Predicting Missing Links in the Human Disease Network

Ted Conklin
Computer Science
CU Boulder
rico6210@colorado.edu

Introduction

For my project I conducted an analysis on the OMIM Human Disease Network. The goal of this project is to answer the research question : Can network analysis be used to predict genetic connections in unkown diseases. The purpose of this project is to discover tendencies and correlation of genetic disorder behaviors based on network proximity, and use network proximity analysis to predict tendencies of diseases that aren't yet in the network. In the past, analyzing genetic diseases had to be done one at a time. However, with the help of data science and network analysis, understanding the genes implicated in a disease, as well as the functional modules affected can be made significantly easier.

The network, which uses data from the genetic deasese database constructed by The Online Mendelian Inheritance in Man (OMIM), is structured as a bipartite graph in which nodes consist of all known genetic disorders, and the specific genes associated with these disorders. The links, or edges, in the network are all gene-disease links, as the network is bipartite with genes and diseases existing as separate sets. A gene and a disease are linked in the network if a mutation in that gene is implicated in the disease. Additionally, not all edges are one-to-one, as a single gene can be implicated in several diseases, and vice-versa. For example, mutations in the TP53 gene have been linked to 11 known cancer disorders. One of the aspects my network analysis aims to discover is : if a new unknown disease is found to be implicated by mutations in the TP53 gene, how likely is this disease in fact a cancer-related disorder ? Additionally, can we predict which functional modules and tissues will be affected by this new disorder ?

1 PROPOSED WORK & METHODS

The plan for this project is to split the work into four sections of development. The first section will involve the collection *and* formatting of data.

1.1.1 Data Collection. The first step of my project was getting the OMIM network data into Python. This involved downloading a GEXF file from Gephi, importing it into Cytoscape, and then exporting the graph as a GraphML file from which the NetworkX Python library can read it and turn it into a mutable graph. The resulting Python network consists of 1419 nodes and 3926 edges.

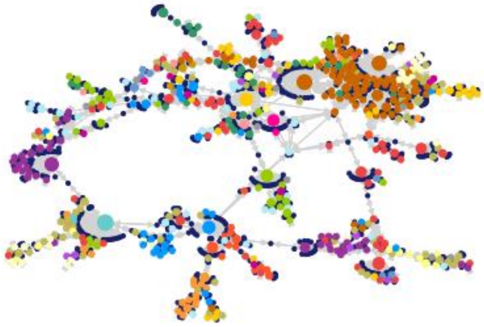


Figure (1.1) – Main Network

In the graph visual above you can see the human disease network in its entirety, with grey lines representing gene-disease links, the dark blue dots representing gene nodes, and the other dots representing genetic disorders. Additionally, the size of a given disease node represents the number of genes implicated in that disease. The color of the disorder nodes indicates the "disorder class" of the disease. In the OMIM dataset, the disorder class refers to the physiological system affected by a given disorder. The data I gathered contains 22 unique disorder classes including:

Bone, Cancer, Cardiovascular, Connective Tissue, Dermatological, Developmental, Ear/nose/throat, Endocrine, Gastrointestinal, Hematological, Immunological, Metabolic, Muscular, Neurological, Nutritional, Ophthamological, Psychiatric, Renal, Respiratory, Skeletal, Multiple, and Unclassified.

Within my analysis I will be using these disorder classes as a method of identifying disease function, as according to source [1] diseases of a given disorder class tend to target the same functional modules and tissues in the human body.

1.1.2 Building the HDN and GDN Networks. The next portion of the data collection component is to restructure the bipartite graph into one-mode representation models. The first one-mode network I built was the Human Disease Network (HDN). This is the network which the study in source [1] analyzed. For this model, the nodes consist of all the genetic disorders in the dataset. For the edges, two diseases in this graph are linked if there exists a gene that is implicated in both diseases.

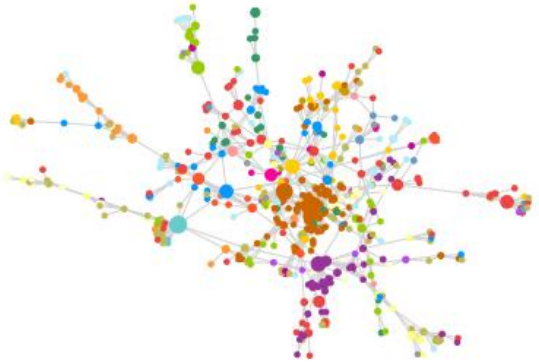
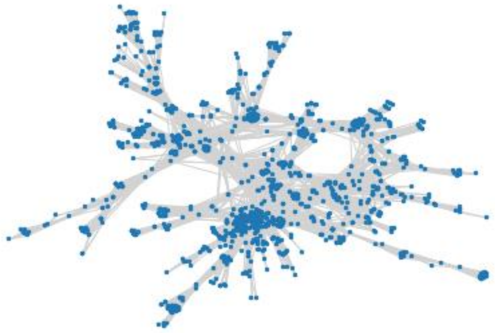


Figure (1.2) – HDN

The figure above displays the Networkx drawing of the HDN. The graph contains 516 disease nodes and 1188 edges. The mean degree is 4.6, with a clustering coefficient of 0.43 and mean geodesic distance of 6.51. Similarly to the complete network, node sizes represent the number of genes connected to the disease and colors represent disorder class. Additionally, the length of the edge between two disease nodes is inversely related to the number of genes shared by the diseases. In other words, two nodes that are close together share many of the same genes, which is common amongst diseases within the same disorder class. Nodes that are connected but far apart typically only share one or two genes. The edges in this figure are a good visual display of network proximity of diseases.

The next network to build was the Gene Disease Network (GDN), a different one-mode representation where the nodes are all genes, and two genes are linked if there exists a disease in which both genes are implicated.



Figure(1.3)-DGN -

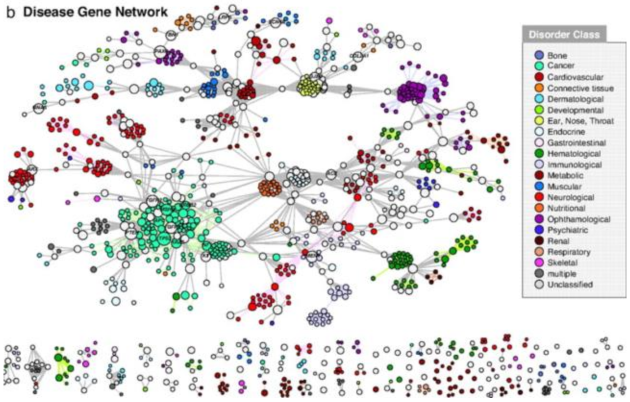


Figure (1.4) – DGN from Source [1]

In figure 1.3 you can see the visual NetworkX graph drawing output from my code. As you may notice the structure of my model closely resembles that of figure 1.4, which is the network model created in the study from source [1].

1.2 Predictive Modeling. The second development phase of this project was to determine the 'predictability' of my networks. To evaluate this I used built three different link predictors: a Jaccard coefficient and degree product predictor, as well as a random uniform predictor that will be used as a baseline for measuring the success of the other two predictors.

To set up this experiment for each network, I first generated replicas of the networks with the same nodes, but only 80% of the edges in the originals were observed in the replicas. The other 20% of edges, selected at random, were added to a missing edge list, which I will denote as 'Y'. Once the partial networks were built, the second step of the experiment was to apply my three predictors on every non-edge in the partial graphs. If my predictors were successful, the true missing edges within the set Y would have relatively high scores from the predictors.

1.3 Model Testing. Since my networks are so large, I needed a way to measure and visualize the success of my predictors. To do this I first sorted my results by each predictor category and computed the True Positive Rate (TPR) and False Positive Rate (FPR) for each. Secondly, I used these two values to compute the AUC score for each predictor and plot an ROC curve to visualize the results.

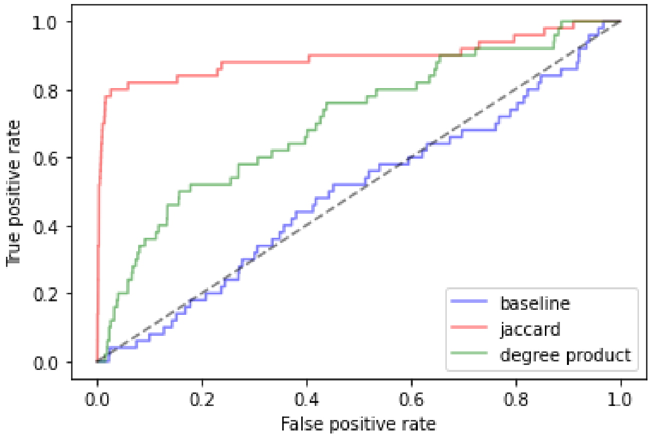


Figure (1.5) – ROC curve of the HDN

1.4 Accuracy Evaluation. As shown in the plot above, both the Jaccard and degree product predictors performed significantly better than random baseline predictor. The baseline predictor yielded a mean AUC score of almost exactly 0.5 over three iterations, which is to be expected because when scores are generated randomly there is no hierarchical structure of edges, meaning false positives and true positives should occur at the same rate. On the other hand, the other AUC scores of the HDN computed by the algorithm were:

Jaccard Predictor: 0.947
Degree Product Predictor: 0.729
Baseline: 0.499

For the main bipartite network, the results were also very good:

Jaccard Predictor: 0.974
Degree Product Predictor: 0.821
Baseline: 0.528

While the AUC scores are likely a sufficient representation of network structure, I wanted to take my structure analysis one step further. A good way to measure the reliability of link predictions within a network is to compare the results to a randomly linked graph with the same nodes. To do this I made another replica of the main network with the same number of nodes and edges. However, I randomly generated all of the edge connections to eliminate the original network structure. In theory, running my prediction algorithms on a random-link graph should result in all of the AUC scores being around 0.5, or equivalent to the baseline.

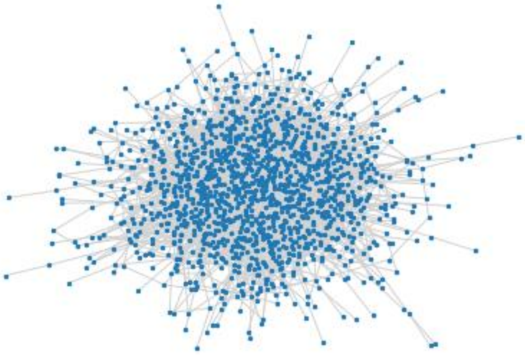


Figure (1.6) – Random-link Network

After running a prediction simulation of the random, the resulting AUC scores were:

Jaccard Predictor: 0.534
Degree Product Predictor: 0.458
Baseline: 0.479

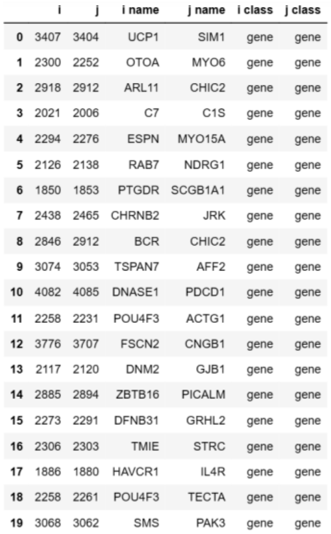
As expected, these scores were around the random baseline AUC of 0.5, proving that my networks are far more predictable than a random-link graph with no modular structure. After calculation, number of connections within the same disorder class was roughly 7 times higher than in the random link graph. This is consistent with the study from source [1] which found an 8-fold difference.

1.5 Analysis. The question to be asked now is: What do these prediction scores mean? The success of our Jaccard predictor tells us that there is sufficient network structure to accurately predict missing links based on Jaccard similarity. For example, if we added a new disease to our full network, based on the gene mutations observed in the disease we can use the Jaccard predictor to predict which other diseases it would be linked to. From these predicted links we can observe the disorder class of the likely connected diseases and infer the disorder class of our new disease. Another type of connection we can predict is:

Based on network proximity of genes, can we predict which functional modules will be affected by mutations in those genes.

2.1 Predicting Missing Links

2.1.1 Predicting Missing Links in Genes. After setting up my networks and analyzing the prediction success of my algorithms, the next step of my project is to apply my algorithms to predict missing links in the human genome data. My first application of my code was to predict these missing gene-gene connections. To do this I ran my predictor simulation function on the GDN network, and sorted my results by the Jaccard scores, from highest to lowest.

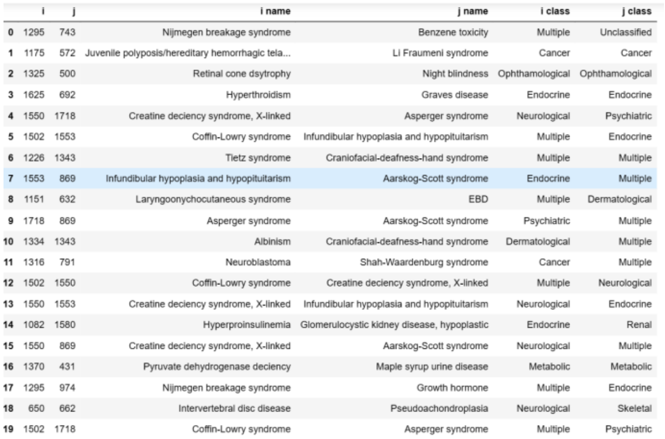


	i	j	i name	j name	i class	j class
0	3407	3404	UCP1	SIM1	gene	gene
1	2300	2252	OTOA	MYO6	gene	gene
2	2918	2912	ARL11	CHIC2	gene	gene
3	2021	2006	C7	C1S	gene	gene
4	2294	2276	ESPN	MYO15A	gene	gene
5	2126	2138	RAB7	NDRG1	gene	gene
6	1850	1853	PTGDR	SCGB1A1	gene	gene
7	2438	2465	CHRNB2	JRK	gene	gene
8	2846	2912	BCR	CHIC2	gene	gene
9	3074	3053	TSPAN7	AFF2	gene	gene
10	4082	4085	DNASE1	PDCD1	gene	gene
11	2258	2231	POU4F3	ACTG1	gene	gene
12	3776	3707	FSCN2	CNGB1	gene	gene
13	2117	2120	DNM2	GJB1	gene	gene
14	2885	2894	ZBTB16	PICALM	gene	gene
15	2273	2291	DFNB31	GRHL2	gene	gene
16	2306	2303	TMIE	STRC	gene	gene
17	1886	1880	HAVCR1	IL4R	gene	gene
18	2258	2261	POU4F3	TECTA	gene	gene
19	3068	3062	SMS	PAK3	gene	gene

Figure (2.1) – gene link predictions

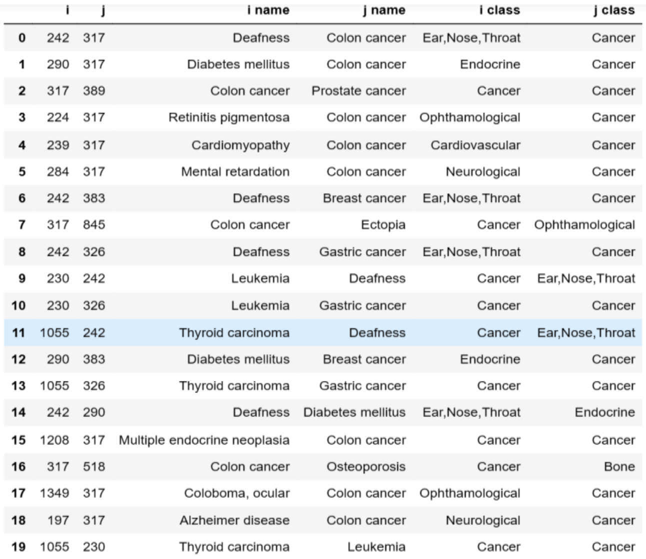
The table in figure 2.1 is a list of the gene connections with the highest Jaccard scores. In other words, this is a list of genes that are most likely, based on disease connections, target the same functional modules. Based on my research these links were very accurate. One significant result was the predicted connection between the OTOA and MYO6 genes. Although not connected in the network, they had a very high Jaccard coefficient, or network proximity, in the Main network. According to source [2] and [3], both genes are implicated in residual hearing loss. All of the other gene links were actually in the GDN, showing that it is possible to predict gene connections by looking solely at their connections with disorders.

2.1.2 *Predicting Missing Links in Diseases*. In addition to predicting missing links between genes, I also wanted to use network proximity in the HDN predict missing links between diseases. To do this I ran a similar analysis to that in part 2.1.1, in which I ranked diseases, that were not linked in the HDN, by Jaccard score and degree product score to predict the most likely links. What I gathered from my results was quite interesting.



	i	j	i name	j name	i class	j class
0	1295	743	Nijmegen breakage syndrome	Benzene toxicity	Multiple	Unclassified
1	1175	572	Juvenile polyposis/hereditary hemorrhagic tela...	Li Fraumeni syndrome	Cancer	Cancer
2	1325	500	Retinal cone dsytrophy	Night blindness	Ophthamological	Ophthamological
3	1625	692	Hyperthroidism	Graves disease	Endocrine	Endocrine
4	1550	1718	Creatine deciency syndrome, X-linked	Asperger syndrome	Neurological	Psychiatric
5	1502	1553	Coffin-Lowry syndrome	Infundibular hypoplasia and hypopituitarism	Multiple	Endocrine
6	1226	1343	Tietz syndrome	Craniofacial-deafness-hand syndrome	Multiple	Multiple
7	1553	869	Infundibular hypoplasia and hypopituitarism	Aarskog-Scott syndrome	Endocrine	Multiple
8	1151	632	Laryngoonychocutaneous syndrome	EBD	Multiple	Dermatological
9	1718	869	Asperger syndrome	Aarskog-Scott syndrome	Psychiatric	Multiple
10	1334	1343	Albinism	Craniofacial-deafness-hand syndrome	Dermatological	Multiple
11	1316	791	Neuroblastoma	Shah-Waardenburg syndrome	Cancer	Multiple
12	1502	1550	Coffin-Lowry syndrome	Creatine deciency syndrome, X-linked	Multiple	Neurological
13	1550	1553	Creatine deciency syndrome, X-linked	Infundibular hypoplasia and hypopituitarism	Neurological	Endocrine
14	1082	1580	Hyperproinsulinemia	Glomerulocystic kidney disease, hypoplastic	Endocrine	Renal
15	1550	869	Creatine deciency syndrome, X-linked	Aarskog-Scott syndrome	Neurological	Multiple
16	1370	431	Pyruvate dehydrogenase deciency	Maple syrup urine disease	Metabolic	Metabolic
17	1295	974	Nijmegen breakage syndrome	Growth hormone	Multiple	Endocrine
18	650	662	Intervertebral disc disease	Pseudoachondroplasia	Neurological	Skeletal
19	1502	1718	Coffin-Lowry syndrome	Asperger syndrome	Multiple	Psychiatric

Figure (2.2) – Jaccard Rankings



	i	j	i name	j name	i class	j class
0	242	317	Deafness	Colon cancer	Ear,Nose,Throat	Cancer
1	290	317	Diabetes mellitus	Colon cancer	Endocrine	Cancer
2	317	389	Colon cancer	Prostate cancer	Cancer	Cancer
3	224	317	Retinitis pigmentosa	Colon cancer	Ophthamological	Cancer
4	239	317	Cardiomyopathy	Colon cancer	Cardiovascular	Cancer
5	284	317	Mental retardation	Colon cancer	Neurological	Cancer
6	242	383	Deafness	Breast cancer	Ear,Nose,Throat	Cancer
7	317	845	Colon cancer	Ectopia	Cancer	Ophthamological
8	242	326	Deafness	Gastric cancer	Ear,Nose,Throat	Cancer
9	230	242	Leukemia	Deafness	Cancer	Ear,Nose,Throat
10	230	326	Leukemia	Gastric cancer	Cancer	Cancer
11	1055	242	Thyroid carcinoma	Deafness	Cancer	Ear,Nose,Throat
12	290	383	Diabetes mellitus	Breast cancer	Endocrine	Cancer
13	1055	326	Thyroid carcinoma	Gastric cancer	Cancer	Cancer
14	242	290	Deafness	Diabetes mellitus	Ear,Nose,Throat	Endocrine
15	1208	317	Multiple endocrine neoplasia	Colon cancer	Cancer	Cancer
16	317	518	Colon cancer	Osteoporosis	Cancer	Bone
17	1349	317	Coloboma, ocular	Colon cancer	Ophthamological	Cancer
18	197	317	Alzheimer disease	Colon cancer	Neurological	Cancer
19	1055	230	Thyroid carcinoma	Leukemia	Cancer	Cancer

Figure (2.3) – Degree Product Rankings

When looking at the two diagrams above you will notice that each table features a predominant disorder class. Based on the Jaccard results, many of the most likely links included nodes that belong to multiple disorder classes, and nodes that belong to disorder classes, with less local clustering and genetic heterogeneity, in the HDN. When looking at the disorder types in the HDN it is clear that nodes belonging to multiple disorder classes act as ‘bridge’ nodes within the network structure.

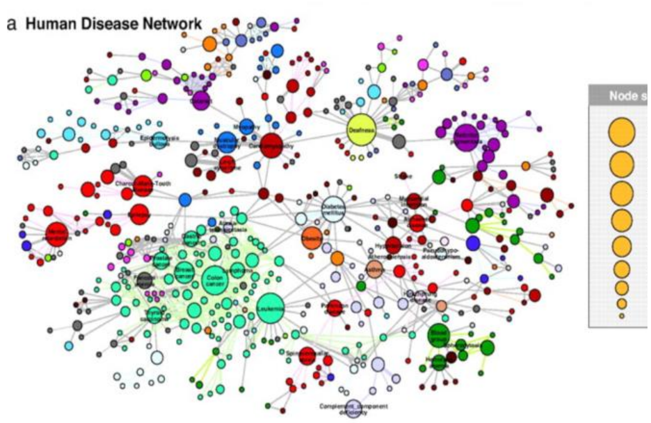


Figure (2.4) – Clear version of HDN

In figure 2.4 you can more clearly visualize the HDN network and the class of nodes in it. In the graphic, disorders of multiple classes have a dark grey color. Two common characteristics of these nodes are the fact that they are well dispersed throughout the graph, and they contain connections to diseases of more than one class, which makes sense because they belong to multiple classes themselves. Multiple class nodes are the ideal representation of bridge nodes in our graph, and it is interesting that the Jaccard predictor heavily favored these nodes as the most likely to contain missing links.

On the other hand, the results from the degree product predictor rankings were quite the opposite. Looking at the table in figure 2.3, all but one of the top twenty most likely missing links contained at least of node of the cancer disorder class. If you look at cancer nodes in the HDN visualizations (represented as green nodes in figure 2.4 and orange nodes in figure 1.2), the cancer cluster is clearly the largest and most tightly connected cluster in the network.

To generalize my findings, the Jaccard link predictor was more likely to predict missing links in scattered bridge nodes while the Degree product predictor was more likely to predict missing links in 'hub' nodes with high degrees. These findings were expected as the degree product algorithm calculates a score based on the degree on each node while Jaccard scores based on proportion of total links shared with the other node.

3 Discussion

After building the models, running my algorithms and analyzing the results, the final step was to construct a network analysis of the Human Disease Network in its entirety.

In section 1.4 I calculated the AUC scores for each of my three predictors on the HDN. Based on the resulting scores over several iterations it was clear that the Jaccard was able to predict the actual missing links with a higher level of accuracy than the degree product predictor. So, what does this tell us about our network and its structure? Based on the analysis in section 2.1.2, the Jaccard predictor ranks link probability based on network proximity as opposed to node degree. Combining these results with those in section 1.4 tells us that the network structure favors local clustering, as opposed to having 'hub' nodes. A hub node in the case of our network would be a disease that shares genes with diseases of many different disorder classes. If these hub nodes were highly present in the model, then we could infer connections between many nodes and these hubs would be present, resulting in a higher degree product AUC score. Instead, the best way to predict a missing link with a new disease in this network would be to look at the Jaccard similarities between our new disease and all other diseases in the network. Rank the most likely links based on the average Jaccard similarity of each node in the 22 respective disorder classes and calculate the probability of our new disease belonging to each class. This is a method I would like to attempt in the future.

As for the conclusions drawn from this project, I was able to define the structure of the network, determine missing link predictability, and predict the most likely missing links in our network. Hopefully network analyses like this can be used to help humans better understand genetic diseases and make new discoveries.

4 Bibliography

1. Barabási, Albert-László, et al. "The Human Disease Network | PNAS." *PNAS*, 3 Apr. 2007, www.pnas.org/doi/10.1073/pnas.0701361104.

2. "Entry - *600970 - Myosin VI; Myo6 - OMIM." *OMIM*, omim.org/entry/600970. Accessed 8 May 2023.

3. Sugiyama, Kenjiro, et al. "Mid-Frequency Hearing Loss Is Characteristic Clinical Feature of Otoa-Associated Hearing Loss." *Genes*, 16 Sept. 2019, www.ncbi.nlm.nih.gov/pmc/articles/PMC6770988/#:~:text=The%20OTOA%20gene%20(Locus%3A%20DNFB22,is%20located%20on%20chromosome%2016p12.

4. Gephi. "Gephi Datasets." *GitHub*, github.com/gephi/gephi/wiki/Datasets. Accessed 8 May 2023.

5. "Index of Complex Networks." *Colorado.Edu*, icon.colorado.edu/?#!/networks. Accessed 8 May 2023.