CS 224W Project Final
Network analysis of Disease Modules using protein based interactions

Margaret Guo, Shawn Xu, Aditya Oza

Abstract— **Within network biology, a disease module is defined as a connected subgraph that contains all molecular determinants of a disease. Many network algorithms attempt to discover these disease models, in hopes of enhancing pathophysiological understanding and therapeutic treatment for any disease of interest. However, biological interactions are often complex, and most disease module detecting algorithms are based off graphs are unsigned, undirected networks that provide limited binary interaction data between two biological entities. In order to increase biological insight into disease mechanisms, we hypothesize that by using richer signed protein-protein interaction data, we can capture more meaningful protein behaviors. Our project builds upon the DIAMOnD algorithm by incorporating it with signed edges which can indicate a diseased versus physiologic state. We identify disease modules using a signed link significance metric based approach for community detection. The validity of our predicted disease modules can be verified against online disease databases. We analyze our algorithm by detecting communities associated with specific diseases such as cancer, cardiomyopathy, and neurological disorders. We discover overlap of disease communities with similar gene characterizations and present case studies for understanding the mechanism of ALS and exploring tissue specific differences in the setting of KRAS-oncogene mutated cancer. Thus, through this project we build, validate, and apply a novel algorithm that can predict disease modules from a signed protein-protein interaction network.**

I. Introduction

The set of all interactions that happen in a human cell is known as human interactome. While, several previous studies have been done on interactome of model organisms - bacteria like E.coli and S.cerevisiae [1], its only in the last decade or so that human interactome have begun to be studied widely. The human interactome contains all forms of biological interactions between permutations of DNA, RNA, and protein. When they function together under physiologic conditions they represent a normal cellular network. However, perturbations within these networks at major hubs (i.e. essential genes) can lead to disruption of entire interactome and can represent the pathogenesis of disease, which is represented in graph theory as a disease module or cluster. Additionally the relationship between diseases can be modeled. The more localized a single disease module is, the higher the similarity of associated genes. Likewise, the shorter the inter-disease module distance metric is the more biologically similar (higher correlated disease symptoms, co-morbidities, etc) two diseases are to one another [2]. Thus, analysis of biological networks can offer deeper insight into disease mechanisms for potential therapeutic and diagnostic advances.

However, biological understanding of disease mechanisms has often been simplified when applied to network theory due to an inability to capture all complex associations between one protein or another. As a result, many of the major protein-protein interaction databases, such as BioGrid [3] and String DB [4] have relegated associations into undirected, unsigned interactions networks. These edges will connect two nodes but provide little information on the nature of the biological interaction. More recently, networks have attempted to add in directionality and created multi-label graphlets that can describe various functionalities of interaction in order to better understand network motifs [5]. Yet, little work has been done to further use and analyze these subgraph motifs to uncover novel disease mechanisms.

In our project, we attempt to understand the paradigm of disease modules in terms of their regulatory and physical interactions by building upon previously existing PPI graphs and augmenting them with signed, directed edges which indicate degrees of positive or negative coexpression also with a directionality of impact. Once we build the network, we define and determine our seed genes that have validated association with a particular disease of interest. Next, we compute the disease modules using a modified DIAMOnD algorithm that incorporates signed link sig-

nificance. We compare the disease module predictions to a gold standard through literature-based validation methods include gene-disease z-scores and significant GO terms. We report the ability of our algorithm to accurately predict disease modules, discover potential new genes of interest, and evaluate mechanistic similarities between different diseases. We hypothesize that by using the much richer signed protein-protein interaction data, we can capture more meaningful protein behaviors.

II. Prior Work

The field of network medicine is an area of growing interest. Barabasi, et. al (2011) [1] states that network modeling of cellular interactions can help identify disease genes, predict disease prognosis and help determine treatment options for certain diseases. Studies presented within this paper form a body of knowledge called network medicine that offers a platform to study genetic human diseases. Additionally, principles of network theory like degree distribution, hubs ,Small World phenomena , modules and Betweenness centrality have been directly applied to biological networks to study their structure.

There have subsequently been algorithms that have focused on detecting disease modules. The DIAMOnD algorithm [9] used three different methods: a link community algorithm based on link similarity and that can capture hierarchical communities, a Louvain method that finds a heuristic to cluster based on modularity, and a Markov Cluster Algorithm that detects dense regions based on random flow to detect communities within a protein-protein interaction (PPI) network. The authors found that by looking at whether a protein has more connections to seed proteins than expected, they could find a significantly different behavior of known disease proteins compared to expected for randomly distributed proteins. As expected, the efficacy of disease module prediction is heavily dependent on the quality of the data (the number of seed proteins that we know). Studies have shown that there must be at least 25 disease-associated genes available in order for a network to able to be able to uncover these disease modules.

Additional studies [2] have used multi-label graphs annotated with different possible interaction types as graphlets, defined as size-2 and size-3 subgraphs that have a specific edge-set hash, and determined that certain graphlets appeared significantly more times in the real network versus a random network.

III. Algorithms and Mathematical Background

A. Community Detection Algorithm

The DIAMOnD paper [9] discovered that due to the sparsity of disease protein subgraphs, one cannot use modularity to find disease modules. Instead, the distinct and predictive patterns of disease proteins can be captured using the idea of link significance. We calculate the significance of these edges with respect to the proteins known to be associated with a disease, i.e. the seeded proteins. The DIAMOnD algorithm describes how to calculate this significance for an undirected graph, where the number of connections of a protein to a seeded disease protein is compared to the expected number of connections for randomly scattered seed proteins:

$$p(k,k_s,s_0) = \frac{\binom{s_0}{k_s}\binom{N-s_0}{k-k_s}}{\binom{N}{k}}$$

$$p-value(k,k_s,s_0) = \sum_{k_i=k_s}^{k} p(k,k_i,s_0)$$

In the above equations, k represents the degree of the protein, k_s is the number of links to seeded proteins, and so is the number of seeded proteins. $p(k,k_s,s_0)$ is the hypergeometric function representing the probability of a node having k_s links to seed proteins out of k total links.

In order to extend the link significance evaluation to a network with signed edges, we make the assumption that a node having positive edges and negative edges are independent events. Then, we can calculate the positive and negative link significance of a node with k^+ positive edges, k_s^+ positive edges to seed nodes, k^- negative edges, and k_s^- negative edges to seed nodes as:

$$p-value^+(k^+,k_s^+,s_0) = \sum_{k_i=k_s^+}^{k+} p(k^+,k_i,s_0)$$

$$p-value^-(k^-,k_s^-,s_0) = \sum_{k_i=k_s^-}^{k-} p(k^-,k_i,s_0)$$

In order to compute the combined p-value, we apply Fisher's method to compute the χ^2 statistic, where $\chi^2 = -2\cdot(\ln\text{p-value}^+ + \ln\text{p-value}^-)$. Then, the combined p-value can be determined from the χ^2 statistic.

We can then follow the same idea of the DIAMOnD algorithm, which iteratively takes the proteins with the highest link significance as the predicted disease

module. Refer to Algorithm 1 for the base DIAMOnD algorithm. For the signed links extension of the DIAMOnD algorithm, the computation for p_i changes to first computing both the positive and negative link significance for each node, then computing the combined p-value using Fisher's method as described above. The rest of the algorithm remains the same.

One drawback of the DIAMOND algorithm is that it does not explicitly terminate. We can deal with this either by explicitly setting the size N of the disease module we wish to obtain, or by employing a statistical signficance cutoff threshold, which we can determine using the Benjamini-Hochberg test, or some similar test.

Algorithm 1 DIAMOND Algorithm

1: Let S be the initial set of seed nodes
2: $p_i \leftarrow p - value(k_i, k_{s,i}, s_0)$ for all nodes i
3: Sort p_i in ascending order
4: $c = \text{argmin}(p_i)$
5: $S \leftarrow S \cup \{i\}$
6: $s_0 \leftarrow s_0 + 1$
7: Repeat 2 through 6 until $\min(p_i)$ is no longer significant

B. Validation

One of the most challenging steps in the analysis of biological networks is validating the findings within a biologically feasible context. Here, we extract associated diseases using the Pharos API [11] and additionally describe our findings using a GO ontology [12] [13]. The Pharos API calculates a z- score [15]. indicating the confidence of a gene-disease association based on the DISEASES methodology [14] which uses text mining of biomedical articles combined with manually compiled comprehensive dictionaries for human gene names, disease names, and their synonyms to find relationships. All significant diseases related to a gene of interest are listed with the calculated z scores. For each algorithm run, we define Z_d for a diseases d to be the total sum of significant z-scores (z) across all genes ($g = i \ldots n$) within our disease modules averaged over the size of our disease module (n) set.

$$Z_d = \frac{\sum_{g=i}^{n} z_g}{n}$$

This number can be interpreted as a proxy for disease relevance to a particular community of genes within the protein-protein network. The significance cutoff for Z_d to be reported is defined as $Z_d^* = \frac{2}{\sqrt{n}}$, or two standardized standard deviations away from null. Because this value is skewed based on the amount of literature research on the particular disease–for instance, a relatively unknown disease will not have as many disease-gene relationships captured by Pharos–we use a bootstrap methods to find find $\bar{Z}_d^{rand}$. To do so, we repeatedly select n random genes to include in a random "disease" module and calculate Z_d^{rand} for that iteration, then average over multiple runs to find our overall baseline set. We then find and report a normalized aggregate z-score, $\tilde{Z}_d$ defined as:

$$\tilde{Z}_d = \frac{Z_d}{\bar{Z}_d^{rand}}$$

All our analysis and interpretation is based off of this normalized value.

Additionally, gene ontology (GO) terms provide a consistent descriptions of gene products that is applicable across databases, species, and experimental protocols. Thus, these terms have been used as standard terminology by biologists for describing biological phenomenon. We use the procedure defined by the Gene Ontology Consortium [13] to determine and report significant GO terms in each disease module.

IV. Data

We have collected data from Pathway Commons 9 which includes 31,693 human proteins and a total of 2264289 edges. These edges include data from Reactome, HPRD, BioGRID, and includes undirected and directed labels between proteins. Table 1 shows the types of interactions given in the network along with their assigned edge type (undirected or directed) and frequency within the graph. For instance, interacts-with label is an undirected association, while controls-expression-of is a directed node. The correlation between genes was calculated from the CoExpressDb database [10] which notes gene correlation. We define a threshold value p by which we can bin the edges into three categories as: +1 (positive interaction), -1 (negative interaction), and 0 (no known interaction). For research purposes, we have used several different threshold levels to determine which one optimizes the balance between sensitivity to biological relationship and specificity to avoid noise. Thus, we have build our model system to perform the next step of disease detection.

To find the seed disease genes for each phenotype. We used OMIM (Online Mendelian Inheritance

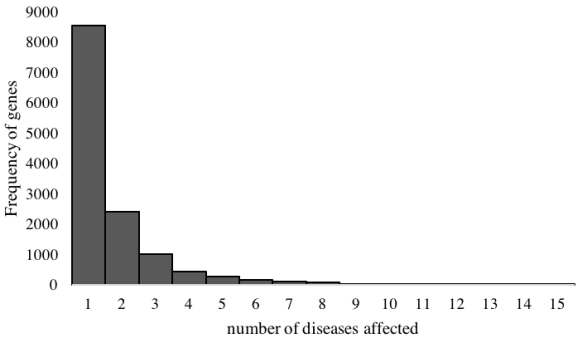

Fig. 1. Frequency distribution of the number of diseases each gene is associated with. Only the genes that have at least one disease-associated gene are shown.

in Man database to extract all genes associated with the phenotypes listed. There are 7426 phenotypes and 13070 associated disease genes. We then counted the number of phenotypes each gene was associated with– a proxy for out-degree of these genes towards diseases of interest–follows a roughly exponential frequency distribution 1.

Dataset Statistics	
Nodes	31693
Edges	2264289
Unique directed edges	2071656
Unique undirected edges	1385672
Zero In degree nodes	5562
Zero Out degree nodes	3469
largest Wrr size	0.9995
largest Srr size	0.7137
Nodes in largest Wrr	31679
Edges in largest Wrr	2264280
Nodes in largest Srr	22620
Edges in largest Srr	2068408
Average clustering coefficient	0.0330
Diameter	7
90% effective diameter	2.9919

TABLE I
PROTEIN-PROTEIN NETWORK STATISTICS

A total of 79 sets of seed genes were used, including those for cancer, cardiomyopathy, amyotrophic lateral sclerosis (ALS), multiple sclerosis (MS), diabetes mellitus (DM), Parkinson's Disease (PD), Alzheimer's Disease (AD), and schizophrenia. These seed genes are derived from the OMIM database. In addition, cancer is further subdivided based on gene expression data from the oncogenic signatures from MolSigDB [16]. MolSigDb contains gene sets that represent significant cellular pathways that are often disregulated in cancer. These gene sets are based on microarray data from NCBI GEO and unpublished databases from the NIH that involve perturbing known oncogenes or tumor suppressor genes implicated in cancer pathogenesis. A list of seed genes and a description of their origin can be found in the Appendix.

V. RESULTS

A. PPI shows Small World behavior

We list the protein-protein network statistics in Table II. We also show a log-log plot detailing the distribution of in degree and out degree of nodes V-A. We see that the degree distribution is roughly linear (or follows a power-law distribution) up to a degree of 10^2 in or out degrees and then falls off. We also note that this network is relatively well connected with the largest weakly connected component size covering 99.95% of the nodes and the largest strongly connected component with 71.37% of the nodes. There are also about 25% of the nodes that are either IN-TENDRILS or OUT-TENDRILS.

90% effective diameter is very small - 2.99 - which shows ultra Small World behavior in this network. Strong connectivity and short diameters also indicate that connectivity significance becomes more important over density of connections in studying protein behavior [1]. This also matches the concept of biological networks in which there are hub proteins, or those with the highest degree, that can represent the most essential genes in the disease network.

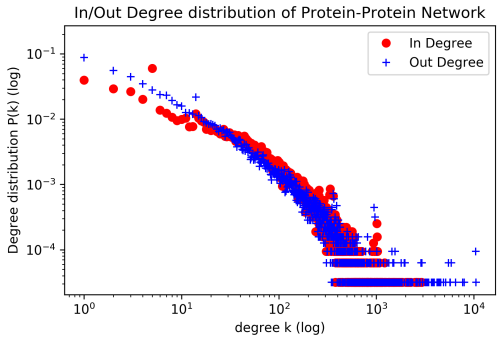

Fig. 2. A log-log plot indicating the degree distribution of our network

B. Predicted diseases with different starting seed nodes

After we created the graphical PPI network, we implemented the DIAMOND algorithm and generated lists of potential target or "hit" gene based on various sets of seed nodes (more details in Appendix A).

Our initial run was through all potential disease genes from the network for a total of 6086 seed genes and extracted a list of predicted genes based on the entire subset. Our initial findings show that the p-values of the chosen disease genes are extremely low (highly significant). The average p-value for the entire graph is around 0.27 while the p-values of the chosen disease nodes are less than 10^{-10}. We note that of these 6086 genes defined as disease-related by OMIM, 5680 of the genes are found in our network. We find that our network most significantly contains diseases as listed in II. A simple explanation for these diseases appearing signficant is that these are the diseases with the highest incidence within the PHAROS dataset. However, we do note that these diseases often cause systemic affects that may affect multiple genes, and in a network sense, impact all parts of the graph. Therefore we keep this in mind while we explore potentially more interesting biological questions.

TABLE II

NORMALIZED Z-SCORES FOR THE SIGNIFICANT GENE MODULES CREATED BY A BROAD SEED OF ALL KNOWN DISEASES GENES

Disease	$\tilde{Z}_d$
Diabetes mellitus	3.32
Obesity	3.15
Kidney disease	2.89
Cancer	2.87
Asthma	2.70
Hypertension	2.52

We next precede to analyze the disease-specific communities namely for: cancer (CA), cardiomyopathy (CM), amyotrophic lateral sclerosis (ALS), immunodeficiency (ID), diabetes mellitus (DM), multiple sclerosis (MS), Parkinson's Disease (PD), Alzheimer's Disease (AD), and schizophrenia (SCZ). We first evaluate the validity of the additional gene hits we include in our community cluster Table III. We show that for all diseases the significant hit genes derived from the algorithm still led to a significant normalized z-score; however the magnitude is much smaller without the seed nodes than with the original nodes. Intuitively, this fits with setup of this model. The genes with the most significant and strongest association to the disease, are labeled as seed genes. These genes often have a critical role in the pathogenesis of the disease. For instance, p53 and ras are known cancer-related genes [17], and have been prolifically studied in the scientific community such that they have been included in clinical databases such as OMIM. The significant

TABLE III

IMPACT OF THE SEED NODES WHEN CALCULATING DISEASE SPECIFIC ASSOCIATIONS FOR EACH OF THE 8 DISEASES. IMPACT OF SEED NODES IS DEFINED AS THE PERCENT CHANGE OF THE Z SCORES UPON ADDING THE SEED NODES TO BE EVALUATED.

Disease	Z_{seed}	$Z_{seed+pred}$	Z_{pred}	% Impact
AD	301.32	21.2	18.85	0.110849057
ALS	492.355	25.18	3.69	0.853455123
CA	12.68	7.39	6.24	0.155615697
CM	215.8	18.65	6.23	0.665951743
ID	111.31	26.64	15.98	0.40015015
MS	53.53	7.1	6.74	0.050704225
PD	246.53	8.05	4.23	0.474534161
SCZ	102.02	13.11	9.72	0.258581236

genes from our network can be interpreted as genes that have similar patterns of gene expression compared to the seed nodes. Thus, these "hit" genes may also be implicated in disease.

As a significant negative, we note that our community detection algorithm for diabetes mellitus (DM), yielded a single signficant gene hit, A1BG, which is a glycoprotein with an unknown function. This is because none of the 27 genes listed on OMIM were found in the network. Future work will be done to optimize the mapping of a gene name to a universal identifier for easier analyses.

C. Algorithm shows significant disease-disease relationships

A benefit of the normalize z-score is that it allows us to discover diseases that share a common risk gene pool set. Simply stated, if there is a strong degree of overlap between gene targets for both disease A and B, then we might say that disease A and disease B could potentially have similar mechanisms of pathogenesis. This is potentially useful for repurposing drugs indicated for a single disease for diseases that share its target gene space. We have attached the significant disease-disease relationships in the Appendix: Section B. We will be analyzing the hits for ALS and cancer in a future section.

Additionally, these gene target characteristics can be observed by measured GO terminology (results not shown but see supplementary files). Previous methods, such as the DIAMOND algorithm make extensive use of GO terminology as part of the validation, we have found that GO terms often lack clear and concise interpretability. For instance, the top three GO process hits for cancer include the terms: "signal transduction",

TABLE IV

MULTIPLE NEUROLOGICAL DISEASE $\tilde{Z}_d$ COMPARISONS

	Alone	ALS, AD, PD	ALS, AD, MS, PD
AD	21.2	18.89	11.27
ALS	25.19	27.2	22.86
MS	7.11	12.22	-
PD	8.05	12.97	4.35

"MAPK cascade" (MAPK is a particular signal amplification pathway), and "positive regulation of cell proliferation." While the last term is a hallmark of cancer progression, the first two terms apply to a broad spectrum of pathways and targets. Thus, we choose to note interesting GO terms that appear significant in our algorithm output gene list without systematically analyzing them.

D. Algorithm is able to distinguish multiple disease centers

Next we test the ability of our community detection algorithm to be able to detect multiple diseases when seed nodes for difference diseases are combined for the input to our algorithm. Here, we explore two combinations: the combination of neurodegenerative diseases AD, ALS, and PD as well as the neurodegenerative disases and MS (a neuro motor disease with autoimmune pathological origins). Table IV shows the normalized Z-scores corresponding to each of the test cases. We note that the combination of ALS, AD, and PD neurodegenerative diseases has little effect on the Z-score. However, the addition of MS seed nodes to the algorithm initialization decreases the strength of the signal. We hypothesize that this is due to the fact that there are two disease clusters of seed nodes within the PathwayCommons PPI graph, and that generating communities based on two disparate loci will decrease the significance and strength of the clustering.

E. A case study with ALS

Now, as we have outlined the basic process of running the algorithm and collecting both normalized Z scores and significant GO processes, we delve into our first case example involving amyotrophic lateral sclerosis (ALS). ALS is a neurodegenerative disease that results in destruction of motor neurons that is reflected by progressive muscular paralysis [18]. Though the mechanism of pathogenesis is not completely known, from the OMIM database we find that, the seed nodes FUS [19], TARDBP [20], PFN1 [21] have been mechanistically proven to be involved in protein aggregation that eventually leads to motor neuron death.

TABLE V

ALS DISEASE MODULE SUBGRAPH STATISTICS

number of nodes:	501
number of edges:	25190
average node degree:	101
avg. local clustering coefficient:	0.76

TABLE VI

ALS DISEASE MODULE SIGNIFICANT GO PROCESSES

pathway ID	**pathway description**	**genes**	**FDR**
GO:0000398	mRNA splicing, via spliceosome	128	2.35e-152
GO:0000278	mitotic cell cycle	199	1.87e-151
GO:1903047	mitotic cell cycle process	183	4.2e-140
GO:0008380	RNA splicing	140	4.53e-139
GO:0006397	mRNA processing	143	1.08e-130

These genes are included in the 82 ALS-relevant genes that initialize the network algorithm. We note that, although we discovered 82 relevant genes from OMIM, only 23, including FUS, TARDBP, and PFN1 were discovered in the network. These genes then went on the yield 500 significant gene target hits, a list of which is attached in supplemental materials. These 500 gene hits include genes such as TBK1, or TANK-1 binding kinase which is involved in innate immunity signaling pathways and has recently been implicated in neuroinflammation that occurs during ALS [22]. Additionally there are several RAB proteins within the list that are GTPase-family member proteins involved axonal transport that is often disrupted in neurodegenerative diseases such as ALS [23]. A visualization of the disease module (Fig.4), as generated by StringDB [24], along with a statistical description of the the graphical properties of the connected graph can be seen. We observe that there seems to be a directional shift of our algorithm hits from the seed nodes, indicating that these new nodes represent different biological processes than our start nodes.

We additionally can see, from our significant go terms (Table VI and Fig. 3) that RNA splicing and membrane trafficking processes are upregulated.

When locking at the normalized Z-scores (see Appendix B). We see that ALS appears as the 4th most significant disease, and that many closely related neurological conditions, such as Joubert syndrome and Frontotemporal dementia appear on this list, indicating that ALS may have some mechanistic similarities with in cerebellar dysfunction and dementia. Interesting to

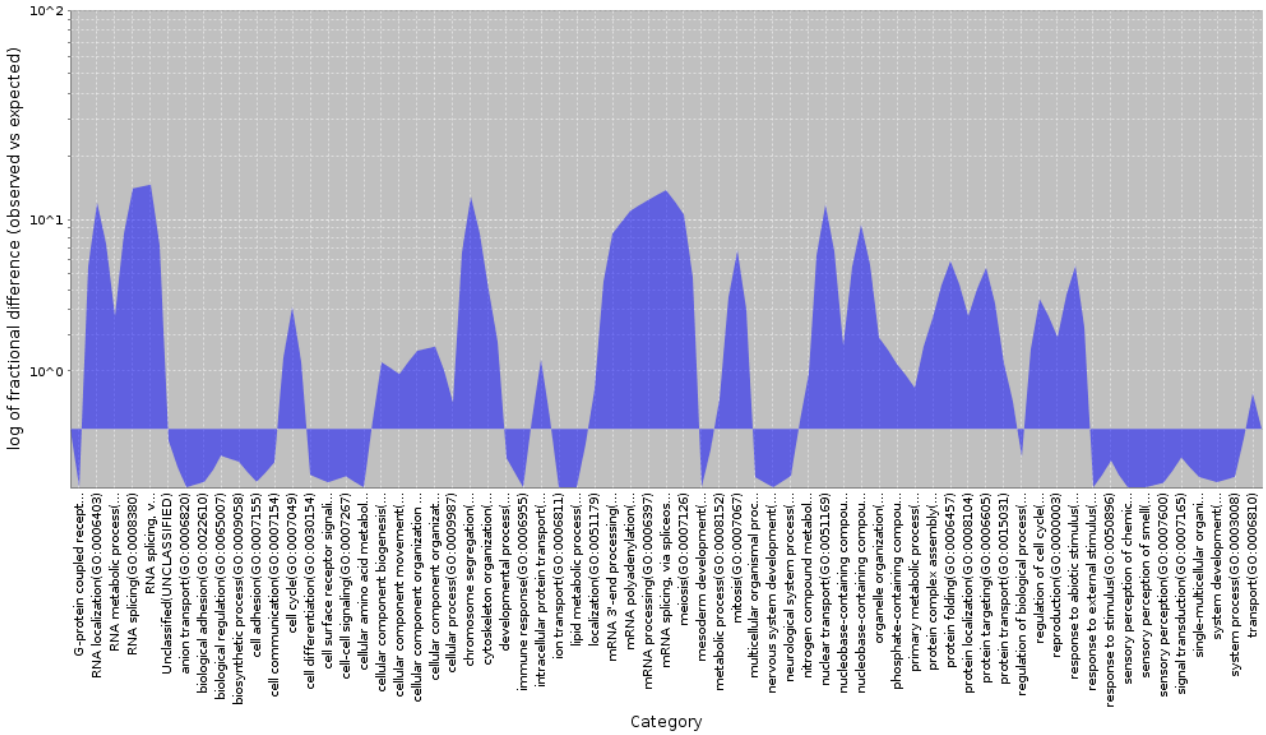

Fig. 3. An overlaid area charter produced by GO that indicates the relative under and overexpression of GO biological processes

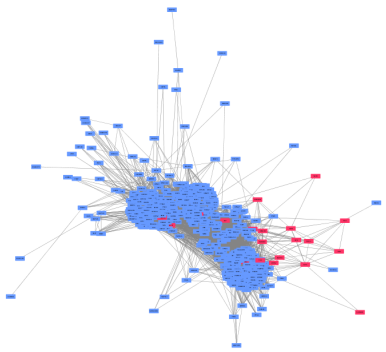

Fig. 4. A graphical representation of the STRINGDB ALS disease module network where the seed nodes (red) and the algorithm hits (blue) are displayed. The length of the edge is inversely related to the strength of the protein-protein relationship.

note that muscular atrophy, which describes the clinical manifestation of ALS is also high on this list. Thus, we can see how this algorithm provides the research investigator ways to probe at potential mechanisms of action as well as explore relevant diseases for future therapeutic and diagnostic uses.

F. Tissue specific KRAS cancer queries

In this final case study we will explore the tissue specific differences between normalized Z-scores based on gene expression data gathered from various sources. Where we use the results of run 13-28, as defined by the Appendix. These seed nodes come from MolSigDB and indicate genes that are up or down regulated in a tissue-specific cell line over-expression the oncogenic form of KRAS. Given that KRAS is an oncogenes, we conjecture that genes upregulated when KRAS mutated as also oncogenes while genes that are downregulated when KRAS either serve a tumor suppressing role or are no long necessary given the oncogenic phenotype of the cell. We look at the normalized Z-scores in the breast, lung, prostate, kidney, and all tissues and find that $\tilde{Z}_{CA}$ for each tissue is slightly different (TABLE VII).

We see that the breast tissue shows a stronger signal for cancer in the oncogenic genes versus thoses who are disrupted by KRAS mutation. On the other hand, the kidney shown lower signal for cancer for the oncogenic genes versus the nononcogenic genes. This can indicated that renal cell carcinoma, or a cancer of the kidney is most likely not KRAS based. In fact, current literature supports the notion that DNA sequencing of renal cell carcinoma cells do not indicate a mutation in KRAS [25]

Future work in this area can involve pulling more gene expression data from the GSEA database or from previously performed microarrays or from the GTeX

TABLE VII

TISSUE-SPECIFIC $\tilde{Z}_{CA}$ FOR UP AND DOWN REGULATED GENES

Tissue Types	UP regulated	DN regulated	Both UP and DN
all	6.72	5.06	4.89
breast	6.9	1.72	6.29
lung	6.37	1.89	5.79
kidney	1.64	5.48	2.04
prostate	5.3	6.01	5.84

portal [26] to future supplement analyses of this single oncogenes.

VI. CONCLUSION

Above we have outlined a process for determining disease modules from a signed, directed network in order to potentially discover novel disease-relevant disease interactors, find disease-disease relationships that share a common risk gene domain, and to begin to characterize tissue-specific effects on directional changes in gene expression.

One of the biggest challenges has been to find be to validate our findings as there is no true gold standard on what constitutes a disease gene and what is not affected. For validation, we extracted associated diseases using the Pharos API [11]. However, there remained a challenge on how to best normalize the z-scores so that would accurately reflect our confidence that the disease-gene set relationship was significant. We explored many different normalization techniques, however this use of a random network also us to take into consideration biases in the graph network that may benefit our community detection process.

Another challenge we have run into is that computation of p-values is expensive. A typical run take on the order of 1 day do computer. The base DIAMOnD algorithm can employ tricks to speed up its computations efficiently, but it is not applicable to our extended algorithm, since we are computing the norm of a combination of differently ranked link significances. Naively, assuming M graph nodes and N iterations, the algorithm would take $O(k \cdot N \cdot M)$ time given k different edge types. One idea we've had in speeding up the computation is to first run one iteration of DIAMOnD for one edge type, and skip computation for nodes below a threshold for the remaining edge types since they will have no chance of being a candidate. In the future, we will try to optimize the runtime of this algorithm

Despite these setback, we have shown that our network is able to predict protective mechanisms of our network via derivation of significant GO terms. Our network can also be used to explain drug treatment mechanism of actions (both known and unknown) by predicting the motifs through our connected network. Finally we believe that these disease module networks can be used to predict previously unknown interactions between proteins to their targeting partners that may be used to uncover novel mechanisms of biological action.

ACKNOWLEDGMENTS

We would like to acknowledge the CS224W course staff for their support during the project.

REFERENCES

[1] Barabasi, A.L, Gulbahce, N., & Loscalzo, J. (2011). An Integrative Systems Medicine Approach to Mapping Human Metabolic Diseases. *Nat Rev Genet*, 12(1), 5668 https://doi.org/10.1038/nrg2918.Network

[2] Sonmez, A. B., & Can, T. (2016). Comparison of tissue/disease specific integrated networks using directed graphlet signatures. *In Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics - BCB 16* (Vol. 18, pp. 533534). BioMed Central https://doi.org/10.1145/2975167.2985674

[3] Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, O'Donnell L, Oster S, Theesfeld C, Sellam A, Stark C, Breitkreutz BJ, Dolinski K, Tyers M. The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* 2016 Dec 14;2017(1)

[4] Szklarczyk, Damian, et al. "The STRING database in 2017: quality-controlled proteinprotein association networks, made broadly arressible." *Nucleic acids research* 45.D1 (2017): D362-D368

[5] Liu, Z.-P., Wang, Y., Zhang, X.-S., & Chen, L. (2012). Network-based analysis of complex diseases. *IET Systems Biology,* 6(1), 22 https://doi.org/10.1049/iet-syb.2010.0052

[6] Esmailian, P., Jalili M. (2015). Community Detection in Signed Networks: the Role of Negative ties in Different Scales

[7] Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., & Barabasi, A.-L. (2015). Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224), 12576011257601 https://doi.org/10.1126/science.1257601

[8] Peel, L., Larremore, D. B., & Clauset, A. (2017). The ground truth about metadata and community detection in networks. *Science Advances*, 3(5), e1602548. https://doi.org/10.1126/sciadv.1602548

[9] Ghiassian, S. D., Menche, J., & Barabasi, A. L. (2015). A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome. *PLoS Computational Biology*, 11(4), e1004120 https://doi.org/10.1371/journal.pcbi.1004120

[10] Okamura et al. (2015) COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems. *Nucleic Acids Res*, 43, D82-6

[11] Nguyen, D.-T., Mathias, S. et al. (2017). "Pharos: Collating Protein Information to Shed Light on the Druggable Genome", *Nucl. Acids Res.*, 45(D1), D995-D1002. DOI: 10.1093/nar/gkw1072

[12] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... & Harris, M. A. (2000). Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1), 25-29.

[13] Gene Ontology Consortium. (2017). Expansion of the Gene Ontology knowledge: base and resources. *Nucleic acids research*, 45(D1), D331-D338.

[14] Pletscher-Frankild, S., Palleja, A., Tsafou, K., Binder, J. X. & Jensen, L. J. (2015) Diseases: Text mining and data integration of diseasegene associations. *Methods* 74, 8389, Text mining of biomedical literature.

[15] Soren Mork, Sune PletscherFrankild, Albert Palleja Caro, Jan Gorodkin, Lars Juhl Jensen (2014). Protein-driven inference of miRNAdisease associations, *Bioinformatics*, 30(1), 392397, https://doi.org/10.1093/bioinformatics/btt677

[16] Subramanian, Tamayo, et al. (2005), *PNAS* 102, 15545-15550.

[17] Ferbeyre, G., De Stanchina, E., Lin, A. W., Querido, E., Mrrurrach, M. E., Hannon, G. J., & Lowe, S. W. (2002). Oncogenic ras and p53 cooperate to induce cellular senescence. *Molecular and cellular biology,* 22(10), 3497-3508.

[18] Wijesekera, L. C., & Leigh, P. N. (2009). Amyotrophic lateral sclerosis. *Orphanet journal of rare diseases*, 4(1), 3.

[19] Lagier-Tourenne, C., Polymenidou, M., Hutt, K. R., Vu, A. Q., Baughn, M., Huelga, S. C., ... & Wancewicz, E. (2012). Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs. *Nature neuroscience,* 15(11), 1488-1497.

[20] Kabashi, E., Valdmanis, P. N., Dion, P., Spiegelman, D., Mrronkey, B. J., Velde, C. V., ... & Pradat, P. F. (2008). TARDBP mutations in individuals with sporadic and familial amyotrophic lateral sclerosis. *Nature genetics*, 40(5), 572-574.

[21] Wu, C. H., Fallini, C., Ticozzi, N., Keagle, P. J., Sapp, P. C., Piotrowska, K., ... & Kost, J. E. (2012). Mutations in the profilin 1 gene cause familial amyotrophic lateral sclerosis. *Nature*, 488(7412), 499-503.

[22] Freischmidt, A., Wieland, T., Richter, B., Ruf, W., Schaeffer, V., Mller, K., ... & Pinto, S. (2015). Haploinsufficiency of TBK1 causes familial ALS and fronto-temporal dementia. *Nature neuroscience*, 18(5), 631-636.

[23] Bucci, C., Alifano, P., & Cogli, L. (2014). The role of rab proteins in neuronal cells and in the trafficking of neurotrophin receptors. *Membranes*, 4(4), 642-677.

[24] Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, Jensen LJ, von Mering C. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 2017 Jan; 45:D362-68.

[25] Gattenlhner, S., Etschmann, B., Riedmiller, H., & Mller-Hermelink, H. K. (2009). Lack of KRAS and BRAF mutation in renal cell carcinoma. *European urology*, 55(6), 1490-1491.

[26] Battle, A., Brown, C. D., Engelhardt, B. E., Montgomery, S. B.,& GTEx Consortium. (2017). Genetic effects on gene expression across human tissues. *Nature*, 550(7675), 204-213.

APPENDIX

A. Description of Seed Nodes Sets

TABLE VIII
MY CAPTION

Run	Name of Run	Description	num seed
0	all seed OMIM	All diseases associated genes arrording to OMIM	6086
1	cancer seed OMIM v2	All genes found to be associated with diseases containing the word ”cancer”	144
2	cardiac seed OMIM v1	All genes found to be associated with diseases containing the word ”cardio”	110
3	P53 DN.V1 DN	Genes down-regulated in NCI-60 panel of cell lines with mutated TP53 [GeneID=7157].	192
4	P53 DN.V1 UP	Genes up-regulated in NCI-60 panel of cell lines with mutated TP53 [GeneID=7157].	194
5	P53 DN.V2 DN	Genes down-regulated in HEK293 cells (kidney fibroblasts) upon knockdown of TP53 [Gene ID=7157] gene by RNAi.	145
6	P53 DN.V2 UP	Genes up-regulated in HEK293 cells (kidney fibroblasts) upon knockdown of TP53 [Gene ID=7157] gene by RNAi.	148
7	KRAS.300 UP.V1 DN	Genes down-regulated in HEK293 cells (kidney fibroblasts) upon knockdown of TP53 [GeneID=7157] gene by RNAi.	143
8	KRAS.300 UP.V1 UP	Genes up-regulated in HEK293 cells (kidney fibroblasts) upon knockdown of TP53 [GeneID=7157] gene by RNAi.	146
9	KRAS.50 UP.V1 DN	Genes down-regulated in four lineages of epithelial cell lines over-expressing an oncogenic form of KRAS [GeneID=3845] gene.	49
10	KRAS.50 UP.V1 UP	Genes up-regulated in four lineages of epithelial cell lines over-expressing an oncogenic form of KRAS [GeneID=3845] gene.	48
11	KRAS.600 .LUNG.BREAST UP.V1 DN	Genes down-regulated in four lineages of epithelial cell lines over-expressing an oncogenic form of KRAS [GeneID=3845] gene.	289
12	KRAS.600 .LUNG.BREAST UP.V1 UP	Genes up-regulated in four lineages of epithelial cell lines over-expressing an oncogenic form of KRAS [GeneID=3845] gene.	288
13	KRAS.600 UP.V1 DN	Genes down-regulated in epithelial lung and breast cancer cell lines over-expressing an oncogenic form of KRAS [Gene ID=3845] gene.	289
14	KRAS.600 UP.V1 UP	Genes up-regulated in epithelial lung and breast cancer cell lines over-expressing an oncogenic form of KRAS [GeneID=3845] gene.	287
15	KRAS.AMP .LUNG UP. V1 DN	Genes down-regulated in epithelial lung cancer cell lines over-expressing KRAS [GeneID=3845] gene.	146
16	KRAS.AMP. LUNG UP. V1 UP	Genes up-regulated in epithelial lung cancer cell lines over-expressing KRAS [GeneID=3845] gene.	144
17	KRAS.BREAST UP. V1 DN	Genes down-regulated in epithelial breast cancer cell lines over-expressing an oncogenic form of KRAS [GeneID=3845] gene.	145
18	KRAS.BREAST UP.V1 UP	Genes up-regulated in epithelial breast cancer cell lines over-expressing an oncogenic form of KRAS [GeneID=3845] gene.	146
19	KRAS.DF.V1 DN	Genes down-regulated in epithelial lung cancer cell lines over-expressing an oncogenic form of KRAS [GeneID=3845] gene.	194
20	KRAS.DF.V1 UP	Genes up-regulated in lung cancer cell lines over-expressing an oncogenic form of KRAS [GeneID=3845] gene.	193

22	KRAS.KIDNEY UP.V1 UP	Genes up-regulated in epithelial lung cancer cell lines over-expressing an oncogenic form of KRAS [GeneID=3845] gene.	145
23	KRAS. LUNG.BREAST UP.V1 DN	Genes down-regulated in epithelial kidney cancer cell lines over-expressing an oncogenic form of KRAS [GeneID=3845] gene.	145
24	KRAS. LUNG.BREAST UP.V1 UP	Genes up-regulated in epithelial kidney cancer cell lines over-expressing an oncogenic form of KRAS [GeneID=3845] gene.	145
25	KRAS.LUNG UP.V1 DN	Genes down-regulated in epithelial breast cancer cell lines over-expressing an oncogenic form of KRAS [Gene ID=3845 gene].	145
26	KRAS.LUNG UP.V1 UP	Genes up-regulated in epithelial lung and breast cancer cell lines over-expressing an oncogenic form of KRAS [GeneID=3845] gene.	141
27	KRAS.PROSTATE UP.V1 DN	Genes down-regulated in epithelial lung cancer cell lines over-expressing an oncogenic form of KRAS [GeneID=3845] gene.	144
28	KRAS.PROSTATE UP.V1 UP	Genes up-regulated in epithelial lung cancer cell lines over-expressing an oncogenic form of KRAS [GeneID=3845] gene.	143
29	RAF UP.V1 DN	Genes down-regulated in epithelial prostate cancer cell lines over-expressing an oncogenic form of KRAS [GeneID=3845] gene.	194
30	RAF UP.V1 UP	Genes up-regulated in epithelial prostate cancer cell lines over-expressing an oncogenic form of KRAS [GeneID=3845] gene.	196
31	AKT UP.V1 DN	Genes down-regulated in MCF-7 cells (breast cancer) positive for ESR1 [GeneID=2099] MCF-7 cells (breast cancer) stably over-expressing constitutively active RAF1 [GeneID=5894] gene.	187
32	AKT UP.V1 UP	Genes up-regulated in MCF-7 cells (breast cancer) positive for ESR1[GeneID=2099]MCF-7 cells (breast cancer) stably over-expressing constitutively active RAF1[GeneID=5894]gene.	172
33	AKT UP MTOR DN.V1 DN	Genes down-regulated in mouse prostate by transgenic expression of human AKT1 gene[GeneID=207]vs controls.	183
34	AKT UP MTOR DN.V1 UP	Genes up-regulated in mouse prostate by transgenic expression of human AKT1 gene [GeneID=207] vs controls.	184
35	VEGF A UP.V1 DN	Genes down-regulated by everolimus[PubChem=6442177]in mouse prostate tissue transgenically expressing human AKT1 gene[GeneID=207]vs untreated controls.	193
36	VEGF A UP.V1 UP	Genes up-regulated by everolimus[PubChem=6442177]in mouse prostate tissue transgenically expressing human AKT1 gene[GeneID=207]vs untreated controls.	196
37	TGFB UP.V1 DN	Genes down-regulated in HUVEC cells (endothelium) by treatment with VEGFA[GeneID=7422].	192
38	TGFB UP.V1 UP	Genes up-regulated in HUVEC cells (endothelium) by treatment with VEGFA[GeneID=7422].	192
39	WNT UP.V1 DN	Genes down-regulated in a panel of epithelial cell lines by TGFB1[GeneID=7040].	170
40	WNT UP.V1 UP	Genes up-regulated in a panel of epithelial cell lines by TGFB1 [GeneID=7040].	180
41	NOTCH DN.V1 DN	Genes down-regulated in C57MG cells (mammary epithelium) by over-expression of WNT1[GeneID=7471]gene.	189
42	NOTCH DN.V1 UP	Genes up-regulated in C57MG cells (mammary epithelium) by over-expression of WNT1 [GeneID=7471] gene.	193
43	MEK UP.V1 DN	Genes down-regulated in MOLT4 cells (T-ALL) by DAPT [PubChem=16219261], an inhibitor of NOTCH signaling pathway.	196
44	MEK UP.V1 UP	Genes up-regulated in MOLT4 cells (T-ALL) by DAPT[PubChem=16219261], an inhibitor of NOTCH signaling pathway.	196
45	IL15 UP.V1 DN	Genes down-regulated in MCF-7 cells (breast cancer) positive for ESR1 [GeneID=2099] MCF-7 cells (breast cancer) stably over-expressing constitutively active MAP2K1 [GeneID=5604] gene.	190
46	IL15 UP.V1 UP	Genes up-regulated in MCF-7 cells (breast cancer) positive for ESR1 [GeneID=2099] MCF-7 cells (breast cancer) stably over-expressing constitutively active MAP2K1 [GeneID=5604] gene.	192
47	IL21 UP.V1 DN	Genes down-regulated in Sez-4 cells (T lymphocyte) that were first starved of IL2 [GeneID=3558] and then stimulated with IL15 [GeneID=3600].	187
48	IL21 UP.V1 UP	Genes up-regulated in Sez-4 cells (T lymphocyte) that were first starved of IL2[GeneID=3558]and then stimulated with IL15[GeneID=3600].	193
49	IL2 UP.V1 DN	Genes down-regulated in Sez-4 cells (T lymphocyte) that were first starved of IL2 [GeneID=3558] and then stimulated with IL21 [GeneID=59067].	196
50	IL2 UP.V1 UP	Genes up-regulated in Sez-4 cells (T lymphocyte) that were first starved of IL2[GeneID=3558]and then stimulated with IL21[GeneID=59067].	192
51	JAK2 DN.V1 DN	Genes down-regulated in Sez-4 cells (T lymphocyte) that were first starved of IL2 [GeneID=3558] and then stimulated with IL2 [GeneID=3558].	173
52	JAK2 DN.V1 UP	Genes up-regulated in Sez-4 cells (T lymphocyte) that were first starved of IL2 [GeneID=3558] and then stimulated with IL2 [GeneID=3558].	188

53	BRCA1 DN.V1 DN	Genes down-regulated in HEL cells (erythroleukemia) after knockdown of JAK2[GeneID=3717]gene by RNAi.	143
54	BRCA1 DN.V1 UP	Genes up-regulated in HEL cells (erythroleukemia) after knockdown of JAK2[GeneID=3717]gene by RNAi.	141
55	PTEN DN.V2 DN	Genes down-regulated in MCF10A cells (breast cancer) upon knockdown of BRCA1 [GeneID=672] gene by RNAi.	144
56	PTEN DN.V2 UP	Genes up-regulated in MCF10A cells (breast cancer) upon knockdown of BRCA1 [GeneID=672] gene by RNAi.	143
57	RELA DN.V1 DN	Genes down-regulated in HCT116 cells (colon carcinoma) upon knockdown of PTEN[GeneID=5728]by RNAi.	141
58	RELA DN.V1 UP	Genes up-regulated in HCT116 cells (colon carcinoma) upon knockdown of PTEN [GeneID=5728] by RNAi.	149
59	RB DN.V1 DN	Genes down-regulated in HEK293 cells (kidney fibroblasts) upon knockdown of RELA [GeneID=5970] gene by RNAi.	126
60	RB DN.V1 UP	Genes up-regulated in HEK293 cells (kidney fibroblasts) upon knockdown of RELA [GeneID=5970] gene by RNAi.	137
61	ERB2 UP.V1 DN	Genes down-regulated in primary keratinocytes from RB1[GeneID=5925]skin specific knockout mice.	197
62	ERB2 UP.V1 UP	Genes up-regulated in primary keratinocytes from RB1 [Gene ID=5925] skin specific knockout mice.	191
63	EGFR UP.V1 DN	Genes down-regulated in MCF-7 cells (breast cancer) positive for ESR1 [Gene ID=2099] and engineered to express ligand-activatable ERBB2 [Gene ID=2064].	196
64	EGFR UP.V1 UP	Genes up-regulated in MCF-7 cells (breast cancer) positive for ESR1 [GeneID=2099] and engineered to express ligand-activatable ERBB2 [GeneID=2064].	193
65	E2F1 UP.V1 DN	Genes down-regulated in MCF-7 cells (breast cancer) positive for ESR1[GeneID=2099]and engineered to express ligand-activatable EGFR[GeneID=1956].	193
66	E2F1 UP.V1 UP	Genes up-regulated in MCF-7 cells (breast cancer) positive for ESR1[GeneID=2099]and engineered to express ligand-activatable EGFR[GeneID=1956].	189
67	RB P107 DN.V1 DN	Genes down-regulated in mouse fibroblasts over-expressing E2F1 [Gene ID=1869] gene.	128
68	RB P107 DN.V1 UP	Genes up-regulated in mouse fibroblasts over-expressing E2F1 [GeneID=1869] gene.	140
69	RB P130 DN.V1 DN	Genes down-regulated in primary keratinocytes from RB1 and RBL1 [GeneID=5925] [GeneID=5933] skin specific knockout mice.	139
70	RB P130 DN.V1 UP	Genes up-regulated in primary keratinocytes from RB1 and RBL1[GeneID=5925][GeneID=5933]skin specific knockout mice.	133
71	dm seed OMIM v1	Genes down-regulated in primary keratinocytes from RB1 and RBL2[GeneID=5925][GeneID=5934]skin specific knockout mice.	27
72	immunodef seed OMIM v1	Genes up-regulated in primary keratinocytes from RB1 and RBL2[GeneID=5925][GeneID=5934]skin specific knockout mice.	207
73	als ms alz park seed OMIM v1	All genes found to be associated with diseases: ALS, MS, Alzheimers', and Parkinsons	152
74	als ms alz seed OMIM v1	All genes found to be associated with diseases: ALS, MS, and Alzheimers'	114
75	als seed OMIM v1	All genes found to be associated with diseases: ALS	82
76	ms seed OMIM v1	All genes found to be associated with diseases: MS	11
77	alz seed OMIM v1	All genes found to be associated with diseases: Alzheimer's	21
78	park seed OMIM v1	All genes found to be associated with diseases: Parkinsons'	38
79	schizo seed OMIM v1	All genes found to be associated with diseases: Schizophrenia	60

B. *Significant disease hits*

Abbreviations: AD: Alzheimer's Disease; ALS: Amyotrophic Lateral Sclerosis; CA: Cancer; CM: Cardiomyopathy; ID: Immunodeficiency; MS: Mutiple Sclerosis; PD: Parkinson's Disease; SCZ: Schizophrenia

Hits for ALS

Disease	$\tilde{Z}_d$
Senior-Loken syndrome	28.87
Joubert syndrome	26.31
Meckel syndrome	25.55
Amyotrophic Lateral Sclerosis	25.18
Frontotemporal dementia	24.2
Muscular atrophy	20.43
Nephronophthisis	20.07
Microcephaly	17.3
Polydactyly	11.02
Intellectual disability	5.44
Cancer	1.73

Hits for AD

Disease	$\tilde{Z}_d$
Angioedema	141.86
Brain disease	99.83
Cholestasis	53.43
Common cold	52.48
Meningitis	51.09
Lipid metabolism disorder	45.38
Amyloidosis	41.79
Bilirubin metabolic disorder	40.02
Leishmaniasis	35.47
Pancreatitis	33.78
Skin disease	33.25
Peritonitis	33.17
Diabetic Retinopathy	32.2
Polycystic Ovary Syndrome	29.92
Fatty liver disease	29.44
Liver disease	27.71
Hyperglycemia	26.43
Eye disease	26.3
Endometriosis	25.39
Metabolic syndrome X	23.72
periodontitis	23.49
Atherosclerosis	23.39
Hyperinsulinism	23.38
Allergic rhinitis	22.63
Toxic encephalopathy	21.43
Malaria	21.35
tuberculosis	21.33
Alzheimer's disease	21.2
Dementia	20.77
Lung disease	19.86
Hepatitis C	19.55
Hepatitis B	19.45
Arthritis	19.29
Vasculitis	17.61
Leukopenia	17.45
Coronary artery disease	17.29
Heart disease	17.24
diabetes mellitus	16.54
Kidney disease	16.36
Peripheral vascular disease	15.77
Influenza	15.33
Dermatitis	15.18
Hypoglycemia	15.06
Osteoporosis	14.59
Hypersensitivity reaction type II disease	14.49

Hits for CA

Disease	$\tilde{Z}_d$
Common cold	48.33
Meningitis	45.86
Brain disease	45.08
Leishmaniasis	39.27
Urticaria	36.28
periodontitis	26.28
Allergic rhinitis	24.95
Pancreatitis	23.97
Lynch syndrome	22.49
Hepatitis C	21.95
Leukopenia	21.54
Vasculitis	21.24
tuberculosis	20.27
Familial adenomatous polyposis	19.83
Severe Combined Immunodeficiency	19
Eosinophilia	18.84
Influenza	18.8
Gastritis	18.49
Polycythemia Vera	17.84
Hepatitis B	16.96
Arthritis	16.68
Malaria	16.56
Hypersensitivity reaction type II disease	15.88
Exanthem	15.49
Human immunodeficiency virus infectious disease	15.38
Inflammatory bowel disease	15.37
Lung disease	14.8
Rheumatoid Arthritis	14.52
Dermatitis	14.21
Adenoma	13.62
Hyperinsulinism	13.49
Hepatitis	13.39
Allergic hypersensitivity disease	13.37
Liver disease	13.29
Hyperglycemia	13.09
Diarrhea	13.03
Tetanus	11.77
Multiple Sclerosis	11.54
Asthma	11.52
Kidney disease	11.07
Metabolic syndrome X	10.93
Anemia	10.65
Atherosclerosis	10.57
Acquired immunodeficiency syndrome	10.5
Thrombocytopenia	10.4
Neutropenia	10.4
Hypothyroidism	10.06
Alcohol dependence	9.98
Heart disease	9.86
diabetes mellitus	9.28
Toxic encephalopathy	9.19
Obesity	8.77

TABLE IX
HITS FOR CM

Disease	$\tilde{Z}_d$
Distal arthrogryposis	30.13
Myopathy	29.09
HIV infectious disease	21.59
tuberculosis	18.91
Allergic rhinitis	18.9
Hypertrophic Cardiomyopathy	18.65
Malignant hyperthermia	13.27
Eosinophilia	13.03
Pain agnosia	12.94
Lung disease	12.74
Nicotine dependence	12.34
Asthma	11.82
Irritable bowel syndrome	11.09
Arthritis	10.61
Toxic encephalopathy	10.46
Dermatitis	10.43
Migraine	10.32
Multiple Sclerosis	10.11
Allergic hypersensitivity disease	9.81
Heart disease	9.47
Schizophrenia	9.1
Hypersensitivity reaction type II disease	8.87
Alcohol dependence	8.78
Acquired immunodeficiency syndrome	8.74
Inflammatory bowel disease	8.05
Liver disease	7.76
Atherosclerosis	7.75
Diarrhea	7.1
Epilepsy	6.99
Coronary artery disease	6.94
Hypertension	6.54
Kidney disease	6.48
Cerebrovascular disease	6.43
Autistic Disorder	6.25
Vascular disease	5.85
Neurodegenerative disease	5.36
diabetes mellitus	4.83
Alzheimer's disease	4.68
Parkinson's disease	4.33
Intellectual disability	3.45
Cancer	3.08
Obesity	3.01

TABLE X
HITS FOR ID

Disease	$\tilde{Z}_d$
Meningitis	45.84
Leishmaniasis	37.6
Pituitary adenoma	34.03
Agammaglobulinemia	27.13
Hepatitis C	26.93
Severe Combined Immunodeficiency	26.65
Hyperprolactinemia	26.27
Allergic rhinitis	24.71
periodontitis	24.7
Pancreatitis	24.03
Leukopenia	23.1
Acromegaly	22.75
Influenza	21.96
tuberculosis	21.72
Anorexia nervosa	20.67
Vasculitis	19.77
Eosinophilia	19.39
Polycythemia Vera	18.82
Mastocytosis	18.24
Malaria	18.06
Hypersensitivity reaction type II disease	17.78
Rheumatoid Arthritis	17.53
Arthritis	17.09
HIV infectious disease	16.94
Hepatitis B	16.66
Inflammatory bowel disease	16.38
Hyperinsulinism	16.36
systemic lupus erythematosus	16.23
Hypogonadism	16.22
Hepatitis	15.7
Exanthem	15.62
Diarrhea	15.58
Dermatitis	15.44
Tetanus	14.3
Allergic hypersensitivity disease	14.21
Multiple Sclerosis	13.9
Lung disease	13.55
Hyperglycemia	13.27
Hypoglycemia	13.14
Asthma	12.98
Candidiasis	12.68
Hypothyroidism	12.59
Liver disease	11.51
Acquired immunodeficiency syndrome	11.4
Metabolic syndrome X	11.31
Kidney disease	11.19
Thrombocytopenia	10.92
Adenoma	10.7
Anemia	10.27
diabetes mellitus	10.09
Neutropenia	10.06
Heart disease	10.03

TABLE XI
HITS FOR MS

Disease	$\tilde{Z}_d$
Severe Combined Immunodeficiency	19.52
Rheumatoid Arthritis	15.91
systemic lupus erythematosus	15.11
Exanthem	14.95
Eosinophilia	12.82
Diarrhea	12.34
tuberculosis	11.80
Hypersensitivity reaction type II disease	11.38
Influenza	10.96
Lung disease	10.79
Arthritis	10.50
Adenoma	9.39
Inflammatory bowel disease	9.26
Heart disease	9.06
Acquired immunodeficiency syndrome	9.03
Hyperglycemia	9.01
Allergic hypersensitivity disease	8.80
Kidney disease	8.80
Neutropenia	8.76
Dermatitis	8.11
Asthma	7.95
Thrombocytopenia	7.79
diabetes mellitus	7.73
Hypertension	7.63
Metabolic syndrome X	7.52
Vascular disease	7.24
Multiple Sclerosis	7.10
Obesity	7.10
Schizophrenia	7.06
Alcohol dependence	6.97
Anemia	6.80
Atherosclerosis	6.53
Liver disease	6.14
Cancer	5.69
Cerebrovascular disease	5.61
Parkinson's disease	5.53
Alzheimer's disease	5.27
Coronary artery disease	4.35
Epilepsy	4.06

TABLE XII
HITS FOR PD

Disease	$\tilde{Z}_d$
Exanthem	12.94
Rheumatoid Arthritis	9.8
Parkinson's disease	8.05
Diarrhea	8
Dementia	7.31
Kidney disease	6.81
Alzheimer's disease	6.37
Lung disease	6.3
Atherosclerosis	5.73
Liver disease	5.69
Thrombocytopenia	5.58
Allergic hypersensitivity disease	5.54
Heart disease	5.47
Arthritis	5.45
Hypersensitivity reaction type II disease	5.22
Anemia	4.85
Inflammatory bowel disease	4.79
diabetes mellitus	4.58
Asthma	4.31
Vascular disease	4.3
Cancer	4.24
Hypertension	3.84
Obesity	3.79
Coronary artery disease	3.75
Cerebrovascular disease	2.97
Intellectual disability	2.73

TABLE XIII
HITS FOR SCZ

Disease	$\tilde{Z}_d$
Distal arthrogryposis	30.36
Myopathy	26.9
Human immunodeficiency virus infectious disease	21.76
Hypertrophic Cardiomyopathy	17.61
tuberculosis	17.04
Malignant hyperthermia	13.37
Nicotine dependence	13.21
Schizophrenia	13.11
Pain agnosia	13.04
Lung disease	12.66
Toxic encephalopathy	12.32
Migraine	12.05
Asthma	11.58
Arthritis	11.45
Irritable bowel syndrome	11.18
Eosinophilia	11.03
Multiple Sclerosis	10.84
Dermatitis	10.51
Influenza	10.48
Alcohol dependence	10.05
Heart disease	9.88
Bipolar Disorder	9.05
Hypersensitivity reaction type II disease	8.94
Acquired immunodeficiency syndrome	8.81
Atherosclerosis	8.45
Inflammatory bowel disease	8.44
Allergic hypersensitivity disease	8.28
Autistic Disorder	8.26
Liver disease	8.24
Major Depressive Disorder	7.5
Epilepsy	7.47
Diarrhea	7.16
Cerebrovascular disease	7.03
Coronary artery disease	6.95
Hypertension	6.58
Kidney disease	6.42
Parkinson's disease	6.16
Alzheimer's disease	6.14
Vascular disease	5.84
diabetes mellitus	5
Neuropathy	4.65
Intellectual disability	3.75
Cancer	3.18
Obesity	2.9