Heterogeneous Network Link Prediction Prioritizes Disease-Associated Genes

~~ dhimmel.com ~~

Daniel Himmelstein[1] and Sergio Baranzini[1,2]

~~ baranzinilab.ucsf.edu ~~

1) Program in Biological & Medical Informatics 2) Department of Neurology

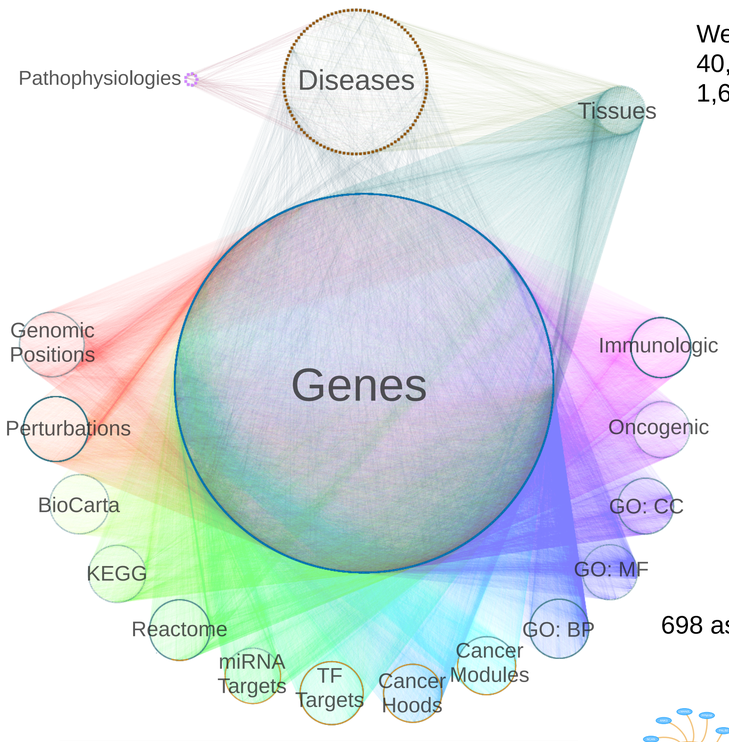
University of California, San Francisco

Abstract

We developed a method to predict the probability that each protein coding gene is associated with each of 29 complex human diseases. Starting with a heterogeneous network (consisting of multiple node and edge types), our method *integrates diverse information sources* and *learns the mechanisms underlying pathogenesis* to make accurate and novel predictions.

Methods

Constructing the heterogeneous network

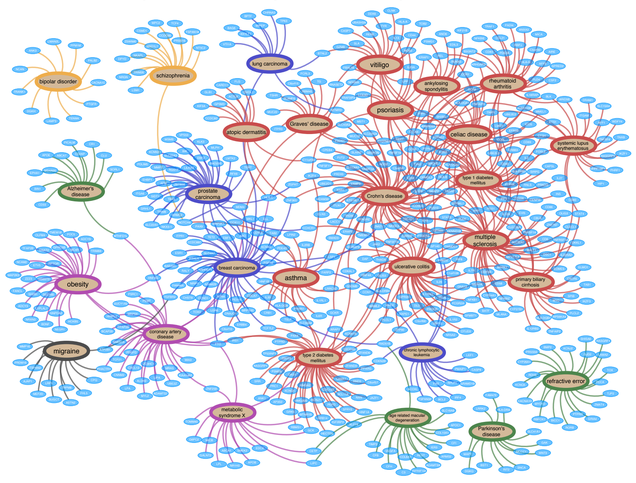


We constructed a heterogeneous network with 40,343 nodes of 18 types (metanodes) and 1,608,168 edges of 19 types (metaedges).

MetaNode	Count	Source
Disease	99	Disease Ontology
Gene	19,116	HGNC (coding)
Tissue	77	BRENDA (BTO)
Pathophysiology	8	manual
Positional	326	MSigDB (C1)
Perturbation	3,402	MSigDB (C2)
BioCarta	217	MSigDB (C2)
KEGG	186	MSigDB (C2)
Reactome	674	MSigDB (C2)
miRNA Target	221	MSigDB (C3)
TF Target	615	MSigDB (C3)
Cancer Hood	427	MSigDB (C4)
Cancer Module	431	MSigDB (C4)
GO Process	825	MSigDB (C5)
GO Component	233	MSigDB (C5)
GO Function	396	MSigDB (C5)
Oncogenic	189	MSigDB (C6)
Immunologic	1,910	MSigDB (C7)

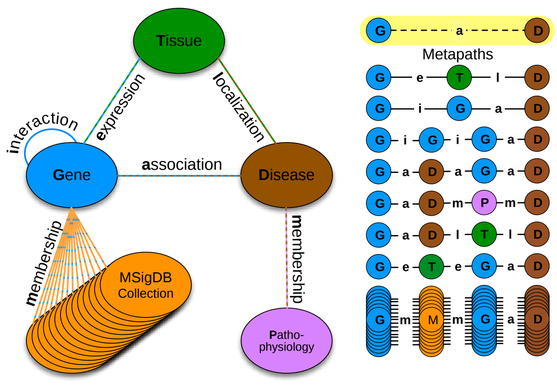
MetaEdge	Count	Source
Disease - association - Gene	938	GWAS Catalog
Disease - membership - Pathophysiology	90	manual
Disease - localization - Tissue	1,086	CoPub 5.0
Gene - expression - Tissue	251,366	GNF BodyMap
Gene - interaction - Gene	97,938	iRefIndex
Gene - membership - Positional	18,343	MSigDB (C1)
Gene - membership - Perturbation	366,211	MSigDB (C2)
Gene - membership - BioCarta	4,456	MSigDB (C2)
Gene - membership - KEGG	12,656	MSigDB (C2)
Gene - membership - Reactome	35,597	MSigDB (C2)
Gene - membership - miRNA Target	33,455	MSigDB (C3)
Gene - membership - TF Target	161,258	MSigDB (C3)
Gene - membership - Cancer Hood	41,913	MSigDB (C4)
Gene - membership - Cancer Module	48,220	MSigDB (C4)
Gene - membership - GO Process	75,155	MSigDB (C5)
Gene - membership - GO Component	34,880	MSigDB (C5)
Gene - membership - GO Function	23,578	MSigDB (C5)
Gene - membership - Oncogenic	30,166	MSigDB (C6)
Gene - membership - Immunologic	370,862	MSigDB (C7)

698 associations extracted from the GWAS Catalog provided experimental positives

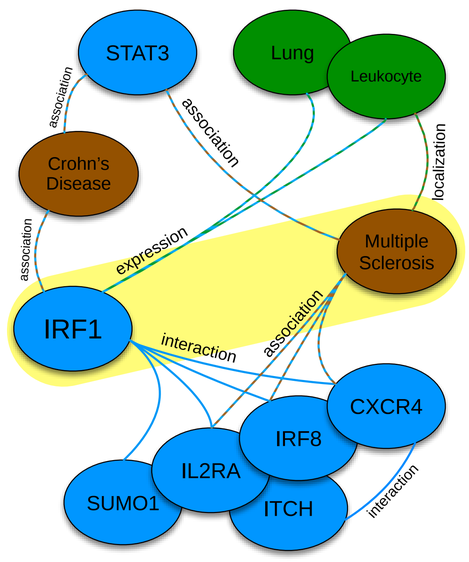


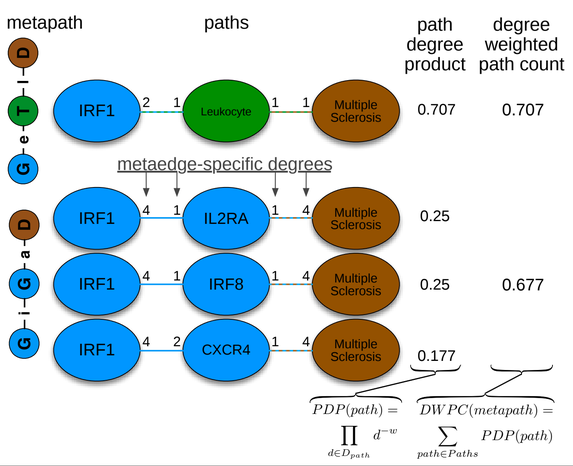
Computing features to quantify network topology

Network topology is decomposed based on metapaths (types of paths originating with a gene and terminating with a disease)[1].



Features quantify the prevalence of a specific metapath. Feature computation for an example subgraph is demonstrated below:

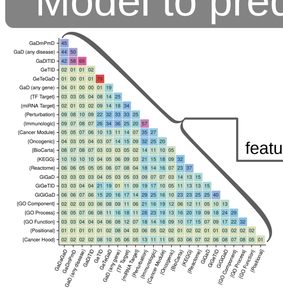




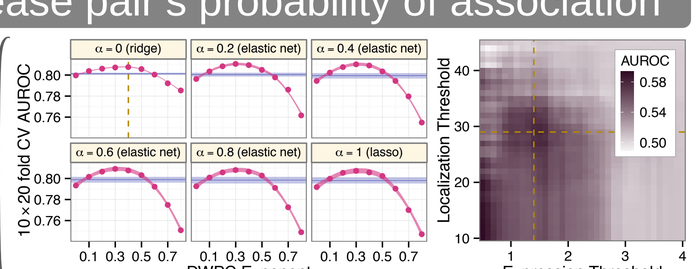
Each feature encodes a relationship with a corresponding biological interpretation:

Path Count	Measures the number of . . .
GaD (any disease)	diseases that the source gene is associated with, ignoring the association with the target disease if present.
GaD (any gene)	genes that the target disease is associated with, ignoring the association with the source gene if present.
DWPC	**Measures the extent that . . .**
GeTID	the source gene is expressed in tissues affected by target disease.
GiGaD	genes associated with the target disease interact with the source gene.
GiGiGaD	genes associated with the target disease interact with genes that interact with the source gene.
GaDaGaD	genes associated with the same diseases as the source gene are associated with the target disease.
GaDmPmD	diseases with the same pathophysiology as the target disease are associated with the source gene.
GaDlTID	diseases affecting the same tissues as the target disease are associated with the source gene.
GeTeGaD	genes expressed in the same tissues as the source gene are associated with the target disease.
GiGeTID	genes interacting with the source gene are expressed in tissues that are affected by the target disease.
{Positional}	genes located in the same cytogenetic band as the source gene are associated with the target disease.
{Perturbation}	genes belonging to the same perterbation signatures as the source gene are associated with the target disease.
{BioCarta}	genes involved in the same BioCarta pathways as the source gene are associated with the target disease.
{KEGG}	genes involved in the same KEGG pathways as the source gene are associated with the target disease.
{Reactome}	genes involved in the same Reactome pathways as the source gene are associated with the target disease.
{miRNA Target}	genes sharing 3'-UTR microRNA binding motifs with the source gene are associated with the target disease.
{TF Target}	genes sharing transcription factor binding sites with the source gene are associated with the target disease.
{Cancer Hood}	genes present in the same expression neighborhoods of cancer-related genes as the source gene are associated with the target disease.
{Cancer Module}	genes belonging to the same cancer modules as the source gene are associated with the target disease.
{GO Process}	genes participating in the same GO Biological Processes as the source gene are associated with the target disease.
{GO Component}	genes belonging to the same GO Cellular Components as the source gene are associated with the target disease.
{GO Function}	genes contributing to the same GO Molecular Functions as the source gene are associated with the target disease.
{Oncogenic}	genes belonging to the same cancer-dysregulated cellular pathways as the source gene are associated with the target disease.
{Immunologic}	genes belonging to the same immunologic signatures as the source gene are associated with the target disease.

Model to predict a gene-disease pair's probability of association



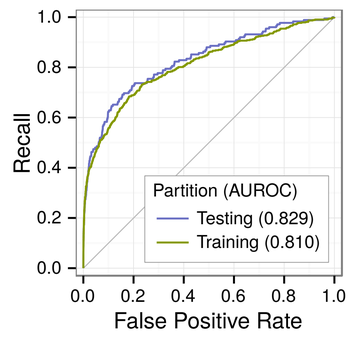
features → Regularized Logistic Regression → optimization

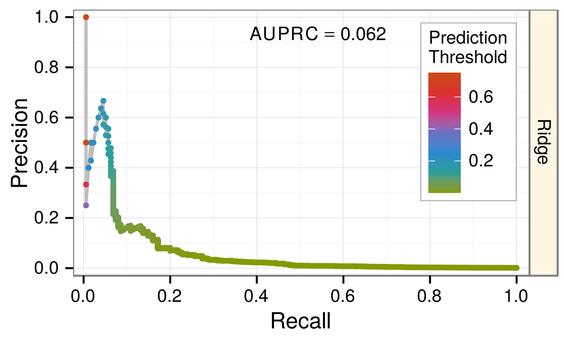


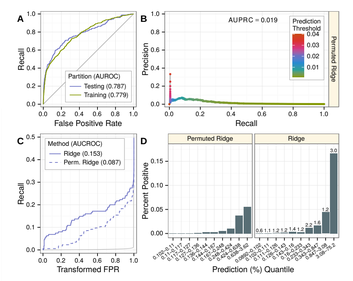
Results

Prioritizing associations withheld for testing

Withholding 30% of gene-disease pairs for testing, our predictions achieved an area under the ROC curve (AUROC) of 0.83 and a 132-fold enrichment in precision at 10% recall.







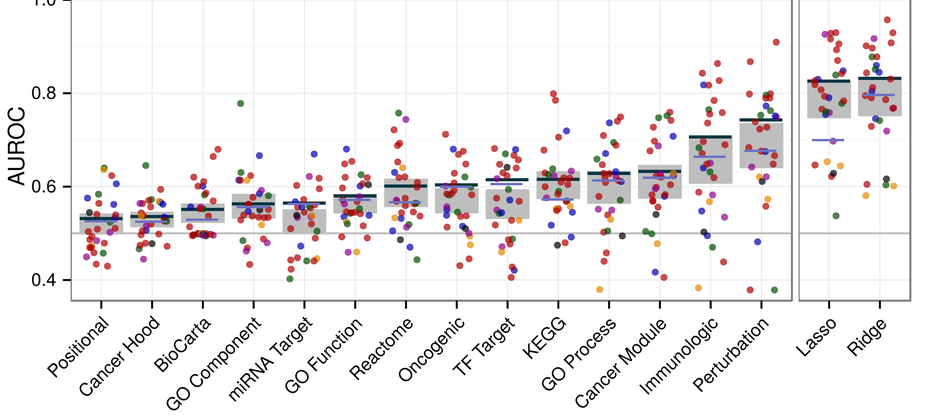
Permuted-network performance highlights that edge-specificity was crucial for top predictions.

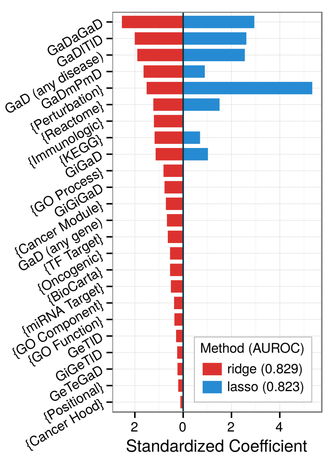
Identifying the mechanisms underlying pathogenesis

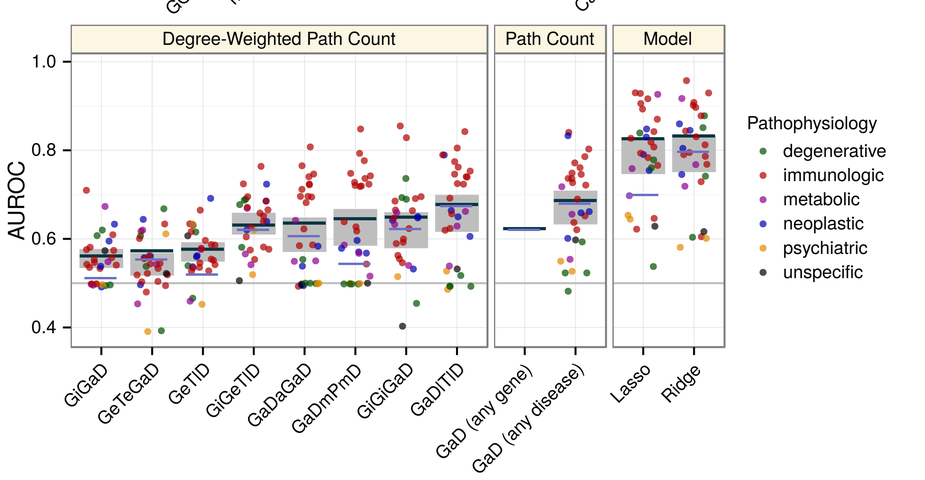
The integrative model outperformed any individual domain.

Influential mechanisms:
- pleiotropy
- **perturbation signatures**
- pathways
- protein interactions

Existing prioritization methods may be limited.

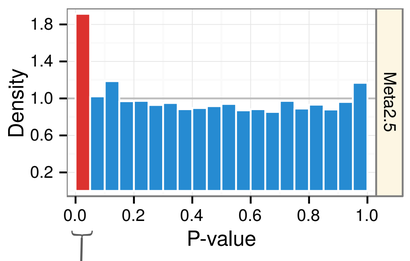




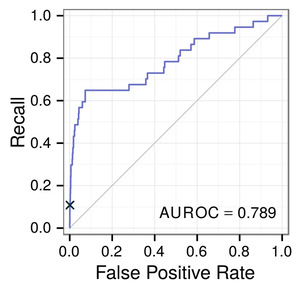


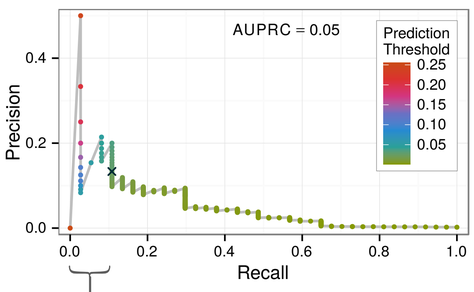
Case study: prioritizing multiple sclerosis associations

A meta-analysis[2] of all MS GWAS prior to WTCCC2[3] showed an enrichment of nominally significant ($p < 0.05$) genes.



We masked the WTCCC2 multiple sclerosis GWAS from our network reducing the number of MS-associated genes from 50 to 13. Despite the low number of seed genes, the 37 novel WTCCC2 genes were ranked highly with AUROC = 0.79.



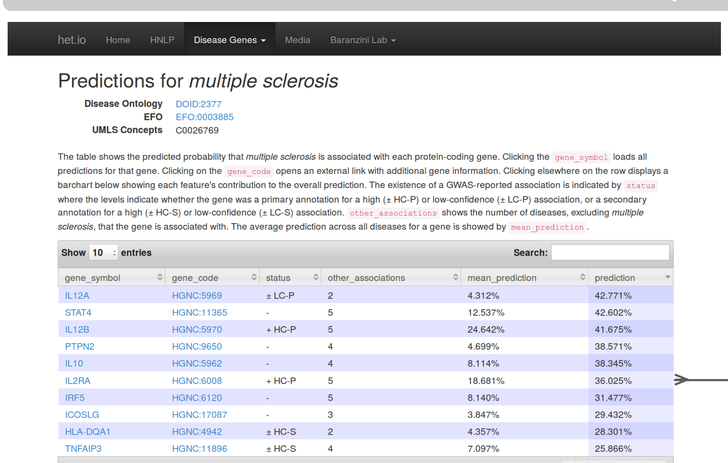


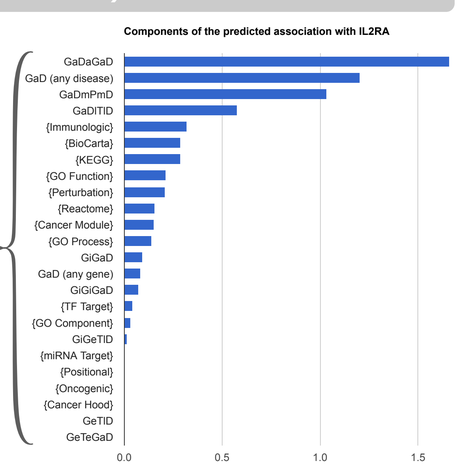
We overlapped this set of experimental candidates with the top network predictions. Four genes were discovered, three of which validated in WTCCC2. The probability of the observed validation rate occurring under random prioritization is 0.01.

Gene	Meta2.5	HNLP	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**

The gene-dense region containing *REL* was uncovered in a recent MS ImmunoChip-based study[4], which reported a long non-coding RNA for the loci.

Online Browser (http://het.io)





References

1) Sun et al (2011) Co-author relationship prediction in heterogeneous bibliographic networks. ASONAM. doi:10.1109/ASONAM.2011.112
2) Patsopoulos et al (2011) Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. Ann Neurol. doi:10.1002/ana.22609
3) Sawcer et al (2011) Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature. doi:10.1038/nature10251
4) Beecham et al (2013). Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat Genet. doi:10.1038/ng.2770

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1144247 to DSH. SEB is a Harry Weaver Neuroscience fellow from the National Multiple Sclerosis Society.