64	■	8.9e−18
7	KEGG	KEGG:04979	Cholesterol metabolism	50	21	■	1.7e−12
8	KEGG	KEGG:03320	PPAR signaling pathway	74	25	■	2.1e−12
9	KEGG	KEGG:05225	Hepatocellular carcinoma	162	36	■	8.2e−12
10	KEGG	KEGG:05219	Bladder cancer	41	18	■	5.3e−11

C6

id	source	term_id	term_name	term_size	intersection_size		p_value
1	KEGG	KEGG:04120	Ubiquitin mediated proteolysis	134	22	■	8.1e−06
2	KEGG	KEGG:04919	Thyroid hormone signaling pathway	118	20	■	1.9e−05
3	KEGG	KEGG:05216	Thyroid cancer	37	11	■	3.5e−05
4	KEGG	KEGG:05215	Prostate cancer	93	17	■	5.3e−05
5	KEGG	KEGG:04137	Mitophagy − animal	70	14	■	1.8e−04
6	KEGG	KEGG:04310	Wnt signaling pathway	162	22	■	2.4e−04
7	KEGG	KEGG:05235	PD−L1 expression and PD−1 checkpoint pathway in cancer	88	15	■	6.3e−04
8	KEGG	KEGG:05200	Pathways in cancer	517	44	■	1.7e−03
9	KEGG	KEGG:05211	Renal cell carcinoma	65	12	■	2.4e−03
10	KEGG	KEGG:01522	Endocrine resistance	90	14	■	3.8e−03

C7

id	source	term_id	term_name	term_size	intersection_size		p_value
1	KEGG	KEGG:00512	Mucin type O−glycan biosynthesis	33	21	■	1.4e−17
2	KEGG	KEGG:04080	Neuroactive ligand−receptor interaction	336	55	■	8.8e−13
3	KEGG	KEGG:04371	Apelin signaling pathway	135	31	■	1.1e−10
4	KEGG	KEGG:04024	cAMP signaling pathway	210	39	■	1.7e−10
5	KEGG	KEGG:04211	Longevity regulating pathway	86	24	■	5.2e−10
6	KEGG	KEGG:00514	Other types of O−glycan biosynthesis	45	17	■	3.3e−09
7	KEGG	KEGG:04152	AMPK signaling pathway	114	26	■	9.7e−09
8	KEGG	KEGG:04926	Relaxin signaling pathway	123	27	■	1.1e−08
9	KEGG	KEGG:04270	Vascular smooth muscle contraction	126	27	■	1.9e−08
10	KEGG	KEGG:04921	Oxytocin signaling pathway	148	29	■	3.9e−08

C8

id	source	term_id	term_name	term_size	intersection_size		p_value
1	KEGG	KEGG:04130	SNARE interactions in vesicular transport	31	28	■	1.8e−29
2	KEGG	KEGG:03022	Basal transcription factors	37	27	■	1.7e−23
3	KEGG	KEGG:03420	Nucleotide excision repair	44	25	■	1.2e−17
4	KEGG	KEGG:05016	Huntington disease	272	50	■	8.8e−11
5	KEGG	KEGG:05200	Pathways in cancer	517	72	■	1.1e−09
6	KEGG	KEGG:04612	Antigen processing and presentation	67	22	■	2.7e−09
7	KEGG	KEGG:04014	Ras signaling pathway	225	42	■	4.1e−09
8	KEGG	KEGG:05330	Allograft rejection	34	15	■	3.4e−08
9	KEGG	KEGG:03440	Homologous recombination	40	16	■	5.2e−08
10	KEGG	KEGG:04015	Rap1 signaling pathway	203	37	■	1.3e−07

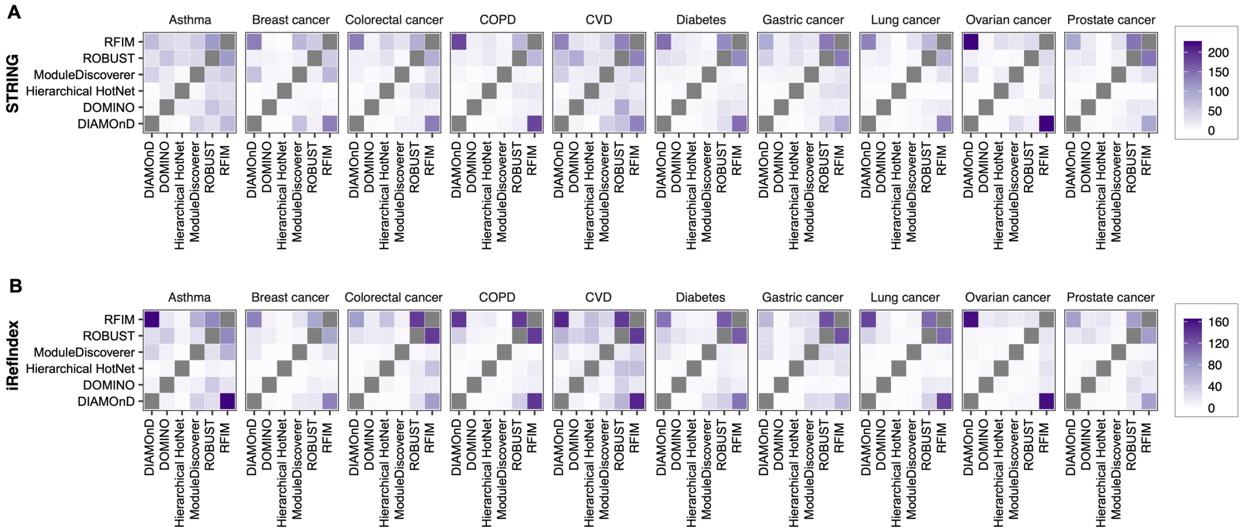
C9

id	source	term_id	term_name	term_size	intersection_size		p_value
1	KEGG	KEGG:03010	Ribosome	127	95	■	3.0e−98
2	KEGG	KEGG:05171	Coronavirus disease − COVID−19	207	65	■	1.3e−33
3	KEGG	KEGG:00830	Retinol metabolism	60	36	■	3.5e−30
4	KEGG	KEGG:00982	Drug metabolism − cytochrome P450	64	31	■	5.7e−22
5	KEGG	KEGG:00980	Metabolism of xenobiotics by cytochrome P450	68	31	■	5.9e−21
6	KEGG	KEGG:03008	Ribosome biogenesis in eukaryotes	88	27	■	6.4e−13
7	KEGG	KEGG:05204	Chemical carcinogenesis − DNA adducts	63	23	■	9.8e−13
8	KEGG	KEGG:05207	Chemical carcinogenesis − receptor activation	202	38	■	6.6e−11
9	KEGG	KEGG:00350	Tyrosine metabolism	33	15	■	1.1e−09
10	KEGG	KEGG:01524	Platinum drug resistance	70	20	■	1.0e−08

C10

id	source	term_id	term_name	term_size	intersection_size		p_value
1	KEGG	KEGG:05200	Pathways in cancer	517	91	■	8.5e−22
2	KEGG	KEGG:04080	Neuroactive ligand−receptor interaction	336	61	■	1.3e−14
3	KEGG	KEGG:04919	Thyroid hormone signaling pathway	118	31	■	2.7e−11
4	KEGG	KEGG:05165	Human papillomavirus infection	315	50	■	1.9e−09
5	KEGG	KEGG:04151	PI3K−Akt signaling pathway	342	50	■	4.1e−08
6	KEGG	KEGG:05146	Amoebiasis	97	24	■	5.7e−08
7	KEGG	KEGG:04935	Growth hormone synthesis, secretion and action	113	26	■	6.0e−08
8	KEGG	KEGG:05222	Small cell lung cancer	90	23	■	6.6e−08
9	KEGG	KEGG:04072	Phospholipase D signaling pathway	139	29	■	7.1e−08
10	KEGG	KEGG:04725	Cholinergic synapse	108	25	■	1.1e−07

Supplemental Fig S5: Subnetworks of genes in the disease module and enriched functional terms using iRefIndex interactome. (A1-A10): Subnetwork of genes in disease modules of asthma, breast cancer, colorectal cancer, COPD, CVD, diabetes, gastric cancer, lung cancer, ovarian cancer and prostate cancer. Genes with *p*-values lower (higher) than 0.001 are colored with red (blue). **(B1-B10):** Enriched functional terms of genes in disease modules of asthma, breast cancer, COPD, CVD, diabetes and lung cancer using gprofier2 (Kolberg et al. 2020, 2). **(C1-C10):** Top-10 enriched KEGG pathways of genes in the disease modules of asthma, breast cancer, colorectal cancer, COPD, CVD, diabetes, gastric cancer, lung cancer, ovarian cancer and prostate cancer.



Supplemental Fig S6: Overlap between disease modules identified by different methods. (A): STRING. **(B):** iRefIndex.