# Exploring Genotypic and Phenotypic Approaches to Aggregating Disease Variants

Ann Cirincione (Mentor: Dr. Olivier Bodenreider)

University of Maryland, Baltimore County, Baltimore, MD
National Library of Medicine, National Institutes of Health, Bethesda, MD

## Abstract

Understanding the effect of human genetic variations on disease can provide insight into phenotype-genotype relationships. While some genetic markers linked to disease susceptibility have been identified, a large number are still unknown. In this study, we have compiled over 80,000 human genetic variants associated with disease phenotypes. To standardize variants with differing terminologies, we first normalized diseases to concepts from the Unified Medical Language System (UMLS). Disease-disease connection networks were constructed through two different approaches, associating two diseases if they consisted of a similar genotype and/or similar phenotype. Novel disease connections resulting from these networks offer explanations for a more complete understanding of the molecular mechanisms underlying disease. In the future, these aggregation approaches will be incorporated with gene-drug associations to link diseases with potentially novel drug connections, enabling new diagnostic and therapeutic interventions.

## Introduction

Current repositories of disease-associated human genetic variants encompass over 4,000 genes and 17,000 disease phenotypes, derived mostly from manual extraction from the literature. Our compilation of these variants spans multiple different databases, including the Online Mendelian Inheritance in Man (OMIM), Clinical Variance database (ClinVar), Universal Protein Resource (UniProt), and Human Gene Mutation Database (HGMD). These databases may represent diseases differently depending on the vocabulary they use as a reference. Therefore, to standardize disease phenotype terms, we normalized diseases to UMLS concepts.

To make new connections between diseases in our database of known disease-variant associations, we aggregated both variants and diseases. Variants were aggregated at the genotypic level, associating two variants if they were mutated on the same gene. Diseases were aggregated at the phenotypic level, associating two diseases if they shared similar Human Phenotype Ontology (HPO) manifestations. Novel disease connections were visualized through constructed networks, providing information on diseases that were not previously known through a single approach. This provides potential new insight into the relationship between phenotype and genotype, as well as the molecular mechanisms involved in disease. This will in turn lead to improved ability to develop diagnostic and therapeutic inventions that will aid in patient treatment.

## Methods

Our compilation of over 80,000 human disease variants unifies specific genetic data from databases including OMIM, ClinVar, UniProt, and HGMD. Each unique variant is defined by a specific gene, protein mutation, and disease description. The first step to aggregating variants was to normalize variant

descriptions to disorder concepts in the UMLS Metathesaurus using exact and normalized string matching functions from the UMLS Terminology Server (UTS) Application Program Interface (API). Terms unable to be normalized through the UTS API in their original form underwent enhanced manual normalization. Manual normalization included splitting terms, expanding stop words, and substituting Roman/Arabic numerals. Enhanced input was then re-run through the UTS API to capture additional normalizations.

After normalization, variants were aggregated at the gene level, grouping those mutated on the same gene using information already available in our database. Diseases were then aggregated at the phenotypic level, grouping those that were semantically similar. To do this, the subset of OMIM diseases normalized to UMLS concepts were first associated with their HPO manifestations. The R package HPOSim was then used to calculate pairwise similarities between manifestations of two diseases. Diseases with a final similarity score in the 75$^{th}$ percentile (sim >= 0.63) were considered semantically similar and therefore aggregated together.

Lastly, networks were constructed using Cytoscape to visualize aggregations through both approaches. First, a bipartite graph was constructed, linking human genetic variants to diseases. Second, a disease-disease connection graph was constructed, linking diseases that had similar manifestations and/or were mapped to the same gene.

## Results

Through UTS API functions and enhanced manual normalization, 87% of human genetic disease variants were normalized to UMLS disorder concepts. Those unable to be normalized will be manually curated in future work. Normalized variants were used to create a bipartite graph linking variants to diseases, resulting in many disease hubs where similar diseases clustered together (Fig 1A-C). The three largest connected components of the graph were analyzed, and the diseases within those components were used to create a disease-disease connection graph (Fig 1D). Diseases were connected if they shared the same gene and/or similar manifestations. This resulted in one main cluster of diseases linked only by similar manifestations, and another cluster linked only by the same gene, with a subset of those clusters linked through both approaches. Those linked only by similar manifestations made up 85% of all connections, those linked only by same gene made up 9%, and those linked through both approaches made up 6%.

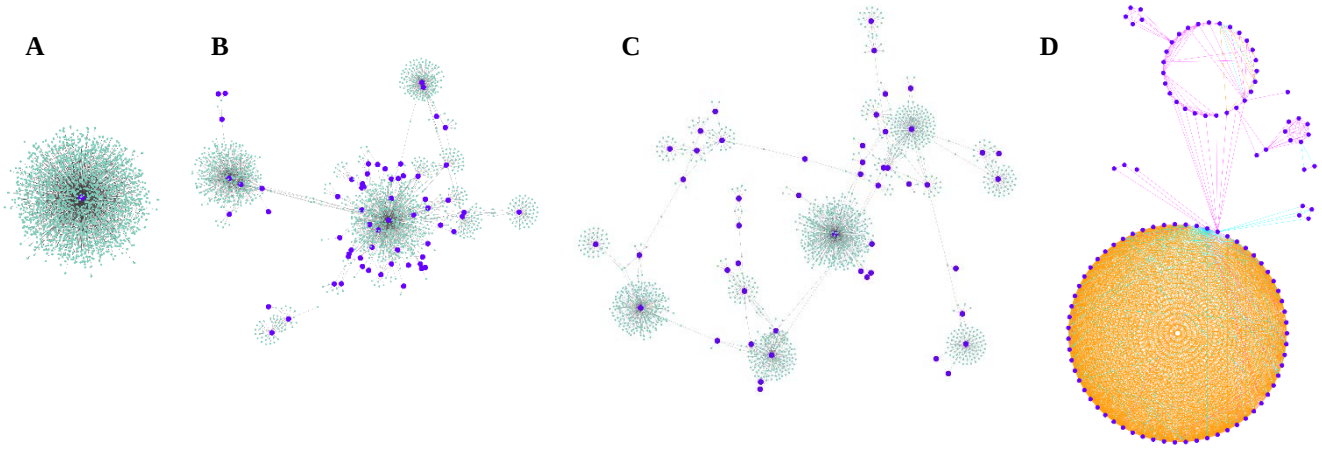


**Figure 1. Network graphs linking diseases to associated human genetic variants and connecting multiple diseases.** The three largest components of the bipartite graph are presented above (A, B, C, in order of size), where green nodes are variants and purple nodes are diseases. The diseases from the bipartite graph are connected in D, either through a genetic approach (pink edges), phenotypic approach (orange edges), or both approaches (blue edges).

## Discussion

The three largest connected components of the bipartite disease-variant graph consisted of three main disease categories: hemophilia, eye diseases, and congenital/developmental diseases. The eye diseases were highly connected through phenotype, and made up the main cluster connected through similar manifestations in the disease-disease connection graph. Additional diseases connected through similar manifestations but not mapped to the same gene include *Mental Retardation, Autosomal Dominant 6* and *Autism, X-Linked, Susceptibility To, 1*. Both *Retinitis Pigmentosa 64 and Cone-rod Dystrophy 16* are mapped to the same gene, but do not have similar manifestations according to our cutoff. *Branched-chain Keto Acid Dehydrogenase Kinase Deficiency* and *Autistic Disorder* were connected through both approaches, mapped to the same gene and with similar manifestations. Diseases connected through one approach and not the other may be explained by those that involve a wide variety of symptoms and/or are associated to more than one gene. Novel disease-disease connections provide insight into the relationship between underlying molecular mechanisms, and can lead to new treatment approaches repurposed between newly-associated diseases, especially with the future addition of gene-drug associations.