Supplementary Text: Reconciling Differential Gene Expression Data with Molecular Interaction Networks

Christopher L. Poirel[1], Ahsanur Rahman[1], Richard R. Rodrigues[1,2]
Arjun Krishnan[2], Jacqueline R. Addesa[1], T. M. Murali[1,3]

1 Supplementary Methods

1.1 Comparison to Previous Approaches

Here we provide a qualitative comparison of network reconciliation methods to techniques that have already been published in the literature. As noted in the main manuscript, BioNet [2] and DEGAS [16] aim to compute small connected subgraphs that are correlated with a disease. Such approaches were motivated by the observation that only part of a pathway is often changed by disease [16] and that interpreting small subnetworks is easier than interpreting larger ones [2]. We differ from these techniques in the following ways:

(i) DEGAS and BioNet formulate problems that are computationally intractable (NP-complete or NP-hard). They solve these problems either by using approximation algorithms or methods to solve integer linear programs. In contrast, we propose comparatively simpler methods that need to solve a (sparse) linear system. Therefore, our algorithms are considerably simpler.

(ii) Rather than directly compute subnetworks, network reconciliation algorithms modify the score of each gene in the network based on their interactions. The modified scores can be smoothly incorporated into a number of systems biology analyses, including active subnetwork discovery.

(iii) Moreover, by computing a new ranking of all genes we retain the possibility of discovering multiple pathways related to the disease of interest.

We do acknowledge that DEGAS incorporates per-gene variations among samples in a more subtle manner than our approaches can.

Other authors have proposed using PageRank to integrate gene expression data with protein interaction networks [9, 10, 18]. PINTA [10] uses the algorithms we have described but evaluates them in a more restricted biological context. They check if knocked-out genes are ranked highly by the algorithms in gene expression data collected following the knock-out. Two applications have used PageRank to study doxorubicin dosage response [9] and to select diagnostic genes in cancers [18, 6]. Our approach extends these previous works in the following ways:

(i) We list three well-motivated criteria to evaluate the algorithms. We demonstrate the dependence of these evaluations on the input parameter q, thereby suggesting suitable values of the parameter.

(ii) We use the reconciliation algorithms to study a diverse set of human diseases. We show that the algorithms can retain differences between diseases while also taking the network structure into account.

[1]Department of Computer Science, Virginia Tech, Blacksburg, VA, USA
[2]Genetics, Bioinformatics, and Computational Biology Ph.D. Program, Virginia Tech, Blacksburg, VA, USA
[3]ICTAS Centre for Systems Biology of Engineered Tissues, Virginia Tech, Blacksburg, VA, USA

(iii) Using a set of seven brain diseases, we demonstrated that our algorithms could reveal genes involved in a relevant biological function, even when those genes were not differentially expressed.

1.2 Computing Weights of Protein Interactions

We used the human MiMI network [15] as the underlying protein interaction network in our analysis. MiMI merges interaction data from over 20 interaction databases. We discarded self-edges and repeated edges from the interaction network, resulting in a network comprised of 11,074 proteins and 77,952 interactions. Although MiMI is a very comprehensive resource, a number of interactions included in it were detected by error-prone high throughput experiments.

Typically, there is little information regarding the confidence of interactions discovered experimentally via high- or low-throughput technologies. As a result, many network-based methods operate on unweighted molecular interaction networks [12, 17, 18] even though these methods may be applicable to weighted networks. However, we anticipate that estimating the quality of each interaction will improve the results of network-based algorithms. A number of approaches have emerged in the literature to estimate the confidence of molecular interactions. For example, predicted interaction networks assign a prediction score to each interaction that is often directly applied as an edge weight [8, 10]. Alternatively, binding affinities between transcription factors and their target binding sites may offer an appropriate estimate of interaction confidence [5] when working with transcriptional regulatory networks. Similarly, weighting edges by the number of mutations observed on a pair of interacting genes [17] has proven appropriate for detecting mutated pathways in cancer. Nevertheless, there is no standard approach for estimating the confidence of an interaction, and implemented approaches are often problem-specific.

We desired an edge weighting method that used purely topological measures of reliability, since we incorporated gene expression data into these methods and used functional annotations to interpret the results. We estimated the reliability of each interaction by its FS-weight [4]. Let $G(V, E)$ denote the interaction network, where V is the set of nodes and E is the set of edges in the network. For each edge $(u, v) \in E$, define the weight of (u, v) as follows:

$$w_{uv} = \left(\frac{2|\hat{N}_u \cap \hat{N}_v|}{|\hat{N}_u - \hat{N}_v| + 2|\hat{N}_u \cap \hat{N}_v| + \lambda_{u,v}} \right) \left(\frac{2|\hat{N}_u \cap \hat{N}_v|}{|\hat{N}_v - \hat{N}_u| + 2|\hat{N}_u \cap \hat{N}_v| + \lambda_{v,u}} \right),$$

where $\hat{N}_u = \{u\} \cup N_u$ is the set containing u and its neighbors, and

$$\lambda_{u,v} = \max \left(0, \frac{2|E|}{|V|} - |\hat{N}_u - \hat{N}_v| - |\hat{N}_u \cap \hat{N}_v| \right).$$

Notice that w_{uv} is large when u and v have many neighbors in common and small when they have diverse neighborhoods. Thus, we assign higher confidence to a pair of nodes that share many common interactors in the interaction network.

1.3 Selection of Functional Enrichment Algorithm

A wide variety of functional enrichment methods are available in the literature [3, 7, 11, 13]. These approaches typically perform a term-by-term analysis, reporting the significance of the relationship between each function and a collection of genes being studied. The disadvantage of these approaches is that they typically return long lists of significantly enriched functions, from which the user must determine which are the most relevant. After applying FuncAssociate [3], GSEA [13], and PAGE [7]

on our datasets, we found it difficult to distinguish top-ranking functions from one another because they annotated similar collections of genes. However, Model-based Gene Set Analysis (MGSA) [1] simultaneously evaluates all gene sets using a Bayesian approach that integrates overlap between gene sets into the enrichment analysis. MGSA attempts to compute a non-overlapping set of pathways that annotate the study set. MGSA computes a posterior probability for each pathway that reflects how well the pathway overlaps with the study set while not overlapping with other pathways with higher posterior probability. We performed all tests for functional enrichment using MGSA. MGSA allows ranges to be set for two primary parameters α and β. Parameter α controls the fraction of unknown false positive genes, while β controls the fraction of unknown false negatives. We set an upper limit of the α and β parameters to 0.3 and 0.5, respectively. Thus, less than 30% of the genes annotated by enriched functions are not in the study set (i.e., top 250 genes from the given ranking), and less than half of the genes from the study set are not annotated by any enriched function. All other parameters were left to their default settings.

1.4 Expression Data for Human Diseases

We used gene expression data from [14], who collected publicly-available expression data for the following 54 different human diseases and for the corresponding normal tissues:

Actinic Keratosis	Idiopathic Thrombocytopenic Purpura
Acute Myeloid Leukemia	Ischemic Cardiomyopathy
Adenocarcinoma of Esophagus	Lung Transplant Rejection
Adenocarcinoma of Lung	Malaria
Alzheimers Disease	Malignant Melanoma
Anaplasmosis	Malignant Neoplasm of Hypopharynx
Bipolar Disorder	Malignant Neoplasm of Prostate
Chronic Obstructive Lung Disease	Malignant Pleural Mesothelioma
Chronic Polyarticular Juvenile Rheumatoid Arthritis	Malignant Tumor of Colon
Chronic Progressive Ophthalmoplegia	Mixed Hyperlipidemia
Clear Cell Carcinoma of Kidney	Myelodysplastic Syndrome
Complex Dental Cavity	Nephroblastoma
Congestive Cardiomyopathy	Obesity
Crohns Disease	Papillary Thyroid Carcinoma
Cystic Fibrosis	Polycystic Ovary Syndrome
Dermatomyositis	Progeria Syndrome
Diabetic Nephropathy	Rett Syndrome
Dilated Cardiomyopathy Secondary to Viral Myocarditis	Rheumatoid Arthritis
Duchenne Muscular Dystrophy	Rotavirus Infection of Children
Endometriosis	Sarcoidosis
Essential Thrombocythemia	Schizophrenia
Glaucoma	Scott Syndrome
Glioblastoma	Senescence
Hereditary Gingival Fibromatosis	Squamous Cell Carcinoma of Lung
Human Immunodefiency Virus Encephalitis	Squamous Cell Carcinoma of Mouth
Huntingtons Disease	Urothelial Carcinoma
Idiopathic Pulmonary Fibrosis	Uterine Leiomyoma

2 Supplementary Results

2.1 Network coherence

Figure S1 reports the number of connected components in the subnetwork induced by the top k genes returned by each algorithm for different values of q. Each point is the average number of connected components across all 54 diseases. The *initial* lines represent the connected components induced by the rankings given solely by the expression data (i.e, $q = 1$); these lines serve as a common reference across the four subplots. For all values of q, the number of connected components increases at every rank cutoff, suggesting that the Vanilla algorithm decreases network coherence among top-ranking genes. However, PageRank, GeneMANIA, and Heat Kernel each decrease the number of connected components drastically as q increases.

Next, we assessed the significance of these connected component counts. We randomly shuffled the gene identifiers in the gene expression data and applied PageRank to the randomized expression data. As in the previous analysis, we computed the number of connected components induced by the top k genes returned by PageRank when we set the parameter q to 1.0, 0.5, and 0.1. We repeated this process 100 times. Figure S2 plots the number of connected components induced by the top k genes for 100 random rankings and the ranking given by applying PageRank to the true expression data. Thus, the blue, cyan, and yellow curves in Figure S2 correspond to the blue, cyan, and yellow curves in the PageRank subfigure (upper right) of Figure S1, and the dashed red lines correspond to the results from 100 random rankings. Figure S2 indicates that the number of connected components induced by the top 1000 genes is lower than the number of connected components given by applying PageRank to any of the 100 random rankings.

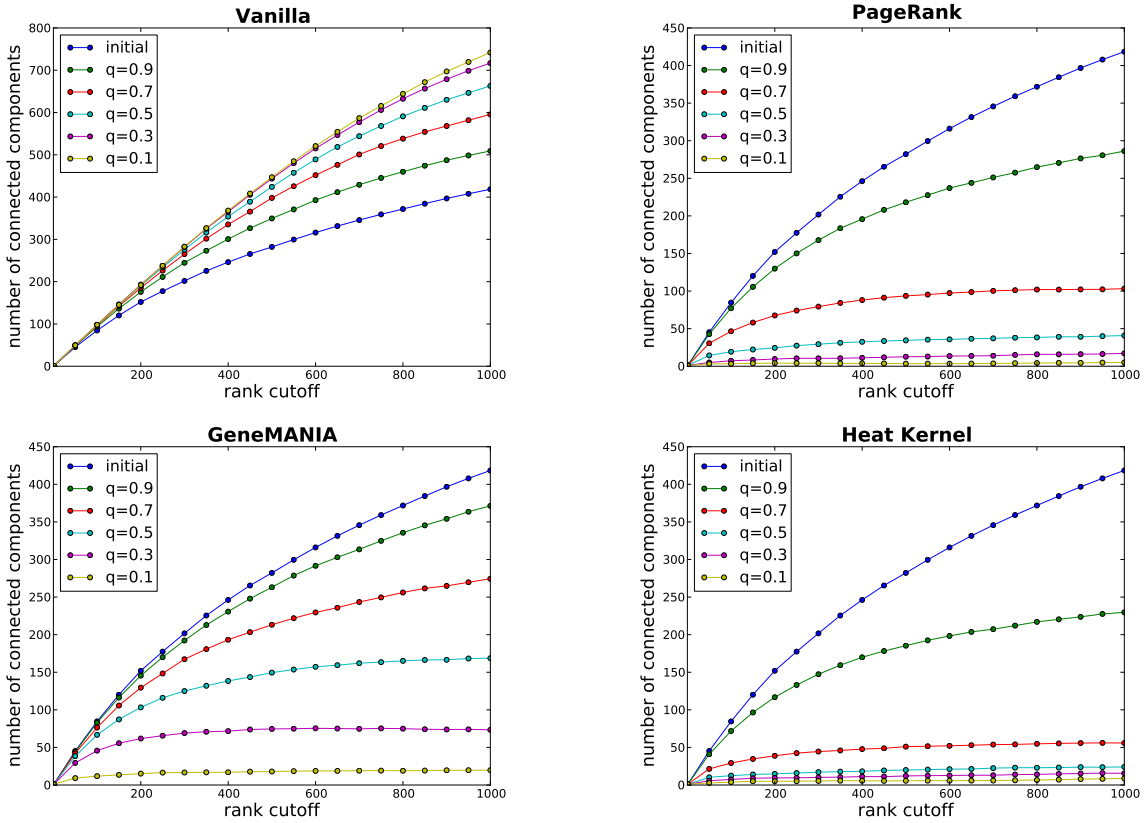


Figure S1: Connected components induced by the top-ranking nodes from each algorithm.

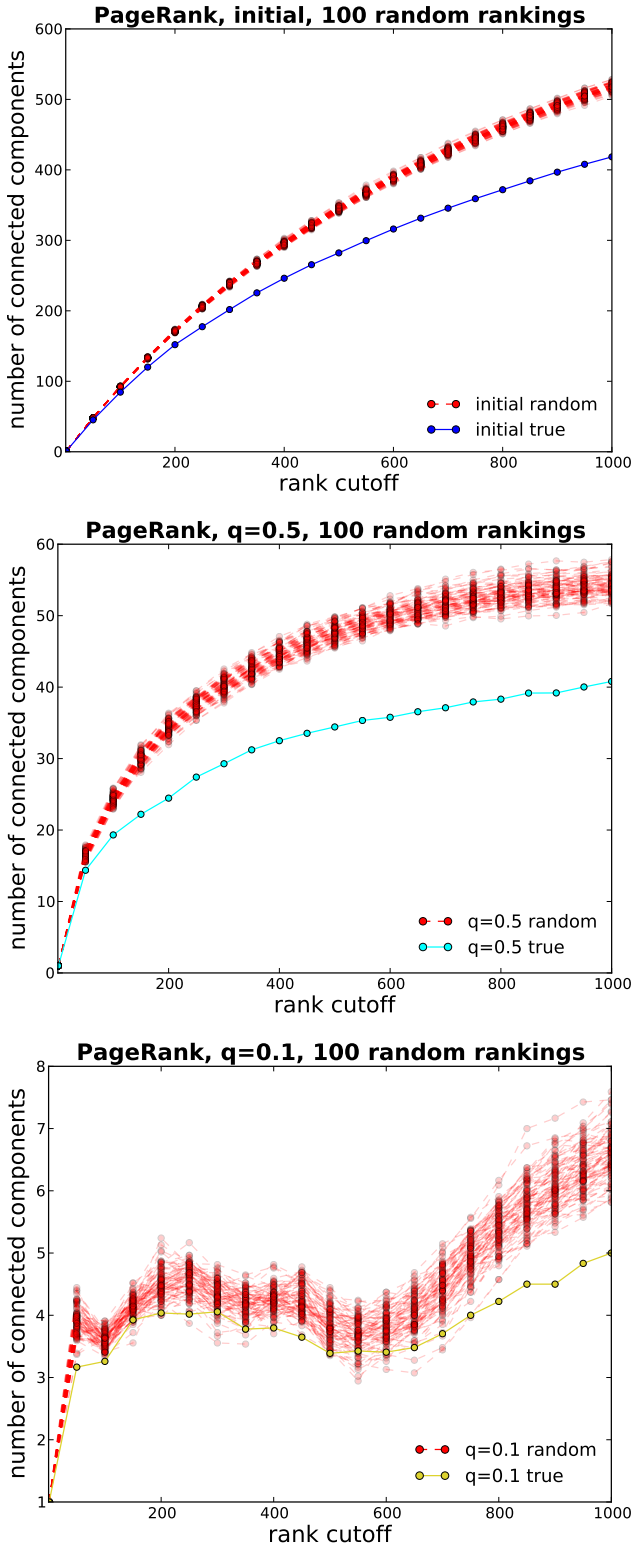


Figure S2: Connected components induced by the top-ranking nodes from PageRank applied to 100 randomized gene expression datasets and to the true gene expression data.

2.2 Similarity of Gene Rankings Between Diseases

For each algorithm, we computed the Jaccard index of the top k, $0 < k \leq 250$, genes after reconciliation between every pair of diseases. Figure S3 illustrates the Jaccard indices for each algorithm using a range of values for the input parameter q. Each point indicates the average Jaccard index across all $\binom{54}{2}$ pairs of diseases. This figure is the same as Figure 2 from the main manuscript but expanded to include results for the Vanilla algorithm.

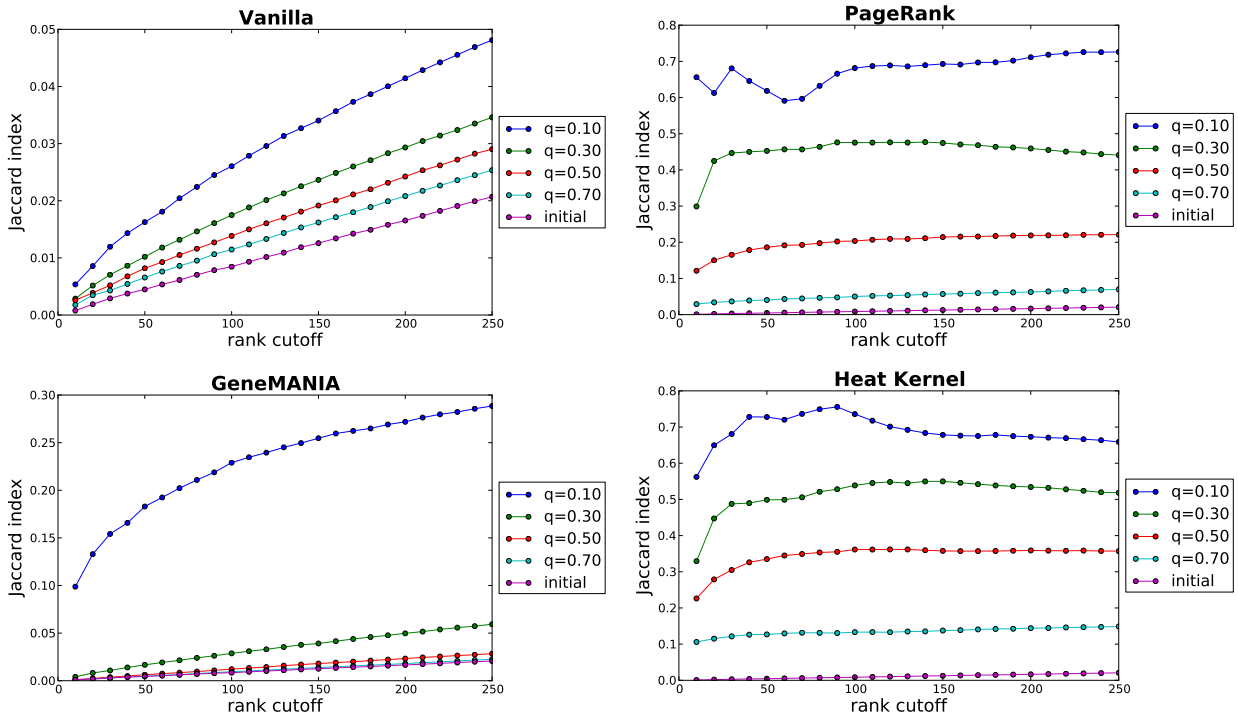


Figure S3: The Jaccard index of the top k genes reported by each algorithm for a pair of different diseases. Each point indicates the average Jaccard index of all $\binom{54}{2}$ pairs of diseases using a particular value of q as input to the algorithm.

2.3 Similarity of Gene Rankings Between Algorithms

For PageRank, Heat Kernel, and GeneMANIA, Figure S4 plots the average Jaccard index across all 54 diseases of the top k genes, $0 < k \leq 1000$, for each pair of algorithms. This plot demonstrates that PageRank and Heat Kernel provide similar rankings for each disease.

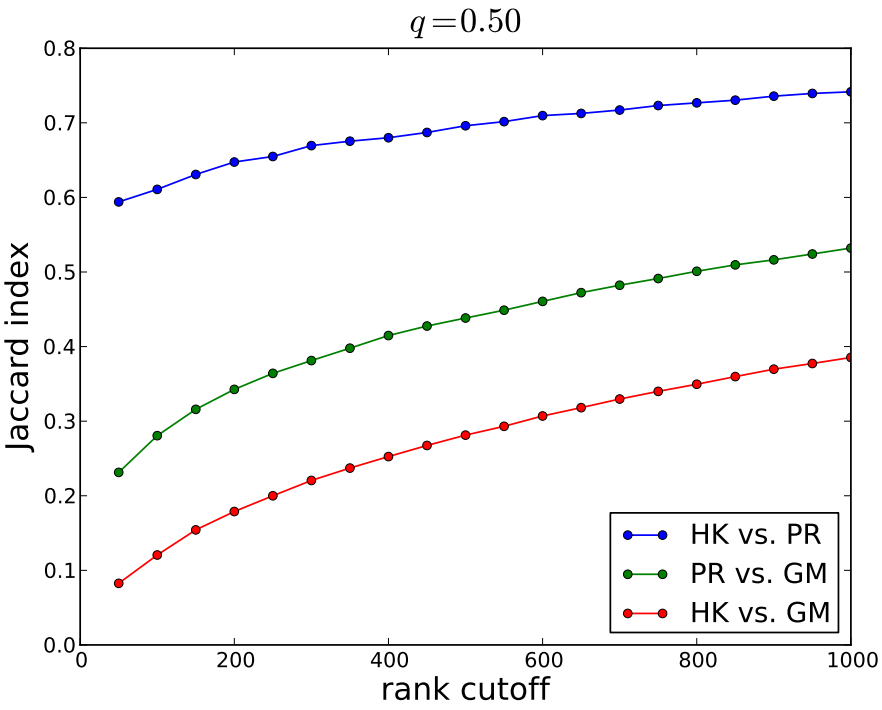


Figure S4: The average Jaccard index of the top-ranking genes between each pair of algorithms returned by MGSA across all 54 diseases.

2.4 Functional Enrichment Between Algorithms

We investigated the similarity between the functional results of different algorithms. In Figure S5, we show the average Jaccard index between the top k functions returned by MGSA for a pair of reconciliation algorithms applied to a single disease. We plot the average across the seven brain disorders. Figure S5 is the same as Figure 3(b) from the main manuscript but expanded to include comparisons with the Vanilla algorithm.

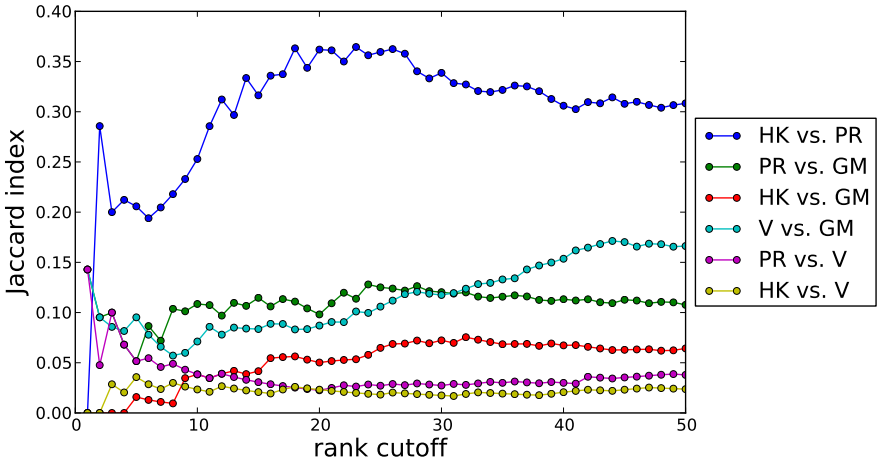
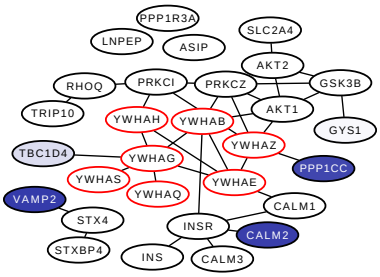


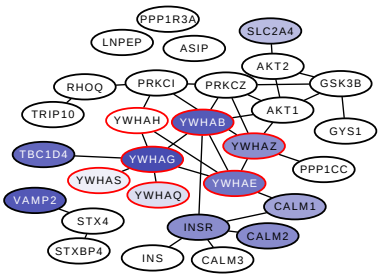
Figure S5: The Jaccard index of the top-ranking functions between all pairs of algorithms returned by MGSA on the seven brain disorders.

2.5 Insulin-Mediated Glucose Transport Subnetworks

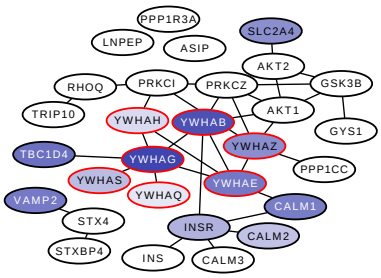
Figure S6 illustrates the subnetworks induced by proteins in the insulin-mediated glucose transport pathway before and after applying each of the four reconciliation algorithms. The networks in Figure S6 are the same as those in Figure 4 from the main manuscript, but we additionally report subnetworks for Heat Kernel, GeneMANIA, and Vanilla.



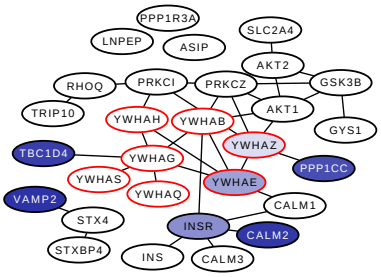
(a) Initial insulin-mediated glucose transport network.



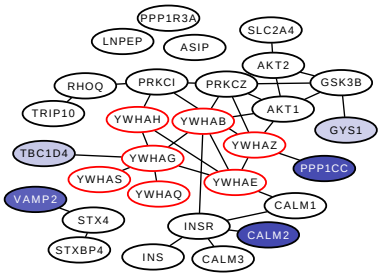
(b) After PageRank.



(c) After Heat Kernel.



(d) After GeneMANIA.



(e) After Vanilla.

Figure S6: A comparison of the subnetwork induced by genes involved in the NCI pathway *insulin-mediated glucose transport* with nodes weighted by (a) differential expression p-values from patients diagnosed with Huntington's disease, and after applying (b) PageRank, (c) Heat Kernel, (d) GeneMANIA, and (e) Vanilla. Blue nodes indicate genes ranked in the top 250, and darker nodes indicate higher ranking. Nodes with a red outline indicate genes in the 14-3-3 family of proteins.

References

[1] S. Bauer, J. Gagneur, and P. N. Robinson. GOing Bayesian: model-based gene set analysis of genome-scale data. *Nucleic acids research*, 38(11):3523–3532, June 2010.

[2] D. Beisser, G. W. Klau, T. Dandekar, T. Muller, and M. T. Dittrich. BioNet: an R-Package for the functional analysis of biological networks. *Bioinformatics*, 26(8):1129–30, 2010.

[3] G. F. Berriz, O. D. King, B. Bryant, C. Sander, and F. P. Roth. Characterizing gene sets with FuncAssociate. *Bioinformatics*, 19(18):2502–4, 2003.

[4] H. N. Chua, W.-K. Sung, and L. Wong. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, 22(13):1623–1630, July 2006.

[5] J. P. Gonçalves, A. P. Francisco, N. P. Mira, M. C. Teixeira, I. Sá-Correia, A. L. Oliveira, and S. C. Madeira. TFRank: network-based prioritization of regulatory associations underlying transcriptional responses. *Bioinformatics*, 27(22):3149–3157, Nov. 2011.

[6] M. Johannes, J. C. Brase, H. Fröhlich, S. Gade, M. Gehrmann, M. Fälth, H. Sültmann, and T. Beißbarth. Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients. *Bioinformatics*, 26(17):2136–2144, Sept. 2010.

[7] S. Y. Kim and D. Volsky. PAGE: Parametric Analysis of Gene Set Enrichment. *BMC Bioinformatics*, 6(1):144+, 2005.

[8] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson. Walking the interactome for prioritization of candidate disease genes. *American journal of human genetics*, 82(4):949–958, Apr. 2008.

[9] K. Komurov, M. A. White, and P. T. Ram. Use of data-biased random walks on graphs for the retrieval of context-specific networks from genomic data. *PLoS Comput Biol*, 6(8), 2010.

[10] D. Nitsch, J. Goncalves, F. Ojeda, B. de Moor, and Y. Moreau. Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC Bioinformatics*, 11(1):460+, Sept. 2010.

[11] C. L. Poirel, C. C. Owens III, and T. M. Murali. Network-based functional enrichment. *BMC Bioinformatics*, 12(Suppl 13):S14, 2011.

[12] Y. Q. Qiu, S. Zhang, X. S. Zhang, and L. Chen. Detecting disease associated modules and prioritizing active genes based on high throughput data. *BMC Bioinformatics*, 11(1):26+, 2010.

[13] A. Subramanian, P. Tamayo, V. Mootha, S. Mukherjee, B. Ebert, M. Gillette, A. Paulovich, S. Pomeroy, T. Golub, E. Lander, and J. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 2005.

[14] S. Suthram, J. T. Dudley, A. P. Chiang, R. Chen, T. J. Hastie, and A. J. Butte. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS Comput Biol*, 6(2):e1000662+, February 2010.

[15] V. G. Tarcea, T. Weymouth, A. Ade, A. Bookvich, J. Gao, V. Mahavisno, Z. Wright, A. Chapman, M. Jayapandian, A. Ozgur, Y. Tian, J. Cavalcoli, B. Mirel, J. Patel, D. Radev, B. Athey, D. States, and H. V. Jagadish. Michigan molecular interactions r2: from interacting proteins to pathways. *Nucleic Acids Res*, 37(Database issue):D642–6, 2009.

[16] I. Ulitsky, A. Krishnamurthy, R. M. Karp, and R. Shamir. DEGAS: De Novo Discovery of Dysregulated Pathways in Human Diseases. *PLoS ONE*, 5(10):e13367, 2010.

[17] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan. Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol*, 6(1):e1000641+, Jan. 2010.

[18] C. Winter, G. Kristiansen, S. Kersting, J. Roy, D. Aust, T. Knösel, P. Rümmele, B. Jahnke, V. Hentrich, F. Rückert, M. Niedergethmann, W. Weichert, M. Bahra, H. J. Schlitt, U. Settmacher, H. Friess, M. Büchler, H.-D. Saeger, M. Schroeder, C. Pilarsky, and R. Grützmann. Google goes cancer: Improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Comput Biol*, 8(5):e1002511+, May 2012.