Creating an ALS diseasome

Maria T. Chavez[1]
[1]Department of Neurology, University of California San Francisco, CA 94158, USA

Recent genetic studies in multiple diseases have identified hundreds of genomic loci harboring risk variants. These variants are shared between diseases at unexpectedly high rates, providing a molecular basis for shared pathogenesis. If properly used, these results could allow us to identify specific pathways underlying disease; explain disease heterogeneity by grouping patients by molecular causes rather than over- all symptomatology; identify drug targets for repositioning; and develop more rational approaches to diagnosis and therapy targeting these molecular defects. Here I present the ALS diseasome (disease network) I created.

Amyotrophic lateral sclerosis (ALS), was first described in 1869 by Dr. Jean Martin Charcot[1]. ALS is a neurodegenerative disorder which causes injury and death of upper motor neurons in the motor cortex, and of lower motor neurons in the brainstem and spinal cord. This results in progressive muscle weakness, atrophy, and spasticity culminating in death from respiratory failure. The average survival rate from symptom onset is approximately 2 to 5 years, although a small proportion of patients have a more favorable prognosis. It is estimated that about 285,200 people currently live with ALS and about 16 of them die each day.

Currently only one drug (Riluzole) is approved for the treatment of ALS. It extends patient's lifespan for a maximum of 3 years. Additionally, one puzzle for understanding ALS is that the known ALS-causing gene products have diverse physiological functions, for example, mRNA processing, Axonal Transport, Endosomal Vesicle trafficking, and Ubiquitination.

Decades-long efforts to map human disease loci, at first genetically and later physically, followed by recent positional cloning of many disease genes and genome-wide association studies, have generated extensive lists of disorder–gene association pairs. In addition, recent efforts have focused on creating maps of the relationships between different disease genes. Most of the successful studies building on this approach have focused on a single disease, using network-based tools to gain a better understanding of the relationship between the genes implicated in a selected disorder [2]

Here I take a conceptually different approach, exploring which diseases are related to the disease of interest (Amyotrophic Lateral Sclerosis) based on shared disease genes, which might help understand pathogenic mechanisms better and find candidates for treating the disease.

Corresponding author: Chavez, MT (materechm@gmail.com)
Keywords: GWAS; shared pathogenesis, diseasome, biological network, genetics

In order to create a diseasome, a larger network of diseases (nodes) connected to genes (edges) is created first, where connections (edges) come from the GWAS catalog[3], and the OMIM database[4]. The Disease Ontology (DO) [5] is used to provide a standardized terminology for disease concepts across both resources. Two processing steps are performed on the resulting network of DO terms annotated with associated genes: (1) closely related concepts are merged by transferring annotations to a single term and removing the other term (for example 'breast cancer' is consolidated with 'breast carcinoma'); and (2) terms of inappropriate generality are removed such as 'immune system disease'.

The edge weight between diseases in the network is calculated as the intersection of genes divided by the union of genes. Edges between diseases with no shared genes are omitted. To verify accuracy of calculations, the results from Cotsapas and Hafler were replicated.7

The network was graphed using cytoscape.

501 conditions from the GWAS catalog and OMIM were found to share genes with ALS. After processing the resulting network of DO terms, the amount of conditions was reduced to 101.

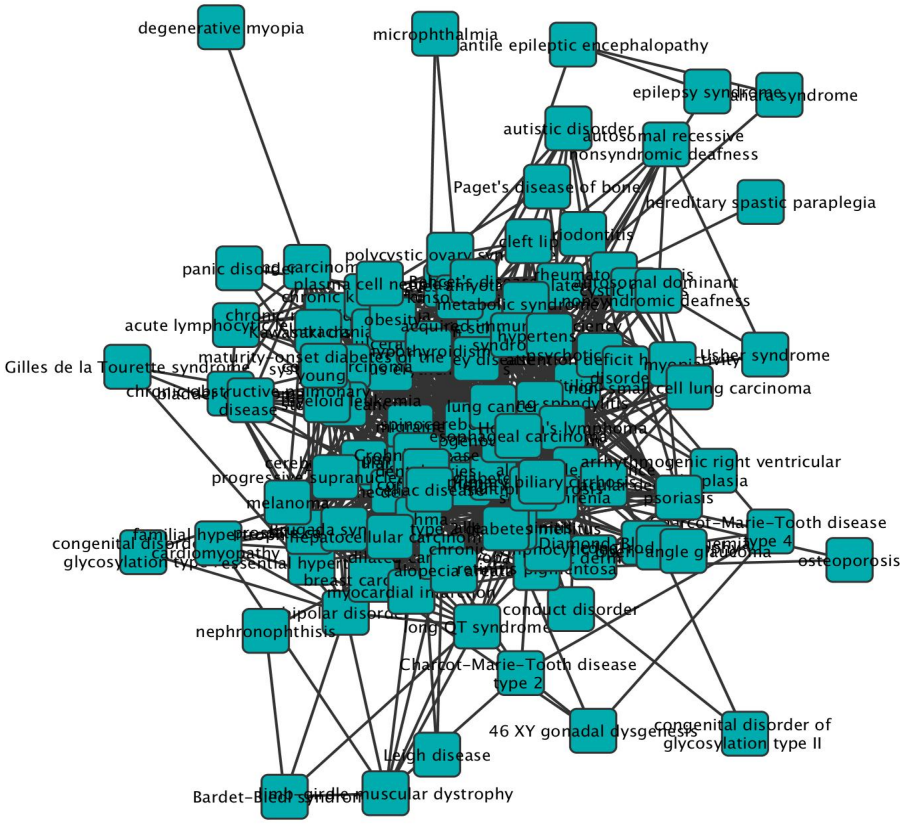

Figure 1: graphical version of the ALS diseasome

source	source_name	target	target_name	proximity
DOID:332	amyotrophic lateral sclerosis	DOID:1441	spinocerebellar ataxia	0.033778892
DOID:332	amyotrophic lateral sclerosis	DOID:14221	metabolic syndrome X	0.033465684
DOID:332	amyotrophic lateral sclerosis	DOID:3908	non-small cell lung carcinoma	0.031858574
DOID:332	amyotrophic lateral sclerosis	DOID:10871	age related macular degeneration	0.028770204
DOID:332	amyotrophic lateral sclerosis	DOID:10652	Alzheimer's disease	0.026946355
DOID:332	amyotrophic lateral sclerosis	DOID:0050541	Charcot-Marie-Tooth disease type 4	0.025954782
DOID:332	amyotrophic lateral sclerosis	DOID:9352	type 2 diabetes mellitus	0.023328564
DOID:332	amyotrophic lateral sclerosis	DOID:4905	pancreatic carcinoma	0.022716122
DOID:332	amyotrophic lateral sclerosis	DOID:10584	retinitis pigmentosa	0.020966546
DOID:332	amyotrophic lateral sclerosis	DOID:5408	Paget's disease of bone	0.020735304

Table 1: proximity to ALS based on percentage of shared genes

The top 10 most proximal diseases to ALS are shown in table 1.

Drug bank was used to find drugs to treat said diseases. 51 FDA approved drugs were found. A list was made with those drugs, followed to a search to determine which drugs had already been tested for ALS. 13 of those had been tested, 8 increased lifespan in SOD1 mice, 1 decreased SOD1 levels but also decreased cell viability, 2 had no effect, 1 produced a short-term benefit in patients, and 1 helped fracture prevention on ALS patients.

Additionally, one drug (Glyburdine), which is approved for treating type II diabetes, has been shown to be more effective than Riluzole in a rat model of cervical spinal cord injury[6] (hasn't been tested on ALS).

Future steps

The gene lists are by no means comprehensive, which is a big weakness of this kind of approach. Databases are improving with time and there are other databases, which could be included to make the approach more comprehensive. Another way of making the diseasome more comprehensive is manually adding genes for diseases you are knowledgeable about. A website could be created to allow researchers to add more genes to the lists of genes associated with diseases.

The drugs found using this approach (that haven't been tested) can be tested on ALS to see if they are a viable treatment option.

Acknowledgments

I thank Laura Deming for motivating me to do this project and reading my research proposals; Daniel Himmelstein for helping process the resulting network; and Ali Yahya and Dan Mane for going over my code and helping me understand why it wasn't working.

References

1. Charcot, JM & Joffory, A. Deux cas d'atrophie musculaire progressive avec lesion de la substance grise et des faisceaux antero-lateraux de la moelle epiniere. Arch Physiol Nerol Pathol **2**, 744-754 (1869)
2. Lim, J et al. A Protein-protein interaction network for human

inherited ataxias and disorders of Purkinje cell degeneration. Cell **4**, 801-814 (2006)
3. Hindorff LA, MacArthur J (European Bioinformatics Institute), Morales J (European Bioinformatics Institute), Junkins HA, Hall PN, Klemm AK, and Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. Accessed December 2013
4. Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), December 2013. World Wide Web URL:http://omim.org/
5. http://disease-ontology.org/
6. Simard, JM et al. Comparative effects of glibenclamide and riluzole in a rat model of severe cervical spinal cord injury. Exp Neurol. 233, 566-574 (2012)
7. Cotsapas, C & Hafler, DA. Immune-mediated disease genetics: the shared basis of pathogenesis. Trends in immunol. 34, 22-26 (2013)

Predicting influential ALS genes

Maria T. Chavez[1]
[1]Stanford Online High School, CA, 94305, USA

Candidate gene prioritization is the process of ranking genes based on how likely it is that they are associated with a disease. Most candidate gene prioritization methods use a list of seed genes to rank the other genes. Genes that are ranked highly are likely to be associated with the disease. Here I used text mining to expand my seed list and calculated measure of influence on gene networks to predict genes associated with Amyotrophic Lateral Sclerosis.

Amyotrophic lateral sclerosis (ALS), was first described in 1869 by Dr. Jean Martin Charcot[1]. ALS is a neurodegenerative disorder which causes injury and death of upper motor neurons in the motor cortex, and of lower motor neurons in the brainstem and spinal cord. This results in progressive muscle weakness, atrophy, and spasticity culminating in death from respiratory failure. The average survival rate from symptom onset is approximately 2 to 5 years, although a small proportion of patients have a more favorable prognosis. It is estimated that about 285,200 people currently live with ALS and about 16 of them die each day.

Mutations in several genes have been found to cause fALS (C9ORF72 is responsible for 30-40% of fALS in the USA, SOD1 cause about 20% of fALS, TARDBP cause 5%, FUS cause 5% and ANG mutations cause 1%). It is estimated that 60% of individuals with fALS have an identified genetic mutation; the cause for the rest remains unknown.

The cause of sporadic ALS is largely unknown, but familial and epidemiological data suggest that genetic and environmental factors contribute to its pathogenesis. No single gene has been associated with increased risk for ALS.

Given that the cause of most ALS cases is unknown and that the disease cannot be explained by simple genotype-phenotype relationships, it is important to work on identifying more disease associated genes.

Many approaches have been dedicated to the discovery of candidate genes, but most require a lot of time and money. Using the data already available in papers and databases to identify genes that are likely associated with a disease and then validating the findings on a disease model might be a more efficient way to find genes associated with diseases.

I collected a list of seed genes associated with ALS from the OMIM and GWAS databases. The result was a list containing 55 genes associated with ALS. I then used seed genes to

Corresponding author: Chavez, MT (materechm@gmail.com)
Keywords: GWAS; biological network, genetics, text mining

initiate pagerank on gene networks (co-expression, physical interaction, predicted interactions, co-localization, shared protein domains, genetic interactions and pathways), to calculate a measure of influence for each gene. 20 genes that were not initially on my seed list were ranked highly, which means they are likely associated with the disease.

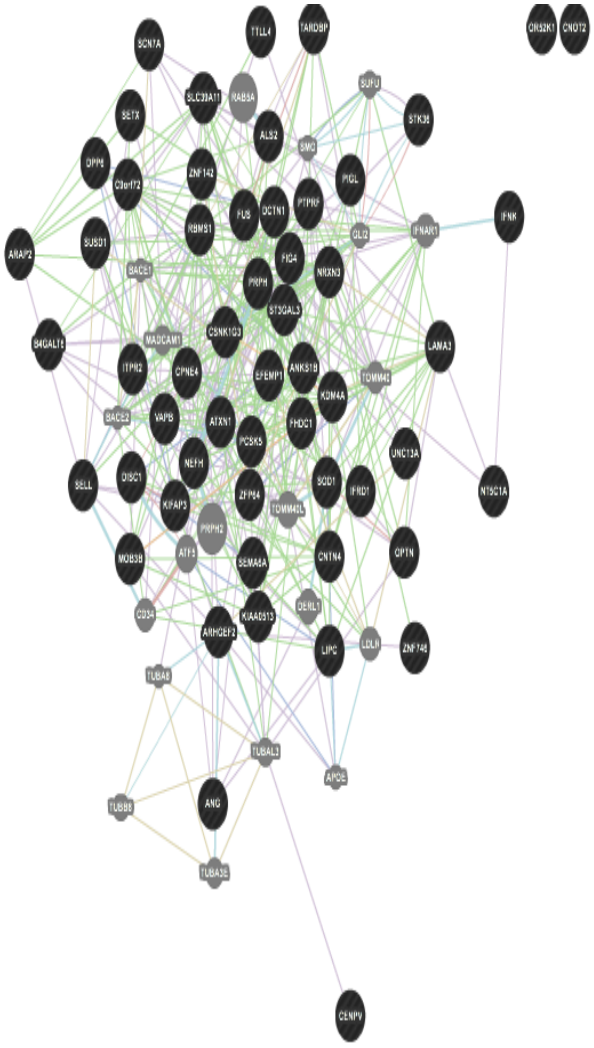

Figure 1: gene networks. Black circles represent seed genes; grey ones are predicted. Purple lines indicate co-expression, red physical interactions, orange predicted interactions, blue co-localization, green genetic interactions and light blue pathways

The seed genes + the newly found genes have different functions; the function with greatest coverage is neuron part. Surprisingly there was not much coverage on functions associated with ALS, like response to oxidative stress and neural death regulation.

Since databases are by no means comprehensive, I decided to use text mining to expand my seed genes list. I included genes with an R scaled score of >= 35 that showed up in >= 3 abstracts. The result was a seed list containing 231 genes. I repeated the page rank step using the new seed list and again 20 genes that were not on my seed list were ranked highly (only one gene, MSMP, was also one of the 20 genes that were found using the shorter seed list)

The top 3 functions with most coverage on the new list of genes were: Perinuclear region of cytoplasm, response to inorganic substance and transition metal ion binding.

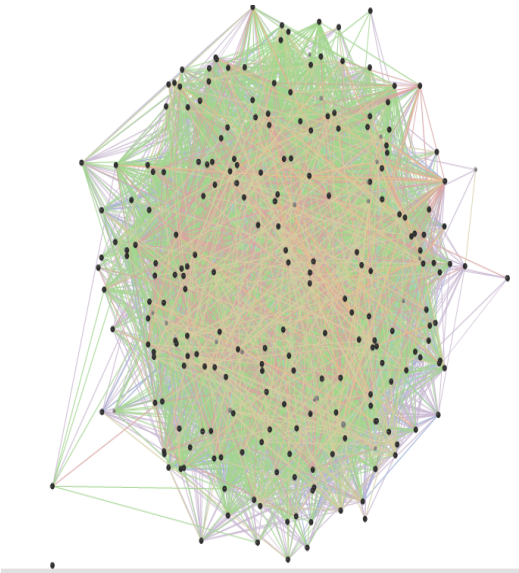

Figure 2: gene networks using new seed list

The entire list of genes and gene functions can be found in supplemental information

Acknowledgments
Thanks to the Thiel Foundation for providing the funding necessary to do this project.

References

1. Charcot, JM & Joffory, A. Deux cas d'atrophie musculaire progressive avec lesion de la substance grise et des faisceaux antero-lateraux de la moelle epiniere. Arch Physiol Nerol Pathol **2**, 744-754 (1869)
2. Lim, J et al. A Protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. Cell **4**, 801-814 (2006)
3. Hindorff LA, MacArthur J (European Bioinformatics Institute), Morales J (European Bioinformatics Institute), Junkins HA, Hall PN, Klemm AK, and Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. Accessed December 2013
4. Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), December 2013. World Wide Web URL:http://omim.org/
5. http://disease-ontology.org/