Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes

Daniel S. Himmelstein[1], Sergio E. Baranzini[1,2,3]*

1 Biological & Medical Informatics, University of California, San Francisco, San Francisco, California, United States of America, 2 Department of Neurology, University of California, San Francisco, San Francisco, California, United States of America, 3 Institute for Human Genetics, University of California, San Francisco, San Francisco, California, United States of America

* sebaran@cgl.ucsf.edu

Abstract

The first decade of Genome Wide Association Studies (GWAS) has uncovered a wealth of disease-associated variants. Two important derivations will be the translation of this information into a multiscale understanding of pathogenic variants and leveraging existing data to increase the power of existing and future studies through prioritization. We explore edge prediction on heterogeneous networks—graphs with multiple node and edge types—for accomplishing both tasks. First we constructed a network with 18 node types—genes, diseases, tissues, pathophysiologies, and 14 MSigDB (molecular signatures database) collections—and 19 edge types from high-throughput publicly-available resources. From this network composed of 40,343 nodes and 1,608,168 edges, we extracted features that describe the topology between specific genes and diseases. Next, we trained a model from GWAS associations and predicted the probability of association between each protein-coding gene and each of 29 well-studied complex diseases. The model, which achieved 132-fold enrichment in precision at 10% recall, outperformed any individual domain, highlighting the benefit of integrative approaches. We identified pleiotropy, transcriptional signatures of perturbations, pathways, and protein interactions as influential mechanisms explaining pathogenesis. Our method successfully predicted the results (with AUROC = 0.79) from a withheld multiple sclerosis (MS) GWAS despite starting with only 13 previously associated genes. Finally, we combined our network predictions with statistical evidence of association to propose four novel MS genes, three of which (*JAK2*, *REL*, *RUNX3*) validated on the masked GWAS. Furthermore, our predictions provide biological support highlighting *REL* as the causal gene within its gene-rich locus. Users can browse all predictions online (http://het.io). Heterogeneous network edge prediction effectively prioritized genetic associations and provides a powerful new approach for data integration across multiple domains.

Author Summary

For complex human diseases, identifying the genes harboring susceptibility variants has taken on medical importance. Disease-associated genes provide clues for elucidating disease etiology, predicting disease risk, and highlighting therapeutic targets. Here, we develop a method to predict whether a given gene and disease are associated. To capture the multitude of biological entities underlying pathogenesis, we constructed a heterogeneous network, containing multiple node and edge types. We built on a technique developed for social network analysis, which embraces disparate sources of data to make predictions from heterogeneous networks. Using the compendium of associations from genome-wide studies, we learned the influential mechanisms underlying pathogenesis. Our findings provide a novel perspective about the existence of pervasive pleiotropy across complex diseases. Furthermore, we suggest transcriptional signatures of perturbations are an underutilized resource amongst prioritization approaches. For multiple sclerosis, we demonstrated our ability to prioritize future studies and discover novel susceptibility genes. Researchers can use these predictions to increase the statistical power of their studies, to suggest the causal genes from a set of candidates, or to generate evidence-based experimental hypothesis.

Introduction

In the last decade, genome-wide association studies (GWAS) have been established as the main strategy to map genetic susceptibility in dozens of complex diseases and phenotypes. Despite the success of this approach in mapping variation in thousands of loci to hundreds of complex phenotypes [1–5], researchers are now confronted with the challenge of maximizing the scientific contribution of existing GWAS datasets, whose undertakings represented a substantial investment of human and monetary resources from the community at large [6].

A central assumption in GWAS is that every region in the genome (and hence every gene) is a-priori equally likely to be associated with the phenotype in question. As a result, small effect sizes and multiple comparisons limit the pace of discovery. However, rational prioritization approaches may afford an increase in study power while avoiding the constraints and expense related to expanded sampling. One such way forward is the current trend of analyzing the combined contribution of susceptibility variants in the context of biological pathways, rather than single SNPs [7]. For example, Yaspan et al described an approach that aggregates variants of interest from a GWAS into biological pathways using genomic randomization to control for multiple testing and minimize type I error [8]. The popular software PLINK also includes an option to evaluate groups of associations at the gene level, thus enabling pathway analysis by computing enriched gene sets [9]. A less explored but potentially revealing strategy is the integration of diverse sources of data to build more accurate and comprehensive models of disease susceptibility.

Several strategies have been attempted to identify the mechanisms underlying pathogenesis and use these insights to prioritize genes for genetic association analyses. Gene-set enrichment analyses identify prevalent biological functions amongst genes contained in disease-associated loci [10,11]. Gene network approaches search for neighborhoods of genes where disease-associated loci aggregate [12,13]. Jia et al. reported dmGWAS, a strategy to integrate association signals from GWAS into the human protein interaction network [14]. A similar approach was developed by our group and tested in two large studies comprising more than 15,000 cases [15]. Literature mining techniques aim to chronicle the relatedness of genes to identify a subset

of highly-related associated genes. For example, Raychaudhuri et al. reported the Gene Relationships Among Implicated Loci (GRAIL) algorithm, an approach to assess relationships among genomic disease regions by text mining of PubMed abstracts [16].

Prioritization strategies generally rely on user-provided loci as the sole input and do not incorporate broader disease-specific knowledge. Typically, the proportion of genome-wide significant discoveries in a given GWAS is low, thus leaving little high-confidence signal for seed-based approaches to build from. To overcome this limitation, here we aimed at characterizing the ability of various information domains to identify pathogenic variants across the entire compendium of complex disease associations. Using this multiscale approach, we developed a framework to prioritize both existing and future GWAS analyses and highlight candidate genes for further analysis.

To approach this problem, we resorted to a method that integrated diverse information domains naturally. Heterogeneous (or multipartite) networks are a class of networks which contain multiple types of entities (nodes) and relationships (edges or links), and provide a data structure capable of expressing diversity in an intuitive and scalable fashion. Most existing techniques available for network analysis have been developed for homogeneous networks [17–19] and are not directly extensible to heterogenous networks. Accordingly, the early network analyses in disease biology concentrated on homogeneous networks. However in the last half-decade, the complexity of biological systems has spurred interest in heterogeneous approaches.

While still a developing field, network-based biological data integration has been pursued using a variety of techniques. Approaches such as GeneMANIA, weight and then project individual data sources onto a single dimension, enabling homogeneous network algorithms to be used to characterize the resulting graphs [20–22]. Other techniques operate on multi-relational (single node type, multiple edge type) networks, for example by taking into account relationships among local clusters and considering the full topology of weighted gene association networks [13,20,23]. Bipartite networks contain two node types and therefore work well for predicting relationships between entities of two different types (such as disease-gene associations or drug-disease indications) following a 'guilt-by-association' paradigm [24–26]. Other approaches incorporate greater-dimension heterogeneous networks as input but conflate types and, while improving predictions compared to simpler approaches, cannot effectively identify influential network components [27,28]. Heterogeneous networks of arbitrary complexity have also been applied for edge prediction without a formalized feature extraction methodology, which requires manual descriptor determination for each new network design [29]. Recently, new types of edge prediction methods were reported that naturally accommodate any size heterogeneous network. These include data fusion by matrix-factorization [30–33] and metapath-based techniques [34,35]. This type of intermediate data fusion can treat all data sources directly (i.e. without transforming data into "disease space") and has been successfully used to infer disease similarities [31] and predict gene function in slime mold and baker's yeast [32]. A metapath-based approach was recently developed by researchers studying social sciences to predict future coauthorship [34] and provides an intuitive framework and interpretable models and results. An advantage of metapath-based approaches is that they preserve the network structure and provide the flexibility to explore a diverse set of descriptors. In this work, we extended this methodology to predict the probability that an association between a gene and disease exists.

Results

Constructing a heterogeneous network to integrate diverse information domains

Using publicly-available databases and standardized vocabularies, we constructed a heterogeneous network with 40,343 nodes and 1,608,168 edges (Fig 1, S2 Data). Databases were selected based on quality, reusability, throughput, and their aggregate ability to provide a diversified, multiscale portrayal of biology (see Resource selection in Methods). The network was designed to encode entities and relationships relevant to pathogenesis. The network contained 18 node types (metanodes) and 19 edge types (metaedges), displayed in Fig 2A. Entities represented by metanodes consisted of diseases, genes, tissues, pathophysiologies, and gene sets for 14 MSigDB collections [36,37] including pathways [38,39], perturbation signatures, motifs [40,41], and Gene Ontology (GO) domains [42] (Table 1). Relationships represented by metaedges consisted of gene-disease association, disease pathophysiology, disease localization, tissue-specific gene expression, protein interaction, and gene-set membership for each MSigDB collection (Table 2).

Gene-disease associations were extracted from the GWAS Catalog [5] by overlapping associations into disease-specific loci. Loci were classified as low or high-confidence based on p-value and sample size of the corresponding GWAS (see Associations in Methods and S1 Fig). When possible, for each loci, the most-commonly reported gene across studies was designated as primary and subsequently considered responsible for the association. Additional genes
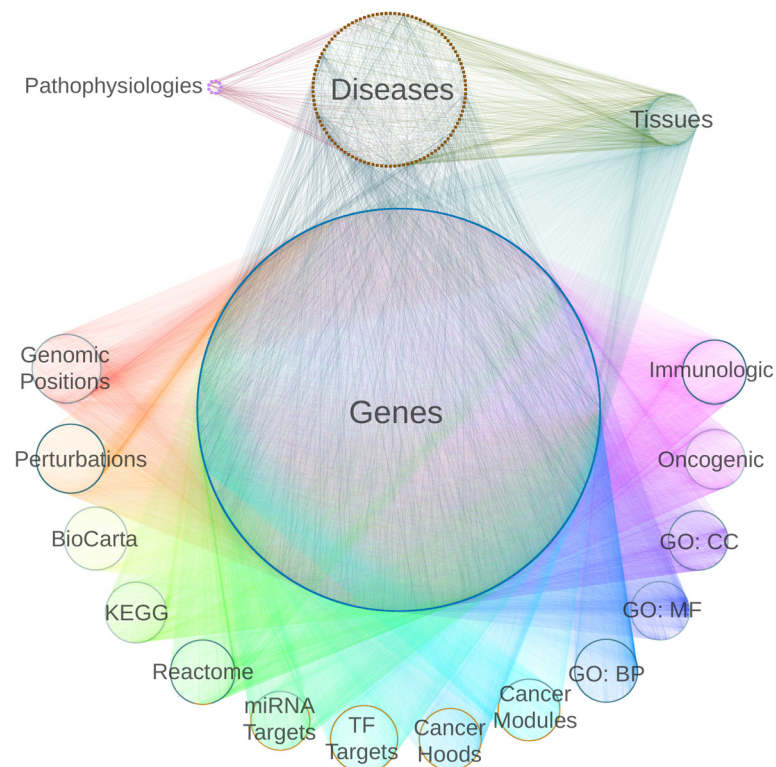
Fig 1. Heterogeneous network integrates diverse information domains. We constructed a heterogeneous network with 18 metanodes (denoted with labels) and 19 metaedges (denoted by color). For each metanode, nodes are laid out circularly. Incorporating type information adds structure to a network which would otherwise appear as an undecipherable agglomeration of 40,343 nodes and 1,608,168 edges.

doi:10.1371/journal.pcbi.1004259.g001
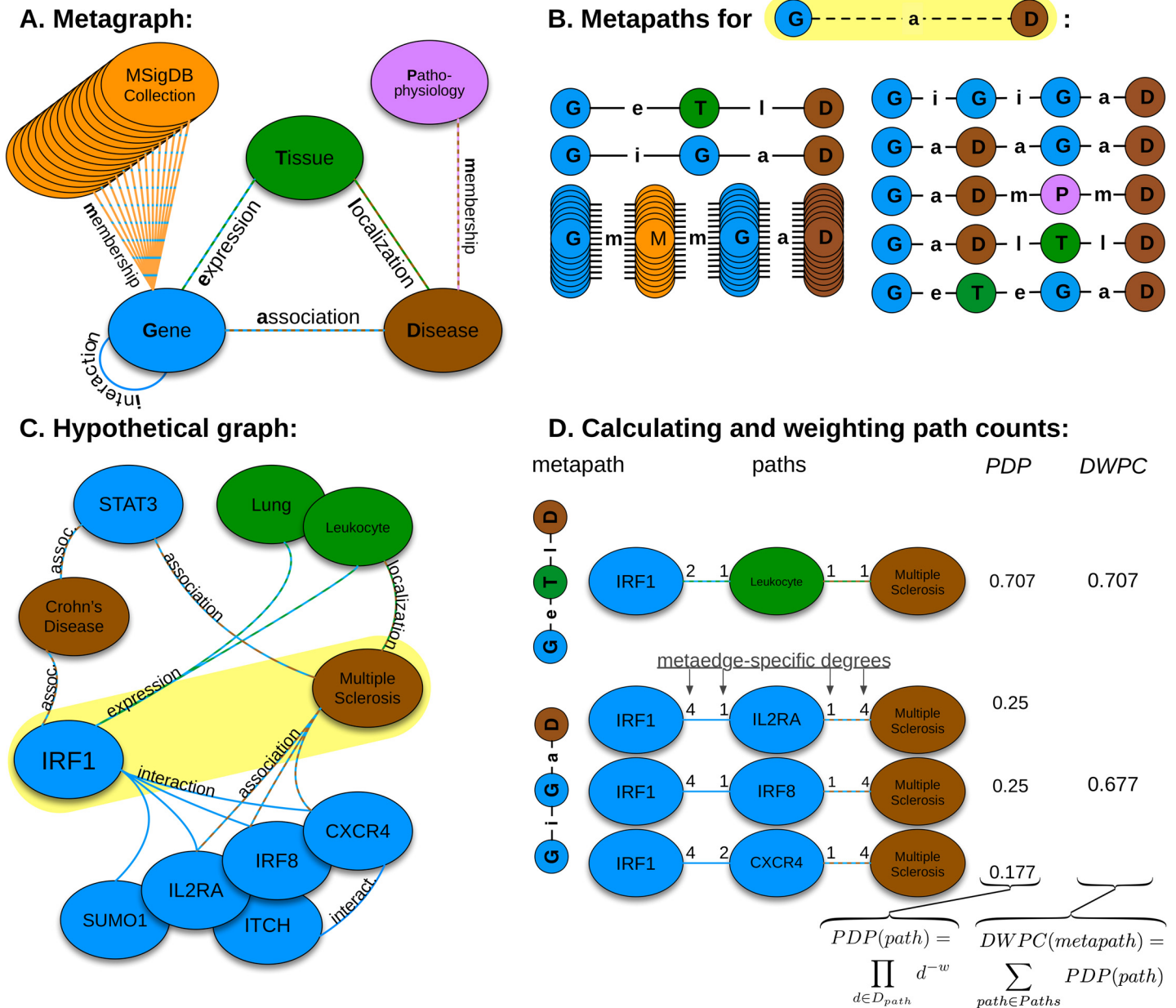
Fig 2. Heterogeneous network edge prediction methodology. A) We constructed the network according to a schema, called a metagraph, which is composed of metanodes (node types) and metaedges (edge types). B) The network topology connecting a gene and disease node is measured along metapaths (types of paths). Starting on Gene and ending on Disease, all metapaths length three or less are computed by traversing the metagraph. C) A hypothetical graph subset showing select nodes and edges surrounding *IRF1* and multiple sclerosis. To characterize this relationship, features are computed that measure the prevalence of a specific metapath between *IRF1* and multiple sclerosis. D) Two features (for the *GeTID* and *GiGaD* metapaths) are calculated to describe the relationship between *IRF1* and multiple sclerosis. The metric underlying the features is degree-weighted path count (*DWPC*). First, for the specified metapath, all paths are extracted from the network. Next, each path receives a path-degree product (*PDP*) measuring its specificity (calculated from node-degrees along the path, D_{path}). This step requires a damping exponent (here $w = 0.5$), which adjusts how severely high-degree paths are downweighted. Finally, the path-degree products are summed to produce the *DWPC*.

reported for the loci were considered secondary. Only high-confidence primary associations were included in the network yielding 938 associations between 99 diseases and 711 genes (S2 Fig visualizes a subset of these associations).

Table 1.

Metanode	Count	Source	References
Disease	99	Disease Ontology	[43]
Gene	19,116	HGNC (coding)	[44]
Tissue	77	BRENDA (BTO)	[45]
Pathophysiology	8	manual	–
Positional	326	MSigDB (C1)	[44]
Perturbation	3,402	MSigDB (C2)	[36,37]
BioCarta	217	MSigDB (C2)	[46]
KEGG	186	MSigDB (C2)	[38]
Reactome	674	MSigDB (C2)	[39]
miRNA Target	221	MSigDB (C3)	[41]
TF Target	615	MSigDB (C3)	[40,41]
Cancer Hood	427	MSigDB (C4)	[47]
Cancer Module	431	MSigDB (C4)	[48]
GO Process	825	MSigDB (C5)	[42]
GO Component	233	MSigDB (C5)	[42]
GO Function	396	MSigDB (C5)	[42]
Oncogenic	189	MSigDB (C6)	[49]
Immunologic	1,910	MSigDB (C7)	[49]

Metanodes. The kind, number of corresponding nodes, and data source for each type of node.

doi:10.1371/journal.pcbi.1004259.t001

Table 2.

Metaedge	Count	Source	References
Disease—association—Gene	938	GWAS Catalog	[5]
Disease—membership—Pathophysiology	90	manual	–
Disease—localization—Tissue	1,086	CoPub 5.0	[50]
Gene—expression—Tissue	251,366	GNF BodyMap	[51]
Gene—interaction—Gene	97,938	iRefIndex	[52]
Gene—membership—Positional	18,343	MSigDB (C1)	[44]
Gene—membership—Perturbation	366,211	MSigDB (C2)	[36,37]
Gene—membership—BioCarta	4,456	MSigDB (C2)	[46]
Gene—membership—KEGG	12,656	MSigDB (C2)	[38]
Gene—membership—Reactome	35,597	MSigDB (C2)	[39]
Gene—membership—miRNA Target	33,455	MSigDB (C3)	[41]
Gene—membership—TF Target	161,258	MSigDB (C3)	[40,41]
Gene—membership—Cancer Hood	41,913	MSigDB (C4)	[47]
Gene—membership—Cancer Module	48,220	MSigDB (C4)	[48]
Gene—membership—GO Process	75,155	MSigDB (C5)	[42]
Gene—membership—GO Component	34,880	MSigDB (C5)	[42]
Gene—membership—GO Function	23,578	MSigDB (C5)	[42]
Gene—membership—Oncogenic	30,166	MSigDB (C6)	[49]
Gene—membership—Immunologic	370,862	MSigDB (C7)	[49]

Metaedges. The kind, number of corresponding edges, and data source for each type of edge.

doi:10.1371/journal.pcbi.1004259.t002

Features quantify the network topology between a gene and disease

To describe the network topology connecting a specific gene and disease, we computed 24 features, each describing a different aspect of connectivity. Each feature corresponds to a type of path (metapath [53]) originating in a given source gene and terminating in a given target disease. The biological interpretation of a feature derives from its metapath (S1 Table), and features simply quantify the prevalence of a specific metapath between any gene-disease pair. To quantify metapath prevalence, we adapted an existing method originally developed for social network analysis (PathPredict) [34], and developed a new metric called degree-weighted path count (*DWPC*, Fig 2D), which we employed in all but two features. The *DWPC* downweights paths through high-degree nodes when computing metapath prevalence. The strength of downweighting depends on a single parameter (w), which we optimized to $w = 0.4$ and that outperformed the top metric resulting from PathPredict (S3 Fig) [34]. We calculated *DWPC* features for the 22 metapaths of length 3 or less that originated with a gene and terminated with disease. Two non-*DWPC* features were included to assess the pleiotropy of the source gene and the polygenicity of the target disease. Referred to as 'path count' features, they respectively equal the number of diseases associated with the source gene and the number of genes associated with the target disease. For all features, paths with duplicate nodes were excluded, and, if present, the association edge between the source gene and target disease was masked.

Machine learning approach to predict the probability of association of gene-disease pairs

Further analysis focused on the 29 diseases with at least ten associated genes (Table 3). The 698 high-confidence primary associations of these 29 diseases were considered positives—gene-disease pairs with positive experimental relationships (as defined in the Associations section of Methods, S2 Fig). The remaining 551,823 (i.e. unassociated) gene-disease pairs were considered negatives. Low-confidence or secondary associations were excluded from either set. We partitioned gene-disease pairs into training (75%) and testing (25%) sets and created a training network with the testing associations removed.

To learn the importance of each feature and model the probability of association of a given gene-disease pair, we used regularized logistic regression which is designed to prevent overfitting and accurately estimate regression coefficients when models include many features. Elastic net regression is a regression method that balances two regularization techniques: ridge (which performs coefficient shrinkage) and lasso (which performs coefficient shrinkage and variable selection) [54]. On the training set, we optimized the elastic net mixing parameter, a single parameter behind the *DWPC* metric, and two edge-inclusion thresholds (S3 Fig). While cross-validated performance was similar across elastic net mixing parameters, ridge demonstrated the greatest consistency (S3 Fig), and thus we proceeded with logistic ridge regression as the primary model for predictions.

Method prioritizes associations withheld for testing

We extracted network-based features for gene-disease pairs from the training network and modeled the training set. We next evaluated performance on the 25% of gene-disease pairs (175 positives, 137,956 negatives) withheld for testing. Our predictions achieved an area under the ROC curve (AUROC) of 0.83 (Fig 3A) demonstrating an excellent performance in retrieving hidden associations. Importantly, we did not observe any significant degradation of performance from training to testing (Fig 3A), indicating that our disciplined regularization approach avoided overfitting and that predictions for associations included in the network

Table 3.

Disease	Pathophysiology	HC-P	HC-S	LC-P	LC-S
Crohn's disease	immunologic	67	179	4	2
multiple sclerosis	immunologic	50	43	38	29
type 2 diabetes mellitus	immunologic	49	49	20	15
breast carcinoma	neoplastic	43	65	2	6
ulcerative colitis	immunologic	40	96	2	3
prostate carcinoma	neoplastic	34	202	3	4
type 1 diabetes mellitus	immunologic	33	56	9	6
rheumatoid arthritis	immunologic	30	27	20	11
coronary artery disease	metabolic	29	43	15	9
obesity	metabolic	28	22	34	18
celiac disease	immunologic	24	32	9	8
systemic lupus erythematosus	immunologic	22	35	14	8
refractive error	degenerative	21	11	2	1
primary biliary cirrhosis	immunologic	20	16	2	0
vitiligo	immunologic	20	27	4	0
age related macular degeneration	degenerative	18	30	11	18
metabolic syndrome X	metabolic	17	11	1	0
asthma	immunologic	17	23	13	4
psoriasis	immunologic	16	14	5	5
schizophrenia	psychiatric	15	27	20	13
chronic lymphocytic leukemia	neoplastic	14	16	3	4
migraine	unspecific	13	15	38	58
Alzheimer's disease	degenerative	12	11	27	18
Graves' disease	immunologic	12	15	1	1
Parkinson's disease	degenerative	12	21	8	13
atopic dermatitis	immunologic	11	15	5	1
bipolar disorder	psychiatric	11	34	26	74
lung carcinoma	neoplastic	10	14	6	6
ankylosing spondylitis	immunologic	10	5	6	6

Diseases. Associations were predicted for 29 diseases with at least 10 positives. For these diseases, the number of high-confidence primary (HC-P), high-confidence secondary (HC-S), low-confidence primary (LC-P), and low-confidence secondary associations (LC-S) that were extracted from the GWAS Catalog is indicated.

doi:10.1371/journal.pcbi.1004259.t003

were not biased by their presence in the network. Furthermore, we observed that at 10% recall (the classification threshold where 10% of true positives were predicted as positives), our predictions achieved 16.7% precision (the proportion of predicted positives that were correct). Since the prevalence of positives in our dataset was 0.13%, the observed precision represents a 132-fold enrichment over the expected probability under a uniform distribution of priors (as in GWAS).

Predicting associations on the complete network

As a next step in our analysis, we recomputed features on the complete network, which now included the previously withheld testing associations. On all positives and negatives, we fit a ridge model (the primary model for predictions) and a lasso model (for comparison). Standardized coefficients ([Fig 4](#)) indicate the effect attributed to each feature by the models. The
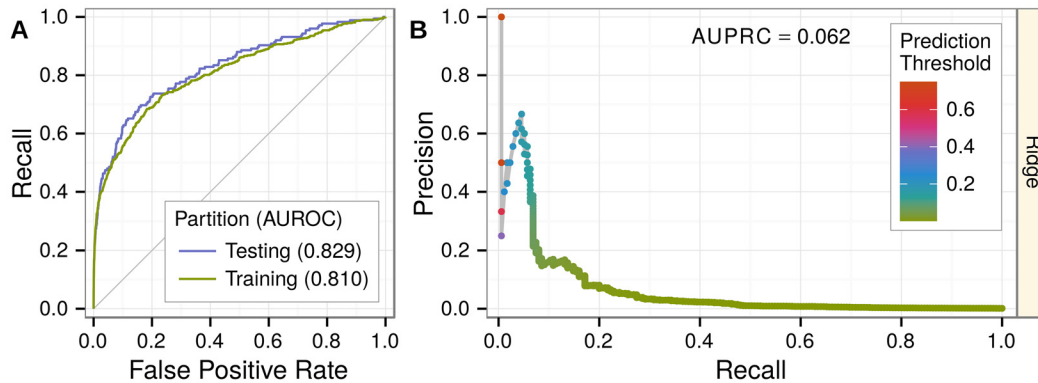
Fig 3. Predicting associations withheld for testing. Performance was evaluated on 25% of gene-disease pairs withheld for testing. A) Testing and training ROC curves. At top prediction thresholds, associated gene-disease pairs are recalled at a much higher rate than unassociated pairs are incorrectly classified as positives. The testing area under the curve (AUROC) is slightly greater than the training AUROC, demonstrating the method's lack of overfitting. Performance greatly exceeds random denoted by gray line. B) The precision-recall curve showing performance in the context of the low prevalence of associated gene-disease pairs (0.13%). Nevertheless, at top prediction thresholds, a high percentage of pairs classified as positives are truly associated. Prediction thresholds, shown as points and colored by value, align with the observed precision at that threshold.

doi:10.1371/journal.pcbi.1004259.g003

lasso highlighted features that captured pleiotropy (4 features), pathways (2), transcriptional signatures of perturbations (1) and protein interactions (1). Despite the parsimony of the lasso, performance was similar between models with training AUROCs of 0.83 (ridge) and 0.82 (lasso). However, since multiple features from a correlated group may be causal, the lasso model risks oversimplifying. Ridge regression disperses an effect across a correlated group of features, providing users greater flexibility when interpreting predictions. From the ridge model, we predicted the probability that each protein-coding gene was associated with each analyzed disease (S1 Data) and built a webapp to display the predictions (http://het.io/disease-genes/browse).
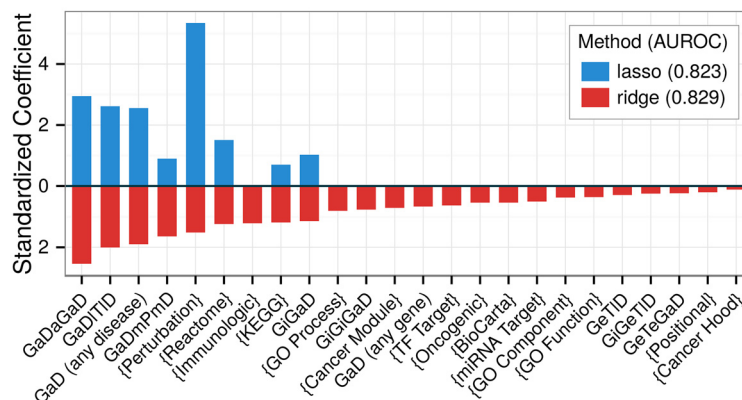
Fig 4. Feature selection identifies a parsimonious yet predictive model. Ridge and lasso models were fit from the complete network. The resulting standardized coefficients (y-axis) assess the effect size of each feature (x-axis). Brackets indicate features from MSigDB-traversing metapaths (*Gm{}mGaD*). The ridge model disperses effects amongst features whereas the lasso concentrates effects. The lasso identifies an 8-feature model with minimal performance loss compared to the ridge model. Besides *KEGG*, gene-set based features were largely captured by *Perturbations*. The lasso retains several measures of pleiotropy as well as the one-step interactome feature (*GiGaD*).

doi:10.1371/journal.pcbi.1004259.g004

Degree-preserving network permutations highlight the importance of edge-specificity for top predictions and ten features

Using Markov chain randomized edge-swaps, we created 5 permuted networks. Since metaedge-specific node degree is preserved, features extracted from the permuted network retain unspecific effects. These effects include general measures a disease's polygenicity and a gene's pleiotropy, multifunctionality, and tissue-specificity. On the first permuted network, we partitioned associations into training and testing sets. Testing associations were masked from the network, features were computed, and a ridge model was fit on the training gene-disease pairs.

Compared to the unpermuted-network model, testing performance was noticeably inferior: the AUROC declined from 0.83 (Fig 3A) to 0.79 and the AUPRC (area under the precision-recall curve) declined from 0.06 (Fig 3B) to 0.02 (S4 Fig). We interpret the modest decline in AUROC but marked reduction in AUPRC as a direct consequence of the permutation's particularly detrimental effect on top predictions (S4 Fig). In other words, edge-specificity was crucial for top predictions, while general effects gleaned from node degree performed reasonably well when ranking the entire spectrum of protein-coding genes for association. A commonly-overlooked finding is that the discriminatory ability of gene networks largely relies on node-degree rather than the edge-specificity [55]. However, we found that for top predictions—which are the only predictions considered by many applications—edge-specificity was critical.

Interestingly, predictions from the permuted-network model displayed a reduced dynamic range with none exceeding 4%, while predictions from the unpermuted-network model exceeded 75% (S4 Fig). Therefore, even though they achieve reasonable AUROC, the permuted-network predictions would have little utility as prior probabilities in a bayesian analysis where dynamic range is crucial. Furthermore, the signal present in permuted-network features was greatly diminished: few features survived the lasso's selection resulting in an average lasso AUROC of 0.70 versus 0.80 for ridge (S5 Fig). Permuting the network significantly reduced the predictiveness of features based on pleiotropy (2 features), protein interactions (2), transcriptional signatures of perturbations (1), tissue-specificity (1), pathways (3), and immunologic signatures (1) (S2 Table). Six of the eight features selected by the lasso and eight of the top ten ridge features (ranked by standardized coefficients) were negatively affected by the permutation. Since our modeling technique preferentially selected/weighted features affected by permutation, we can infer that network components where edge-specificity matters underlie a large portion of predictions.

Feature importance identifies the mechanisms underlying associations

We assessed the informativeness of each feature by calculating feature-specific AUROCs. Feature-specific AUROCs universally exceeded 0.5, indicating that network connectivity, regardless of type, positively discriminates associations. However, performance varied widely by feature and within feature from disease to disease (Fig 5). Top performing domains consisted of transcriptional signatures of perturbations (AUROC = 0.74), immunologic signatures (0.70), and pleiotropy (0.68, 0.67, 0.64, 0.63). Notably, the models greatly outperformed any individual feature, highlighting the importance of an integrative approach.

Features whose metapaths originate with an association (GaD) metaedge measure pleiotropy (S1 Table). The four pleiotropic features were among the top performing features that did not rely on set-based gene categorization (Fig 5). Of the four features, *GaD (any disease)* had the highest AUROC despite its lack of disease-specificity, reflecting both the sparsity of disease-specific features and the existence of genetic overlap between seemingly disparate diseases. *GaDmPmD* and *GaDaGaD* performed best for immunologic diseases and were affected by
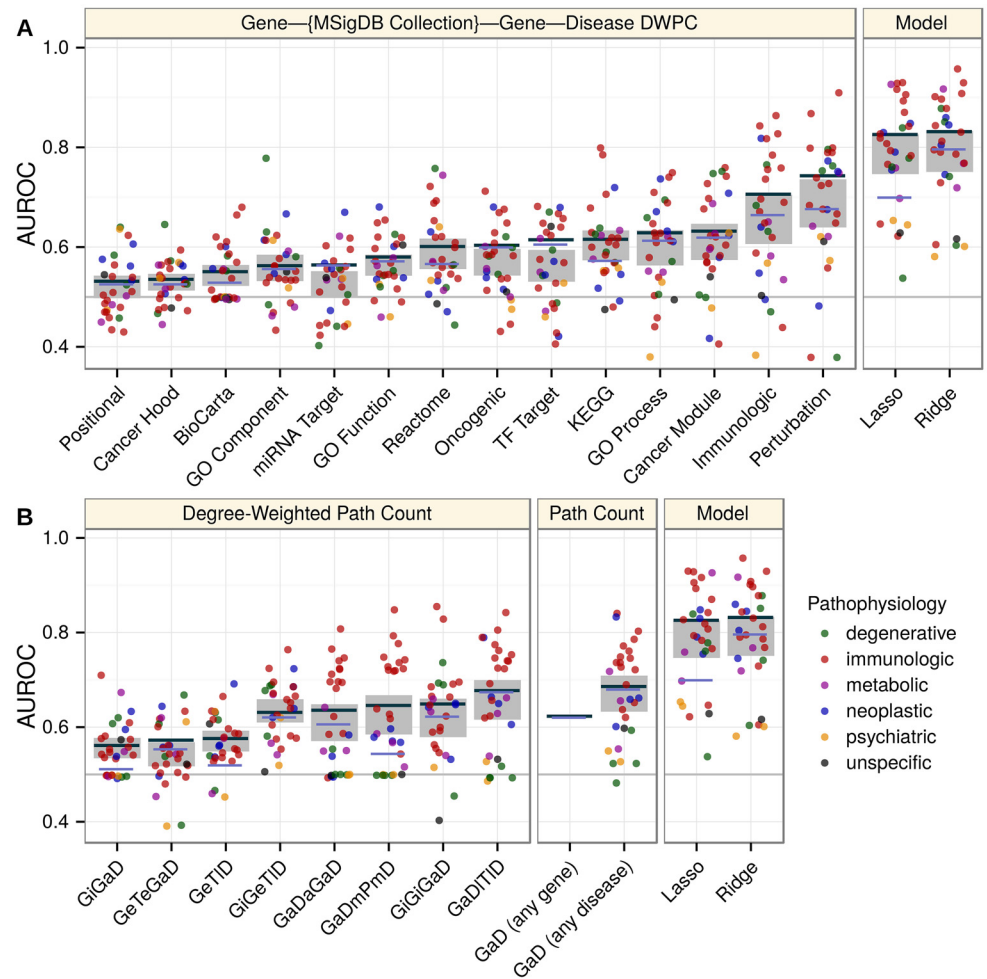
Fig 5. Decomposing performance shows the superiority of the integrative model and compares individual features. Disease, feature, and model-specific performance on the complete network. The AUROC (y-axis) was calculated for each classifier (x-axis). In addition to the ridge and lasso models (rightmost panels), each feature was considered as a classifier. Line segments show the classifier's global performance (average across permuted networks shown in violet as opposed to dark grey). Points indicate disease-specific performance and are colored by the disease's pathophysiology. Grey rectangles show the 95% confidence interval for mean disease-specific performance. A) Features from metapaths that traverse an MSigDB collection. B) Features from non-MSigDB-traversing metapaths. Metapaths are abbreviated using first letters of metanodes (uppercase, Table 1) and metaedges (lowercase, Table 2). Feature descriptions are provided in S1 Table.

permutation, indicating that genetic overlap was greatest between immunologic diseases. On the other hand, the performance of *GaDlTlD* did not decrease after permutation indicating disease colocalization was not a primary driver of genetic overlap (S5 Fig and S2 Table).

We also observed that the lasso regression model discarded the majority of features with a minimal performance deficit, suggesting redundancy among features. Indeed, pairwise feature correlations showed moderate collinearity among features (S6 Fig). Collinearity was especially pervasive with respect to the *Perturbations* feature, explaining its threefold increase in standardized coefficient in the lasso versus ridge model. The disappearance of all but one other MSigDB-based feature in the lasso model indicated that *Perturbations*—the feature traversing chemical and genetic transcriptional signatures of perturbations—exhausted meaningful gene-

set characterization. In other words, the faulty molecular processes behind pathogenesis align with and are encapsulated by the processes perturbed by chemical and genetic modifications. The *Immunologic signatures* feature—traversing gene-sets characterizing "cell types, states, and perturbations within the immune system"—was highly predictive and correlated with *Perturbations*. As expected this feature performed best for diseases with an immune pathophysiology. The one well-performing neoplastic disease ([Fig 5]) was chronic lymphocytic leukemia, a hematologic cancer with a strong immune component [56]. Additionally, the performance of both the *Perturbation* and *Immunologic* features was affected by permutation indicating information beyond the extent of a gene's multifunctionality was encoded.

Existing network-based gene-prioritization methods, frequently rely solely on protein-protein interactions. Our results supported incorporating protein interactions as the two interactome-based features were discriminatory (AUROCs = 0.65, 0.56) and affected by permutation. However, when compared to the integrative models or other top-performing features, performance of features that relied solely on the interactome was severely limited. Pathways, another founding resource for many approaches, proved important with *KEGG* selected by the lasso and all three pathway resources (AUROCs = 0.61 for *KEGG*, 0.60 for *Reactome*, 0.55 for *BioCarta*) affected by permutation. The *GeTlD* feature—measuring to what extent a gene is expressed in tissues affected by the disease in question—peaked in performance around AUROC = 0.58 ([S3 Fig]), was affected by permutation, and required no preexisting knowledge of associated genes. In other words, while approaches based on tissue-specificity may have limited predictive ability on their own, they are broadly applicable (i.e. less susceptible to knowledge bias) and provide orthogonal information that could enhance the overall performance of a model.

Gene set robustness

For each type of gene set, we evaluated the effect of increased sparsity on performance by randomly subsampling gene set nodes or edges and measuring the resulting AUROC of the affected feature ([S7 Fig]). Robustness refers to a gene set collection's ability to withstand a high extent of masking with little performance deficit. Several of the top gene sets had this property, especially GO processes (where supersets are common), which may indicate nodal redundancy. Contrastingly, the MSigDB gene set with the fewest nodes, KEGG, experienced a more immediate and linear decline in performance. Since KEGG avoids duplication and is stringently and manually curated, this finding is expected. To investigate whether the high predictivity of certain gene set collections was due only to size, we compared performance when subsampling nodes to the KEGG level (crosses in [S7 Fig]). The two top performing collections, perturbations and immunologic signatures, which also happen to be large, continued to perform better than the majority of complete collections. While performance benefited from increasing densities, a resource's sparsity often reflects an intrinsic property of the underlying information type. Therefore, when identifying influential mechanisms of pathogenesis, we preferred unadjusted comparisons using the complete network.

Case study: Prioritizing multiple sclerosis associations

The WTCCC2 multiple sclerosis (MS) GWAS tested 465,434 SNPs for 9,772 cases and 17,376 controls and identified over 50 independently associated loci [57]. Since the GWAS Catalog excludes targeted arrays (such as ImmunoChip), this study remains the largest MS GWAS in the Catalog. To evaluate our method's ability to prioritize associations identified in a future study, we masked the WTCCC2 MS study from the GWAS Catalog and created a pre-WTCCC2 network. The number of high-confidence primary MS associations was thus reduced
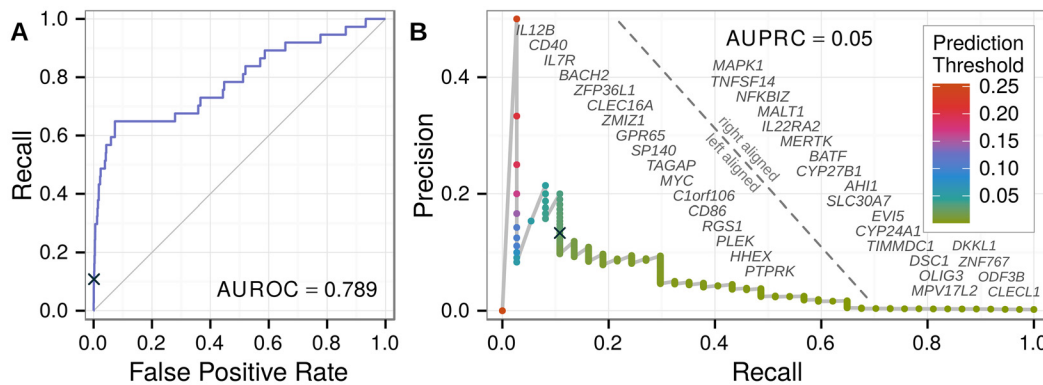
Fig 6. Prioritizing multiple sclerosis associations identified by a masked GWAS. From a network with the WTCCC2 MS associations omitted, we predicted probabilities of association for all potentially novel genes. The 37 novel genes identified by the WTCCC2 GWAS were considered positives, and the resulting performance was plotted. The ROC (A) and precision-recall (B) curves show performance, with AUCs in line with the testing performance across all diseases. A prediction threshold (black cross) that resulted in high performance was selected as the discovery threshold for further analysis. As the classification threshold decreases along the precision-recall curve, the advent of each true positive is denoted by its gene symbol.

from 50 to 13, with the 37 novel genes identified by WTCCC2 available to evaluate performance. On the pre-WTCCC2 network, we extracted features, fit a ridge model, and predicted each gene's probability of association with MS. Amongst all 18,993 potentially novel genes, the 37 WTCCC2 genes were ranked highly (AUROC = 0.79, Fig 6).

Prioritizing statistical candidates with network-based predictions identifies novel multiple sclerosis genes

Finally, we designed a framework for discovering and validating novel MS genes that incorporates our network-based predictions. Meta2.5 is a meta-analysis of all MS GWAS prior to the WTCCC2 study [58]. We calculated genewise p-values for Meta2.5 using VEGAS [59] and observed a large enrichment in nominally significant ($p < 0.05$) genes, suggesting multiple potential associations (S9 Fig). We combined this set of experimental candidates with the top predictions from the pre-WTCCC2 network to discover genes with both strong statistical and biological evidence of association (S12 Data). To ensure novelty, we excluded genes from GWAS-established MS loci and the extended MHC region. We chose a threshold (S3 Table) for network-based predictions that performed well in prioritizing the genes identified by WTCCC2 (Fig 6).

This strategy discovered four genes, three of which—*JAK2*, *REL*, *RUNX3*—achieved Bonferroni validation on VEGAS-converted WTCCC2 p-values (Table 4). The probability of the observed validation rate occurring under random prioritization is 0.01 (S3 Table), demonstrating that incorporating our network-based predictions as a prior increased study power. *JAK2* displays overexpression in MS-affected Th17 cells [60] and was implicated in an interactome-based prioritization of GWAS [15]. *RUNX3*, a transcription factor influencing T lymphocyte development, has been associated with celiac disease [61] and ankylosing spondylitis [62] and was hypermethylated in systemic lupus erythematosus patients [63]. The region containing *REL* was uncovered in a recent MS ImmunoChip-based study with 14,498 cases [64]. For the gene-dense region containing *REL*, the ImmunoChip study reported a long non-coding RNA, *LINC01185*, overlapping the lead-SNP, rs842639. However, since greater than 80% of the genome shows evidence of transcription [65], the probability of incidental overlap with long non-coding RNA is high. *REL*, however, is an essential transcription factor for lymphocyte development [66] and plays a critical role in autoimmune inflammation [67]. Hence, gene

Table 4.

Gene	Meta2.5	HNEP	WTCCC2
JAK2	0.047	0.102	**0.0015**
REL	0.001	0.040	**0.0003**
SH2B3	0.012	0.034	0.0130
RUNX3	0.016	0.025	**0.0073**

Multiple sclerosis gene discovery. Four genes showed nominal statistical evidence of association (Meta2.5 column) and exceeded the network prediction threshold (HNEP column). Three genes achieved Bonferroni validation (bold) in an independent GWAS (WTCCC2 column).

doi:10.1371/journal.pcbi.1004259.t004

prioritization through integrative analyses offers not only to streamline loci discovery but also subsequent causal gene identification.

Discussion

In this work, we developed a framework to predict the probability that each protein-coding gene is associated with each of 29 complex diseases. Our predictions draw on a diverse set of pathogenically-relevant relationships encoded in a heterogeneous network. The predictions successfully prioritized associations hidden from the network. Using MS as a representative example, we were able to combine our predictions with statistical evidence of association to increase study power and identify three novel susceptibility genes in this disease. The disease-specific performance (measured by the AUROC) for MS was exceeded by twelve other diseases suggesting that our predictions have broad applicability for prioritizing genetic association analyses. Prioritization can range from a genome-wide scale to a single loci where this approach can highlight the causal gene from several candidates within the same association block. For researchers focused on a specific disease, these predictions can be used to propose genes for experimental investigation. Inversely, researchers focused on a specific gene can use this resource to find suggestions for relevant complex disease phenotypes.

Most previous explorations of the factors underlying pathogenicity have focused on a single domain such as tissue-specificity [68], protein interactions [69], pathways [7], or disease similarity [70]. The method presented here integrates disparate data sources, learns their importance, and unifies them under a common framework enabling comparison. Therefore, we can conclude that perturbation gene sets—the core of our top-performing feature—are an under-utilized resource for disease-associated gene prioritization. Not only did perturbations encompass other set-based gene categorizations, but they greatly outperformed features based on protein interactions, pathways, and tissue-specificity, which form the basis of several prominent prioritization techniques. In addition to characterizing the overall importance of each feature, our online prediction browser visually decomposes an individual prediction into its components.

We observed a prominent influence of pleiotropy, consistent with previous studies that identified pervasive overlap of susceptibility loci across complex diseases [71], especially those of autoimmune nature [72]. Since many existing prioritization techniques are agnostic to the compendia of GWAS associations, they fail to adequately leverage pleiotropy. Unlike approaches initiated from a user-provided gene list, our study only provides predictions for 29 diseases. By not relying on user-provided input, our predictions can serve as independent priors for future analyses. By predicting probabilities, we provide an extensible and interpretable assessment of association that circumvents the limitations inherent to frequentist analyses

[73]. Many approaches return no assessment for the majority of genes which fall outside of their set of predicted positives. Here, we overcome this issue and provide a comprehensive and genome-wide output by returning a probability of association for each protein-coding gene.

High-throughput biological data is frequently noisy and incomplete [74]. Combining orthogonal resources can help overcome these issues. Accordingly, we found that our integrative model outperformed any individual domain. While this method has shown encouraging performance, some limitations are worth noticing. For example, many biological networks preferentially cover well-studied vicinities [75]. Knowledge biases that span multiple, presumably-orthogonal resources could diminish the benefits of integration. Here, several of the literature-derived domains were removed by the lasso, suggesting redundancy. In addition, biases in network completeness can lead to high-quality predictions for well-studied vicinities and low-quality predictions for poorly-studied vicinities. The permutation analysis provided evidence of this disparity: edge-specificity was critical for top predictions yet only moderately beneficial for the remainder. Subsequently, we caution users to avoid overinterpreting predictions for poorly-characterized genes. To help place predictions in context, the online browser provides a gene's mean prediction across all diseases and a disease's mean prediction across all genes. However, we recognize that false negatives will continue to persist in our predictions, and users should be mindful of this limitation when interpreting results. As more systematic and unbiased resources become available [74], high-quality predictions will emerge for more network vicinities.

We reason that the desirable qualities of our predictions are the consequence of the heterogenous network edge prediction methodology. The approach is versatile (most biological phenomena are decomposable into entities connected by relationships), scalable (no theoretical limit to metagraph complexity or graph size), and efficient (low marginal cost to including an additional network component). We have extended the previous metapath-based framework set forth by PathPredict [34], by: 1) incorporating regularization allowing coefficient estimation for more features without overfitting; 2) designing a framework for predicting a metaedge that is included in the network; 3) developing an improved metric for assessing path specificity; and 4) implementing a degree-preserving permutation. Metapath-based heterogeneous network edge prediction provides a powerful new platform for bioinformatic discovery.

Methods

Ethics statement

This study was approved by the UCSF institutional review board on human subjects under protocol #10–00104.

Heterogeneous networks

We created a general framework and open source software package for representing heterogeneous networks. Like traditional graphs, heterogeneous networks consist of nodes connected by edges, except that an additional meta layer defines type. Node type signifies the kind of entity encoded, whereas edge type signifies the kind of relationship encoded. Edge types are comprised of a source node type, target node type, kind (to differentiate between multiple edge types connecting the same node types), and direction (allowing for both directed and undirected edge types). The user defines these types and annotates each node and edge, upon creation, with its corresponding type. The meta layer itself can be represented as a graph consisting of node types connected by edge types. When referring to this graph of types, we use the prefix 'meta'. Metagraphs—called schemas in previous work [34,35]—consist of metanodes connected by metaedges. In a heterogeneous network, each path, a series of edges with common

intermediary nodes, corresponds to a metapath representing the type of path. A path's meta-path is the series of metaedges corresponding to that path's edges. The possible metapaths within a heterogeneous network can be enumerated by traversing the metagraph. We implemented this framework as an object-oriented data structure in python and named the resulting package *hetio*. Users are free to browse, use, or contribute to the software, through the online repository (https://github.com/dhimmel/hetio).

Network construction

The included resources, and hence the metaedges and metanodes composing our network, were selected empirically based on a balance among the following properties: 1) **quality**—relevance to human pathogenesis; high accuracy and an optimal trade-off between false positives and false negatives. In some cases, quality concerns prevented the inclusion of a desired metaedge. For example, we omitted ontology-based disease similarly [76] due to an inaccurate Disease Ontology hierarchy [43], and we omitted disease comorbidity due to high measurement error for uncommon diseases [77]. For included metaedges, we attempted to select the highest quality resource in that domain. 2) **reusability**—easily retrievable and parsable; mapped to controlled vocabularies; well documented; amenable to reproducible (scripted) analysis; free of prohibitive reuse stipulations. 3) **throughput**—broad domain-specific coverage generated using systematic platforms that minimize bias. While genetic interactions have previously proven informative [31], their sparse characterization in humans was deemed unfavorable for our approach. 4) **diversified, multiscale portrayal of biology**—capturing, in aggregate, many aspects of pathophysiology across multiple levels of biological complexity. Levels of the hierarchical architecture of biological complexity include the genome, transcriptome, proteome, interactome, metabolome, cell and tissue organization, and phenome. Balancing these considerations, we integrated as many resources as possible within our computational runtime constraints.

Nodes. Protein-coding genes (S3 Data) were extracted from the HGNC database [44]. Resources were mapped to HGNC terms via gene symbol (ambiguous symbols were resolved in the order: approved, previous, synonyms) or Entrez identifiers. Disease nodes (S6 and S7 Data) were taken from the Disease Ontology (DO) [43]. Due to the limited number of diseases with GWAS, relevant disease references were manually mapped to the DO (S11 Data). Tissues were taken from the BRENDA Tissue Ontology (BTO) [45]. Only tissues with profiled expression were included enabling manual mapping. Nodes for the 14 MSigDB metanodes were directly imported from the Molecular Signature Database version 4.0 [36,37]. All MSigDB collections were included except those that were supersets of other collections. For example, 'C3: motif gene sets' was the union of two disjoint collections ('C3: microRNA targets' and 'C3: transcription factor targets') and was therefore excluded. Diseases were classified manually into 10 categories according to pathophysiology. The 'idiopathic' and 'unspecific' categories were not included as pathophysiology nodes, since they do not signify meaningful similarities between member diseases.

Associations. Disease-gene associations were extracted from the GWAS Catalog [5], a compilation of GWAS associations where $p < 10^{-5}$. First, associations were segregated by disease. GWAS Catalog phenotypes were converted to Experimental Factor Ontology (EFO) terms using mappings produced by the European Bioinformatics Institute. Associations mapping to multiple EFO terms were excluded to eliminate cross-phenotype studies. We manually mapped EFO to DO terms (now included in the DO as cross-references) and annotated each DO term with its associations.

Associations were classified as either high or low-confidence, where exceeding two thresholds granted high-confidence status. First, $p \leq 5^*10^{-8}$ corresponding to $p \leq 0.05$ after

Bonferroni adjustment for one million comparisons (an approximate upper bound for the number of independent SNPs evaluated by most GWAS). Second, a minimum sample size (counting both cases and controls) of 1,000 was required, since studies below this size are underpowered [78]—i.e. any discovered associations are more likely than not to be false—for the majority of true effect size distributions commonly assumed to underlie complex disease etiology [73].

Lead-SNPs were assigned windows—regions wherein the causal SNPs are assumed to lie—retrieved from the DAPPLE server [12]. Windows were calculated for each lead-SNP by finding the furthest upstream and downstream SNPs where $r^2 > 0.5$ and extending outwards to the next recombination hotspot. Associations were ordered by confidence, sorting on following criteria: high/low confidence, p-value (low to high), and recency. In order of confidence, associations were overlapped by their windows into disease-specific loci (S4 Data). By organizing associations into loci, associations from multiple studies tagging the same underlying signal were condensed (S1 Fig). A locus was classified as high-confidence if any of its composite associations were high-confidence and low-confidence otherwise.

For each disease-specific loci, we attempted to identify a primary gene. The primary gene was resolved in the following order: 1) the mode author-reported gene; 2) the containing gene for an intragenic lead-SNP; 3) the mode author-reported gene for an intragenic lead-SNP (in the case of overlapping genes); 4) the mode author-reported gene of the most proximal up and downstream genes. Steps 2–4 were repeated on each association composing the loci, in order of confidence, until a single gene resolved as primary. Loci where ambiguity was unresolvable or where no genes were returned did not receive a primary gene. All non-primary genes—genes that were author-reported, overlapping the lead-SNP, or immediately up or downstream from the lead-SNP—were considered secondary.

Accordingly, four categories of processed associations were created: high-confidence primary, high-confidence secondary, low-confidence primary, and low-confidence secondary (S5 Data). We assume that our primary gene annotation for each loci represents the single causal gene responsible for the association. To investigate the validity of this assumption, we evaluated the performance of our predictions separately using each category of association as positives (S8 Fig). For both confidence levels, primary associations outperformed secondary associations suggesting our method succeeded at categorizing causal genes as primary. However, for high-confidence secondary associations, the AUROC equaled 0.74, which could result from multiple causal genes per loci or categorizing sole causal genes as secondary. The performance decline from high to low confidence associations was severe, pointing to a preponderance of falsely identified loci in the GWAS Catalog when $p > 5 \times 10^{-8}$ or sample size drops below 1000.

Protein interactions.   Physical protein-protein interactions (S8 Data) were extracted from iRefIndex 12.0, a compilation of 15 primary interaction databases [52]. The iRefIndex was processed with ppiTrim to convert proteins to genes, remove protein complexes, and condense duplicated entries [79].

Tissue-specific gene expression.   Tissue-specific gene expression levels (S9 Data) were extracted from the GNF Gene Expression Atlas [51]. Starting with the GCRMA-normalized and multisample-averaged expression values, 44,775 probes were converted to 16,466 HGNC genes and 84 tissues were manually mapped and converted to 77 BTO terms. For both conversions, the geometric mean was used to average expression values. The log base 10 of expression value was used as the threshold criteria for *GeT* edge inclusion.

Disease localization.   Disease localization was calculated for the 77 tissues with expression profiles (S10 Data). Literature co-occurrence was used to assess whether a tissue is affected by a disease. We used CoPub 5.0 to extract R-scaled scores between tissues and diseases measuring whether two terms occurred together in Medline abstracts more than would be expected by

chance [50]. DO terms for diseases with GWAS and BTO tissues with expression profiles were manually mapped to the 'biological identifier' terminology used by CoPub. The R-scaled score was used as the threshold criteria for *TlD* edge inclusion.

Feature computation metrics. The simplest metapath-based metric is path count (*PC*): the number of paths, of a specified metapath, between a source and target node. However, *PC* does not adjust for the extent of graph connectivity along the path. Paths traversing high-degree nodes will account for a large portion of the *PC*, despite high-degree nodes frequently representing a biologically broad or vague entity with little informativeness. The previous work evaluated several metrics that include a *PC* denominator to adjust for connectivity and reported that normalized path count (*NPC*) performed best [34]. The denominator for *NPC* equals the number of paths from the source to any target plus the number of paths from any target to the source.

$$NPC_m(s, t) = \frac{PC_m(s, t)}{\sum_{t_i \in T_m} PC_m(s, t_i) + \sum_{s_i \in S_m} PC_m(s_i, t)},$$

where *m* is the metapath, *s* is the source node, *t* is the target node, S_m is the set of nodes corresponding to the source metanode of *m*, and T_m is the set of nodes corresponding to the target metanode of *m*. We adopt the any source/target concept to compute the two *GaD* features. However, dividing the *PC* by a denominator is flawed because each path composing the *PC* deserves a distinct degree adjustment. If two paths—one traversing only high-degree nodes and one traversing only low-degree nodes—compose the *PC*, the network surrounding the high-degree path will monopolize the *NPC* denominator and overwhelm the contribution of the low-degree path despite its specificity. Therefore, we developed the degree-weighted path count (*DWPC*) which individually downweights each path between a source and target node. Each path receives a path-degree product (*PDP*) calculated by: 1) extracting all metaedge-specific degrees along the path (D_{path}), where each edge composing the path contributes two degrees; 2) raising each degree to the $-w$ power, where $w \geq 0$ and is called the damping exponent; 3) multiplying all exponentiated degrees to yield the *PDP*.

$$PDP(path) = \prod_{d \in D_{path}} d^{-w}$$

The *DWPC* equals the sum of *PDPs*.

$$DWPC_m(s, t) = \sum_{path \in Paths_m(s,t)} PDP(path)$$

See Fig 2C and 2D for a visual description of the *DWPC*.

Machine learning approach

PathPredict relied on basic logistic regression to predict coauthorship status from features corresponding to nine distinct metapaths [34]. However, faced with fewer positives to train our model and a large number of features, we adopted a regularized approach, which aims to contain the overfitting tendencies inherent to regression. Regularization penalizes complexity, a trademark of overfitting. We chose the elastic net technique of regularization [54], which is efficiently implemented for logistic regression by the R glmnet package [80].

Regularized logistic regression requires a parameter, λ, setting the strength of regularization. We optimized λ separately for each model fit. Using 10-fold cross-validation and the

"one-standard-error" rule to choose the optimal λ from deviance, we adopted a conservative approach designed to prevent overfitting [80].

On the training set of gene-disease pairs, we optimized the elastic net mixing parameter (α), the *DWPC* damping exponent (*w*), and two edge inclusion thresholds. First, we optimized α and *w* on the 20 features whose metapaths did not include threshold-dependent metaedges. For each combination of α and *w*, we calculated average testing AUROC using 20-fold cross-validation repeated for 10 randomized partitionings. After setting α and *w*, we jointly optimized the two edge-inclusion thresholds using the AUROC for the *GeTlD* feature, whose metapath is composed from the two edges requiring thresholds (S3 Fig).

We adopt standardized coefficients as a measure of feature effect size. Standardized coefficients refer to the coefficients from logistic regression when all features have been transformed to *z*-scores. Standardization provides a common scale to assess feature effect, both within and across models [81].

Degree-preserving permutation

Starting from the complete network, a permuted network was created by swapping edges separately for each metaedge. Edge swaps were performed by switching the target nodes for two randomly selecting edges [82]. For each metaedge, the number of attempted swaps was ten times the corresponding edge count. We adopted a Markov Chain strategy where additional rounds of permutation were initiated from the most-recently permuted network [82]. A training network was generated from the first permuted network by masking 25% of the associations for testing. Testing performance for the permuted training network model is shown in S4 Fig. When contrasting this performance with the unpermuted-network model, we employed the Condensed-ROC curve to magnify the importance of top predictions [83]. Using the exponential transformation with a magnification factor of 460—the value which maps a FPR of 0.01 to 0.99—we concentrated on the top 1% of predictions (S4 Fig). A one-sided unpaired DeLong test [84] was used to assess whether feature-specific testing AUROCs from the unpermuted network exceeded those from the first permuted network (S2 Table).

Gene set subsampling

We performed a subsampling analysis for 15 gene sets—the 14 MSigDB gene sets and tissues—to assess the effect of sparsity on feature-specific performance (S7 Fig). Two without-replacement subsampling schemes were investigated: node masking and edge masking. For a specific gene set and scheme, we masked a percentage of the gene set and calculated the corresponding feature's AUROC. We evaluated a range of percentages and performed ten subsampling repetitions for each percentage.

Multiple sclerosis gene discovery

We excluded 588 genes from the discovery phase of the multiple sclerosis analysis. First we excluded genes in the extended MHC region (spanning from *SCGN* to *SYNGAP1* on chromosome 6 [85]) due to the complex pattern of linkage characterizing this region containing several highly-penetrant MS-risk alleles [57]. Second, we excluded putative MS genes: high-confidence primary genes from the GWAS Catalog and reported genes for the WTCCC2-replicated loci. We omitted genes in linkage disequilibrium with the putative genes by excluding: 1) consecutive sequences of nominally significant genes (using the WTCCC2-VEGAS p-values) that included a putative gene; and 2) high-confidence secondary genes from the GWAS catalog. Post exclusion, 1211 genes were nominally significant in Meta2.5, four of which exceeded the network-based discovery threshold. Using a hypergeometric test for overrepresentation, we

calculated the probability of randomly selecting 4 of the 1211 genes and Bonferroni validating at least 3 of the 4 on WTCCC2 (S3 Table).

Data availability

See S1–S12 Datasets for the supporting data and S13 Data for vector figures. The website provides additional resources (http://het.io/disease-genes/downloads/) as well as an interface for browsing results (http://het.io/disease-genes/browse/). Project related code is available from the github repository (https://github.com/dhimmel/hetio).

Supporting Information

S1 Fig. Extracting disease-specific loci and associated genes from the GWAS Catalog. First, associations from the GWAS Catalog are segregated by disease. In this hypothetical example, we show a selection of multiple sclerosis associations. Each association consists of a lead SNP (point) and window (line segment) within which the causal SNP is expected to reside. In order of significance, association windows are overlapped into disease-specific loci (blue bars). For loci identified by multiple studies, the most commonly reported gene is considered primary and the remainder are considered secondary. A loci is classified as high-confidence if any of its associations exceed the threshold shown by the dashed red line.
(PDF)

S2 Fig. Bipartite network of gene-disease associations. Gene-disease associations were extracted from the GWAS Catalog. Here we show the 698 high-confidence primary associations for the 29 diseases with at least 10 associations. Diseases (large nodes) and their incident edges are colored according to disease pathophysiology. Node positions were manually adjusted for clarity after an initial force-directed layout. The network highlights pervasive pleiotropy as well as the overlap of susceptibility genes among autoimmune diseases.
(PDF)

S3 Fig. Parameter optimization. Using the training network, optimal parameter values (yellow dashed lines) were chosen. A) Using average cross-validated AUROC to assess performance, six elastic net mixing parameters were evaluated. For each mixing parameter value α, 10 feature metrics were evaluated: the *DWPC* for 9 weighting exponents (*w*, magenta with a 99.99% loess confidence band) and the *NPC* (violet with a 99.99% confidence interval). The *DWPC* with $w = 0.4$ outperformed the *NPC*, the best metric from previous work, as well as the path count which equals the *DWPC* when $w = 0$. Performance variability was minimized when $α = 0$. B) Edge-inclusion thresholds for two metaedges were jointly optimized. Expression threshold refers to the minimum microarray intensity required for a tissue-specific expression (*GeT*) edge. Localization threshold refers to the minimum literature co-occurrence score required for a disease localization (*TlD*) edge. Treating the *DWPC* ($w = 0.4$) for the *GeTlD* metapath as a classifier, the AUROC was calculated at each pairwise threshold combination. The optimal thresholds were chosen as the center of a stable, high-performing, and computationally-feasible section of the solution space.
(PDF)

S4 Fig. Performance of the degree-preserving permutation. Testing performance is contrasted between ridge models for the permuted-network and unpermuted-network. A) Testing and training ROC curves for the permuted-network model. B) Testing precision-recall curve for the permuted-network model. C) Testing CROC curves for the permuted-network and unpermuted-network models. The FPR has been scaled to focus on the first 1% placing greater emphasis on top predictions. While both models vastly outperform random (grey line), the

unpermuted-network model provides far superior top predictions. D) For both networks, gene-disease pairs were stratified by deciles of the predicted probabilities for positives. The x-axis labels show the predicted probabilities (as percentages) composing each decile. For each strata, the percent of positive pairs (precision) is plotted. The fold change over permuted is denoted for the unpermuted deciles.
(PDF)

S5 Fig. Disease, feature, and model-specific performance across permuted-network models. Disease, feature, and model-specific AUROCs were calculated separately for each of 5 permuted networks and averaged. The figure is analogous to Fig 5, except all measures refer to permuted-network performance. Disease-specific performance tends towards the mean, as disease-specific information has been altered by permutation. For features ending with an association (*GaD*) metaedge, global performance exceeds disease-specific performance. These features capture disease polygenicity, which improves the ranking of gene-disease pairs only if multiple diseases are included. Performance of the lasso model is affected, since the signals become too weak and few features survive regularization.
(PDF)

S6 Fig. Pairwise feature correlation. Pearson's correlation coefficients (shown by color and as a percent) were calculated for all pairwise feature combinations. Features were ordered using Ward's hierarchical clustering. Moderate collinearity is pervasive across features. The four pleiotropy-focused features form a tight cluster (top left). *Perturbations* and *Immunologic signatures* are correlated with many other features, including several other MSigDB features.
(PDF)

S7 Fig. Gene set subsampling. The effect of sparsity on feature performance is displayed for each gene set. The provided gene set information refers to node number (n), edge number (e), and mean degree (d). Crosses show performance for each gene after being subsampled to 186 nodes, corresponding to the number of pathways in KEGG—the MSigDB gene set in our network with the fewest nodes. The 95% loess confidence bands show expected performance across the entire range of node and edge masking percentages.
(PDF)

S8 Fig. Performance of the predictions on the four categories of associations. Keeping unassociated gene-disease pairs as negatives, ROC curves were calculated separately for each category of association as positives. Predictions from the complete-network ridge model were used as the classifier. For both high and low-confidence associations, primary gene annotations received higher predictions than secondary gene annotations. High-confidence associations received considerably higher predictions than low-confidence associations suggesting a high frequency of false positives amongst low-confidence associations.
(PDF)

S9 Fig. Excess of nominally significant genewise p-values in Meta2.5. The histogram of genewise p-values from Meta2.5, a meta-analysis of multiple sclerosis GWAS preceding the WTCCC2 study. If no associations are present, uniformly distributed p-values (grey line) would be expected. Instead, we observed an excess of nominally significant genes ($p \leq 0.05$, red) indicating a set of genes likely enriched for true associations.
(PDF)

S1 Table. Features. The 24 features computed for each gene-disease pair and the aspect of network topology described.
(PDF)

S2 Table. Feature-specific performance before and after network permutation. Ten features (bold) showed a significant ($p < 0.05$, one-sided DeLong test) decrease in performance. (PDF)

S3 Table. Multiple sclerosis gene discovery statistics. The upper section details the high-performing network prediction threshold. The lower section details the hypergeometric test for overrepresentation of validating genes. (PDF)

S1 Data. Predictions. Predicted probabilities of association between all genes (rows) and diseases (columns). (GZ)

S2 Data. Network tables. A table of network edges (sif format) and a table of node attributes. (ZIP)

S3 Data. Protein-coding genes. Information on the included protein-coding genes, derived from the HGNC database. (TXT)

S4 Data. Processed GWAS Catalog loci. Loci-disease associations. The file includes the gene resolution information for each loci including the studies and SNPs underlying the association. (TXT)

S5 Data. Gene-disease associations. All gene-disease associations extracted from the GWAS catalog for the four categories of association. (TXT)

S6 Data. Complex diseases. An extended version of Table 3 including all diseases with at least one GWAS-Catalog-extracted association. The manual pathophysiology classification is included. (TXT)

S7 Data. Disease Ontology modifications. Ten DO terms that appeared in the GWAS Catalog were redundant with other terms. Seven were removed. Three were merged with recipient terms by removing the term and transferring the associations. (TXT)

S8 Data. Protein interactions. Protein-protein interactions processed from iRefIndex using ppiTrim. (GZ)

S9 Data. Tissue-specific gene expression. A processed version of the GNF BodyMap providing a gene's (row, HGNC symbols) expression value for each of 77 tissues (columns, BRENDA Tissue Ontology IDs). (GZ)

S10 Data. Disease localization. Literature co-occurrence scores between diseases and tissues computed using CoPub 5.0. (TXT)

S11 Data. Terminology mappings. All mappings that were manually performed. Specifically, tissue and disease mappings to CoPub 'Biologic Identifiers', tissue mappings to GNF BodyMap samples, disease mappings to the EFO terms appearing in the GWAS Catalog, and disease

pathophysiologies.
(ZIP)

S12 Data. Multiple sclerosis analysis. For each gene (row), the genewise Meta2.5 and WTCCC2 p-values and network-based predictions are reported.
(TXT)

S13 Data. Vector images. PDF formatted versions of the figures.
(ZIP)

Author Contributions

Conceived and designed the experiments: DSH SEB. Performed the experiments: DSH. Analyzed the data: DSH SEB. Contributed reagents/materials/analysis tools: SEB. Wrote the paper: DSH SEB.

References

1. (2010) On beyond GWAS. Nat Genet 42: 551. doi: 10.1038/ng0710-551 PMID: 20581872

2. Goldstein DB (2009) Common genetic variation and human traits. N Engl J Med 360: 1696–1698. doi: 10.1056/NEJMp0806284 PMID: 19369660

3. Hirschhorn JN (2009) Genomewide association studies—illuminating biologic pathways. N Engl J Med 360: 1699–1701. doi: 10.1056/NEJMp0808934 PMID: 19369661

4. Kraft P, Hunter DJ (2009) Genetic risk prediction—are we there yet? N Engl J Med 360: 1701–1703. doi: 10.1056/NEJMp0810107 PMID: 19369656

5. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42: D1001–1006. doi: 10.1093/nar/gkt1229 PMID: 24316577

6. Wade N (2010) A decade later, genetic map yields few new cures. The New York Times New York.

7. Wang K, Li M, Hakonarson H (2010) Analysing biological pathways in genome-wide association studies. Nat Rev Genet 11: 843–854. doi: 10.1038/nrg2884 PMID: 21085203

8. Yaspan BL, Bush WS, Torstenson ES, Ma D, Pericak-Vance MA, Ritchie MD, et al. (2011) Genetic analysis of biological pathway data through genomic randomization. Hum Genet 129: 563–571. doi: 10.1007/s00439-011-0956-2 PMID: 21279722

9. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575. doi: 10.1086/519795 PMID: 17701901

10. Holmans P, Green EK, Pahwa JS, Ferreira MA, Purcell SM, Sklar P, et al. (2009) Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. Am J Hum Genet 85: 13–24. doi: 10.1016/j.ajhg.2009.05.011 PMID: 19539887

11. Segre AV, Consortium D, investigators M, Groop L, Mootha VK, Daly MJ, et al. (2010) Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. PLoS Genet 6. doi: 10.1371/journal.pgen.1001058 PMID: 20714348

12. Rossin EJ, Lage K, Raychaudhuri S, Xavier RJ, Tatar D, Benita Y, et al. (2011) Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS genetics 7: e1001273. doi: 10.1371/journal.pgen.1001273 PMID: 21249183

13. Tasan M, Musso G, Hao T, Vidal M, MacRae CA, Roth FP (2015) Selecting causal genes from genome-wide association studies via functionally coherent subnetworks. Nat Methods 12: 154–159. doi: 10.1038/nmeth.3215 PMID: 25532137

14. Jia P, Zheng S, Long J, Zheng W, Zhao Z (2011) dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. Bioinformatics 27: 95–102. doi: 10.1093/bioinformatics/btq615 PMID: 21045073

15. Consortium IMSG (2013) Network-based multiple sclerosis pathway analysis with GWAS data from 15,000 cases and 30,000 controls. American journal of human genetics 92: 854–865. doi: 10.1016/j.ajhg.2013.04.019 PMID: 23731539

16. Raychaudhuri S, Plenge RM, Rossin EJ, Ng AC, International Schizophrenia C, Purcell SM, et al. (2009) Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP

associations and rare deletions. PLoS Genet 5: e1000534. doi: 10.1371/journal.pgen.1000534 PMID: 19557189

17. Jungnickel D, SpringerLink (Online service) (2013) Graphs, networks, and algorithms. Algorithms and computation in mathematics,. 4. ed. Berlin, Heidelberg: Springer,. pp. 1 online resource (xx, 675 p.) ill. doi: 10.1007/978-3-642-32278-5

18. Lu LY, Zhou T (2011) Link prediction in complex networks: A survey. Physica a-Statistical Mechanics and Its Applications 390: 1150–1170. doi: 10.1016/j.physa.2010.11.027

19. Tong HH, Faloutsos C, Pan JY (2006) Fast random walk with restart and its applications. Icdm 2006: Sixth International Conference on Data Mining, Proceedings: 613–622. doi: 10.1109/ICDM.2006.70

20. Goncalves JP, Francisco AP, Moreau Y, Madeira SC (2012) Interactogeneous: Disease Gene Prioritization Using Heterogeneous Networks and Full Topology Scores. Plos One 7. doi: 10.1371/journal.pone.0049634 PMID: 23185389

21. Valentini G, Paccanaro A, Caniza H, Romero AE, Re M (2014) An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods. Artificial Intelligence in Medicine 61: 63–78. doi: 10.1016/j.artmed.2014.03.003 PMID: 24726035

22. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, et al. (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Research 38: W214–W220. doi: 10.1093/nar/gkq537 PMID: 20576703

23. Davis DA, Chawla NV (2011) Exploring and Exploiting Disease Interactions from Multi-Relational Gene and Phenotype Networks. Plos One 6. doi: 10.1371/journal.pone.0022670 PMID: 21829475

24. Davis D, Lichtenwalter R, Chawla NV (2012) Supervised methods for multi-relational link prediction. Social Network Analysis and Mining 3: 127–141. doi: 10.1007/s13278-012-0068-6

25. Guo XL, Gao L, Wei CS, Yang XF, Zhao Y, Dong AG (2011) A Computational Method Based on the Integration of Heterogeneous Networks for Predicting Disease-Gene Associations. Plos One 6. doi: 10.1371/journal.pone.0024171 PMID: 21912671

26. Wang W, Yang S, Li J (2013) Drug target predictions based on heterogeneous graph inference. Pac Symp Biocomput: 53–64. doi: 10.1142/9789814447973_0006 PMID: 23424111

27. Li Y, Li J (2012) Disease gene identification by random walk on multigraphs merging heterogeneous genomic and phenotype data. BMC Genomics 13 Suppl 7: S27. doi: 10.1186/1471-2164-13-S7-S27 PMID: 23282070

28. Li Y, Patra JC (2010) Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. Bioinformatics 26: 1219–1224. doi: 10.1093/bioinformatics/btq108 PMID: 20215462

29. Radivojac P, Peng K, Clark WT, Peters BJ, Mohan A, Boyle SM, et al. (2008) An integrated approach to inferring gene-disease associations in humans. Proteins 72: 1030–1037. doi: 10.1002/prot.21989 PMID: 18300252

30. Gligorijevic V, Janjic V, Przulj N (2014) Integration of molecular network data reconstructs Gene Ontology. Bioinformatics 30: i594–600. doi: 10.1093/bioinformatics/btu470 PMID: 25161252

31. Zitnik M, Janjic V, Larminie C, Zupan B, Przulj N (2013) Discovering disease-disease associations by fusing systems-level molecular data. Sci Rep 3: 3202. doi: 10.1038/srep03202 PMID: 24232732

32. Zitnik M, Zupan B (2014) Matrix factorization-based data fusion for gene function prediction in baker's yeast and slime mold. Pac Symp Biocomput: 400–411. doi: 10.1142/9789814583220_0038 PMID: 24297565

33. Zitnik M, Zupan B (2015) Data Fusion by Matrix Factorization. IEEE Transactions on Pattern Analysis and Machine Intelligence 37: 41–53. doi: 10.1109/TPAMI.2014.2343973

34. Sun Y, Barber R, Gupta M, Aggarwal CC, Han J (2011) Co-author Relationship Prediction in Heterogeneous Bibliographic Networks. 121–128. doi: 10.1109/ASONAM.2011.112

35. Sun Y, Han J (2012) Mining Heterogeneous Information Networks: Principles and Methodologies. Synthesis Lectures on Data Mining and Knowledge Discovery 3: 1–159. doi: 10.2200/S00433ED1V01Y201207DMK005

36. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP (2011) Molecular signatures database (MSigDB) 3.0. Bioinformatics 27: 1739–1740. doi: 10.1093/bioinformatics/btr260 PMID: 21546393

37. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102: 15545–15550. doi: 10.1073/pnas.0506580102 PMID: 16199517

38. Kanehisa M (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 28: 27–30. doi: 10.1093/nar/28.1.27 PMID: 10592173

39. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, et al. (2009) Reactome knowledge-base of human biological pathways and processes. Nucleic Acids Res 37: D619–622. doi: 10.1093/nar/gkn863 PMID: 18981052

40. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, et al. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 34: D108–110. doi: 10.1093/nar/gkj143 PMID: 16381825

41. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, et al. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature 434: 338–345. doi: 10.1038/nature03441 PMID: 15735639

42. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29. doi: 10.1038/75556 PMID: 10802651

43. Schriml LM, Arze C, Nadendla S, Chang YW, Mazaitis M, Felix V, et al. (2012) Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res 40: D940–946. doi: 10.1093/nar/gkr972 PMID: 22080554

44. Gray KA, Daugherty LC, Gordon SM, Seal RL, Wright MW, Bruford EA (2013) Genenames.org: the HGNC resources in 2013. Nucleic Acids Res 41: D545–552. doi: 10.1093/nar/gks1066 PMID: 23161694

45. Gremse M, Chang A, Schomburg I, Grote A, Scheer M, Ebeling C, et al. (2011) The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. Nucleic Acids Res 39: D507–513. doi: 10.1093/nar/gkq968 PMID: 21030441

46. BioCarta.

47. Brentani H, Caballero OL, Camargo AA, da Silva AM, da Silva WA Jr., Dias Neto E, et al. (2003) The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. Proc Natl Acad Sci U S A 100: 13418–13423. doi: 10.1073/pnas.1233632100 PMID: 14593198

48. Segal E, Friedman N, Koller D, Regev A (2004) A module map showing conditional activity of expression modules in cancer. Nat Genet 36: 1090–1098. doi: 10.1038/ng1434 PMID: 15448693

49. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, et al. (2009) NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res 37: D885–890. doi: 10.1093/nar/gkn764 PMID: 18940857

50. Fleuren WW, Verhoeven S, Frijters R, Heupers B, Polman J, van Schaik R, et al. (2011) CoPub update: CoPub 5.0 a text mining system to answer biological questions. Nucleic Acids Res 39: W450–454. doi: 10.1093/nar/gkr310 PMID: 21622961

51. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 101: 6062–6067. doi: 10.1073/pnas.0400782101 PMID: 15075390

52. Razick S, Magklaras G, Donaldson IM (2008) iRefIndex: a consolidated protein interaction database with provenance. BMC Bioinformatics 9: 405. doi: 10.1186/1471-2105-9-405 PMID: 18823568

53. Sun Y, Han J, Yan X, PS Y. PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks; 2011. pp. 992–1003.

54. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67: 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

55. Gillis J, Pavlidis P (2011) The impact of multifunctional genes on "guilt by association" analysis. PLoS One 6: e17258. doi: 10.1371/journal.pone.0017258 PMID: 21364756

56. Chiorazzi N, Rai KR, Ferrarini M (2005) Chronic lymphocytic leukemia. N Engl J Med 352: 804–815. doi: 10.1056/NEJMra041720 PMID: 15728813

57. Consortium IMSG, Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, et al. (2011) Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 476: 214–219. doi: 10.1038/nature10251 PMID: 21833088

58. Patsopoulos NA, Esposito F, Reischl J, Lehr S, Bauer D, Heubach J, et al. (2011) Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. Annals of neurology 70: 897–912. doi: 10.1002/ana.22609 PMID: 22190364

59. Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, et al. (2010) A versatile gene-based test for genome-wide association studies. American journal of human genetics 87: 139–145. doi: 10.1016/j.ajhg.2010.06.009 PMID: 20598278

60. Conti L, De Palma R, Rolla S, Boselli D, Rodolico G, Kaur S, et al. (2012) Th17 cells in multiple sclerosis express higher levels of JAK2, which increases their surface expression of IFN-gammaR2. J Immunol 188: 1011–1018. doi: 10.4049/jimmunol.1004013 PMID: 22219326

61. Dubois PC, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, et al. (2010) Multiple common variants for celiac disease influencing immune gene expression. Nat Genet 42: 295–302. doi: 10.1038/ng.543 PMID: 20190752

62. Evans DM, Spencer CC, Pointon JJ, Su Z, Harvey D, Kochan G, et al. (2011) Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. Nat Genet 43: 761–767. doi: 10.1038/ng.873 PMID: 21743469

63. Jeffries MA, Dozmorov M, Tang Y, Merrill JT, Wren JD, Sawalha AH (2011) Genome-wide DNA methylation patterns in CD4+ T cells from patients with systemic lupus erythematosus. Epigenetics 6: 593–601. doi: 10.4161/epi.6.5.15374 PMID: 21436623

64. Beecham AH, Patsopoulos NA, Xifara DK, Davis MF, Kemppinen A, Cotsapas C, et al. (2013) Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nature genetics 45: 1353–1360. doi: 10.1038/ng.2770 PMID: 24076602

65. Hangauer MJ, Vaughn IW, McManus MT (2013) Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS Genet 9: e1003569. doi: 10.1371/journal.pgen.1003569 PMID: 23818866

66. Gilmore TD, Kalaitzidis D, Liang MC, Starczynowski DT (2004) The c-Rel transcription factor and B-cell proliferation: a deal with the devil. Oncogene 23: 2275–2286. doi: 10.1038/sj.onc.1207410 PMID: 14755244

67. Hilliard BA, Mason N, Xu L, Sun J, Lamhamedi-Cherradi SE, Liou HC, et al. (2002) Critical roles of c-Rel in autoimmune inflammation and helper T cell differentiation. J Clin Invest 110: 843–850. doi: 10.1172/JCI15254 PMID: 12235116

68. Lage K, Hansen NT, Karlberg EO, Eklund AC, Roque FS, Donahoe PK, et al. (2008) A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. Proc Natl Acad Sci U S A 105: 20870–20875. doi: 10.1073/pnas.0810772105 PMID: 19104045

69. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL (2007) The human disease network. Proc Natl Acad Sci U S A 104: 8685–8690. doi: 10.1073/pnas.0701361104 PMID: 17502601

70. van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JA (2006) A text-mining analysis of the human phenome. Eur J Hum Genet 14: 535–542. doi: 10.1038/sj.ejhg.5201585 PMID: 16493445

71. Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, Manolio T, et al. (2011) Abundant pleiotropy in human complex diseases and traits. Am J Hum Genet 89: 607–618. doi: 10.1016/j.ajhg.2011.10.004 PMID: 22077970

72. Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, Wallace C, et al. (2011) Pervasive sharing of genetic effects in autoimmune disease. PLoS genetics 7: e1002254. doi: 10.1371/journal.pgen.1002254 PMID: 21852963

73. Stephens M, Balding DJ (2009) Bayesian statistical methods for genetic association studies. Nat Rev Genet 10: 681–690. doi: 10.1038/nrg2615 PMID: 19763151

74. Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, et al. (2009) An empirical framework for binary interactome mapping. Nat Methods 6: 83–90. doi: 10.1038/nmeth.1280 PMID: 19060904

75. Gillis J, Ballouz S, Pavlidis P (2014) Bias tradeoffs in the creation and analysis of protein-protein interaction networks. J Proteomics 100: 44–54. doi: 10.1016/j.jprot.2014.01.020 PMID: 24480284

76. Seco N, Veale T, Hayes J. An intrinsic information content metric for semantic similarity in WordNet; 2001. pp. 1089.

77. Hidalgo CA, Blumm N, Barabasi AL, Christakis NA (2009) A dynamic network approach for the study of human phenotypes. PLoS Comput Biol 5: e1000353. doi: 10.1371/journal.pcbi.1000353 PMID: 19360091

78. Sawcer S (2008) The complex genetics of multiple sclerosis: pitfalls and prospects. Brain 131: 3118–3131. doi: 10.1093/brain/awn081 PMID: 18490360

79. Stojmirovic A, Yu YK (2011) ppiTrim: constructing non-redundant and up-to-date interactomes. Database (Oxford) 2011: bar036. doi: 10.1093/database/bar036 PMID: 21873645

80. Friedman J, Hastie T, Tibshirani R (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 33: 1–22. PMID: 20808728

81. Schielzeth H (2010) Simple means to improve the interpretability of regression coefficients. Methods in Ecology and Evolution 1: 103–113. doi: 10.1111/j.2041-210X.2010.00012.x

82. A Ramachandra R, Rabindranath S (1996) A Markov Chain Monte Carlo Method for Generating Random (0, 1)-Matrices with Given Marginals. Sankhya Indian J Stat Ser A 58: 225–242.

83. Swamidass SJ, Azencott CA, Daily K, Baldi P (2010) A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval. Bioinformatics 26: 1348–1356. doi: 10.1093/bioinformatics/btq140 PMID: 20378557

84. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44: 837–845. doi: 10.2307/2531595 PMID: 3203132

85. Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, et al. (2004) Gene map of the extended human MHC. Nat Rev Genet 5: 889–899. doi: 10.1038/nrg1489 PMID: 15573121