CS224w Final Report: Development and comparison of tissue-specific causal gene regulatory networks

Nicole Ferraro and Adam Lavertu

December 9, 2017

1 Introduction

Our understanding of genetics has progressed beyond the original formulation of one gene → one protein → one function. A complex architecture underlies the process from genotype to phenotype, and upwards of thousands of genes may be involved in any given trait. Causality in genetic regulation has primarily been assessed via targeted experimentation, though with the wealth of large scale transcriptomic data being generated, methods to infer causal relationships directly from high-throughput datasets are desirable. With recent work focused on incorporating genetics into therapeutic decision-making, an ability to not only identify genes associated with a disease, but also genes involved in the mechanism of disease is essential for effective treatment development. Several studies have implemented network representations to attempt to elucidate this complex regulation architecture. However, these analyses have largely ignored directionality. In cases where directionality has been assessed, the analyses lack incorporation of how this structure varies across tissues, or across disease states. As availability of large-scale datasets of gene expression data increases, our ability to represent and analyze the full extent of the relationships present in that data, both between genes and genes and external phenotypes, will be essential in effectively harnessing those datasets to improve our understanding of human health and disease. We present a network-based approach to examine how gene expression networks vary across tissues, as well as an analysis of regulatory directionality in known gene sets and their variation across tissues. Finally, we assess how these regulatory networks differ between healthy and cancer states in tissue-matched samples.

2 Related Work

Understanding and analyzing tissue specific gene interaction networks is an important component of prioritizing genes potentially implicated in a phenotype of interest, as changes to any given gene can potentially impact many others. Previous work has made progress in this goal. One recent study aimed to determine such networks by sourcing the structure of each tissue network from Gene Ontology functional relationship information and incorporated tissue specificity, which genes were included in a particular tissue network, from the Human Protein Reference Database [1]. They created novel tissue specific gene networks from tissue-specific gene expression data and demonstrated the biological relevance of these networks, and made these networks publicly available through the online GIANT resource. This has also been explored in examining both transcription and splicing regulation in a tissue-specific manner [2]. From these tissue-specific networks of transcription and splicing, hub genes relative to connections to total expression, isoform ratio, or a combination can be identified which could indicate transcription factors likely involved in splicing. These tissue-specific networks were able to provide biological insight, but largely lacked any indication as to the directionality of proposed effects. Methods have been proposed to address this causality [3, 4, 5], using methods such as Mendelian Randomization and Bayesian Inference. Previously described Mendelian Randomization approaches [5] require both genotype and gene expression data to infer relationships, and given that gene expression varies greatly among different tissues, allowing them to perform various functions, any causal relationships learned are likely inconsistent depending on the origin of the expression data. Thus, we seek here to leverage tissue-specific gene expression data and Bayesian Inference approaches to construct directed networks of known gene sets across tissues, and demonstrate that there is variation in this regulation that may contribute to the observed variation in tissue function.

3 Data Collection and Processing

3.1 GIANT networks

To initially explore undirected relationships among genes, we analyzed the structure of tissues within the GIANT network collection [1], that are most similar to the tissues we use for the downstream analysis. GIANT networks for lung, skin, and blood were downloaded from the GIANT website. We chose to analyze the "top" networks, which are networks that contain only interactions (edges) between genes (nodes) for which there is evidence of a tissue specific function. Essentially, this means we are looking at only the gene interactions that are specific to a particular tissue type. The downloaded networks were then processed to remove edges that had a posterior probability of <0.5 and unconnected nodes, as these are uninformative for this analysis.

3.2 GTEx gene expression data

To construct tissue-specific networks, we obtained tissue-specific gene expression data from the Genotype-Tissue Expression project, or GTEx [6]. All samples were collected post-mortem. We focus here on four tissues of interest, chosen based on sample size and relevance to cancer gene expression data for downstream comparison. These include whole blood and lung, as well as both sun exposed and sun unexposed skin samples. Among all tissues, there are 22030 unique genes and 438 total samples. Each tissue contains roughly the same number of genes, with whole blood containing the fewest, and the sample proportions vary, with 338 samples in blood, and 278, 302, and 196 in lung, exposed skin, and unexposed skin respectively. The expression data for each tissue was processed according to a previously described protocol [7]. Briefly, RPKM values, or reads per kilobase of transcript per million mapped reads, were log transformed and standardized by gene, after filtering for very low expression.

3.3 TCGA gene expression data

In order to compare the directed networks learned in healthy samples from the GTEx project, we also obtained publicly available expression data from cancer samples in matched tissues from The Cancer Genome Atlas [8]. We analyzed expression data for three cancer types - lung (Lung Squamous Cell Carcinoma), blood (Acute Myeloid Leukemia), and skin (Skin Cutaneous Melanoma). For each tissue, we obtained measurements across 60,483 transcripts in the form of FPKM counts, or fragments per kilobase of transcript per million mapped reads, which is comparable to the RPKM values used in the GTEx data processing and then filtered for very low expression, resulting in 17,722 (blood), 18,295 (lung), and 16,809 (skin) measured transcripts. We again log-normalized and standardized the measurements by gene. Our sample sizes in this case are 188 (blood), 552 (lung), and 473 (skin).

3.4 Gene set data

In order to understand how known pathways vary across tissues, we focus on established gene sets curated by the Broad Institute's Molecular Signatures Database, or MSigDB. This resource includes sets of genes sourced from biological experiments, external databases including KEGG, REACTOME and the Gene Ontology, positional gene sets, and computationally predicted sets. We chose to focus on a subset of these sets derived from databases of known molecular pathways and curated from domain expert knowledge. This resulted in an analysis of 1329 gene sets in total, which together encompass 8903 unique genes, and range in size from 6 to 1028 genes, with a mean of 50.88 ± 77.96 genes in a set.

4 Results

4.1 GIANT tissue network comparison

4.1.1 Network statistics

Using the iGraph package[9] in R, we calculated summary statistics for the blood, lung, and skin tissue networks(Table 1). We can see from the statistics that the lung and skin networks have

a similar number of nodes and edges, while the blood network has a considerably higher number of edges. Interestingly, despite the difference in the number of edges, all three networks have similar clustering coefficients and network diameters. The greater number of edges in the blood network makes biological sense as blood hosts a broad range of biological processes that each involve interactions between many different genes. This idea of many modular processes is supported by the similarity in clustering coefficients between the networks, despite the increased size of the blood network.

Table 1: Summary statistics for GIANT tissue networks

Tissue	# nodes	# edges	diameter	edge density	clust. coeff.
blood	7,143	214,736	11	0.00842	0.356
lung	4,533	51,686	14	0.00503	0.328
skin	4,287	49,955	11	0.00544	0.354

4.1.2 Degree distributions

In addition to the summary statistics generated above, we sought to compare the degree distributions of each of the tissue networks (Figure 1), after pruning using different probability thresholds. We can see that the shape of the degree distributions in log-log space are similar and that they all follow a power law distribution, as we would expect from biological networks[10]. In addition to this, we see that the scale-free networks are robust to the probability pruning threshold, having similar degree distributions and values of α after pruning at various thresholds. For the remainder of the analyses, we used the network pruned at 0.5 probability, as this gave similar network structure to the more conservative pruning thresholds and contains more information.

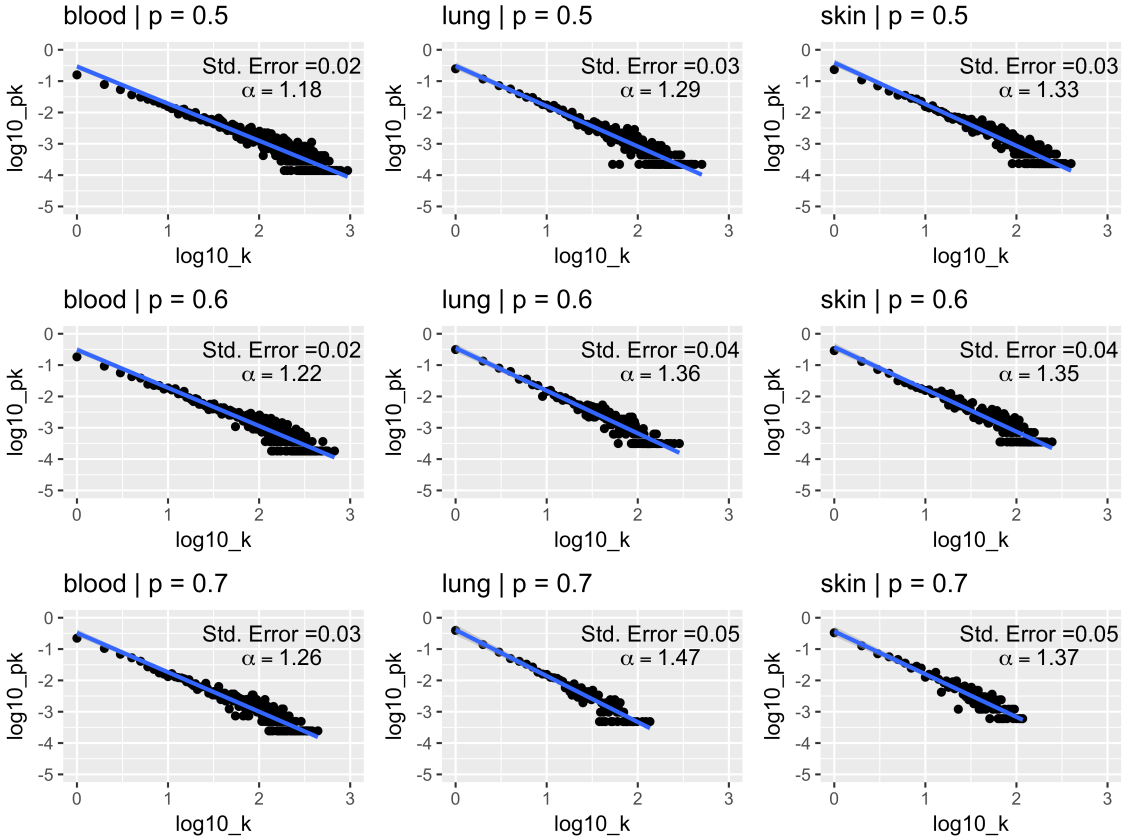

Figure 1: **Degree distribution of the tissue networks after different probability pruning.** The edges of the blood, lung, and skin GIANT tissue networks were pruned using different probability thresholds and then fit with a linear regression, so that alpha could be calculated. Pruning within the range above was shown not to dramatically affect the degree distributions, based on visual inspection

4.1.3 PageRank analysis

Genes are known to have various functions based on the context in which they're expressed or not expressed. To investigate the key genes across our GIANT tissue networks we performed PageRank analysis on each of the networks and ranked genes by their PageRank scores (Table 2).

Table 2: Top genes in each tissue by PageRank

blood	PgRnkSc	lung	PgRnkSc	skin	PgRnkSc
CSE1L	0.0064	BRCA1	0.0064	RPS6	0.0152
MRPL3	0.0058	PSMD13	0.0060	BRCA1	0.0125
DKC1	0.0051	RPL4	0.0058	PSMD14	0.0121
MCM3	0.0045	BUB1B	0.0056	RPL4	0.0104
TNFAIP3	0.0044	PSMD14	0.0054	MCM7	0.0100

PageRank worked well in this context and gave an intuitive result, wherein the majority of top genes for each tissue type are related to DNA replication and repair. DNA replication and repair are tightly regulated processes in cells across many tissue types. So the high PageRank scores of these genes are indicative of the fact that their protein products interact within many key hub genes as well as being tightly regulated themselves. It's notable that the two solid tissues, skin and lung share many of the same top ranked genes while blood has no overlap with them. PageRank could be a useful tool in the context of gene networks built from differential expression induced by specific conditions. As, based on the results above, we would expect PageRank to identify genes that are integral to the processes induced by the differential exposure.

4.1.4 PageRank sensitivity analysis

To investigate the sensitivity of the PageRank gene ranking to the probability threshold of the network, we tracked the top 100 genes by PageRank score as the probability used for pruning was increased (Figure 2). The initial 100 genes were the top 100 genes by PageRank score in the GIANT network pruned using >0.50 edge probability as the cutoff. The probability threshold was then incremented by 0.01 from 0.5 to 0.9 and the proportion of the top 100 genes that remained in the top 100 was tracked. The PageRank score demonstrates reasonable sensitivity to the probability threshold, with a threshold increase of 0.1 only removing $\approx 20\%$ of the top 100 genes. Note that our metric doesn't take into account where in the PageRank ranking those genes were, so those at the bottom may have been the most affected.

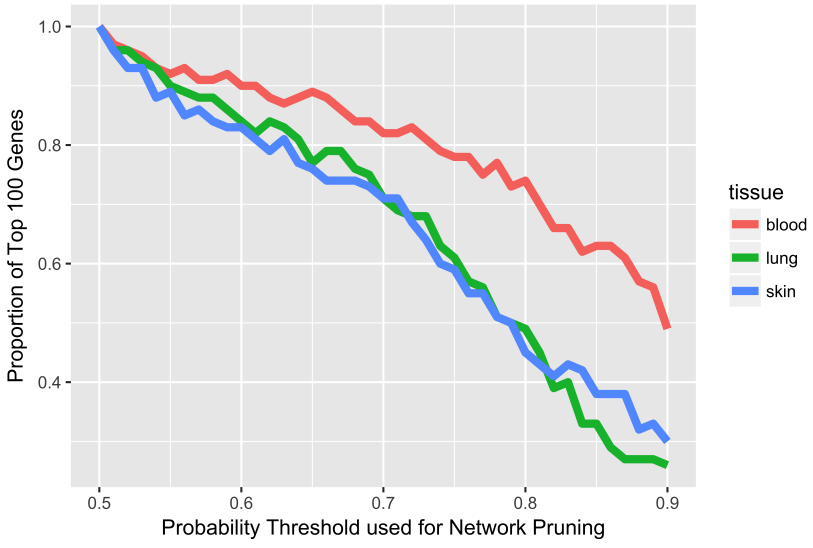

Figure 2: **Robustness of PageRank to pruning.** The top 100 genes for each tissue were ranked by PageRank score and their inclusion in the top 100 genes was tracked as each network was pruned using an increasing probability threshold.

4.2 MSigDB networks across tissues

In order to obtain greater granularity in understanding how specific pathways vary across tissues, we used a max-min hill climbing (MMHC) approach [11], described below, to assign directional edges to the 1329 gene sets from MSigDB using expression data from each of the individual tissues. This resulted in 4 networks per set, one for each tissue. We compared the number of edges discovered in each set across tissues, normalized to the number of possible edges that could be added to a network, defined as $2\binom{n}{2}$ with n as the number of genes in the set, and scaled by a factor of 2 as any given pair of genes could have an edge in either direction. We see consistency in the relative number of edges added for a given gene set across all tissues, with about 8 percent of possible edges appearing in a network (Table 3). Additionally, we assessed how unique those edges are for each network, so we also report the mean proportion of unique edges per network, normalized to the total number of edges in that network for that tissue. We find that across all tissues, the majority of edges are unique to that tissue and are not included in any other networks. We attempted to assess strongly connected components within these networks, but found none, which appears a limitation of the MMHC algorithm.

Table 3: Summary statistics over all MSigDB gene sets across 4 tissues

Tissue	Mean # edges	Mean # unique edges
Whole Blood	0.089 ± 0.052	0.766 ± 0.139
Lung	0.079 ± 0.045	0.720 ± 0.165
Skin Exposed	0.079 ± 0.044	0.680 ± 0.159
Skin Unexposed	0.071 ± 0.041	0.662 ± 0.168

We further assessed which of the gene sets had the most, and least, stable set of edges across all four networks. In order to reduce redundancy in the functions, we subset to only KEGG pathways for this analysis. We determined consistency by calculating the number of edges within a gene set that appeared in all four tissue networks, divided by the total number of unique edges present across all tissues. The top 10 most stable networks, ordered by this proportion, were found to primarily pertain to immune function while the least consistent were entirely disease-specific gene sets (Table 4). We note that, while this result is not unexpected given the relevance of immune functions to many tissues and specific diseases will have primary impact in the tissue they affect, we may be capturing more highly expressed genes in what appear to be more consistent networks, as increased noise is expected on the lower end of expression.

Table 4: Gene sets with the most and least consistent directed edges across all four tissues

Top 10 Stable Sets	**Description**
M13950	Asthma
M13103	Autoimmune thyroid disease
M18615	Allograft rejection
M13519	Graft-versus-host disease
M12617	Type I diabetes mellitus
M17946	Valine, leucine and isoleucine biosynthesis
M615	Intestinal immune network for IgA production
M12294	Viral myocarditis
M4085	Primary immunodeficiency
M7330	Glycosaminoglycan biosynthesis - heparan sulfate
Bottom 10 Stable Sets	**Description**
M13486	Huntington's disease
M16848	Epithelial cell signaling in Helicobacter pylori infection
M19877	Endometrial cancer
M1835	Glioma
M523	Thyroid cancer
M17807	Basal cell carcinoma
M15798	Melanoma
M19096	Bladder cancer
M19888	Acute myeloid leukemia
M19818	Non-small cell lung cancer
M16376	Arrhythmogenic right ventricular cardiomyopathy (ARVC)

Table 5: Top 5 gene sets by mean PageRank score across tissues

Top Blood Sets	**Description**
M16853	DNA replication
M12039	One carbon pool by folate
M13515	Mismatch repair
M10680	Proteasome
M13519	Graft-versus-host disease
Top Lung Sets	**Description**
M16853	DNA replication
M12039	One carbon pool by folate
M13515	Mismatch repair
M10680	Proteasome
M13519	Graft-versus-host disease
Top Skin Sets	**Description**
M12039	One carbon pool by folate
M16853	DNA replication
M13515	Mismatch repair
M16894	Complement and coagulation cascades
M7098	ECM-receptor interaction

We then looked at genes with high PageRank scores in the tissue-wide networks to understand how they interacted within functional networks. We hypothesized that these genes are likely to play a central role in a directed causal network. We selected the top 10 genes by PageRank score in each tissue and then filtered for MSigDB gene sets that contain those genes and took the gene set that appeared most frequently in each tissue. This gene set wound up being the same for both blood and skin, which was annotated as "Genes involved in Immune System", while the top network for lung is "Genes involved in Cell Cycle". Considering the ubiquitous nature of both the immune system and cell cycle throughout tissues in the body, it follows that genes involved in these functions would have higher PageRank scores. We then assessed the mean PageRank score per KEGG gene set (Table 5). We find that the top sets are the same for both lung and blood, suggesting that gene regulation is potentially more conserved between these two tissues, as the same genes have high influence, but there is also overlap with skin, with the exception of the last two top sets, where we see functions that may be more skin-specific.

4.3 MSigDB to GIANT comparison

We investigated the consistency of the evidence for edges between the MSigDB networks, derived from GTEx, and the GIANT networks, derived from Gene Expression Omnibus (GEO data). For each tissue, we found the size of the intersection of edges in each MSigDB network versus those in the GIANT networks pruned at 0.5 edge probability. Edges in both of these networks are the result of amount of evidence found in their respective data sources, although the method used to construct those edges was slightly different with MMHC being more conservative in edge creation. For each MSigDB network, the size of the MSigDB-GIANT edge set intersection, the size of the MSigDB edge set, and the size of the GIANT tissue edge set were used to calculate p-values using a hyper-geometric distribution. The most significant MSigDB sets in each tissue were generally related to core cellular processes, such as metabolism or DNA replication. The inter-tissue similarity was evaluated by comparing the significant MSigDB sets between the different tissues (Figure 3). Blood had the highest number of significant gene sets and also the largest amount of pairwise overlap with the other tissues. The two skin networks had higher overlap with each other than with lung, but overlapped the most with blood. The MSigDB sets present in only the skin tissues were related to skin molecular attributes such as sweat production and tight junctions. Future work would include further investigation of these intersections.

4.4 Comparison of networks between healthy and cancer tissues

We then repeated the MMHC algorithm for network construction across KEGG gene sets using cancer expression data. We compared the degree distribution across all KEGG gene sets in tissue networks from GTEx vs. TCGA (Figure 4). We expect consistency in the distribution given what

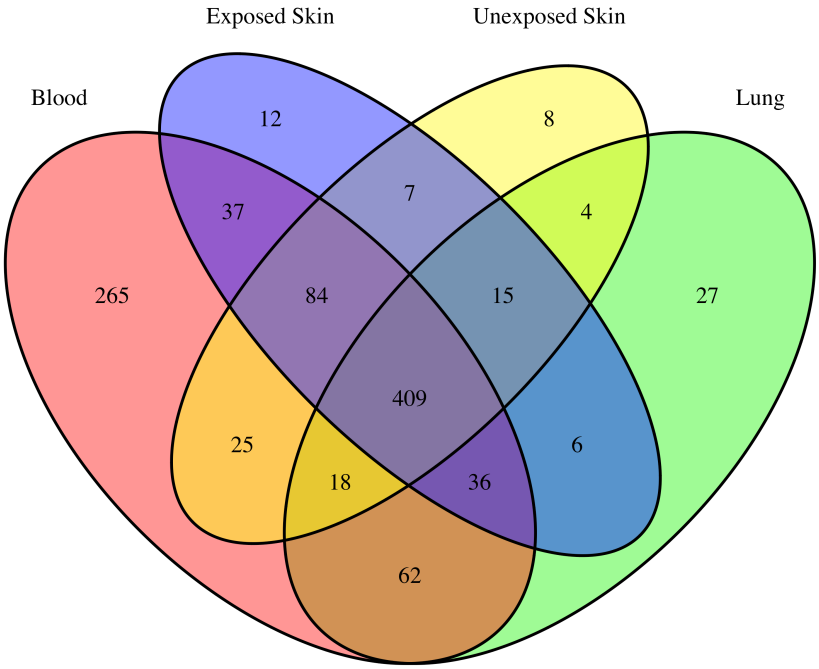

Figure 3: **Significant MSigDB gene set overlap** Venn diagram showing the number of significantly edge enriched MSigDB gene sets that overlap between each of the MSigDB tissues. Significance was determined based on overlap between directional edges in the MSigDB network and the associated GIANT tissue network. For example, the 409 indicates that 409 MSigDB gene sets were significantly enriched in all 5 tissues when looking for evidence of an edge in both GTEx and GIANT.

has been shown in biological networks. We estimate a value for the slope of the degree distribution on the log-log scale for both the disease and health networks by fitting a linear model and find the slope to be consistently around -2 for all tissues, ranging from -1.93 (TCGA blood) to -2.15 (GTEx exposed skin).

We do not see a difference in the degree distribution between the datasets, and so we next looked at which networks were most altered when learned from healthy vs. disease expression data. We assessed the edge overlap in all KEGG networks between the GTEx networks and the TCGA networks in matched tissues (Figure 5).

For each tissue, there were around 13 gene sets that saw no overlap between healthy and disease networks, and this increased to 16 sets for unexposed skin. We had hypothesized that exposed skin would have more in common with skin cancer tissue than unexposed skin, and we find some support for this, as both the average and maximum overlap between edges was higher for the network trained on data from exposed skin than unexposed skin. Because many networks had little overlap between the two datasets, we choose two networks to manually inspect based on their annotated involvement with cancer, with include gene set M12868 (Pathways in Cancer) and M19818 (Non-small cell lung cancer). We examine M12868 in blood and M19818 in lung (Figure 6). For M12868, we report clustering coefficients of 0.023 and 0.028 for healthy and disease, respectively, and average node betweenness of 485.10 and 126.38. For M19818, we report 0.036 and 0.076 for clustering coefficients, and 36.59 and 27.44 average node betweenness, again for healthy and disease respectively.

5 Discussion

A network based approach can be useful in elucidating differences in expression between tissues of varying origins. Upon our examination of the GIANT tissue specific networks, we found their scale-free structure and PageRank (PR) scores to be robust to the selection of posterior probability for edge retention. This is a comforting finding, as when this cutoff is select in many research applications the cutoff is selected using some heuristic. The PR evaluation of the networks indicated

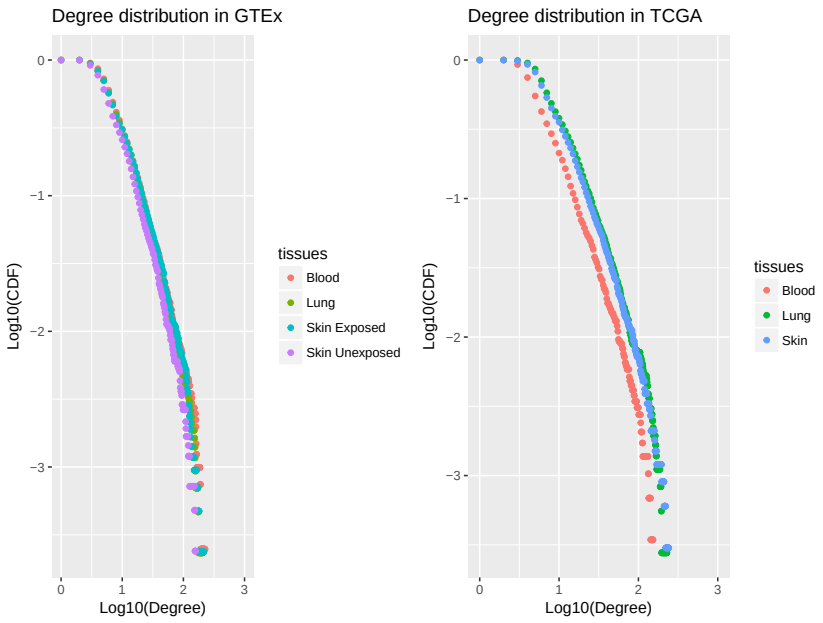

Figure 4: **Degree distribution of nodes across all networks from GTEx vs. TCGA data.** The degree of each node in each network in GTEx and TCGA were combined and plotted above on a log-log scale for each tissue. The x-axis represents the degree and the y-axis is the cumulative probability.

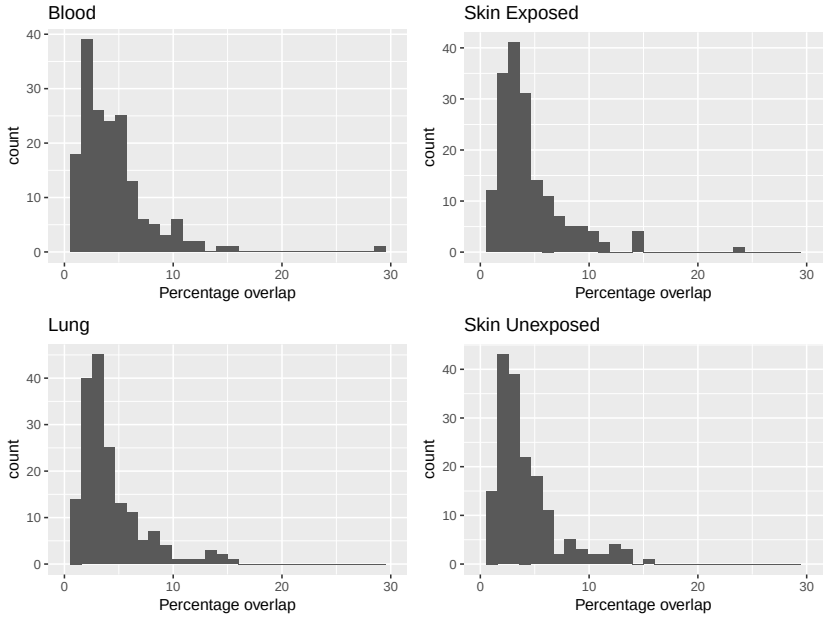

Figure 5: **Proportion of edges per gene set appearing in networks derived from healthy vs. cancer data.** Histogram of the percentage of edge overlap in all KEGG networks learned from healthy (GTEx) vs. disease (TCGA) expression data.

similar genes in both of the solid tissues (lung and skin), while different genes were indicated in the blood PR. It's possible this is the result of our evaluation only consisting of the top 10 or is the result of real differences in the networks of fluid versus solid tissues. The majority of the genes in both cases were related to DNA replication and repair, a heavily regulated cellular process. This makes sense upon consideration of how the PR algorithm works. When assigning directionality to edges, we find that the majority of possible connections within predefined gene sets are not present, as only 7 percent of possible edges are added, indicating that there are a few strong relationships

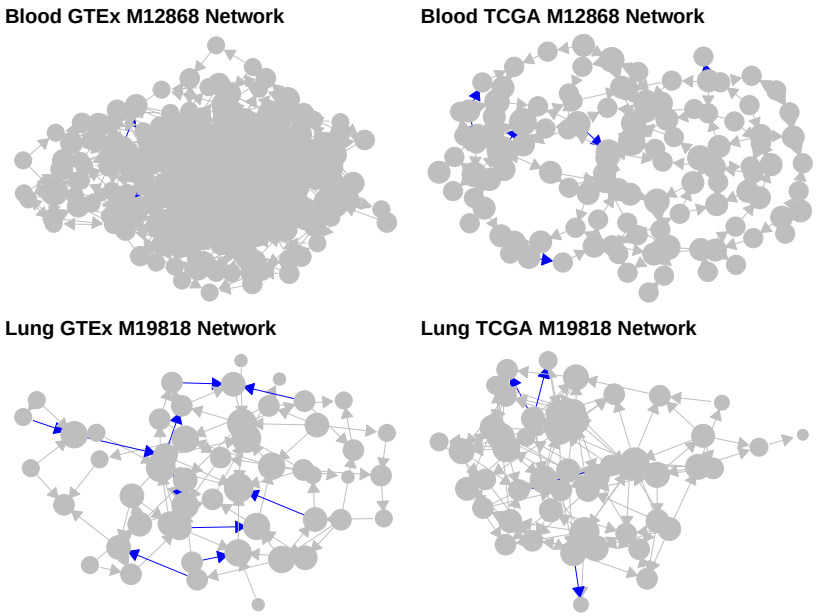

Figure 6: **Comparison of cancer-specific directed networks between healthy and disease.** Directed networks for pecific cancer pathways M12868 and M19818 in healthy and disease tissues, with shared edges highlighted in blue. For M12868, nodes with degree < 3 are removed for visibility. Some edges are obscured in areas of high connectivity.

potentially driving the downstream function associated with that module, or that expression is not the correct modality to capture additional relationships. We would expect the integration of additional molecular data, such as protein abundance, to better inform causal gene regulatory relationships. We examined genes with high PageRank scores from a network that considered all genes within a specific tissue, and find that these genes are associated with modules that relate to basic cellular function, such as DNA replication and mismatch repair, and see high overlap between tissues, indicating that PageRank score could be a useful metric for a gene's impact on cellular function. Finally, when comparing networks between healthy and disease tissue, we find no change in the degree distribution across networks, and thus expect similar connectivity in healthy and disease. However, when we further examine two gene sets that are annotated as associated with cancer, one set, pathways in cancer, shows less connectivity in the blood network derived from TCGA data,as it has a lower average betweenness as compared to the GTEx network for this module, indicating fewer paths between nodes, though the clustering coefficient is roughly equal. Whereas when we examine a module associated with lung cancer, we see around the same amount of overlap in edges, but also a similar level of connectivity, as measured by both the clustering coefficients and node betweenness of the networks. While the difference in connectivity could be an artifact of the different datasets, given that we see no change in the degree distributions between the datasets, the genes in the M12868 module, while already associated with cancer, could represent an area of future work in specifically assessing changes in regulatory relationships between these genes.

6 Methods

6.1 Max min hill climbing

Max min hill climbing is an algorithm for learning a Bayesian network structure [11]. Given a set of nodes, the algorithm learns the skeleton of the network and then uses a greedy hill-climbing search to determine edge direction. When learning the structure, a Bayesian approach requires that any variable be independent of any subset of non-descendant variables when conditioning on its parents. Overall, we assess the conditional independence of variables in order to decide if a relationships exists. The MMHC algorithm employs a related algorithm, max min parents and children (MMPC) to learn the structure. The algorithm for MMPC is shown below, as presented in [11]. It uses a heuristic to add connections to a node, which selects an external variable to connect

to the current node which maximizes the minimum association with the target node relative to its parents and children. After running MMPC for each variable, MMHC will perform a greedy search to find the highest-scoring directed acyclic graph (DAG) for the data, which optimizes the Bayesian Dirichlet equivalent uniform, or BDeu score [4]. This score seeks to maximize the posterior probability of a graph given the input data and assumes a uniform prior over all possible DAGs. We used the bnlearn R package [12].

Algorithm 1: `MMPC`

Input: target variable T; data D
Output: parents and children of T
CPC = 0
while *CPC unchanged* **do**
  [F, *assocF*] = MaxMinHeuristic(T;CPC)
  if *assocF* $\neq$ *0* **then**
    CPC = CPC $\cup$ F
for $X \in CPC$ **do**
  if $\exists S \subseteq CPC$, *s.t. Ind(X;T|S)* **then**
    CPC = CPC $\setminus$ {X}
return (CPC)

Algorithm 2: `MaxMinHeuristic`

Input: target variable T; CPC
Output: max value and variable associated with T
$assocF = max_{X \in V} MinAssoc(X; T \mid CPC)$
F = $argmax_{X \in V} MinAssoc(X; T \mid CPC)$
return (F, *assocF*)

Algorithm 3: `MMHC`

Input: data D
Output: DAG of variables in D
for *variable* $X \in D$ **do**
  PC_X = MMPC(X,D)
//Perform greedy hill-climb search from empty graph
//Only add edge Y $\rightarrow$ X if Y $\in PC_X$
return (highest-scoring DAG)

References

[1] Greene et al. Understanding multicellular function and disease with human tissue-specific networks. *Nature genetics*, 47(6):569–576, jun 2015.

[2] Ashis Saha, Yungil Kim, Ariel D H Gewirtz, Brian Jo, Chuan Gao, Ian C McDowell, Barbara E Engelhardt, and Alexis Battle. Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome research*, oct 2017.

[3] Elias Chaibub Neto, Aimee T Broman, Mark P Keller, Alan D Attie, Bin Zhang, Jun Zhu, and Brian S Yandell. Modeling Causality for Pairs of Phenotypes in System Genetics. *Genetics*, 193(3):1003–1013, mar 2013.

[4] D Heckerman, D Geiger, and D M Chickering. Learning Bayesian Networks - the Combination of Knowledge and Statistical-Data. *Machine Learning*, 1995.

[5] Md. Bahadur Badsha and Audrey Q Fu. Learning causal biological networks with generalized Mendelian randomization. *bioRxiv*, aug 2017.

[6] The GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nature genetics*, 45(6):580–585, jun 2013.

[7] Xin Li, Yungil Kim, Emily K. Tsang, ... Hall Ira M. Davis, Joe R., Alexis Battle, and Stephen B. Montgomery. The impact of rare variation on gene expression across tissues. *Nature*, 2017.

[8] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

[9] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[10] Daniel M. Busiello, Samir Suweis, Jorge Hidalgo, and Amos Maritan. Explorability and the origin of Network Sparsity in Living Systems. *Scientific Reports*, (September):1–8, 2016.

[11] Ioannis Tsamardinos, Laura Brown, and Constantin Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 2006.

[12] Marco Scutari. Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22, 2010.