Network Analysis of the Human Reference Interactome

Cameron MacKenzie

Abstract:

The Human Reference Interactome (HuRI) is a large dataset representing over 50,000 protein-protein interactions. In this project, I aimed to examine the structure of the HuRI network, with a particular focus on identifying features of potential translational and clinical relevance. I identified sets of disease-relevant genes by scraping PubMed abstracts and used this information to look for structure in the HuRI network. While I was surprised to find a lack of clustering among disease-relevant gene sets, I discovered that removing disease-specific genes from the network tended to affect the connectivity of other disease-related genes to each other more than the connectivity of other genes in the network as a whole. This project provided me with an exciting foundation for continuing to explore potential clinical insights from large biological datasets.

1 Introduction

The acceleration of technological capabilities in recent decades has rapidly decreased the cost of high-throughput experimental techniques and enabled widespread collection of vast biological datasets, many of which are publicly available. Nevertheless, generating clinically meaningful results from these datasets remains challenging. The Human Reference Interactome (HuRI) is a high-throughput dataset representing protein-protein interactions between over 8,000 human proteins (Luck et al. 2020). In this project, I examined the structure of the network of genes represented by the HuRI dataset. I also explored sub-networks of genes corresponding to particular disorders including autism, diabetes, and lupus. Specifically, I set out to understand whether patterns might exist in the arrangement of disease-specific genes within the broader network and the extent to which perturbations to these genes might affect the connectivity of others. While my work barely scratches the surface of the information to be gleaned from the HuRI network, it provides me a useful starting point for understanding the benefits and limitations of approaching clinical problems with large omics datasets.

2 Methods

I used the NetworkX package in Python to generate a network with nodes representing genes and edges representing interactions given by the HuRI dataset between proteins encoded by each gene. To identify sets of disease-specific genes, I used the PubMed Entrez API to access large numbers of journal article abstracts matching a search term and counted references to each gene among the abstracts. For example, to identify which of the >8,000 genes in the HuRI dataset are associated with autism, I queried all PubMed articles since the year 2000 matching the search term "autism," then tallied the number of abstracts containing a match to each gene. Matches were identified either as an exact, case-insensitive match to the full name of a gene (e.g. "Fragile X Messenger Ribonucleoprotein 1") or to specific phrases containing a gene abbreviation, with whitespace padding for specificity (e.g. " gene FMR1 ", " FMR1 gene ", " gene, FMR1 ", or " FMR1 expression "). While this method excludes many potential matches and may not entirely exclude all possible false positives, manual investigation of the top few genes identified for each disease indicated the method was indeed identifying disease-relevant genes.

3 Results

Preliminary analysis of the HuRI network revealed 73 disconnected subgraphs, with one of these (the “main network”) containing 8,109 genes and each of the 72 others containing no more than three genes (Fig. 1A-B). For simplicity, I discarded the small subgraphs and considered only the features of the main network. Degree-rank plots revealed that while more than half the genes in the network have fewer than 10 connections, some have several hundred connections (Fig. 1C-D). I also analyzed the betweenness centrality of each node, which was correlated with the node’s degree with a Pearson’s coefficient of 0.76 (Fig. 1E).

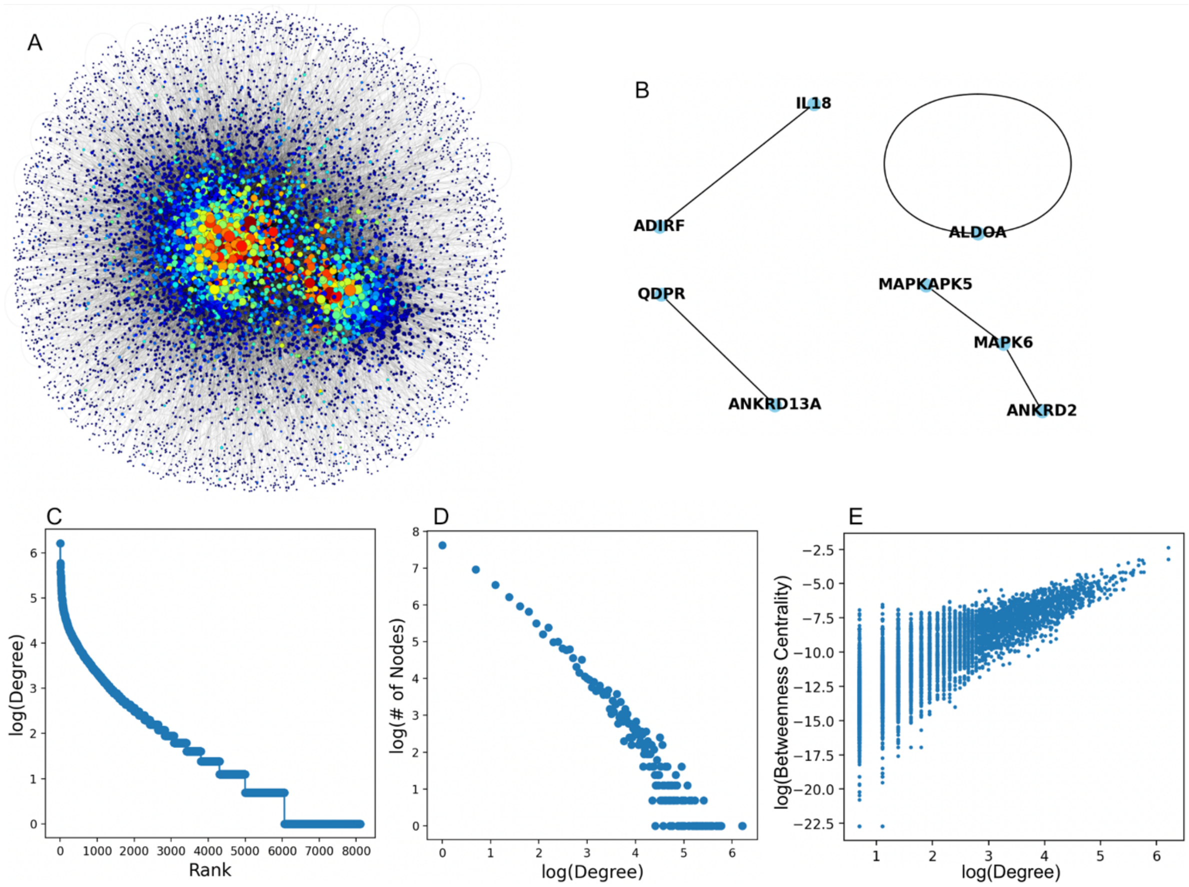


Figure 1. (A) Visualization of HuRI protein interaction network, with node diameter representing degree and color representing betweenness centrality (red: higher, blue: lower). (B) Examples of small, disconnected sub-graphs present in the HuRI dataset. (C) Degree-rank plot of all nodes in the main network, with degree plotted on a logarithmic scale. (D) Plot showing the number of nodes of each degree in the main network, plotted on a log-log scale. (E) Scatterplot showing the relationship between each node’s degree and betweenness centrality, plotted on a log-log scale.

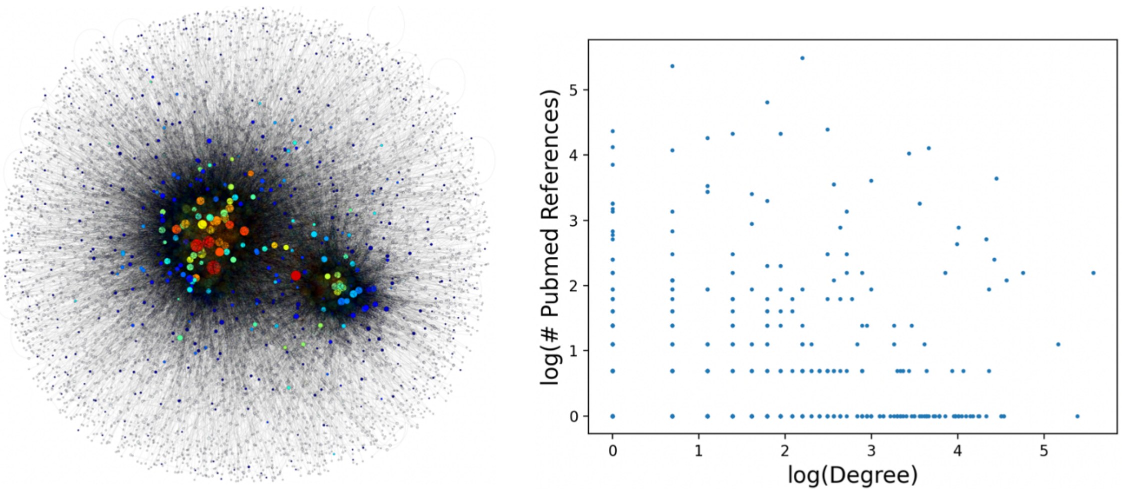


Figure 2. (Left) Visualization of the network with autism-related genes emphasized. (Right) Scatterplot showing the relationship between a gene's degree in the network and the number of times the gene was referenced in PubMed abstracts since 2000 matching the search term "autism."

Next, I analyzed autism-specific genes within the network. I used PubMed to identify 536 autism-related genes within the main network, spanning a wide range of centralities (Fig. 2-left). While I had suspected that the most-referenced genes would likely correspond to those with more central roles in the network, I found essentially no correlation between degree and number of PubMed abstracts referencing the gene (Fig. 2-right). I also expected to find that the autism-related genes might be clustered together in the network, so I was surprised to find that the mean ± S.D. path length between autism-related genes was 3.89±0.93, marginally higher than the average path length between all genes in the network, 3.84±0.91. I found this to be the case for 34 other disorders as well, with the average path length between disease-related genes differing very minimally from the average path length between all genes (Supp. Fig. 1). A summary of the PubMed search results for all 35 disorders considered is given in Table 1.

Next, I wanted to understand the extent to which the removal of one or a few nodes might disrupt the broader connectivity of the network. I started by removing the highest-degree node, CYSRT1, with 498 connections, which slightly increased the average path length of the network to 3.86±0.91. Next, I found that removing the most-referenced autism-related gene, FMR1, which only had 9 connections, also increased the average path length between autism-related genes to 3.90±0.93. While these increases were very subtle, I was surprised that such small changes in the network could have a detectable effect at all.

I extended this perturbation analysis by examining the effects of removing the single most-referenced gene and the twenty most-referenced genes (together) for each disorder. While these perturbations had almost no effect on the average path length between all genes in the network, they did tend to increase the average path length between each set of corresponding disease-related genes (Fig. 3). Specifically, removing the single most-referenced gene increased the average disease-related gene path length by 0.00015±0.00508 (not significant, $p>0.05$, by t-test comparison to 0 mean), while removal of the twenty most-referenced gene increased the average disease-related gene path length by 0.02137±0.053046 (significant, $p<0.05$, by t-test comparison to 0 mean). In short, these results suggest that perturbation of a disease-specific gene may affect the connectivity of the genes associated with the same disease more than connectivity of the genes in the broader network.

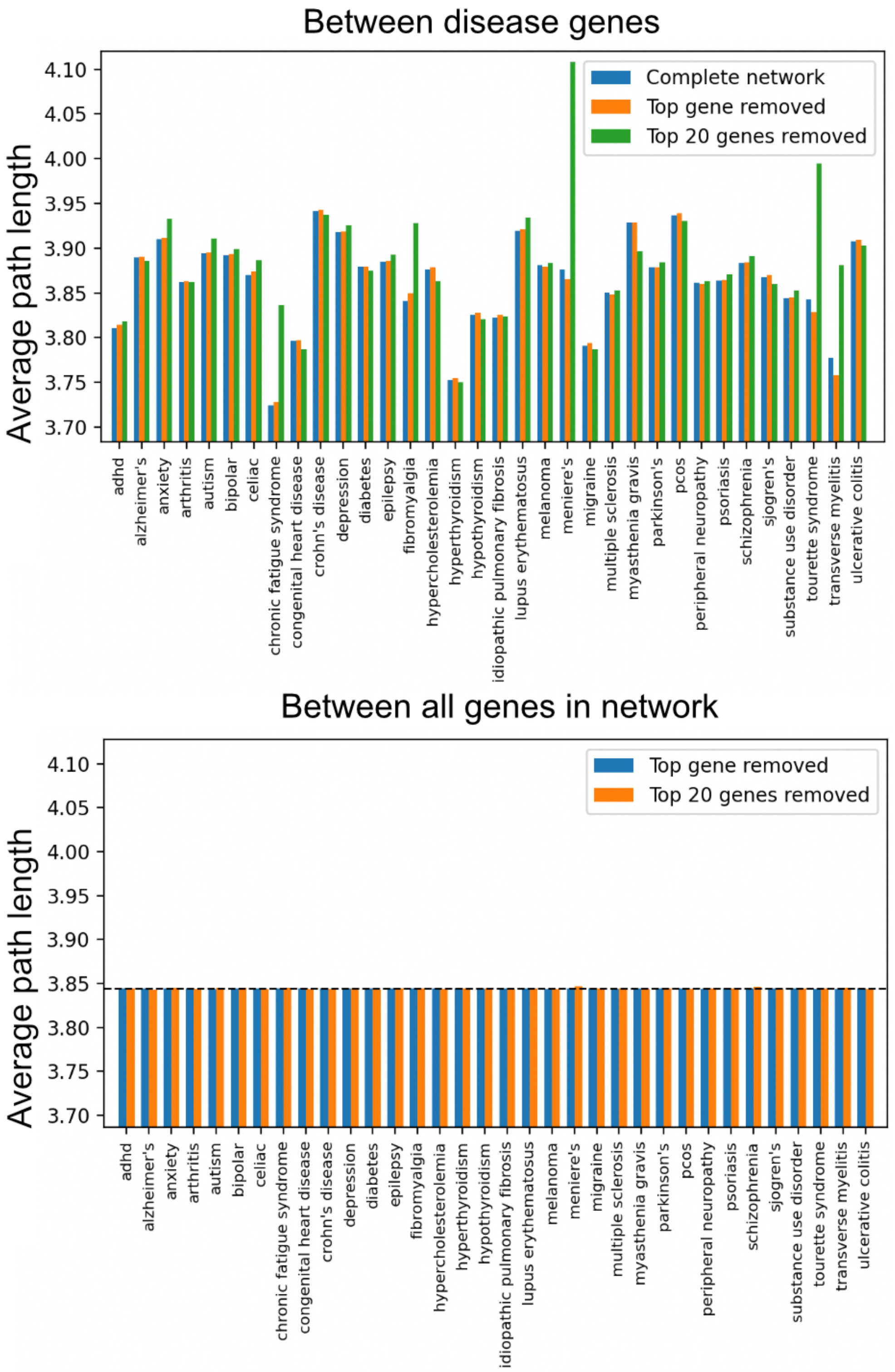


Figure 3. (Top) Comparison of average path lengths between disease-related genes (blue) vs. path lengths between these genes following the removal of the single most-referenced (orange) or twenty most-referenced (green) genes for each disorder. (Bottom) Comparison of average path lengths among all genes in the network following the removal of the single most-referenced (blue) or twenty most-referenced (orange) genes for each disorder.

Table 1. Summary of PubMed search results for each disorder

Disorder	Article Count	Genes in Network	Disorder	Article Count	Genes in Network
Alzheimer’s	234,989	1,075	Lupus erythematosus	60,294	468
Anxiety	344,106	521	Melanoma	135,069	1,220
Arthritis	308,327	1,135	Meniere’s	4,884	36
ADHD	54,383	187	Migraine	40,428	153
Autism	86,838	536	Multiple sclerosis	103,020	567
Bipolar	81,368	445	Myasthenia gravis	12,387	103
Celiac	27,962	137	Parkinson’s	166,341	773
Chronic fatigue syndrome	9,626	67	PCOS	19,064	260
Congenital heart disease	137,426	508	Peripheral neuropathy	138,204	716
Crohn’s disease	58,671	387	Postural orthostatic tachy.	1,460	11
Depression	558,886	902	Psoriasis	54,692	440
Diabetes	906,510	2,132	Schizophrenia	130,431	615
Epilepsy	144,985	796	Sjogren’s	18,672	224
Fibromyalgia	14,035	61	Substance use disorder	228,189	439
Hypercholesterolemia	36,668	280	Tourette syndrome	5,363	40
Hyperthyroidism	24,621	190	Transverse myelitis	7,677	53
Hypothyroidism	32,787	233	Ulcerative colitis	49,963	413
Idiopathic pulm. fibrosis	16,885	276	**Total**	**4,225,211**	**8,109**

4 Discussion

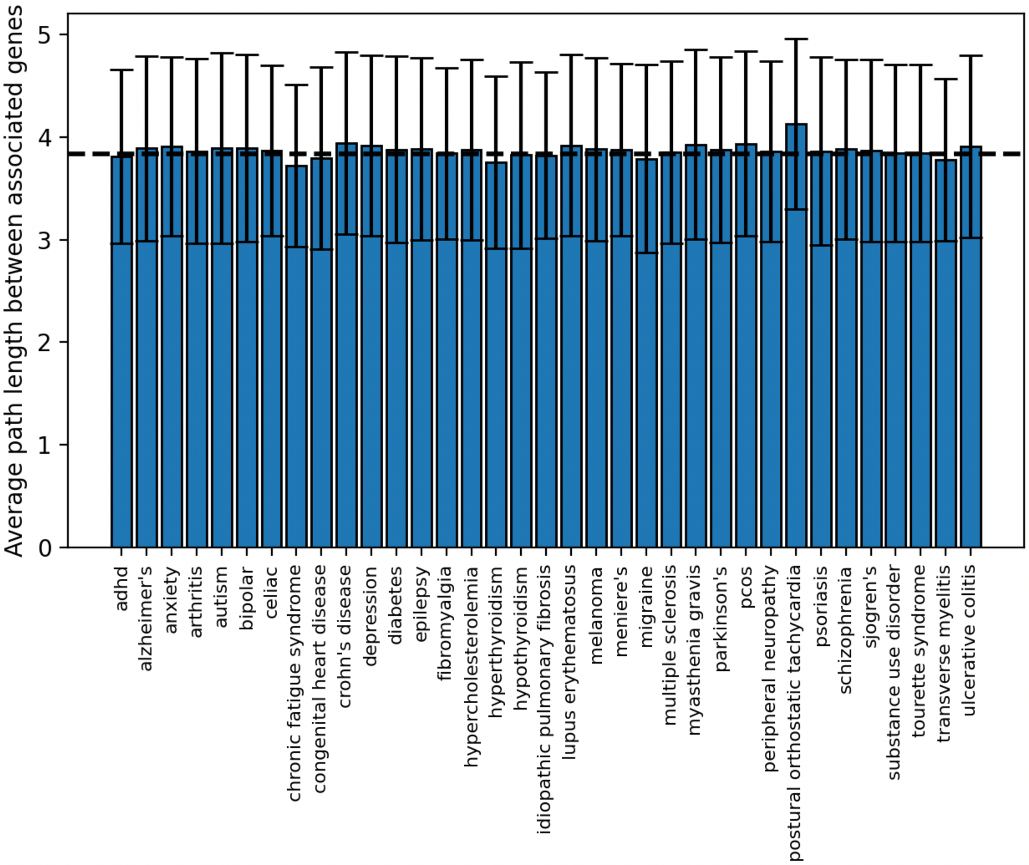
In this project, I aimed to carry out a preliminary exploration of the HuRI protein-protein interactome network. Large datasets such as this one provide promising sources of information for understanding the complex interactions that underlie human disease, but translating big data into clinically relevant information is not trivial. Entering this project, I thought that if I could show that disease-relevant genes tended to be closer together in the network, then other genes close to those in these clusters might serve as promising targets for yet-unknown biomarkers or targets for clinical intervention. I was disappointed to find no such easy result, with the average path length between disease-related genes virtually indistinguishable between average path length between all genes. However, I was surprised to discover that that removal of disease-related genes from the network seems to have a stronger effect on other genes related to the same disease. This seems to suggest that those genes whose connectivity is most affected by disease-specific perturbations may be more likely to be tied to a specific disorder themselves. Supporting this hypothesis requires significant future work, and a possible interesting extension of this project could involve the addition of protein-RNA and protein-DNA interactions, as these provide another significant source of connectivity between the genes underlying diseases. For example, the autism-related FMR1 gene encodes an RNA-binding protein (Richter & Zhou 2021); inclusion of RNA binding partners would likely significantly increase the centrality of FMR1 in the network. In summary, this project has provided me with a starting point for thinking about clinical translation of omics data.

References

Luck K, Kim DK, Lambourne L, et al. A reference map of the human binary protein interactome. *Nature*. 2020;580(7803):402-408.

Richter JD, Zhao X. The molecular biology of FMRP: new insights into fragile X syndrome. *Nat Rev Neurosci*. 2021;22(4):209-222.

Supplemental Materials



Supplemental Figure 1. Average path length between each set of disease-relevant genes. Each bar represents the mean ± standard deviation.