Characterization of Tissue-Specific Functional Networks and Genome-Wide Association Study Genes

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Jacquelyn Katrina Kuriger-Laber

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Chad L. Myers, Heather H. Nelson

January 2016

Acknowledgements

I would like to thank everyone who provided me support during my time in the BICB program. I am especially grateful for Dr. Chad Myers, whose belief in his students and unfailing kindness coupled with his wealth of knowledge and new ideas is always inspiring, and Dr. Heather Nelson whose unfailing advocacy and outstanding mentoring always encouraged me to do what is best for myself, even if it was directly opposed to what was best for her. I am also very grateful to Dr. Logan Spector who not only agreed to be on my committee but who also was kind enough to help me find the data I used for this project. I also would like to thank Dr. Rui Kuang, both for his patience as an instructor and for being willing to be on my committee.

I also owe a great deal of thanks to all the members of the Myers Lab, particularly big thank you to Dr. Ben VanderSluis, Rob Schafer, Wen Wang, Scott Simpkins, Jean-Michel Michno, Elizabeth Koch, and Joe Jeffers helping me go from feeling lost in front of a computer to feeling like I was ready and able to take on new challenges. Additionally, I want to offer my sincere appreciation to Gabe Al-Ghalith and also to the RISS staff at the UMN Supercomputer Institute for always being willing to help when I was feeling stuck.

And last but not least I would like to acknowledge all the support I have received from my family. I am especially grateful to my brother Bill, who was available for emergency programing questions at any and all hours, and my daughters Michelle and Madison for all the emotional support and encouragement they have given me over the last two years. And most of all, I cannot adequately thank my husband, Terry Laber, for taking care of everything and carrying so many of my responsibilities so that I could make this journey.

Dedication

To the women in my life

Jackie, Michelle, Madison, Tantie, Lacey, Isla, Emma

You inspire me

Abstract

Present-day biological research has generated a vast body of data related to variation in the human genome, but in many cases the biological role of this variation is unknown or only partially understood. In an effort to integrate the diverse body of experimental genetic and genomic data, the systems biology community pioneered computational approaches to infer gene functional networks. These networks provide a powerful platform to investigate genomic findings at a functional level. Recently, systems biologists designed a second generation of functional networks that reflect tissue-specificity in gene functional interactions. In order to develop network-enabled methods for interpreting tissue-specific roles of genetic variants identified by population-based genotyping studies, we examine both characteristics of these tissue-specific functional networks and the topology of genome-wide association study (GWAS) variant-related genes in these networks. We find significant variation in quality across a collection of networks commonly used by the community and suggest informative metrics that can be used to identify well-performing networks. Finally, we show that trait-associated genes from GWAS studies have non-random topology in the tissue-specific networks and that this must be taken into account when applying network-enabled methods to the interpretation of genomic data.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction and Background

1A. Project Motivation

Genetic variation is the suspected cause of many human diseases. Genomic research has yielded a vast amount of experimental data, but despite fully sequencing the human genome and subsequent years of in-depth genomic studies, there is still a great deal not understood about how genetic variation affects disease. A wide variety of experimental approaches have been used to investigate human disease, and the resulting data from each offers a small insight into the disease process. However, most of this data focuses on one process or pathway, and the complexity of this data creates difficultly in interpretation on a system-wide scale.

The systems biology community developed functional gene networks to provide a method for integrating this existing knowledge. By incorporating diverse data types into a single functional model, functional networks provide a framework that gives context to genetic discovery. They also work as a tool that can use existing knowledge to provide insight and lead to new observations about the role of genetic variation in disease.

1B. Genome-wide Association Studies (GWAS)

A genome-wide association study (GWAS) is a genotyping test designed to find genetic variation that is associated with a phenotype of interest, often a disease. If a genetic variant is found to be overrepresented in groups with that phenotype, that variant is said to be associated with the phenotype. GWAS is

usually performed on a large number of samples, and this allows detection of variations that have low penetrance or convey slightly increased risk. Testing is done by genotyping thousands or millions of single nucleotide polymorphisms (SNPs) that are chosen to provide genome-wide coverage. Results of GWAS are given for each SNP as odds ratios and p-values that report the probability the SNP is associated with the tested phenotype. Because of this large number of statistical tests, p-values must be adjusted to correct for multiple testing error.

Many of the SNPs tested are tagging SNPs. That means testing is designed to utilize linkage disequilibrium (LD) and tagging SNPs give genotype information not simply for that one nucleotide but also for other variants in the region that are known to be commonly linked. This design allows more efficient genomic coverage, so that in regions where many SNPs are present and LD occurs, genotyping a subset of these SNPs yields genotype information for the entire region.

Some genotyped SNPs correspond to protein-coding mutations and may be responsible for functional change. But given that SNPs genotyped in GWAS are often simply representative of that region of the genome, it is known that these SNPs often are not the mutation functionally responsible for the phenotype. Instead the associated SNP may be in LD with a causal mutation. SNPs functionally responsible for the phenotype association can be acting by one of several different mechanisms. Those that cause a protein-coding change might have a functional impact on the activity of the protein product, but others may

occur in a promoter region or other untranslated region of a gene, also, non-coding SNPs can occur in enhancer regions or other regulatory regions that affect expression of either local or distant genes. All of these factors often make it difficult to determine how associated SNPs are related to the phenotype.

1C. Functional gene networks

Functional gene networks summarize gene expression, protein-protein interaction, cell localization, sequence, and genetic interaction data into a functional model using Bayesian data integration, as shown in Figure 1C.1.

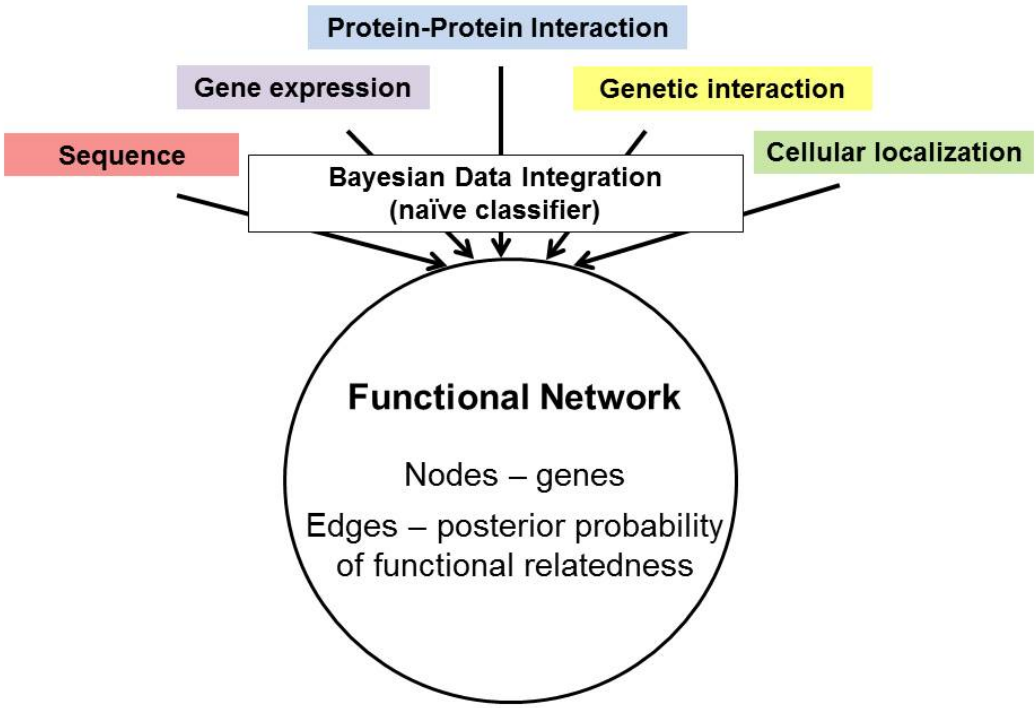


Figure 1C.1 Integrating diverse data into a functional network
Diverse data sets including but not limited to those in the cartoon are combined via Bayesian data integration into a functional network that both reduces the noise of individual data types and offers a system-wide view of genetic interaction.

This approach allows for a summarization of heterogeneous data types in the framework of biological context while also reducing the noise of individual data sets (Myers and Troyanskaya, 2007). However, creating functional gene networks for complex organisms and especially for humans presents unique challenges. One challenge is the assumption when using naïve classifiers that the data are independent. Creation of these complex organism functional networks requires the use of a large number of datasets; however with increasing numbers of datasets, this assumption of independence becomes more problematic. Huttenhower et al. (2009) used Bayesian regularization of the naïve classifier parameters to weight the data sets appropriately and overcome lack of independence allowing them to create human functional networks or maps. The resulting functional maps emphasize either gene-level information, biological processes, or disease, facilitating use tailored to the question of interest. HEFalMp, a web-based tool for exploration of genes or topics of interest in these human functional networks is available online at hefalmp.princeton.edu.

An additional challenge is added when recognizing that in complex organisms the functional relationship between genes will not be constant system-wide, but instead will vary between tissues. Recently, systems biologists used known tissue-specific functional interactions and tissue-specific annotations from the Human Protein Reference Database (Keshava Prasad, et al. 2009) and BRENDA Tissue Ontology (Gremse, 2011) to assess datasets for tissue-specific functional relevance. They weighted relevant data and integrated it to create the

first human tissue-specific functional networks (Greene et al. 2015), including 144 tissue or cell-type specific networks and a global (all tissues) network. GIANT, a network analysis tool for analyzing genes of interest in any or all of these tissues, is available online. Also, the tissue-specific networks are available for download. Both are located at giant.princeton.edu.

A functional gene network resulting from these approaches consists of nodes representing the genes and edges between genes representing their functional relationship. Edge weights in the network correspond to the probability that each gene pair is functionally related. Because of the complexity of a functional network, it is common to display or discuss only a small portion of that network that is of particular interest, i.e. a subnetwork, as seen in Figure 1C.2.

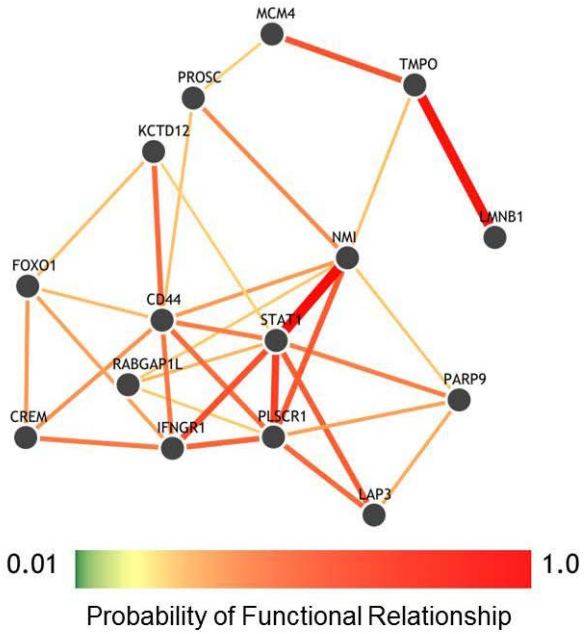


Figure 1C.2 Representation of a portion of a functional network
Each gene is an individual node. The probability of a functional relationship between two genes is represented by an edge whose weight is that probability. Edge weight is often represented by the width of the edge or by color; both representations are used here.

1D. Using functional gene networks to explore genomic data

Genes or gene sets of interest can be explored using a network analysis approach. One particularly interesting data type for network analysis is GWAS data. Although GWAS can yield a great deal of genomic information and has the power to detect weakly penetrating variants associated with a phenotype, the information in this form is of limited value, since variations can be identified as associated with the phenotype, but the information is lacking in context. Also, the large number of SNPs tested result in a high likelihood of finding random associations, yet when using multiple testing correction, it is clear that some real associations are certainly thrown out. Functional gene networks can be used to address these issues.

Exploring a gene or gene sets in functional networks can result in the discovery of subnetworks of interest that point toward a relevant pathway or additional candidate genes that may be involved. Functional network analysis of GWAS genes or other gene lists of interest can yield many types of additional information such as new candidate genes, members of a pathway or other functionally related genes. Additionally, network analysis overcomes the need for an arbitrary cutoff of a specified p-value, instead network analysis of GWAS findings can be used as an amplifier for functional coherence of GWAS results.

Several experiments have shown the power of functional network analysis of gene sets or genomic data. Huttenhower et al. (2009) used the HEFalMp human functional network to perform process specific network analysis of known

autophagy genes ATG7, BECN1, and MAP1LC3B and three test genes LAMP2, RAB11A, and VAMP7 in the context of autophagy. The result was two groups of genes, a group of known autophagy genes and a group of vesicular and transport genes. They performed experimental validation on six genes from that result, AP3B1, ATP6AP1, BLOC1S1, LAMP2, VAMP7 and RAB11A, and found experimental evidence supporting a role in autophagy for five of the six genes.

Another example of the power of functional network analysis shows its use in predicting gene expression. Greene et al. (2015) used functional network analysis of the human blood vessel network to predict gene expression connected to IL-1β. Of the twenty genes predicted to be most functionally related to IL-1β, eighteen were confirmed experimentally to have upregulated expression. This same group also showed that network exploration of GWAS hypertension genes showed better performance of hypertension-associated genes in the relevant heart, kidney, and liver networks than in the global network (Greene et al. 2015).

In another method, a group of cell-type specific networks were created by integrating cell-type specific gene expression data and protein-protein interaction data. Analysis of gene-gene relationships of disease-associated genes in these networks showed they could map diseases to the cell-type affected (Cornish et al. 2015). In summary, using varied networks created under different protocols, researchers have shown that network analysis of genomic data is a flexible and powerful tool.

1E. Overview

In Chapter 2 we detail our use of network analysis to investigate genes associated with acute lymphoblastic leukemia (ALL) in some of the tissue-specific networks generated by Greene et al. (2015). In Chapter 3 we will demonstrate the how the topology of GWAS trait-associated genes varies from the overall topology of genes in tissue-specific networks, and discuss the implications of our findings. These findings lead us to examine the characteristics of these networks in Chapter 4, where we detail variation in quality of the tissue-specific networks and demonstrate metrics that can be used to determine network performance.

Chapter 2: Acute Lymphoblastic Leukemia (ALL) GWAS genes in tissue-specific networks

2A. Examination of ARID5B using the web-based GIANT tool

ARID5B is a gene that is well-documented as having a strong association with ALL in GWAS (Papaemmanuil et al. 2009, Yang et al. 2009, Xu et al. 2013). As a preliminary investigation of ALL genes in tissue-specific networks, we examined ARID5B using the GIANT online tool. We explored functional relationships of ARID5B in the cell type of interest (hematopoietic stem cell), a more mature cell type (lymphocyte), an unrelated tissue (brain), and in the global network. As seen in Figure 2A, functional connections in the cell type of interest are very different, being more numerous and having stronger edge weights than connections in mature or unrelated cell types.

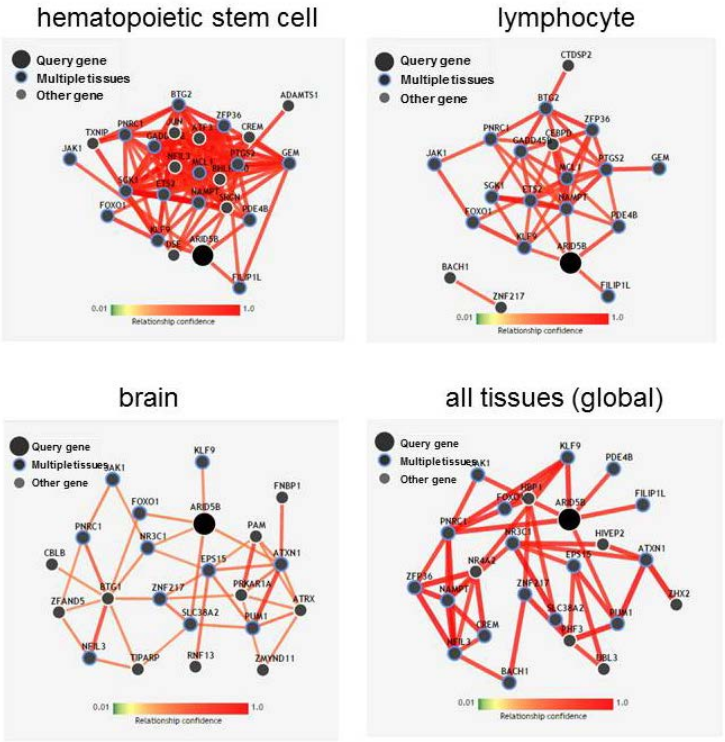


Figure 2A Tissue-specific network analysis of ARID5B
Functional connections in the ARID5B subnetwork of the hematopoietic stem cell network have higher probabilities and are more numerous than those found in the lymphocyte, brain or all tissues network.

2B. Memphis ALL GWAS data

We searched for publically available data for ALL GWAS, with the intent of doing tissue-specific network analysis of GWAS p-values. Our goal was to find a complete set of unadjusted p-values, however publications only listed the top associated genes and we did not find a publically available data set. We contacted Dr. Jun J. Yang at St Jude Children's Research Hospital in Memphis, TN who provided unpublished data containing the 680,000 unadjusted p-values for the GWAS of childhood ALL corresponding to his 2013 publication (Xu et al. 2013).

We used the bioDBnet Database to Database Conversion tool (Mudunuri et

al. 2009) available at http://biodbnet.abcc.ncifcrf.gov to convert SNP rsID numbers to genes. We discarded all intergenic SNPs from the data and collapsed multiple SNPs associated with a single gene retaining the most significant p-value. Although some intragenic SNPs are likely not to be causal variants, we did not attempt to impute genotypes or identify causal variants.

To prepare the ALL associated p-values for network analysis, we calculated transformed scores equal to the square root of the negative log of the p-values. We used these scores and associated genes for network analysis.

2C. Filtering networks

The tissue-specific networks developed by Greene et al. (2015) contain probability weighted edges between each of 25,825 human genes; therefore each network has more than 330 million edges. Due to the computational challenges of working with such a large number of edges, and because most of these edges represent a very small probability of a functional relationship, we sorted the networks by decreasing edge weight and filtered them by retaining only a specified number of the top weighted edges. These filtered networks are designated by the number of edges remaining in the network (i.e. a 250K network contains the top 250,000 edges). For the majority of the analyses discussed in this paper, 250K networks were used. Additionally 50K, 100K, 1M, and 10M filtered networks were generated, and when they are used it will be specifically noted. Table 2C lists the minimum edge weight remaining in all of these filtered networks.

Table 2C Minimum edge weights in filtered tissue-specific networks

tissue network name	50K	100K	250K	500K	1M	10M
adipose tissue	0.742	0.660	0.538	0.446	0.363	0.161
adrenal cortex	0.674	0.580	0.457	0.372	0.296	0.134
adrenal gland	0.383	0.338	0.284	0.248	0.217	0.137
all tissues	0.879	0.798	0.646	0.512	0.385	0.125
amygdala	0.402	0.358	0.302	0.262	0.224	0.128
aorta	0.812	0.758	0.677	0.609	0.536	0.288
artery	0.614	0.528	0.413	0.333	0.264	0.127
astrocyte	0.954	0.937	0.907	0.875	0.831	0.557
b lymphocyte	0.713	0.605	0.448	0.340	0.250	0.110
basal ganglion	0.353	0.315	0.268	0.235	0.204	0.125
basophil	0.994	0.991	0.983	0.973	0.957	0.778
blood plasma	0.553	0.473	0.368	0.296	0.235	0.118
blood platelet	0.698	0.589	0.435	0.327	0.240	0.109
blood	0.717	0.620	0.476	0.369	0.276	0.111
blood vessel	0.658	0.556	0.415	0.320	0.241	0.110
bone marrow	0.541	0.448	0.333	0.260	0.201	0.106
bone	0.745	0.636	0.476	0.362	0.267	0.113
brain	0.504	0.434	0.341	0.277	0.222	0.114
bronchial epithelial cell	0.904	0.849	0.745	0.644	0.530	0.187
bronchus	0.863	0.794	0.672	0.564	0.451	0.159
cardiac muscle	0.827	0.753	0.628	0.521	0.412	0.150
cartilage	0.674	0.593	0.482	0.401	0.327	0.154
caudate nucleus	0.417	0.369	0.308	0.266	0.226	0.127
caudate putamen	0.564	0.521	0.463	0.418	0.373	0.228
cecum	0.909	0.878	0.826	0.775	0.712	0.422
central nervous system	0.513	0.441	0.345	0.279	0.222	0.114
cerebellar cortex	0.820	0.761	0.666	0.583	0.492	0.203
cerebellum	0.329	0.288	0.238	0.206	0.177	0.113
cerebral cortex	0.340	0.301	0.253	0.220	0.190	0.119
chondrocyte	0.823	0.762	0.666	0.582	0.491	0.196
choroid	0.981	0.962	0.909	0.834	0.725	0.282
cochlea	0.99996	0.9999	0.9998	0.9997	0.999	0.989
colon	0.442	0.367	0.279	0.226	0.183	0.105
cornea	0.814	0.716	0.553	0.423	0.309	0.118
corpus callosum	0.456	0.401	0.332	0.283	0.237	0.124
corpus luteum	0.931	0.887	0.800	0.710	0.602	0.216
corpus striatum	0.378	0.337	0.284	0.247	0.213	0.126
culture condition cd8 cell	0.973	0.960	0.935	0.905	0.862	0.567
dendritic cell	0.842	0.803	0.743	0.689	0.628	0.377
dentate gyrus	0.960	0.924	0.836	0.730	0.595	0.177
diencephalon	0.410	0.371	0.321	0.284	0.248	0.149
duodenum	0.802	0.708	0.564	0.456	0.359	0.155
ear	0.980	0.972	0.956	0.937	0.909	0.683
embryo	0.496	0.424	0.334	0.274	0.222	0.115
eosinophil	0.842	0.774	0.654	0.545	0.431	0.149
epidermis	0.523	0.436	0.333	0.268	0.217	0.116
esophagus	0.606	0.551	0.475	0.417	0.360	0.184
eye	0.402	0.344	0.276	0.232	0.195	0.119
fetus	0.489	0.408	0.310	0.248	0.196	0.106

Table 2C (Con't.)

tissue network name	50K	100K	250K	500K	1M	10M
forebrain	0.338	0.300	0.254	0.222	0.192	0.121
frontal lobe	0.490	0.448	0.394	0.353	0.313	0.191
gastrointestinal tract	0.442	0.367	0.280	0.226	0.183	0.105
glia	0.904	0.875	0.827	0.780	0.722	0.434
granulocyte	0.722	0.612	0.453	0.342	0.251	0.108
hair follicle	0.724	0.626	0.487	0.386	0.297	0.125
heart	0.449	0.378	0.293	0.238	0.191	0.105
hematopoietic stem cell	0.709	0.597	0.436	0.325	0.235	0.106
hepatocyte	0.959	0.935	0.885	0.827	0.748	0.357
hippocampus	0.388	0.344	0.289	0.249	0.214	0.125
hypophysis	0.516	0.463	0.394	0.343	0.295	0.165
hypothalamus	0.838	0.797	0.735	0.678	0.615	0.365
ileum	0.930	0.905	0.859	0.813	0.754	0.454
intestine	0.445	0.369	0.280	0.225	0.181	0.105
jejunum	0.955	0.940	0.913	0.883	0.842	0.573
keratinocyte	0.564	0.484	0.388	0.327	0.274	0.142
kidney	0.475	0.396	0.301	0.240	0.189	0.105
large intestine	0.422	0.354	0.274	0.225	0.185	0.107
lens	0.989	0.983	0.969	0.951	0.924	0.659
leukocyte	0.751	0.650	0.493	0.376	0.275	0.107
liver	0.459	0.382	0.290	0.232	0.185	0.104
locus ceruleus	0.999	0.998	0.996	0.993	0.987	0.875
lung	0.504	0.421	0.320	0.254	0.199	0.104
lymph node	0.468	0.406	0.326	0.271	0.224	0.119
lymphocyte	0.690	0.592	0.450	0.348	0.261	0.109
macrophage	0.684	0.583	0.443	0.344	0.259	0.111
mammary epithelium	0.809	0.748	0.650	0.565	0.473	0.195
mammary gland	0.567	0.455	0.325	0.248	0.191	0.108
mast cell	0.843	0.789	0.700	0.620	0.531	0.239
medulla oblongata	0.724	0.679	0.614	0.559	0.500	0.284
megakaryocyte	0.662	0.556	0.412	0.313	0.233	0.109
midbrain	0.396	0.354	0.300	0.261	0.225	0.128
monocyte	0.722	0.612	0.450	0.338	0.247	0.109
mononuclear phagocyte	0.715	0.604	0.443	0.332	0.243	0.109
muscle	0.413	0.348	0.273	0.225	0.184	0.107
myometrium	0.864	0.806	0.705	0.612	0.508	0.195
natural killer cell	0.898	0.859	0.796	0.738	0.670	0.388
nephron	0.521	0.445	0.350	0.285	0.230	0.118
nervous system	0.525	0.452	0.355	0.288	0.230	0.115
neuron	0.758	0.697	0.606	0.531	0.453	0.212
neutrophil	0.701	0.590	0.436	0.332	0.246	0.108
nucleus accumbens	0.893	0.846	0.759	0.675	0.579	0.257
occipital lobe	0.600	0.548	0.475	0.418	0.361	0.190
occipital pole	0.953	0.935	0.901	0.863	0.810	0.483
osteoblast	0.886	0.811	0.668	0.538	0.407	0.130
ovarian follicle	0.886	0.827	0.721	0.623	0.514	0.186
ovary	0.512	0.433	0.333	0.266	0.211	0.107
oviduct	0.713	0.633	0.518	0.428	0.343	0.145

Table 2C (Con't)

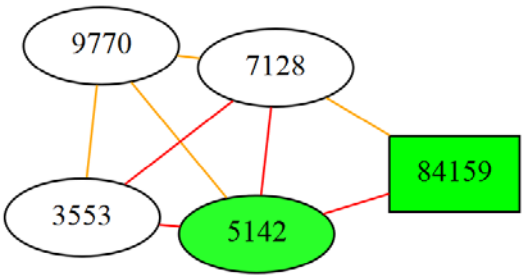
tissue network name	50K	100K	250K	500K	1M	10M
pancreas	0.414	0.352	0.278	0.230	0.189	0.106
pancreatic islet	0.777	0.698	0.583	0.492	0.403	0.170
parietal lobe	0.997	0.996	0.993	0.989	0.981	0.871
peripheral nervous system	0.882	0.815	0.694	0.583	0.464	0.165
placenta	0.555	0.452	0.331	0.256	0.196	0.105
podocyte	0.977	0.967	0.946	0.922	0.887	0.623
pons	0.992	0.988	0.980	0.969	0.951	0.762
prostate gland	0.435	0.366	0.285	0.234	0.191	0.108
renal glomerulus	0.607	0.525	0.416	0.341	0.275	0.136
renal tubule	0.715	0.655	0.565	0.492	0.414	0.182
retina	0.409	0.369	0.320	0.286	0.254	0.162
salivary gland	0.513	0.434	0.339	0.279	0.229	0.123
serum	0.441	0.384	0.314	0.267	0.227	0.133
skeletal muscle	0.398	0.344	0.279	0.234	0.195	0.107
skin fibroblast	0.641	0.563	0.457	0.379	0.308	0.138
skin	0.500	0.417	0.317	0.253	0.202	0.110
small intestine	0.379	0.316	0.246	0.203	0.168	0.106
smooth muscle	0.676	0.596	0.482	0.396	0.315	0.131
spermatid	0.886	0.833	0.736	0.643	0.536	0.200
spermatocyte	0.903	0.864	0.796	0.728	0.645	0.292
spermatogonium	0.903	0.865	0.797	0.729	0.646	0.292
spinal cord	0.325	0.287	0.242	0.210	0.182	0.113
spleen	0.493	0.426	0.339	0.278	0.225	0.108
stomach	0.385	0.338	0.281	0.243	0.208	0.125
substantia nigra	0.396	0.353	0.298	0.260	0.224	0.130
subthalamic nucleus	0.491	0.428	0.348	0.292	0.242	0.122
t lymphocyte	0.638	0.541	0.406	0.314	0.237	0.110
tear gland	0.604	0.535	0.444	0.379	0.320	0.167
telencephalon	0.338	0.298	0.250	0.217	0.187	0.118
temporal lobe	0.363	0.324	0.275	0.240	0.207	0.125
testis	0.514	0.425	0.319	0.252	0.198	0.105
thalamus	0.450	0.401	0.341	0.297	0.256	0.145
thymocyte	0.793	0.711	0.582	0.478	0.378	0.137
thyroid gland	0.447	0.382	0.305	0.255	0.212	0.121
tonsil	0.709	0.651	0.569	0.504	0.438	0.232
tooth	0.838	0.762	0.631	0.520	0.407	0.148
trachea	0.600	0.536	0.450	0.386	0.324	0.156
trophoblast	0.900	0.814	0.645	0.499	0.368	0.129
umbilical cord	0.653	0.558	0.427	0.337	0.259	0.113
umbilical vein endothelial cell	0.779	0.681	0.526	0.411	0.309	0.119
urinary bladder	0.893	0.859	0.802	0.748	0.684	0.408
uroepithelium	0.869	0.780	0.624	0.495	0.376	0.148
uterine cervix	0.658	0.566	0.446	0.362	0.289	0.133
uterine endometrium	0.784	0.663	0.478	0.350	0.250	0.113
uterus	0.522	0.424	0.310	0.242	0.190	0.107
vascular endothelial cell	0.697	0.598	0.457	0.357	0.272	0.112
vascular endothelium	0.727	0.615	0.452	0.339	0.247	0.108
vermiform appendix	0.924	0.897	0.850	0.804	0.747	0.459

2D. Exploring ALL GWAS genes in tissue-specific networks using neXus

NeXus is a subnetwork discovery algorithm developed by Deshpande et al. (2010) to find conserved subnetworks across species. It was originally used to work with parallel differential expression studies and used to identify conserved subnetworks between mouse and human. NeXus has a single species option for its discovery algorithm, which we used to investigate ALL GWAS genes in filtered hematopoietic stem cell and all tissues networks with 100K and 250K edges. There are three adjustable parameters required by neXus: clustering coefficient threshold (cc), network score cutoff (sco), and depth first search cutoff (dfs). We used a range of values, ranging cc between 0.1 and 0.5, dfs from 0.3 to 0.7 and sco from 0.1 to 0.5, and combined these parameters in several different combinations.

With this analysis, we hoped to reprioritize the ALL GWAS candidate genes and find putative subnetworks of functional sets of genes that were enriched for genes involved in the progression of ALL. However, regardless of parameters used, we obtained only very small networks from the hematopoietic stem cell network and no subnetworks from the all tissue network. These small networks, shown in Figure 2D.1, centered on the two genes that had the strongest ALL GWAS scores, ARID5B (84159) and PDE4B (5142). To confirm the findings were not due to noise, we performed a randomization analysis using neXus. As shown in the confidence information for the networks in Figure 2D.1, these subnetworks did not have high confidence when compared to random networks.

Repeated testing of neXus-generated subnetworks showed that randomization of input data prior to network discovery always resulted in more subnetworks than were found using the real data. An example of these results is shown in Figure 2D.2. This underperformance of real compared to random occurred for both the 100K and 250K filtered hematopoietic stem cell networks and regardless of neXus parameter combination used.

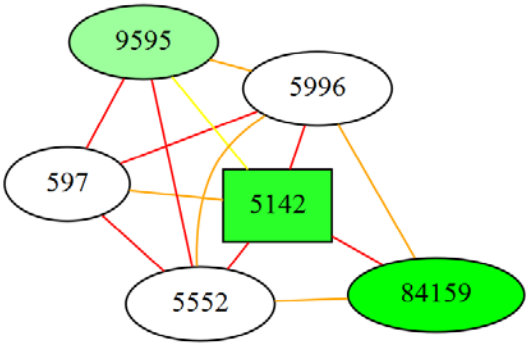


Confidence Information

Average fold change (normalized)	2.47738
Number of real subnetworks at this score	3
Number of random subnetworks at this score	3.00 +/- 0.80

Human genes' details

Gene	Name	Actual insertion score
5142		287.592200438
84159		833.281536453
3553		0
9770		0
7128		0



Confidence Information

Average fold change (normalized)	2.50203
Number of real subnetworks at this score	3
Number of random subnetworks at this score	3.00 +/- 0.80

Human genes' details

Gene	Name	Actual insertion score
597		0
5142		287.592200438
84159		833.281536453
5996		0
5552		0
9595		13.8080533771

Figure 2D.1 Subnetworks generated by use of the neXus algorithm
These subnetworks were generated using network analysis of ALL GWAS genes in the hematopoietic stem cell network. Confidence information shows low confidence in these networks in comparison to networks discovered after randomizing the data.

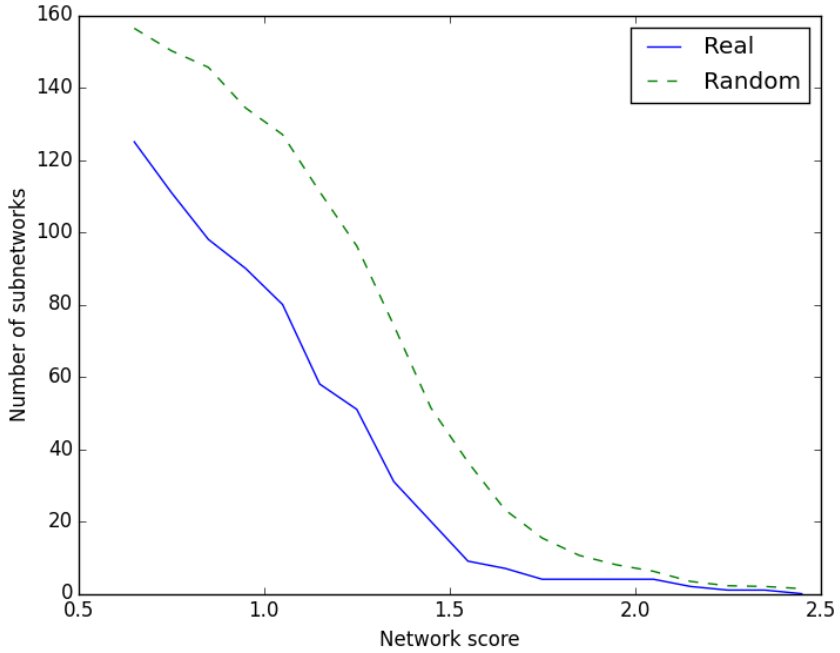


Figure 2D.2 Subnetworks obtained with real vs. randomized data
Fewer subnetworks are found by neXus using real data than using random data regardless of the network score considered.

The underperformance of real against random caused concern because if real data were not able to find useful networks, we would expect it to look similar to random data, not to underperform random data. To test the validity of the randomization algorithm, we manually randomized the ALL GWAS scores and ran the neXus algorithm using the 'real' script. Figure 2D.3 shows that manually randomized data performed similarly to the data randomized by the algorithm, suggesting the problem was in our data and not in the software.

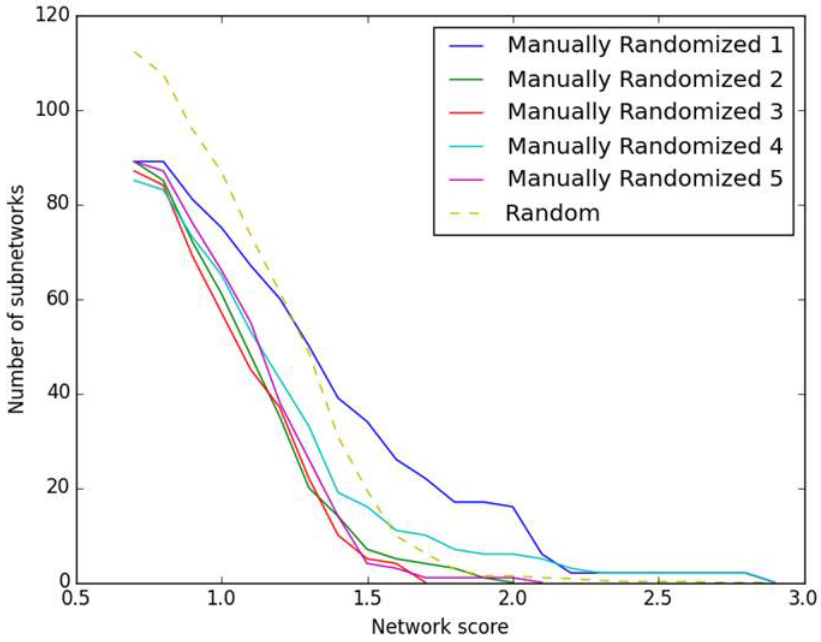


Figure 2D.3 Subnetworks obtained with manually randomized data vs. with data randomized by the neXus software
Manually randomized data gave results that were very similar to those given by using the randomization provided by neXus.

2E. Hypothesis: GWAS genes may be underconnected in functional networks

We observed that genes with strong p-values in the ALL GWAS seemed to have few edges remaining in the filtered functional network. This appeared likely to be the reason that our real data was underperforming randomized data. It also led to the hypothesis that this might be due to the nature of functional connectedness of genes that cause disease. We hypothesized this might be suggestive of a larger trend, and that GWAS trait-associated genes might be less connected in tissue-specific networks than overall network genes.

Chapter 3: Exploration of the topology of GWAS genes in human tissue-specific functional networks

3A. Data used: Catalog of Published GWAS

To determine if GWAS genes displayed unusual topology in tissue-specific functional networks, we used data downloaded from the Catalog of Published GWAS provided by the National Human Genome Research Institute, http://www.ebi.ac.uk/gwas. This data consisted of 2200 total studies that covered 1135 different phenotypes, which are referred to as traits. These traits included both qualitative phenotypes, such as disease, and quantitative phenotypes, such as height. As with the ALL GWAS data set, we converted SNPs to genes and discarded all intergenic SNPs.

After conversion to gene level IDs, the GWAS catalog data included 4253 unique genes. Most genes were associated with very few traits, 86% of genes were associated with no more than three traits and 58% were only associated with one trait. Figure 3A.1 shows a histogram of the number of genes associated with individual traits. Almost all genes are associated with ten or fewer traits, and although a few genes were associated with many traits, none were associated with more than forty-four traits. A list of genes found most frequently in the GWAS catalog is given in Table 3A.1.

Table 3A.1 Genes associated with at least 15 traits in the GWAS Catalog

Gene ID	Number of Traits	Gene Symbol	Name	Location	Type
2646	44	GCKR	glucokinase (hexokinase 4) regulator	2p23	protein-coding
28	29	ABO	ABO blood group (transferase A and transferase B)	9q34.2	protein-coding
64478	29	CSMD1	CUB and Sushi multiple domains 1	8p23.2	protein-coding
3992	27	FADS1	fatty acid desaturase 1	11q12.2-q13.1	protein-coding
341	25	APOC1	apolipoprotein C-I	19q13.2	protein-coding
8882	22	ZPR1	ZPR1 zinc finger	11q23.3	protein-coding
79068	22	FTO	fat mass and obesity associated	16q12.2	protein-coding
1012	22	CDH13	cadherin 13	16q23.3	protein-coding
3077	21	HFE	hemochromatosis	6p21.3	protein-coding
10665	21	C6orf10	chromosome 6 open reading frame 10	6p21.3	protein-coding
54715	20	RBFOX1	RNA binding protein, fox-1 homolog (C. elegans) 1	16p13.3	protein-coding
100048912	19	CDKN2B-AS1	CDKN2B antisense RNA 1	9p21.3	ncRNA
10452	17	TOMM40	translocase of outer mitochondrial membrane 40 homolog (yeast)	19q13	protein-coding
164656	17	TMPRSS6	transmembrane protease, serine 6	22q12.3	protein-coding
9415	16	FADS2	fatty acid desaturase 2	11q12.2	protein-coding
283450	15	HECTD4	HECT domain containing E3 ubiquitin protein ligase 4	12q24.13	protein-coding
2524	15	FUT2	fucosyltransferase 2 (secretor status included)	19q13.3	protein-coding
5789	15	PTPRD	protein tyrosine phosphatase, receptor type, D	9p23-p24.3	protein-coding

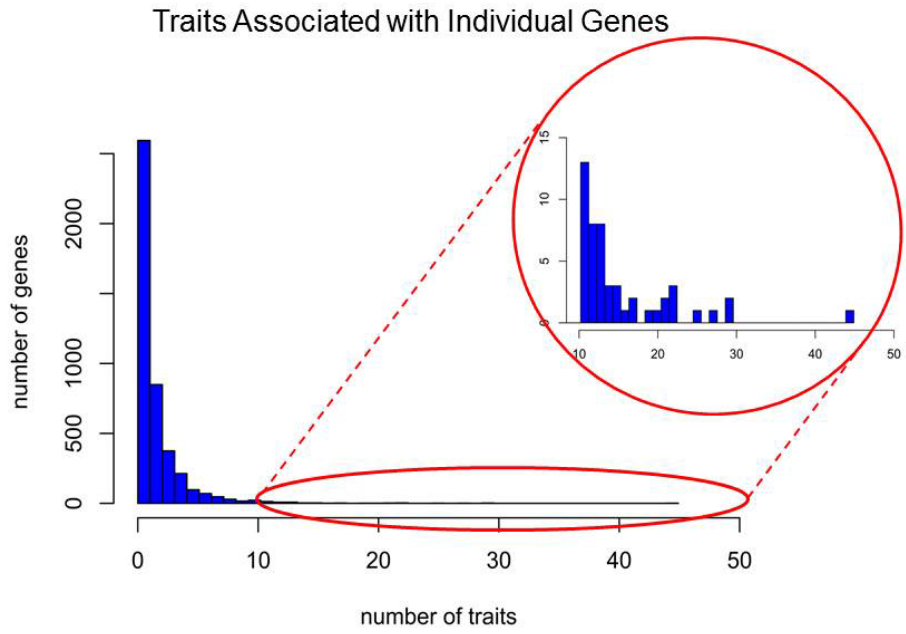


Figure 3A.1 Number of traits associated with genes in the GWAS Catalog
Most genes have a small number of associated traits. The expanded region shows the few genes with many associated traits.

When examining the trait to gene ratio from the perspective of number of genes per trait, we found most GWAS catalog traits are associated with a small number of genes, 75% of the traits have 15 or fewer associated genes. Because of limited statistical power, we excluded all traits with fewer than five associated genes, leaving 475 traits included in our analysis. Figure 3A.2 is a histogram of the number of genes associated with the remaining traits; it shows that most traits used in our analysis have five genes. It also shows there are a few traits associated with a large number of genes. The traits with the most associated genes are obesity-related traits, height, and IgG glycosylation which all have more than 150 associated genes. Table 3A.2 lists traits with fifteen or more associated genes and gives the number of genes corresponding to each trait.

Table 3A.2 Traits with at least thirty associated genes in the GWAS Catalog

Trait	Number of genes associated
Obesity-related traits	418
Height	259
IgG glycosylation	165
Blood metabolite levels	101
Schizophrenia	98
Type 2 diabetes	80
Multiple sclerosis	77
Rheumatoid arthritis	76
Crohn's disease	69
Menarche (age at onset)	66
HDL cholesterol	66
LDL cholesterol	62
Cholesterol, total	60
Prostate cancer	59
Breast cancer	58
QT interval	57
Inflammatory bowel disease	56
Metabolite levels	55
Cognitive performance	54
Coronary heart disease	54
Bipolar disorder	52
Bipolar disorder and schizophrenia	51
Amyotrophic lateral sclerosis (sporadic)	49
Type 1 diabetes	49
Platelet counts	49
Bone mineral density	48
Ulcerative colitis	48
Body mass index	47
Attention deficit hyperactivity disorder	43
Alzheimer's disease (cognitive decline)	42
Blood pressure	42
Systemic lupus erythematosus	41
PR interval in Tripanosoma cruzi seropositivity	41
Triglycerides	40
Parkinson's disease	40
Glucose homeostasis traits	37
Pulmonary function	36
Urate levels	36
Autism, bipolar & depressive disorders, ADHD and schizophrenia	36
Migraine	35
Blood metabolite ratios	34
Alzheimer's disease (late onset)	33
Educational attainment	32
Response to amphetamines	32
Red blood cell traits	31
Celiac disease	31
Hypertension	30

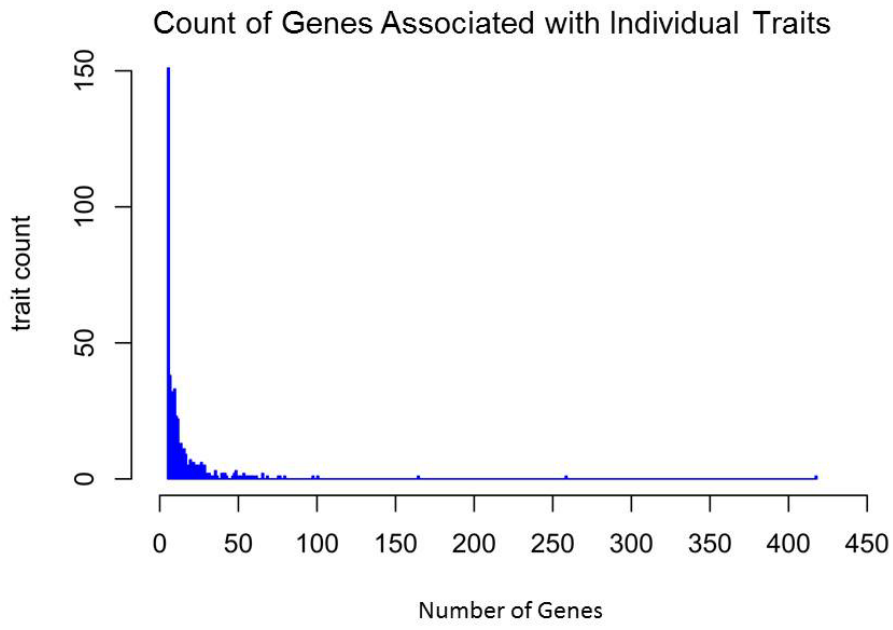


Figure 3A.2 Number of genes associated with traits used in GWAS Catalog analysis After excluding all traits with fewer than five associated genes, most traits still have a very small number of associated genes. Three traits with more than 150 associated genes are present.

3B. Analysis of degree for GWAS candidate genes

We analyzed degree in 250K filtered tissue-specific functional networks. Degree was determined by assigning all edges remaining in filtered network a value of one, and all edges removed by filtration from the network given a value of zero. This reduced the complexity of the network so that every edge was binarized to an unweighted edge with a value of 0 or 1, then summing the edges for each gene. The resulting degree for each gene corresponded to its number of functional interactions remaining in the filtered network.

For each trait, degree was calculated for all associated genes. A Wilcoxon rank sum test was used to determine if the degree values for the trait-associated genes came from a significantly different distribution than degree values

observed in the overall network and an associated p-value, false discovery rate (FDR) adjusted p-value, and z-score were obtained for each trait. This testing was done for each of the 475 traits in all 145 tissue-specific networks.

Many of these traits have significantly different gene degrees than those seen in the overall networks, 16.2% of traits were found to have significant p-values after 20% FDR correction. In comparison, when we performed the same calculations on 250 sets of randomly selected network genes, we found only 0.03% of these sets had significant p-values after 20% FDR correction. We noted that most of the traits found to be significant after 20% FDR correction had higher mean degree values than the mean degree found the overall network. When examining all traits, without regard to significance, we found that approximately half (47.9%) had mean degree values that were less than the mean degree for the overall network, but for the traits meeting the 20% FDR cutoff, only 9.7% had mean degree less than the network mean. These findings caused us to suspect possible bias in the methodology and led us to search for another network characterization metric which is discussed in section 3C. A summary of statistics related to degree is shown in Table 3B.

Table 3B Summary of statistical testing of degree in 250K filtered tissue-specific networks

	Traits with ≥ 5 associated genes	Traits with ≥ 20 associated genes	Random genes
pvals < 0.05	28.1%	75.6%	8.5%
FDRs < 0.2	16.2%	57.8%	0.03%
trait mean < network mean	47.9%	33.6%	58.0%
significant pvals with trait mean < network mean	14.5%	21.7%	51.5%
significant FDRs with trait mean < network mean	9.7%	14.3%	18.2%

Using the z-scores obtained from analysis of degree, we clustered on both tissue networks and traits to look for structure in the data. The result of this clustering is shown as a heat map in Figure 3B. We were unable to find any indication that the size of the z-score for a given trait was different in relevant tissues than that trait's z-score in the tissue networks overall.

The clustering is predominantly driven by traits, and the clusters seem to correlate with the number of trait-associated genes. For example, the left-most section of the heat map consists of the following twenty traits: height, obesity-related traits, blood metabolite levels, metabolite levels, blood metabolite ratios, metabolic traits, bipolar disorder, platelet counts, LDL cholesterol, metabolic syndrome, coronary heart disease, myopia, mean platelet volume, corneal structure, immune response to smallpox, coronary artery disease or ischemic stroke, coronary artery disease or large artery stroke, neutrophil count, Alzheimer's disease, and IgG glycosylation. More than half of these traits have thirty or more associated genes (55%), and all but one of these traits (coronary artery disease or ischemic stroke) have at least twelve associated genes. This clustering driven by number of trait-associated genes is apparently caused by the trend for traits with many associated genes to have large z-scores, which can also be seen in Table 3B. In traits with at least 20 associated genes, 57.8% of the FDR adjusted values are significant.

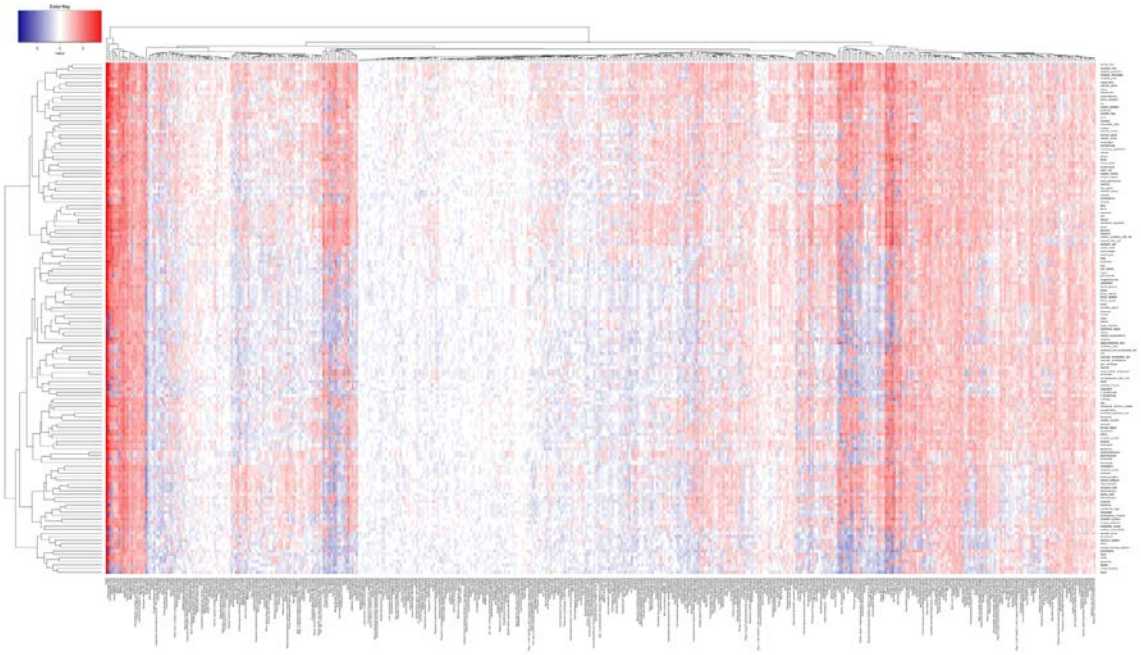


Figure 3B Heat map of z-scores of degree for GWAS genes associated with 475 traits in 145 tissue-specific networks
Traits are clustered along the x-axis and tissues are clustered along the y-axis

3C. Adjusted weighted degree (AWD) as improved methodology

We developed an alternative measure of degree that preserved edge weight as a means to overcome the methodological bias we observed using the binarized version of degree. This offered the advantage of maintaining the probability of functional relationships instead of reducing edges to binary value based on an arbitrary cutoff. For this analysis, we used the full networks instead of a filtered network. Adjusted weighted degree (AWD) was calculated as follows. All of the edge weights were adjusted by subtracting the prior probability (0.1) from the posterior probability (i.e. edge weight). Any edges with a posterior probability less than or equal to 0.1 were assigned a value of zero. The AWD for each gene was the sum of the adjusted edge weights for that node.

Statistical analysis was done as with degree, Wilcoxon rank sum testing was used to determine if the AWD values for the trait-associated genes were significantly different than overall values in the network. A summary of results is given in Table 3C. AWD shows less statistically significant findings than degree, but statistical significance trends even more strongly toward traits with increased mean AWD over that of the total network. Similar to what we found in our analysis of degree, significant differences in AWD are seen more often in traits with larger numbers of associated genes. This is likely a function of the increased statistical power of additional values used in testing.

Table 3C Summary of statistical testing of AWD in tissue-specific networks

	Traits with ≥ 5 associated genes	**Traits with ≥ 20 associated genes**	**Random genes**
pvals < 0.05	20.5%	55.7%	9.4%
FDRs < 0.2	9.8%	35.1%	0.12%
trait mean < network mean	20.0%	35.0%	56.2%
significant pvals with trait mean < network mean	1.14%	0.82%	48.2%
significant FDRs with trait mean < network mean	0.015%	0.024%	32.6%

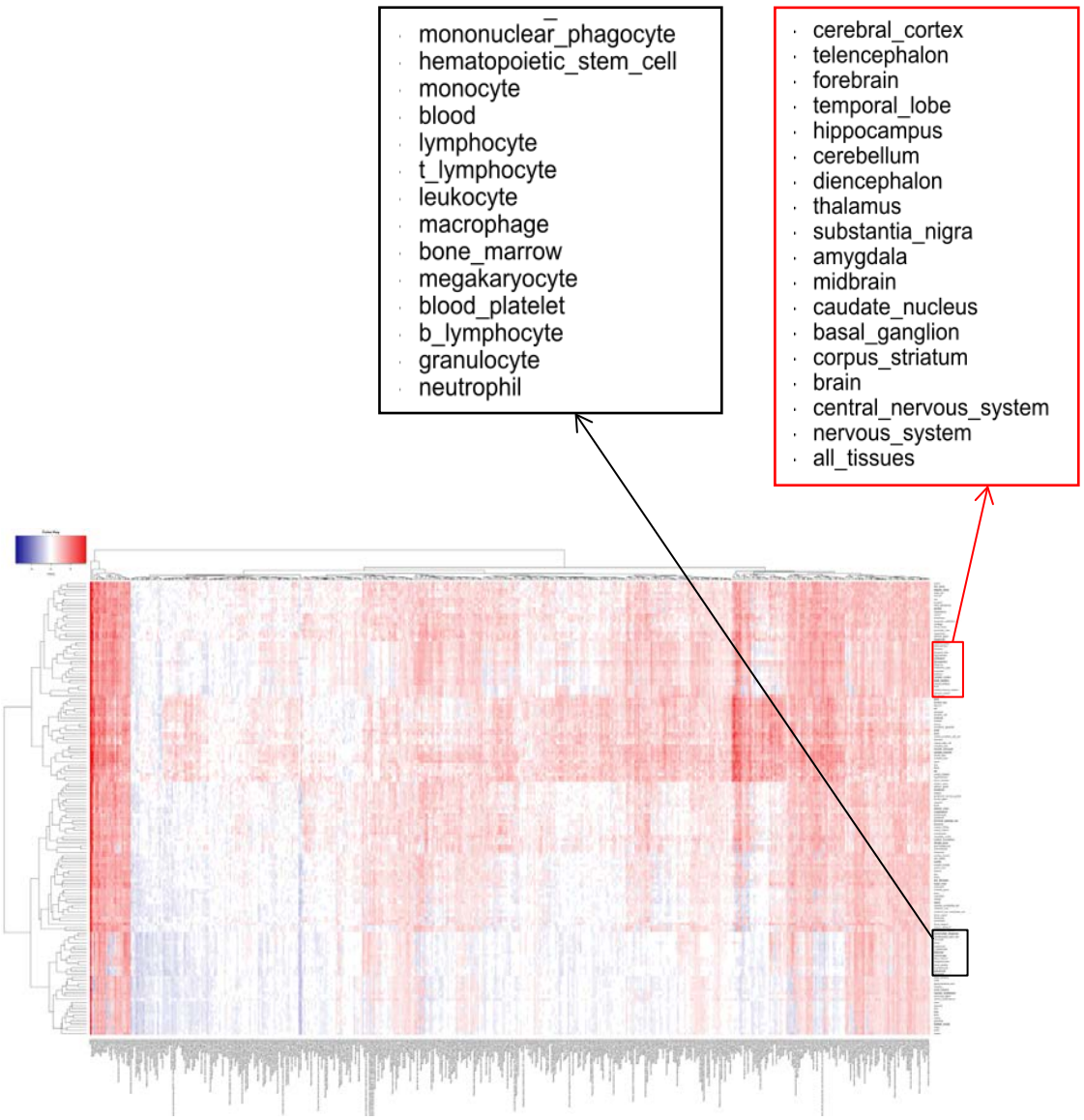


Figure 3C Heat map of z-scores of adjusted weighted degree of GWAS genes associated with 475 traits in 145 tissue-specific networks
Traits are clustered along the x-axis and tissues are clustered along the y-axis. Portions of the lists of tissue networks are expanded for legibility.

Figure 3C, shows a heat map of z-scores for trait AWD in tissue-specific networks clustered both on traits and tissues. As with degree, we were unable to identify any association between the z-scores for individual traits and likely tissue types of interest. The x-axis clustering of AWD is very similar to that seen with degree in that it still shows a predominant cluster of traits that have a large number of associated genes. We did note more clustering on tissues than was observed with degree, and unlike the clustering that we saw in the degree value heat maps, some of the tissue clusters on the AWD heat map appear to be comprised of similar tissue type. In Figure 3C two of these clusters are emphasized, one consists of cell types found in blood and lymph. The other contains the brain, central nervous system, and several different regions of the brain. It should be noted that our later research findings suggest factors other than tissue type might drive clustering, and this will be discussed in Chapter 5.

3D. Clustering coefficient, average neighbor degree, and betweenness

We completed our characterization of network topology by calculating three additional metrics: clustering coefficient, average neighbor degree and betweenness. All three metrics were calculated using a 250K filtered network.

Clustering coefficient was calculated for a gene node by first determining its neighborhood, i.e. the genes that had an existing edge to the gene of interest. The value of the clustering coefficient was then determined by taking the number of existing edges between all of the neighborhood genes and dividing by the total number of possible edges that could be in the neighborhood.

Average neighbor degree calculation also required first determining the neighborhood of the gene node. The degree for each of those neighbor genes was determined, and the average neighbor degree for the gene was the average of these values.

To calculate betweenness, first we determined the shortest paths between all nodes in the entire network. Betweenness for a gene was then equal to the number of times that gene was contained in all the shortest paths.

As with the previous metrics clustering coefficient, average neighbor degree, and betweenness were calculated for all genes associated with a trait. Then a Wilcoxon rank sum test was used to determine if the values for the trait-associated genes were significantly different than those in the overall network. Associated p-values, false discovery rate (FDR) adjusted p-values, and z-scores were obtained for each trait. A summary of these statistics is given in Table 3D.

For all three metrics, we observed that many trait-associated gene sets were significantly different from the overall networks, and that the number of significantly different trait-associated gene sets was much greater than the number found to be significantly different using randomly chosen network genes. Average neighbor degree showed the largest number of significant traits, 16.9% after FDR correction, and clustering coefficient the smallest number, 7.8%. Also, as seen with degree and adjusted weighted degree, all three metrics showed most of the significantly different traits after 20% FDR correction had higher mean metric values than the mean metric value for the overall network.

Table 3D Summary of results for clustering coefficient, average neighbor degree and betweenness

	Traits with ≥ 5 associated genes	Traits with ≥ 20 associated genes	Random genes
cc[1] pvals < 0.05	18.7%	45.9%	8.9%
cc[1] FDRs < 0.2	7.8%	25.5%	0.23%
cc[1] trait mean < network mean	44.1%	28.3%	58.3%
cc[1] significant pvals with trait mean < network mean	3.8%	4.9%	29.2%
cc[1] significant FDRs with trait mean < network mean	1.3%	1.9%	2.4%
AND[2] pvals < 0.05	28.6%	76.0%	9.8%
AND[2] FDRs < 0.2	16.9%	59.8%	0.18%
AND[2] trait mean < network mean	38.8%	19.3%	60.8%
AND[2] significant pvals with trait mean < network mean	5.3%	7.6%	45.9%
AND[2] significant FDRs with trait mean < network mean	3.2%	4.4%	1.5%
btwns[3] pvals < 0.05	24.4%	63.9%	7.6%
btwns[3] FDRs < 0.2	11.6%	42.0%	0.039%
btwns[3] trait mean < network mean	59.9%	48.8%	75.2%
btwns[3] significant pvals with trait mean < network mean	28.9%	35.5%	47.4%
btwns[3] significant FDRs with trait mean < network mean	22.9%	29.0%	50.0%

[1]clustering coefficient
[2]average neighbor degree
[3]betweenness

Z-scores for trait clustering coefficient, average neighbor degree and betweenness values in tissue-specific networks were clustered both on traits and tissues. The three resulting heat maps are shown in Figure 3D. As with previous metrics, we did not see that z-score values for traits differed with tissues of interest. Heat maps for all three metrics show very little clustering based on tissue. Also similar to the clustering observed with degree z-scores, the trait clustering appears to be driven by the number of trait-associated genes. The heat map for average neighbor degree seen in Figure 3D(B) also shows more intensity which corresponds to the greater number of significant values found using this metric as shown in Table 3D.

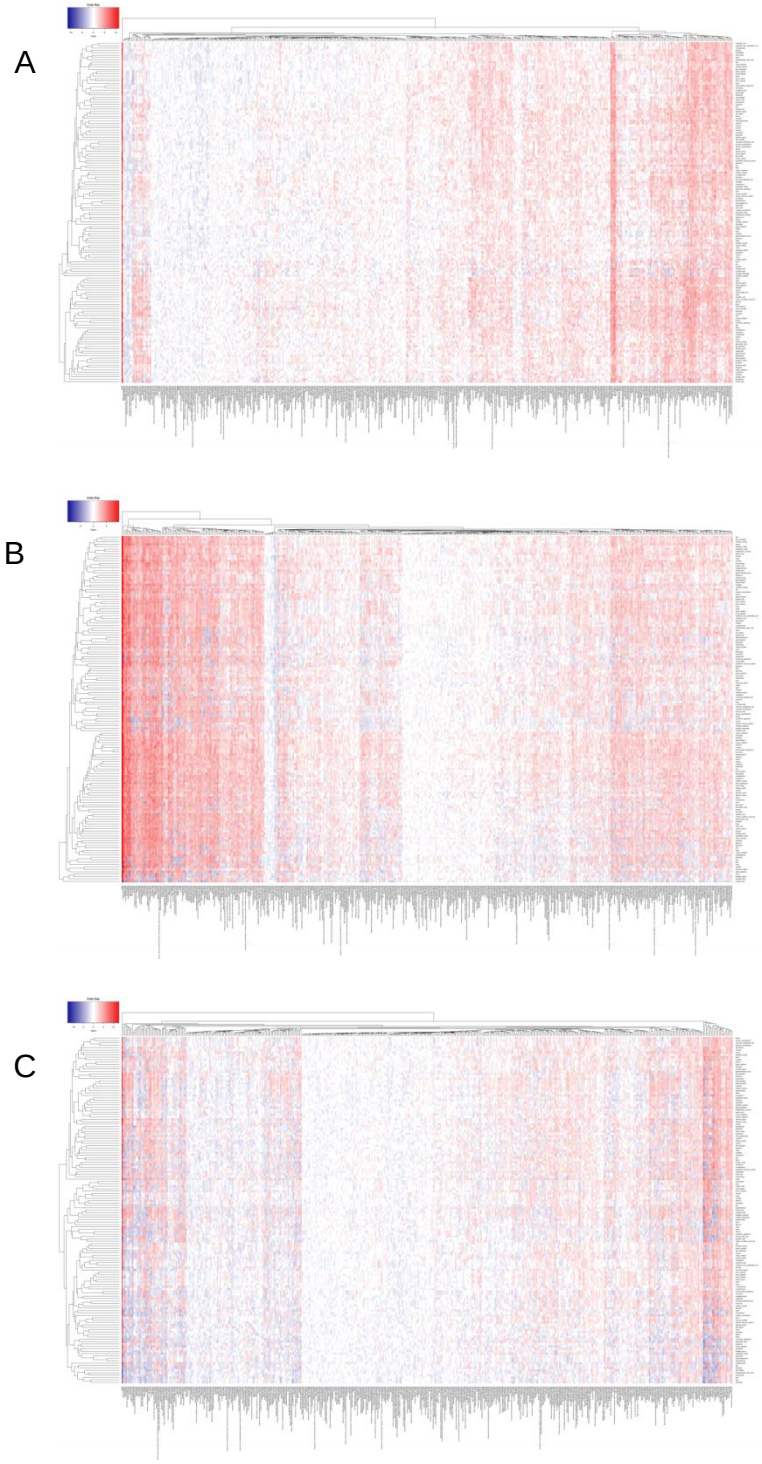


Figure 3D Heat maps of z-scores of GWAS genes associated with 475 traits in 145 tissue-specific networks
A) Clustering Coefficient B) Average neighbor degree C) Betweenness. Traits are clustered along the x-axis and tissues are clustered along the y-axis.

3E. Summary of topology of GWAS genes in tissue-specific networks

We performed a survey of the topology of trait-associated genes from the Catalog of Published GWAS in tissue-specific functional networks using five network metrics. The analyses of four of the metrics: degree, clustering coefficient, average neighbor degree, and betweenness used 250K filtered networks while the fifth metric, adjusted weighted degree, was analyzed in full networks. The only metric that showed a significant tissue-related contribution in clustering of z-scores was adjusted weighted degree. This was also the only metric that used complete networks instead of filtered networks. Additional discussion of these findings is included in Chapter 5.

For all five metrics, Wilcoxon rank sum testing for trait-associated genes against overall network genes found significant differences more often than was seen using randomly chosen network genes. Also for all metrics, most significantly different gene sets were found to have trait mean values greater than the network mean values. Although the findings with randomly selected gene sets trend slightly toward more significant findings with greater mean trait value, this trend was not as strong as observed with trait-associated genes. This indicates that it is not simply due to bias in the methodology, instead trait-associated gene sets are in fact more likely to be significantly different when the trait mean value is greater than the network mean value.

This appears to conflict with our initial observation that ALL GWAS genes were underconnected in the hematopoietic stem cell network, but this

observation was made using only genes that remained in the network after filtering to 250,000 edges. Since all of our measurements of significance using GWAS catalog trait-associated genes were compared against all network genes, the two results are not necessarily in conflict. Additional analysis is needed to determine if GWAS genes overall are significantly different from genes retaining edges in a 250K network.

The finding that significantly different GWAS gene sets tend to have greater than the average mean metric values can be interpreted in terms of biological relevance. For degree, adjusted weighted degree and betweenness it suggests that GWAS genes are more likely than average to be part of functional pathways in these networks. For average neighbor degree, the biological relevance is similar to that of degree, but more specifically suggests that these genes have a functional relationship with genes that are more likely than average to be part of functional pathways in these networks. The biological relevance of the findings related to clustering coefficient suggests that these genes are more likely to be part of a cluster of genes representing a distinct biological process. All of these concepts would be consistent with the idea that variation in these genes would result in measureable phenotypic change.

Chapter 4: Tissue-Specific Functional Network Edge Distribution

4A. General Observations

All of the 145 tissue-specific functional networks are comprised of posterior probabilities of the likelihood of a functional relationship between 25825 genes.

These probabilities are represented as edge weights; this means that the networks have approximately 333.5 million edges. Figure 4A shows four representative histograms of edge weights in tissue-specific networks. The histograms of all 145 networks share several common features. First, there is a marked peak at probability = 0.1. This appears to correspond to the prior probability initially set for all edges, and most of the network edges do not have any evidence to cause them to shift away from 0.1 when the posterior probability is determined. Another common feature is that for the edges that show a different posterior probability, most of them shift to the left, i.e. have posterior probabilities less than 0.1. The final common feature is that there are very few edges with weights greater than 0.2.

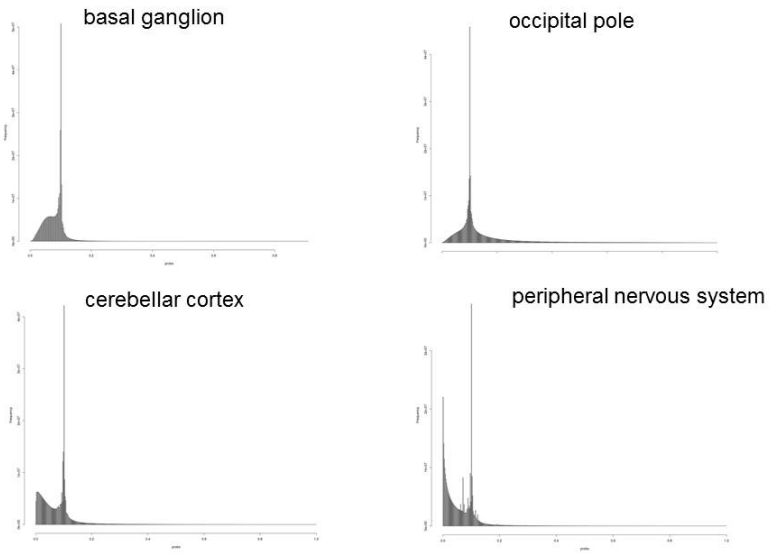


Figure 4A Representative histograms of tissue-specific functional network edges
These four tissue-specific networks show the features common to histograms from all 145 tissues, a marked peak at 0.1, most of the edges that shift moving to the left of the 0.1 peak, and very few edges with values > 0.2.

4B. Gold standard edge weights in tissue-specific networks

The tissue-specific networks include edges that are known edges, i.e. pairs of genes that are known to have a functional relationship in the tissue. Each tissue had a different set of these known functional relationship edges that we will call positive gold standards, and these gold standards were used to train the classifier when creating the functional networks. Due to our observation that some of these edges had low weights, we decided to perform a systematic analysis of their characteristics. Initially, we created histograms of edge weights for each set of tissue-specific positive gold standards. The histograms of the positive gold standards showed that their edge weights varied widely between tissues. For some tissues, the positive gold standard edge distribution was similar to what we would expect as shown in Figure 4B.1. In these histograms, we can see that most of the networks' gold standard edges have weights greater than 0.1.

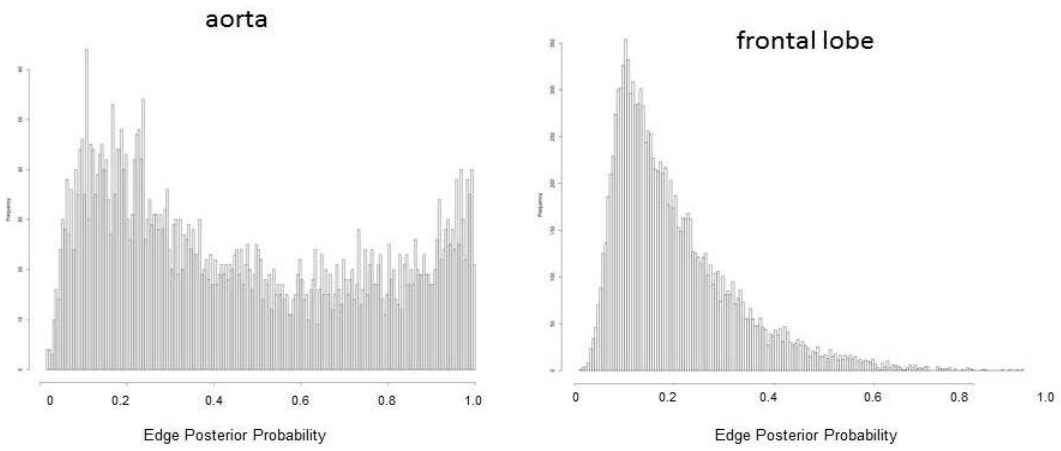


Figure 4B.1 Examples of gold standard edge distribution indicating good performance

However, some networks had positive gold standard edges that were quite different, displaying predominantly very high probabilities. In figure 4B.2, we see that the many of the positive gold standard edge weights in these networks are at or near 1.0, suggesting that these networks may exhibit overfitting.

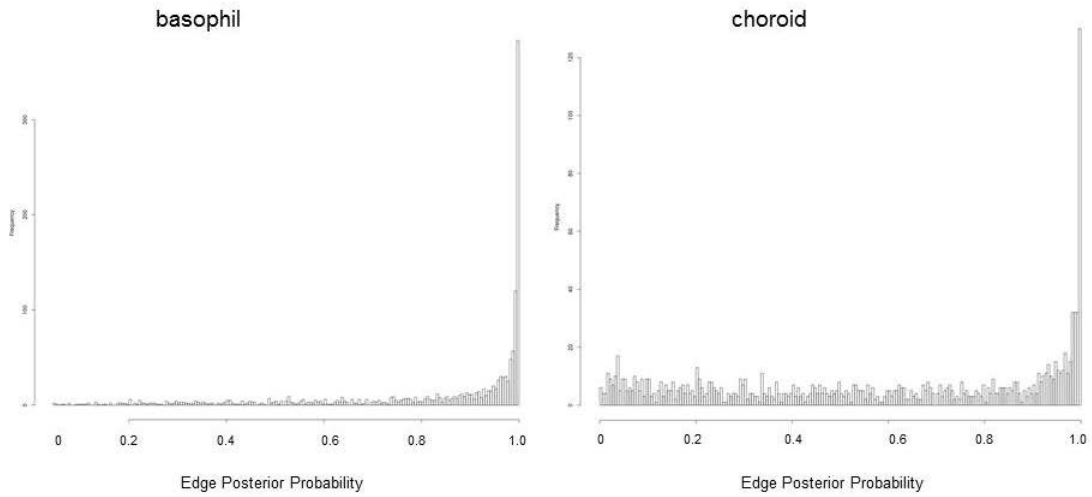


Fig4B.2 Examples of gold standard edge distribution that suggest overfitting

Finally, we also observed that some networks had gold standard edges with predominantly very low edge weights. Figure 4B.3 shows examples of two tissues that have many positive gold standard edge weights of less than 0.1. This suggests that the evidence used to create these networks incorrectly down-weighted the probability of a functional connection since their posterior probability is less than the prior probability. In these networks, most of the gold standard edges have such low scores that their signal is hidden beneath the strong signal of other edges at the prior probability of 0.1.

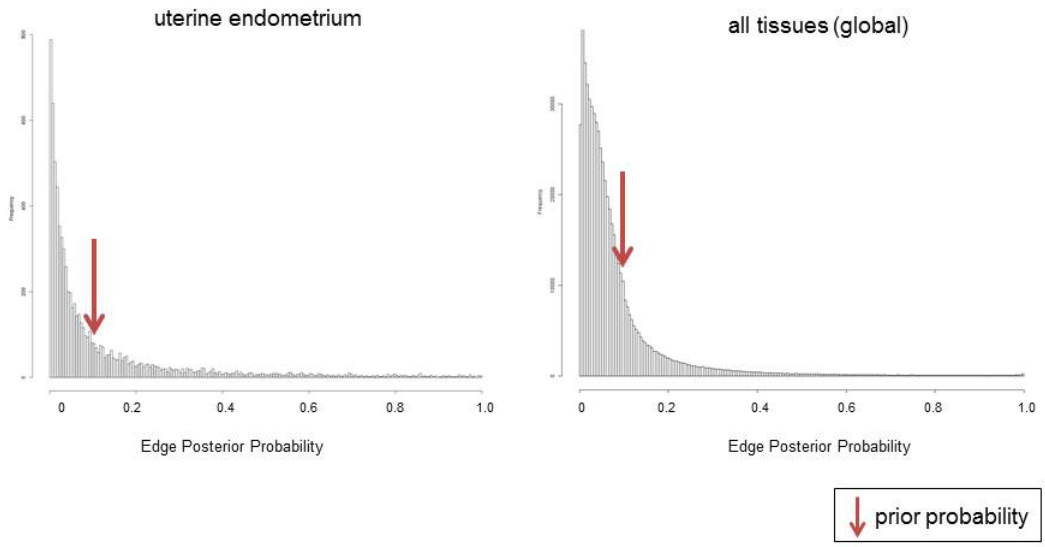


Fig 4B.3 Examples of gold standard edge distribution indicating poor performance
Gold standard edge weights are predominantly less than the prior probability score of 0.1

4C. Gold standards and area-under-the-curve (AUC) statistics

We measured the performance of the positive gold standards against all other edges in their network using an AUC statistic. This statistic was determined by dividing the edges for the network into two groups, the positive gold standards and all others. For each tissue, a Wilcoxon rank sum test of the positive gold standard edges against all others was performed and the test statistic (W value) was obtained. The W value was then divided by the number of all possible pairs to determine the AUC statistic. As shown in Table 4C, 26% of networks had gold standards that underperformed overall network edges. We also noticed a strong correlation between the number of edges in the gold standard and the AUC statistic, shown in Figure 4C.

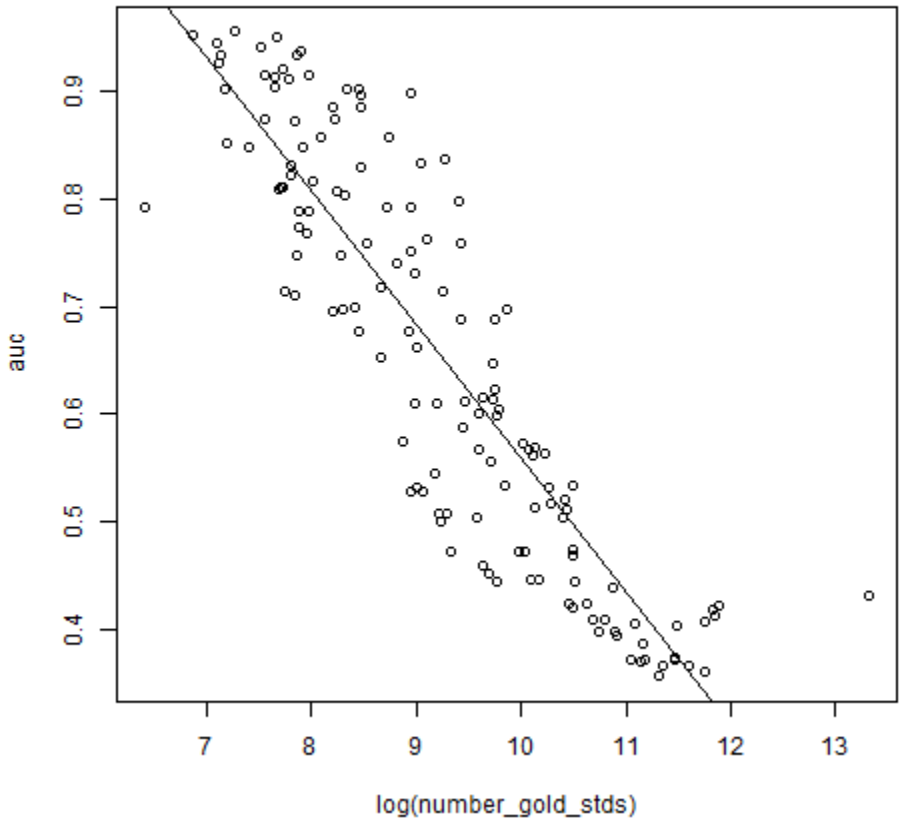


Figure 4C Correlation between positive gold standard AUC and the number of edges in the positive gold standard.
There is a strong negative correlation between the AUC and number of edges in the standard. Pearson's product correlation is -0.89 with a 95% confidence interval of -0.92 to -0.85.

We contacted the authors of the publication (Greene et al., 2015) to discuss our findings. They suggested that the correlation between AUC and number of gold standards was likely driven by overfitting in tissues where there was a small standard size. They provided a list of negative gold standards and requested that we evaluate these as well. We measured the performance of the positive gold standards against the negative standard again using an AUC statistic. The

negative gold standards used for our calculations were not tissue-specific and were used as negatives for all networks. We also calculated an AUC value assessing the performance of the positive gold standard against a random subset of 10,000 other network edges. The AUC statistics against negative edges and against random edges are both shown in Table 4C. The statistics obtained for the performance of the positive gold standards against a random set of genes are very similar to those obtained against all other edges in the network. The AUC statistics measuring performance against negative gold standards show that positives exhibit higher prediction values than negative examples for all tissue-specific networks.

Table 4C Relationship between number of edges and performance of the gold standard positives

Number of positive gold standard edges for each network and AUC scores for the positive gold standards against negative gold standard edges, against 10,000 random network edges, and against all other network edges.

tissue	gold std num.	gold.all.auc	gold.rand.auc	gold.neg.auc
adipose tissue	6687	0.740	0.740	0.808
adrenal cortex	3918	0.747	0.747	0.845
adrenal gland	16742	0.647	0.647	0.733
all tissues	604038	0.431	0.431	0.584
amygdala	16693	0.614	0.614	0.708
aorta	4796	0.885	0.885	0.856
artery	8075	0.661	0.661	0.769
astrocyte	2144	0.949	0.949	0.909
b lymphocyte	15173	0.459	0.459	0.655
basal ganglion	25280	0.569	0.569	0.672
basophil	1439	0.955	0.959	0.932
blood	126331	0.408	0.408	0.579
blood plasma	28966	0.517	0.517	0.657
blood platelet	10081	0.508	0.508	0.679
blood vessel	22834	0.472	0.472	0.632
bone	8206	0.532	0.532	0.680
bone marrow	34444	0.425	0.425	0.598
brain	136990	0.419	0.419	0.576
bronchial epithelial cell	2431	0.822	0.822	0.869
bronchus	2859	0.767	0.767	0.834
cardiac muscle	2589	0.747	0.747	0.810
cartilage	8048	0.731	0.731	0.796
caudate nucleus	17821	0.605	0.605	0.705
caudate putamen	10662	0.836	0.836	0.791
cecum	4806	0.896	0.896	0.846
central nervous system	140823	0.413	0.413	0.573
cerebellar cortex	2199	0.810	0.810	0.854
cerebellum	32654	0.504	0.504	0.653
cerebral cortex	33171	0.521	0.521	0.652
chondrocyte	3825	0.807	0.807	0.861
choroid	1298	0.902	0.903	0.913
cochlea	607	0.792	0.916	0.810
colon	45983	0.399	0.399	0.574
cornea	5789	0.652	0.652	0.766
corpus callosum	14614	0.601	0.601	0.716
corpus luteum	1650	0.848	0.848	0.889
corpus striatum	22260	0.573	0.573	0.679
culture condition cd8 cell	2596	0.934	0.934	0.897
dendritic cell	7717	0.897	0.897	0.841
dentate gyrus	1316	0.852	0.852	0.905
diencephalon	19235	0.697	0.697	0.725
duodenum	3671	0.696	0.696	0.799
ear	1907	0.915	0.915	0.846
embryo	18782	0.533	0.533	0.668
eosinophil	5740	0.717	0.717	0.789

Table 4C (Con't.)

tissue	gold std num.	gold.all.auc	gold.rand.auc	gold.neg.auc
epidermis	14473	0.505	0.505	0.655
esophagus	5059	0.759	0.759	0.798
eye	28439	0.533	0.533	0.653
fetus	127993	0.361	0.361	0.537
forebrain	35912	0.533	0.533	0.655
frontal lobe	12186	0.798	0.798	0.775
gastrointestinal tract	71838	0.372	0.372	0.551
glia	2700	0.936	0.936	0.887
granulocyte	17451	0.446	0.446	0.631
hair follicle	4029	0.697	0.697	0.787
heart	95409	0.373	0.373	0.553
hematopoietic stem cell	52998	0.440	0.440	0.625
hepatocyte	1838	0.941	0.941	0.940
hippocampus	24395	0.563	0.563	0.673
hypophysis	10376	0.714	0.714	0.766
hypothalamus	3716	0.873	0.873	0.845
ileum	2099	0.912	0.912	0.854
intestine	68702	0.370	0.370	0.550
jejunum	2256	0.921	0.921	0.852
keratinocyte	8036	0.609	0.609	0.728
kidney	107980	0.368	0.368	0.557
large intestine	49147	0.410	0.410	0.574
lens	1231	0.925	0.935	0.888
leukocyte	98026	0.404	0.404	0.591
liver	81685	0.357	0.357	0.550
locus ceruleus	952	0.952	0.954	0.933
lung	95723	0.375	0.375	0.560
lymph node	23538	0.568	0.568	0.683
lymphocyte	41308	0.425	0.425	0.606
macrophage	16397	0.557	0.557	0.710
mammary epithelium	2517	0.872	0.872	0.889
mammary gland	16009	0.453	0.453	0.636
mast cell	3265	0.856	0.856	0.861
medulla oblongata	8372	0.833	0.833	0.775
megakaryocyte	10824	0.508	0.508	0.676
midbrain	17444	0.599	0.599	0.693
monocyte	35690	0.470	0.470	0.653
mononuclear phagocyte	35690	0.475	0.475	0.658
muscle	24139	0.447	0.447	0.607
myometrium	2993	0.816	0.816	0.845
natural killer cell	3643	0.885	0.885	0.859
nephron	12810	0.611	0.611	0.734
nervous system	146248	0.422	0.422	0.576
neuron	2728	0.847	0.847	0.872
neutrophil	11187	0.472	0.472	0.656
nucleus accumbens	1917	0.874	0.874	0.891
occipital lobe	7704	0.791	0.791	0.791
occipital pole	2874	0.915	0.915	0.860

Table 4C (Con't.)

tissue	gold std num.	gold.all.auc	gold.rand.auc	gold.neg.auc
osteoblast	2537	0.711	0.711	0.814
ovarian follicle	2244	0.811	0.811	0.867
ovary	53810	0.399	0.399	0.576
oviduct	4756	0.829	0.829	0.878
pancreas	70365	0.387	0.387	0.561
pancreatic islet	2657	0.788	0.788	0.857
parietal lobe	1243	0.934	0.934	0.865
peripheral nervous system	2458	0.830	0.830	0.879
placenta	85220	0.368	0.368	0.556
podocyte	2406	0.910	0.910	0.878
pons	1198	0.944	0.944	0.897
prostate gland	43332	0.410	0.410	0.578
renal glomerulus	8988	0.762	0.762	0.829
renal tubule	4086	0.803	0.803	0.829
retina	12262	0.759	0.759	0.768
salivary gland	9749	0.609	0.609	0.741
serum	12340	0.688	0.688	0.768
skeletal muscle	54563	0.395	0.395	0.572
skin	36652	0.446	0.446	0.613
skin fibroblast	4506	0.700	0.700	0.787
small intestine	36041	0.421	0.421	0.591
smooth muscle	7521	0.678	0.678	0.757
spermatid	2235	0.811	0.811	0.855
spermatocyte	2926	0.788	0.788	0.819
spermatogonium	2926	0.788	0.788	0.819
spinal cord	25234	0.513	0.513	0.659
spleen	65111	0.406	0.406	0.578
stomach	15368	0.616	0.616	0.713
substantia nigra	17226	0.623	0.623	0.707
subthalamic nucleus	12495	0.588	0.588	0.714
t lymphocyte	21398	0.473	0.473	0.662
tear gland	6133	0.792	0.792	0.823
telencephalon	33714	0.511	0.511	0.649
temporal lobe	27805	0.564	0.564	0.669
testis	62270	0.373	0.373	0.556
thalamus	16958	0.689	0.689	0.729
thymocyte	4678	0.676	0.676	0.791
thyroid gland	14681	0.567	0.567	0.691
tonsil	6279	0.858	0.858	0.839
tooth	2627	0.773	0.773	0.838
trachea	7636	0.751	0.751	0.782
trophoblast	2320	0.715	0.715	0.828
umbilical cord	9595	0.545	0.545	0.685
umbilical vein endothelial cell	7107	0.575	0.575	0.707
urinary bladder	4148	0.902	0.902	0.874
uroepithelium	2103	0.904	0.904	0.945
uterine cervix	4472	0.699	0.699	0.799

Table 4C (Con't.)

tissue	gold std num.	gold.all.auc	gold.rand.auc	gold.neg.auc
uterine endometrium	7639	0.528	0.528	0.682
uterus	25793	0.447	0.447	0.617
vascular endothelial cell	8549	0.529	0.529	0.672
vascular endothelium	10307	0.501	0.501	0.663
vermiform appendix	4704	0.902	0.902	0.849

4D. Conclusion: which networks have acceptable performance

Several different factors needed to be taken into account to determine which tissue-specific networks were exhibiting adequate performance. During our discussions with the authors of the tissue-specific functional network publication (Greene et al., 2015), one author, Arjun Krishnan, a member of Dr. Olga Troyanskaya's lab at Princeton, indicated that only a subset of the networks "seem to have a reasonable amount of tissue-specific functional signal" and provided a list of those 105 networks. This information is indicated in table 4D.2. By eliminating all networks not included on this list, we eliminated all networks that exhibited signs of overfitting. However, this list of 105 networks still contained many networks that showed other performance problems.

The AUC statistics shown in Table 4C indicated that the positive gold standard was outperforming the negative gold standard in all networks, but showed that the positive gold standard did not outperform other edges in many of the networks, indicating additional problems with those networks. We looked at cutoffs used in making filtered networks and determined what percentage of positive gold standard edges remained in each filtered network, these

percentages are shown in Table 4D.1. This perspective makes it clear that several of the networks will not retain many functional edges of interest after filtering. Most notably, liver, kidney, placenta, lung, fetus, heart, testis, skeletal muscle, pancreas, intestine, and gastrointestinal tract all have less than 10% of the positive gold standard edges retained in a filtered network with 10 million edges. Also noteworthy is that the 250K hematopoietic stem cell network that we used for our neXus network analysis of ALL GWAS only contains 3.7% of the gold standard edges. This evidence showing that the majority of gold standard positives are missing from even filtered networks caused us the question if these networks are meaningful. Since the networks are missing many of the positive examples the classifier was given during training, we would not be confident that it would be able to properly classify other positive edges found in the data.

Table 4D.1 Percentage of positive gold standard edges in filtered tissue-specific networks

Tissue Network	top50K	top100K	top250K	top500K	top1M	top10M
adipose tissue	14	18	23	27	32	52
adrenal cortex	15	19	26	31	36	54
adrenal gland	4.4	6.2	10	13	17	36
all tissues	0.4	0.6	1.2	2.0	3.3	17
amygdala	0.7	1.2	2.4	3.9	6.9	29
aorta	19	23	29	34	38	61
artery	4.0	6.2	10	15	21	41
astrocyte	22	25	31	37	43	71
b lymphocyte	1.1	2.0	3.5	5.7	8.1	18
basal ganglion	0.3	0.6	1.4	2.5	4.7	23
basophil	28	34	41	45	52	75
blood	0.6	1.0	2.0	3.2	4.9	16
blood plasma	1.1	1.8	3.5	5.6	8.8	26
blood platelet	1.8	2.8	4.9	8.0	11	24
blood vessel	1.2	2.0	3.7	5.7	8.7	22
bone	2.8	4.5	7.5	10	15	29
bone marrow	0.8	1.3	2.4	3.9	6.0	15
brain	0.1	0.1	0.4	0.9	2.2	14
bronchial epithelial cell	18	21	27	33	39	63
bronchus	12	16	22	27	32	56
cardiac muscle	19	23	28	31	36	55
cartilage	11	15	21	25	30	51
caudate nucleus	0.4	0.7	1.6	3.1	5.7	28
caudate putamen	4.6	6.5	10	14	19	45
cecum	21	24	29	33	38	58
central nervous system	0.1	0.1	0.4	0.9	2.1	14
cerebellar cortex	22	26	31	36	41	58
cerebellum	0.5	1.0	2.1	3.7	6.0	21
cerebral cortex	0.2	0.5	1.3	2.4	4.4	20
chondrocyte	24	28	34	38	43	63
choroid	16	22	30	36	44	74
cochlea	15	20	28	33	38	58
colon	0.3	0.5	1.1	1.9	3.2	11
cornea	4.0	6.9	12	17	24	43
corpus callosum	0.6	1.1	2.3	3.8	6.4	27
corpus luteum	19	24	32	37	43	67
corpus striatum	0.3	0.6	1.3	2.6	4.7	23
culture condition cd8 cell	22	27	33	38	44	68
dendritic cell	10	12	18	22	28	54
dentate gyrus	22	29	38	43	49	70
diencephalon	1.1	1.8	3.3	5.3	8.5	32
duodenum	21	26	32	37	41	52
ear	13	17	22	27	33	58
embryo	1.3	2.1	4.1	6.5	10	26
eosinophil	5.1	7.6	12	16	22	47
epidermis	1.4	2.6	5.2	7.7	11	25
esophagus	9.4	12	16	20	24	46
eye	1.3	2.2	4.2	6.5	10	27

Table 4D.1 (Con't.)

Tissue Network	top50K	top100K	top250K	top500K	top1M	top10M
fetus	0.1	0.2	0.6	1.0	1.8	8.1
forebrain	0.3	0.6	1.4	2.7	4.7	21
frontal lobe	3.4	5.1	8.3	11	16	42
gastrointestinal tract	0.2	0.3	0.7	1.3	2.4	9.9
glia	19	22	28	33	40	67
granulocyte	1.5	2.3	3.9	5.7	8.4	19
hair follicle	16	20	25	30	35	51
heart	0.1	0.2	0.6	1.1	1.9	8.4
hematopoietic stem cell	1.3	2.1	3.7	5.6	8.3	19
hepatocyte	42	48	56	60	66	82
hippocampus	0.2	0.5	1.3	2.5	4.6	23
hypophysis	7.2	10	14	17	22	42
hypothalamus	12	14	18	21	26	52
ileum	9	12	17	21	26	53
intestine	0.2	0.3	0.7	1.4	2.4	9.6
jejunum	15	19	23	28	34	57
keratinocyte	5.7	7.9	12	15	18	34
kidney	0.1	0.1	0.4	0.8	1.6	7.6
large intestine	0.3	0.5	1.1	1.9	3.3	13
lens	32	35	40	44	49	69
leukocyte	0.6	1.0	2.0	3.3	5.0	15
liver	0.1	0.2	0.4	0.8	1.6	7.1
locus ceruleus	25	30	39	45	53	75
lung	0.1	0.2	0.4	0.9	1.6	8.0
lymph node	1.4	2.2	4.2	6.4	10	29
lymphocyte	0.8	1.3	2.5	3.8	5.9	17
macrophage	2.4	3.7	6.3	9.3	13	28
mammary epithelium	28	33	40	45	51	72
mammary gland	0.9	1.6	3.3	5.5	8.3	18
mast cell	15	17	23	27	32	58
medulla oblongata	5.8	7.8	11	15	19	45
megakaryocyte	1.7	2.8	5.1	8.0	11	24
midbrain	0.7	1.2	2.4	3.9	6.3	26
monocyte	1.7	2.6	4.6	6.6	10	20
mononuclear phagocyte	1.7	2.8	4.8	6.9	10	21
muscle	1.2	1.9	3.3	5.0	7.3	18
myometrium	29	33	39	43	48	64
natural killer cell	17	20	25	30	36	56
nephron	2.2	3.5	6.5	10	14	34
nervous system	0.1	0.1	0.4	0.9	2.2	15
neuron	19	22	28	33	38	62
neutrophil	2.0	3.2	5.2	7.7	11	22
nucleus accumbens	20	24	31	37	43	67
occipital lobe	6.2	8.7	13	17	22	47
occipital pole	16	18	23	28	33	62
osteoblast	8	11	16	21	26	48
ovarian follicle	11	15	22	28	35	61
ovary	0.2	0.3	0.6	1.1	2.1	11

Table 4D.1 (Con't.)

Tissue Network	top50K	top100K	top250K	top500K	top1M	top10M
oviduct	21	26	35	42	49	69
pancreas	0.1	0.2	0.5	0.9	1.7	9.5
pancreatic islet	26	29	35	39	44	59
parietal lobe	12	14	19	24	29	58
peripheral nervous system	34	39	46	51	55	70
placenta	0.2	0.4	0.8	1.4	2.3	7.7
podocyte	34	37	42	46	50	68
pons	21	24	29	34	39	67
prostate gland	0.2	0.4	1.1	2.0	3.4	14
renal glomerulus	6.3	9.2	15	21	28	54
renal tubule	9	13	17	23	28	56
retina	3.6	5.2	8.3	12	16	40
salivary gland	7.6	10	15	19	24	38
serum	3.9	5.8	9.5	13	18	41
skeletal muscle	0.1	0.2	0.4	0.7	1.4	8.6
skin	0.6	1.1	2.5	4.4	7.1	20
skin fibroblast	5.6	8.1	12	16	22	46
small intestine	0.3	0.6	1.5	2.5	4.2	14
smooth muscle	3.9	6.1	10	14	19	44
spermatid	37	42	47	51	56	68
spermatocyte	25	28	33	36	40	56
spermatogonium	25	28	33	37	40	56
spinal cord	0.4	0.8	1.5	2.7	4.5	19
spleen	0.1	0.2	0.4	0.8	1.6	10
stomach	3.0	4.4	7.1	9.2	12	31
substantia nigra	0.7	1.2	2.4	3.9	6.5	28
subthalamic nucleus	0.5	1.0	2.3	4.1	7.0	28
t lymphocyte	1.2	2.0	3.4	5.0	7.5	20
tear gland	12	15	19	23	29	52
telencephalon	0.2	0.5	1.2	2.4	4.3	20
temporal lobe	0.2	0.4	1.1	2.2	4.1	22
testis	0.3	0.4	0.7	1.2	2.0	8.5
thalamus	0.9	1.5	2.9	4.8	8.1	32
thymocyte	10	13	17	21	26	44
thyroid gland	2.7	4.5	7.3	11	14	31
tonsil	15	18	23	28	33	57
tooth	13	18	25	30	37	59
trachea	8.5	11	16	21	27	50
trophoblast	7.9	12	17	22	28	49
umbilical cord	2.5	4.1	7.3	10	14	29
umbilical vein endothelial cell	3.9	5.7	10	13	18	35
urinary bladder	17	20	26	31	37	62
uroepithelium	33	43	53	60	66	80
uterine cervix	16	19	25	29	34	50
uterine endometrium	1.8	3.0	6.2	10	14	29
uterus	0.5	1.1	2.5	4.5	7.2	18
vascular endothelial cell	2.8	4.0	6.7	10	13	28
vascular endothelium	1.8	3.0	5.6	8.6	12	25
vermiform appendix	22	25	30	34	38	58

Using the information provided by the Troyanskaya lab and the AUC values measuring positive gold standard performance against all edges, we established specific criteria required for network performance to be considered acceptable. First, only networks that had been indicated by Arjun Krishnan to be performing well in the Troyanskaya lab were considered. In many of those 105 networks, we had observed that the majority of gold standard positives were being downweighted by the classifier. We established a requirement of an AUC statistic ≥0.52 in order to exclude those networks and retain only the networks that provided reasonable confidence they would be likely to properly classify positive edges. Sixty-one tissue-specific functional networks met these criteria, as shown in Table 4D.2.

Table 4D.2 Criteria to assess network performance and networks meeting the criteria

tissue	in passing list	gold.all.auc	acceptable
adipose tissue	YES	0.740	YES
adrenal cortex	YES	0.747	YES
adrenal gland	YES	0.647	YES
amygdala	YES	0.614	YES
aorta	YES	0.885	YES
artery	YES	0.661	YES
basal ganglion	YES	0.569	YES
bone	YES	0.532	YES
bronchial epithelial cell	YES	0.822	YES
bronchus	YES	0.767	YES
cardiac muscle	YES	0.747	YES
cartilage	YES	0.731	YES
caudate nucleus	YES	0.605	YES
cerebellar cortex	YES	0.810	YES
cerebral cortex	YES	0.521	YES
chondrocyte	YES	0.807	YES
cornea	YES	0.652	YES
corpus callosum	YES	0.601	YES
corpus luteum	YES	0.848	YES
corpus striatum	YES	0.573	YES
dendritic cell	YES	0.897	YES
dentate gyrus	YES	0.852	YES
diencephalon	YES	0.697	YES
duodenum	YES	0.696	YES
embryo	YES	0.533	YES
eye	YES	0.533	YES
forebrain	YES	0.533	YES
frontal lobe	YES	0.798	YES
hippocampus	YES	0.563	YES
hypophysis	YES	0.714	YES
hypothalamus	YES	0.873	YES
keratinocyte	YES	0.609	YES
macrophage	YES	0.557	YES
medulla oblongata	YES	0.833	YES
midbrain	YES	0.599	YES
nephron	YES	0.611	YES
neuron	YES	0.847	YES
nucleus accumbens	YES	0.874	YES
occipital lobe	YES	0.791	YES
osteoblast	YES	0.711	YES
ovarian follicle	YES	0.811	YES
pancreatic islet	YES	0.788	YES
renal glomerulus	YES	0.762	YES
renal tubule	YES	0.803	YES
salivary gland	YES	0.609	YES
serum	YES	0.688	YES
skin fibroblast	YES	0.700	YES
smooth muscle	YES	0.678	YES

Table 4D.2 (Con't.)

tissue	in passing list	gold.all.auc	acceptable
substantia nigra	YES	0.623	YES
subthalamic nucleus	YES	0.588	YES
tear gland	YES	0.792	YES
temporal lobe	YES	0.564	YES
thalamus	YES	0.689	YES
thymocyte	YES	0.676	YES
thyroid gland	YES	0.567	YES
tooth	YES	0.773	YES
trophoblast	YES	0.715	YES
umbilical cord	YES	0.545	YES
umbilical vein endothelial cell	YES	0.575	YES
uterine cervix	YES	0.699	YES
vascular endothelial cell	YES	0.529	YES
all tissue	NO	0.431	NO
astrocyte	NO	0.949	NO
b lymphocyte	YES	0.459	NO
basophil	NO	0.955	NO
blood	YES	0.408	NO
blood plasma	YES	0.517	NO
blood platelet	YES	0.508	NO
blood vessel	YES	0.472	NO
bone marrow	YES	0.425	NO
brain	YES	0.419	NO
caudate putamen	NO	0.836	NO
cecum	NO	0.896	NO
central nervous system	YES	0.413	NO
cerebellum	YES	0.504	NO
choroid	NO	0.902	NO
cochlea	NO	0.792	NO
colon	YES	0.399	NO
culture condition cd8 cell	NO	0.934	NO
ear	NO	0.915	NO
eosinophil	NO	0.717	NO
epidermis	YES	0.505	NO
esophagus	NO	0.759	NO
fetus	YES	0.361	NO
gastrointestinal tract	YES	0.372	NO
glia	NO	0.936	NO
granulocyte	YES	0.446	NO
hair follicle	NO	0.697	NO
heart	YES	0.373	NO
hematopoietic stem cell	YES	0.440	NO
hepatocyte	NO	0.941	NO
ileum	NO	0.912	NO
intestine	YES	0.370	NO
jejunum	NO	0.921	NO
kidney	YES	0.368	NO
large intestine	YES	0.410	NO

Table 4D.2 (Con't.)

tissue	in passing list	gold.all.auc	acceptable
lens	NO	0.925	NO
leukocyte	YES	0.404	NO
liver	YES	0.357	NO
locus ceruleus	NO	0.952	NO
lung	YES	0.375	NO
lymph node	NO	0.568	NO
lymphocyte	YES	0.425	NO
mammary epithelium	NO	0.872	NO
mammary gland	YES	0.453	NO
mast cell	NO	0.856	NO
megakaryocyte	YES	0.508	NO
monocyte	YES	0.470	NO
mononuclear phagocyte	YES	0.475	NO
muscle	YES	0.447	NO
myometrium	NO	0.816	NO
natural killer cell	NO	0.885	NO
nervous system	YES	0.422	NO
neutrophil	YES	0.472	NO
occipital pole	NO	0.915	NO
ovary	YES	0.399	NO
oviduct	NO	0.829	NO
pancreas	YES	0.387	NO
parietal lobe	NO	0.934	NO
peripheral nervous system	NO	0.830	NO
placenta	YES	0.368	NO
podocyte	NO	0.910	NO
pons	NO	0.944	NO
prostate gland	YES	0.410	NO
retina	NO	0.759	NO
skeletal muscle	YES	0.395	NO
skin	YES	0.446	NO
small intestine	YES	0.421	NO
spermatid	NO	0.811	NO
spermatocyte	NO	0.788	NO
spermatogonium	NO	0.788	NO
spinal cord	YES	0.513	NO
spleen	YES	0.406	NO
stomach	NO	0.616	NO
t lymphocyte	YES	0.473	NO
telencephalon	YES	0.511	NO
testis	YES	0.373	NO
tonsil	NO	0.858	NO
trachea	NO	0.751	NO
urinary bladder	NO	0.902	NO
uroepithelium	NO	0.904	NO
uterine endometrium	NO	0.528	NO
uterus	YES	0.447	NO
vascular endothelium	YES	0.501	NO
vermiform appendix	NO	0.902	NO

Chapter 5: Summary

5A. GWAS genes exhibit non-random topology in functional networks

Using five different metrics: degree, adjusted weighted degree, clustering coefficient, average neighbor degree, and betweenness we found that GWAS trait-associated gene sets were more likely to be significantly different from the network than randomly chosen network gene sets. In addition for all metrics, we found that these significantly different gene sets usually had increased mean metrics when compared to the network mean. This suggests that these genes are more likely than average to be part of or have functional relationships with biological pathways or processes.

Additionally, this finding has an important technical implication for potential bias in randomization analyses. Randomization analysis is frequently used to assess the statistical significance of network analysis discoveries. There is reason for concern that the non-random topology of the genes of interest could cause incorrect results when using randomization analysis. In particular, if the genes of interest have higher than average values for the relevant metric, randomization analysis would be likely to support findings that were not truly significant. Because of this, our findings indicate the use of topology-preserving randomization is recommended when performing network randomization analyses.

5B. Effect of network edge distribution on statistical findings

Statistical tests found more significant differences for all metrics when the

trait mean was higher than the network mean. This trend was also seen, but to a much lesser extent, in randomly chosen genes. Examination of the networks edge distribution suggests, to the extent that this is seen in randomly chosen gene sets, it is possible that it is an artifact created by skewed edge distributions. However due to the very small number of random gene sets found to be significantly different, more testing would be necessary to confirm this. It is clear that the trend toward increased mean metric seen in the GWAS gene sets is stronger than that seen in random genes and therefore not due to the skewed edge distributions.

Also, in the context of network edge weight distribution, we revisited the apparent tissue-network-driven clustering seen in the heat map of adjusted weighted degree statistics, Figure 3C. Comparing how the clustering correlates with gold vs. all edges AUC scores, it is apparent that the clustering coincides with AUC scores better than it does with the concept of “similarity” between the tissues. We had already observed that adjusted weighted degree captured network signal better than other metrics. We have also shown that distribution of positive functional edges, i.e. network signal, varies widely among networks. It then follows that we should expect adjusted weighted degree to be the network most likely to show differentiation between networks based on performance.

Following our investigation of network edge distribution and its relation to performance, we eliminated the networks that failed to meet our performance criteria and examined our metrics for the 475 GWAS trait in the sixty-one

remaining networks. Heat maps from this analysis are shown in Figure 5B. Similar to what was observed when using all networks, trait-driven clustering patterns correlating to number of trait-associated genes are still present in the heat maps containing only these sixty-one tissues. We still did not find any relationship between relevance of tissue networks and strength of z-scores for a given trait.

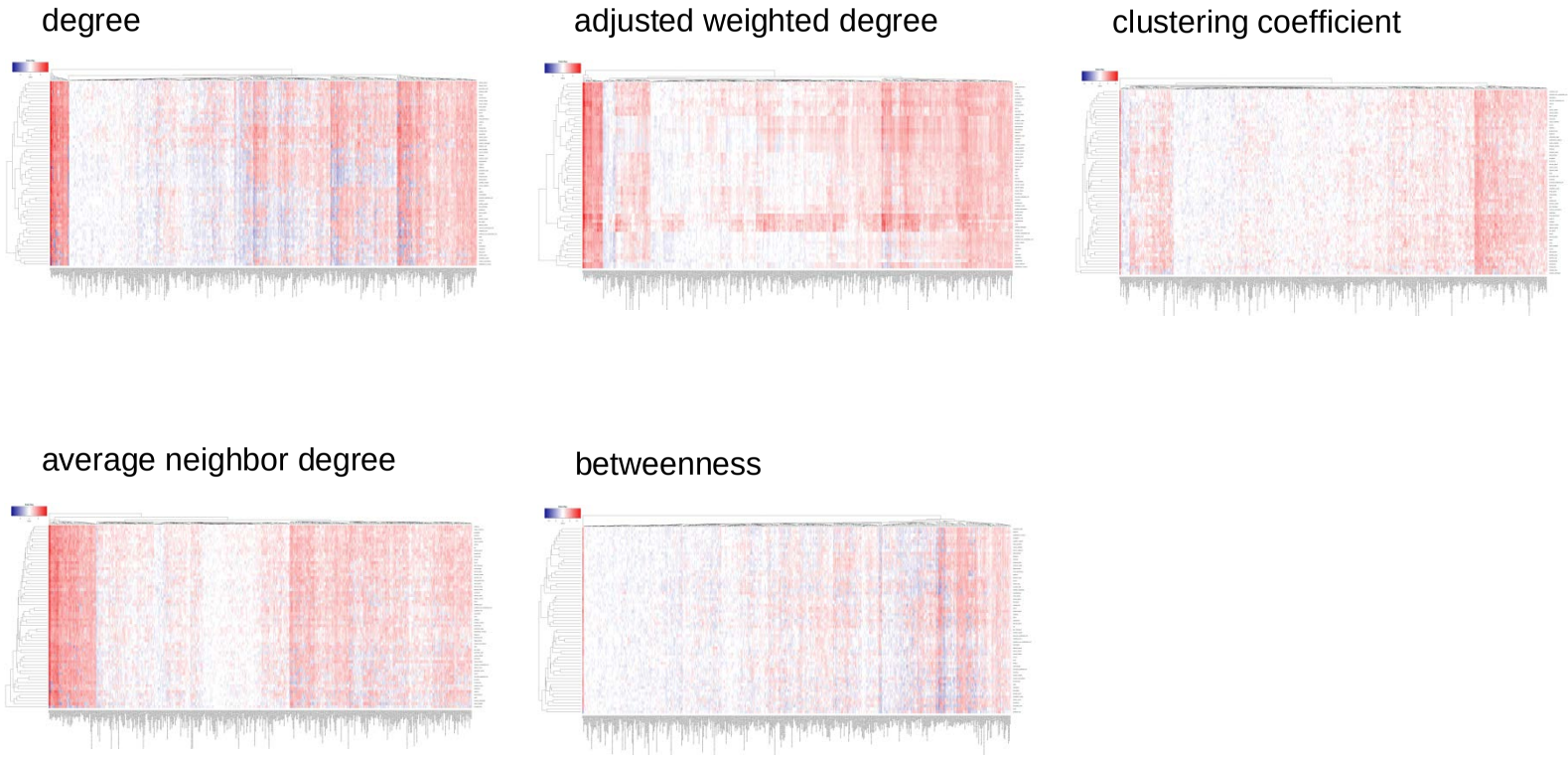


Figure 5B Heat maps of five metrics of GWAS trait-associated genes in 61 tissue-specific networks.
The topology of genes associated with 475 traits were assessed in 61 acceptably performing tissue-specific functional networks chosen for meeting criteria of acceptable network performance.

5C. Summary of tissue-specific functional network performance

When discussing our observations of problematic network performance for the positive gold standards with the authors of the tissue specific functional network paper (Greene et al. 2015), they provided insight into the likely cause of this problem. Their initial attempts to integrate appropriate data into tissue-specific networks resulted in networks whose functional signal was relatively general and lacked tissue specificity. To overcome this, they reclassified positive examples of functional relationships that occurred in multiple unrelated tissues as part of their set of negative edges used for training the model. This appears to have caused misclassification of many positive edges, likely due to a situation where the classifier was given many examples of edges that were assigned as negatives but yet had a good deal of data supporting functional relationships. This would also explain the correlation between number of gold standards and AUC score, because the authors varied the number of negative edges used the train the model proportionally with the number of gold standard positive edges. The classifier's ability to recognize a positive functional edge is apparently decreased as it is given more examples of ubiquitous positive edges classified as negative edges.

Unfortunately, while this possible explanation may be useful knowledge for the creation of future networks, it does not provide us a means to correct the current networks. The low edge weight of the gold standards remains a concern,

even in the networks that are exhibiting relatively good performance; many gold standard positive edges in these networks have very low edge weights and suggest that many positives in the data will be missed. Because of what equates to only partial coverage of current knowledge about functional relationships, the “acceptable” networks may work relatively well for generating novel findings, but will perform poorly if trying to recreate specific relationships or interrogate relationships based on previously known information. Knowledge of their limitations will be important for their successful use.

Bibliography

Cornish AJ, Filippis I, David A, Sternberg MJ. “Exploring the cellular basis of human disease through a large-scale mapping of deleterious genes to cell types.” *Genome Medicine*. 2015. 7(1).

Deshpande R, Sharma S, Verfaillie CM, Hu WS, Myers CL. “A Scalable Approach for Discovering Conserved Active Subnetworks across Species.” *PLOS Computational Biology*. 2010. 6(12). e1001028.

Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, Zhang R, Hartmann BM, Zaslavsky E, Sealfon SC, Chasman DI, FitzGerald GA, Dolinski K, Grosser T, Troyanskaya OG. “Understanding multicellular function and disease with human tissue-specific networks.” *Nature Genetics*. 2015. 47(6). 569-76.

Gremse M, Chang A, Schomburg I, Grote A, Scheer M, Ebeling C, Schomburg D. “The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources.” *Nucleic Acids Research*. 2011. 39. D507–D513.

Huttenhower C, Haley EM, Hibbs MA, Dumeaux V, Barrett DR, Coller HA, Troyanskaya OG. “Exploring the human genome with functional maps.” *Genome Research*. 2009. 19(6).1093-106.

Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A. “Human Protein Reference Database—2009 update.” *Nucleic Acids Research*. 2009. 37. D767–D772.

Mudunuri U, Che A, Yi M, Stephens RM. “bioDBnet: the biological database network.” *Bioinformatics*. 25 (2009). 555-556.

Myers CL, and Troyanskaya, O.G.. “Context-sensitive data integration and prediction of biological networks.” *Bioinformatics.* 2007. 23: 2322–2330

Papaemmanuil E, Hosking FJ, Vijayakrishnan J, Price A, Olver B, Sheridan E, Kinsey SE, Lightfoot T, Roman E, Irving JA, Allan JM, Tomlinson IP, Taylor M, Greaves M, Houlston RS. “Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia.” *Nature Genetics*. 2009. 41(9). 1006-10.

Xu H, Yang W, Perez-Andreu V, Devidas M, Fan Y, Cheng C, Pei D, Scheet P, Burchard EG, Eng C, Huntsman S, Torgerson DG, Dean M, Winick NJ, Martin PL, Camitta BM, Bowman WP, Willman CL, Carroll WL, Mullighan CG, Bhojwani D, Hunger SP, Pui CH, Evans WE, Relling MV, Loh ML, Yang JJ. “Novel susceptibility variants at 10p12.31-12.2 for childhood acute lymphoblastic leukemia in ethnically diverse populations.” *Journal of the National Cancer Institute*. 2013. 105(10). 733-42.

Yang JJ, Cheng C, Yang W, Pei D, Cao X, Fan Y, Pounds SB, Neale G, Treviño LR, French D, Campana D, Downing JR, Evans WE, Pui CH, Devidas M, Bowman WP, Camitta BM, Willman CL, Davies SM, Borowitz MJ, Carroll WL, Hunger SP, Relling MV. “Genome-wide interrogation of germline genetic variation associated with treatment response in childhood acute lymphoblastic leukemia.” *The Journal of the American Medical Association*. 2009. 301(4). 393-403.