Identification of Molecular Properties to Cluster Genetic Markers of Cardiovascular Disease

Ann G Cirincione, Kaylyn L. Clark, Maricel G. Kann

Department of Biological Sciences
University of Maryland, Baltimore County
Baltimore, MD 21250

Abstract

Cardiovascular disease (CVD) is the leading cause of death worldwide in both men and women. It is especially prevalent in the United States, where it is responsible for one out of every four deaths. Genetic markers linked to susceptibility to CVD have been identified, however a large number are still unknown. We have mapped 15% of disease variants from the Human Gene Mutation Database (HGMD), 62% from the Clinical Variance database (ClinVar), and 28% from the Universal Protein Resource (UniProt) to the Online Mendelian Inheritance in Man (OMIM), which has direct links to the Human Phenotype Ontology (HPO). From these variants, we will extract the subset linked to CVD, as well as increase coverage of variants mapped to OMIM. We aim to use machine learning techniques and algorithms to cluster new genetic variants to molecular properties of known CVD markers in order to better diagnose and treat individual patients. For this purpose we have compiled a list of molecular properties including protein domain interactions, common gene ontologies, and metabolic pathways. In the future, these methods will be used to match individual genome data to corresponding CVD clusters, developing new diagnostic tools to personalize and optimize diagnosis of CVD.

Background

Human Disease Variants

- Naturally occurring mutations result in variations in an individual's genome
- Some of these genetic variants have been linked to susceptibility to specific phenotypes and/or diseases (Figure 1)

Gene	Mutation	Phenotype/Disease
MYH7	p.THR1351MET	Cardiomyopathy, hypertrophic(HGMD)

Figure 1. Human disease variant example.[1] Sample gene/variant entry in our human disease variants database; includes the gene name, specific amino acid mutation, disease with which the mutation is associated, and source it was mapped from.

Functional Enrichment

- KEGG: computational representation of a biological system, linking genes to metabolic pathways (Figure 2)
- GO: ontology model that links genes to functions

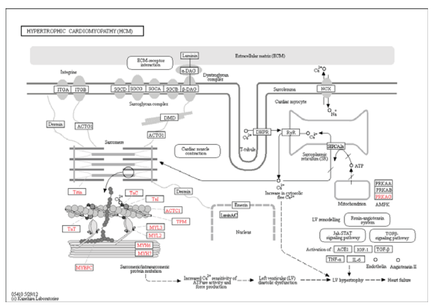

Figure 2. Sample KEGG pathway.[2] Mapping genes and systems involved in hypertrophic cardiomyopathy.

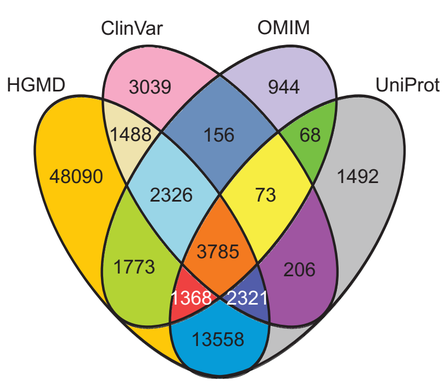

Figure 3. Comparing coverage of human disease variants.[1] Comparison of unique gene/variant entries and their overlap from HGMD[3], ClinVar[4], OMIM[5], and UniProt[6] databases.

Methods

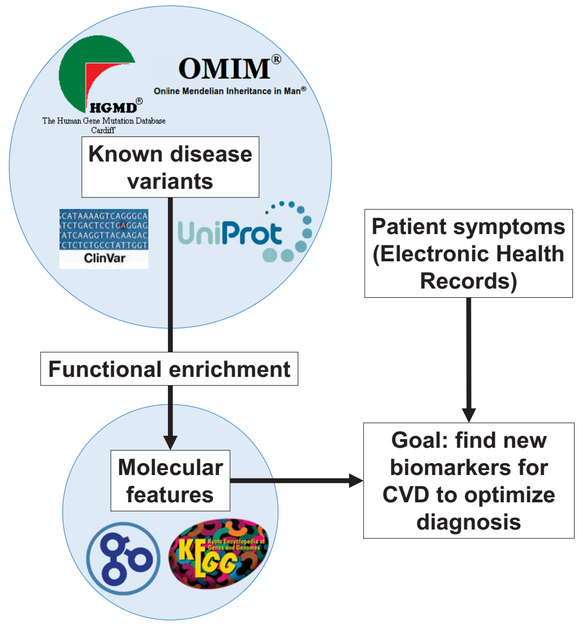

Figure 4. Overall work flow. Known human disease variants are first run through a functional enrichment to find molecular features (GO and KEGG associations) that link gene and phenotype. Combining these features with patient symptom data will allow us to better classify patients and personalize/optimize diagnosis.

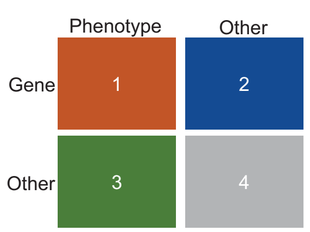

Figure 5. Fisher's Exact Test.
1) Gene associated with phenotype
2) Gene associated with other phenotypes
3) Phenotype associated with other genes
4) All other gene associations

Acknowledgements

This work was funded, in part, by NIH, R01LM009722 to M.G.K. (co-investigator). Grant #1446406: "On Ramps" to Full Professor: Institutional Support for Post-Family Leave Faculty Research Reintegration.

A special thanks to Adetola Abdulkadir, Brianna Richardson, Thomas Coard, Elizabeth Hobbs, Nikhil Yesupriya, Dr. Thomas Peterson, and Dr. Maricel Kann.

Results

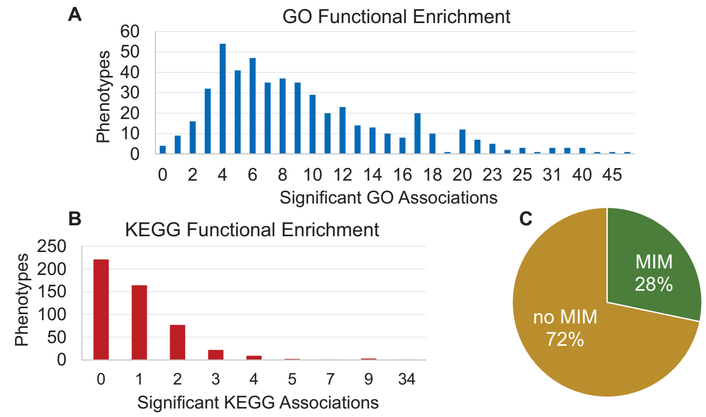

Figure 6. Functional enrichment results and classification of database sources. (A) GO functional enrichment of 500 unique phenotypes grouped by MIM number, showing the frequency of amount of significant GO associations ($p \leq 0.01$ (Figure 5)). (B) KEGG functional enrichment using same methodology as (A). (C) Classification of all disease variants by those that are mapped to a MIM number and those that are not.

Future Work

- Obtain complete coverage of variants mapped to MIM
- Complete functional enrichment for additional unique phenotypes, specifically a CVD subset
- Incorporate EHR and use machine learning techniques to reclassify patients using genomic data in addition to phenotype

References

[1] Peterson TA et al. "Towards precision medicine: advances in computational approaches for the analysis of human variants." J Mol Biol. 2013 Nov 1;425(21):4047-63. doi: 10.1016/j.jmb.2013.08.008. Epub 2013 Aug 17. Review.
[2] Kanehisa, Minoru et al. "KEGG as a Reference Resource for Gene and Protein Annotation." *Nucleic Acids Research* 44.Database issue (2016): D457–D462.*PMC*.
[3] Stenson et al (2003), The Human Gene Mutation Database (HGMD®): 2003 Update. *Hum Mutat* (2003) 21:577-581.
[4] Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), World Wide Web URL: http://omim.org/
[5] Landrum, Melissa J. et al. "ClinVar: Public Archive of Interpretations of Clinically Relevant Variants." *Nucleic Acids Research* 44.Database issue (2016): D862–D868. *PMC*.
[6] The UniProt Consortium UniProt: a hub for protein information Nucleic Acids Res. 43: D204-D212 (2015)
[7] The Gene Ontology Consortium. Gene Ontology Consortium: going forward. (2015) Nucleic Acids Res 43 Database issue D1049–D1056.