Using network clustering to predict copy number variations associated with health disparities

Substantial health disparities exist between African Americans and Caucasians in the United States. Copy number variations (CNVs) are one form of human genetic variations that have been linked with complex diseases and often occur at different frequencies among African Americans and Caucasian populations. Here, we aimed to investigate whether CNVs with differential frequencies can contribute to health disparities from the perspective of gene networks. We inferred network clusters from human gene/protein networks based on two different data sources. We then evaluated each network cluster for the occurrences of known pathogenic genes and genes located in CNVs with different population frequencies, and used false discovery rates to rank network clusters. This approach let us identify five clusters enriched with known pathogenic genes and with genes located in CNVs with different frequencies between African Americans and Caucasians. These clustering patterns predict two candidate causal genes located in four population-specific CNVs that play potential roles in health disparities

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.584v1 | CC-BY 4.0 Open Access | rec: 3 Nov 2014, publ: 3 Nov 2014

Using network clustering to predict copy number variations associated with health disparities.

Yi Jiang[*]

Department of Computer Science and Engineering, University of Tennessee at Chattanooga, TN

Hong Qin[*§]

Department of Biology, Spelman College, Atlanta, GA

Li Yang

Department of Computer Science and Engineering, University of Tennessee at Chattanooga, TN

* Co-first authors.

§ Corresponding author. Phone: (404) 270-5757, Fax: (404) 270-5725, Email: hqin@spelman.edu

Abstract

Substantial health disparities exist between African Americans and Caucasians in the United States. Copy number variations (CNVs) are one form of human genetic variations that have been linked with complex diseases and often occur at different frequencies among African Americans and Caucasian populations. Here, we aimed to investigate whether CNVs with differential frequencies can contribute to health disparities from the perspective of gene networks. We inferred network clusters from human gene/protein networks based on two different data sources. We then evaluated each network cluster for the occurrences of known pathogenic genes and genes located in CNVs with different population frequencies, and used false discovery rates to rank network clusters. This approach let us identify five clusters enriched with known pathogenic genes and with genes located in CNVs with different frequencies between African Americans and Caucasians. These clustering patterns predict two candidate causal genes located in four population-specific CNVs that play potential roles in health disparities.

Keywords:

Health disparities, Copy Number Variations (CNVs), gene network, clustering, gene-disease association, Gene Ontology (GO).

List of Key Abbreviations:

CNV: Copy number variation

SNP: Single nucleotide polymorphism

PPIN: Protein-protein interaction network

HPRD: Human protein reference database

PPI: Protein-protein interaction

AA: African American

MCL: Markov Cluster Algorithm

FDR: false discovery rate

GO: Gene ontology

OMIM: Online Mendelian Inheritance in Man

dbSNP: Single Nucleotide Polymorphism Database

SERCA1: Sarco/endoplasmic reticulum Ca^{2+}-ATPase 1

Introduction

Health disparities refer to differences in the disease distribution and/or health outcomes across racial and ethnic groups. In United States, health disparities in African Americans are found in life expectancy, death rates, and health measures (National Center for Health Statistics 2013). In addition to social determinants, such as socio-economical status, health care access and cultural practices, human genetic variations play a significant role in health disparities. Genetic variations at different frequencies among populations can lead to differences in disease susceptibility. Studies on genetic variations and disease association are greatly advanced by the completion of the International HapMap Project and new genome sequencing techniques (Ramos & Rotimi 2009).

Genome-wide association studies (GWAS) are currently an effective approach to identify diseases-associated genetic variations (Hirschhorn & Daly 2005; Wang et al. 2005). Although GWAS have revealed many disease-associated single nucleotide polymorphisms (SNPs), GWAS are often limited to individual genetic variations and often do not address complex gene interactions. Moreover, associated SNPs are often located in haplotype blocks that contain more than one gene. To address these limitations, human gene networks have been used to improve GWAS detection of genes associated with complex diseases, such as the comorbidity analysis (Sharma et al. 2013), an improved guilt-by-association method (Baranzini et al. 2009; Lee et al. 2011), and a distance-based scoring method using seeded diseases genes (Liu et al. 2012).

Copy number variations (CNVs) are duplications or deletions of genomic segments that can contain one or more genes (McCarroll & Altshuler 2007). CNVs have been associated with complex diseases such as autism (Gilman et al. 2011; Glessner et al. 2009). Computational tools and methods have been developed, such as the CNV annotator (Zhao & Zhao 2013) and NETBAG (Gilman et al. 2011), to address the potential roles of CNVs in human diseases. Recently, it is reported that CNVs can occur at different frequencies between African Americans and Caucasians (McElroy et al. 2009), and naturally the question about the potential roles of CNVs in health disparity is raised.

Here, we aim to investigate the clustering of pathogenic genes and genes in CNVs with different population frequencies in two human gene/protein networks, in order to better understand health disparities between African Americans and Caucasians. The current human gene/protein networks contain thousands of interacting molecules (Barabasi et al. 2011; Vidal et al. 2011). We will partition gene networks into clusters and use these clusters to predict potential diseases associated with population-specific CNVs, based on the rationale that interacting genes often share similar functions (Pizzuti et al. 2012).

Materials and Methods

Our overall work flow is shown in Figure 1. To identify potential diseases associated with CNVs, our basic idea is to identify gene interaction clusters that involve genes in population-specific CNVs. The diseases associated with a CNV-gene's interacting genes are potential diseases associated with this CNV. Specifically, we first obtained two human gene/protein networks and partitioned them into gene clusters. We then performed statistical tests on each cluster to estimate its significances of containing pathogenic genes and genes in population-specific CNVs. Finally, we ranked gene clusters based on false discovery rates (FDRs). High-ranked clusters were enriched both for pathogenic genes and for genes in CNVs with differential frequencies between African-Americans and Caucasians. These clusters were then searched for enriched Gene Ontology (GO) terms and related disease phenotypes.

Network clustering

We obtained two human gene/protein networks, one from Human Protein Reference Database (HPRD) (Mishra et al. 2006; Peri et al. 2003; Prasad et al. 2009) and another from MultiNet (Khurana et al. 2013). The HPRD network (referred to as HPRDNet) contains only physical protein-protein interactions (PPIs), whereas MultiNet is a unified network including PPI, phosphorylation, metabolic, signaling, genetic and regulatory networks. The two networks share 8468 genes (89.6% of HPRDNet and 58.6% of MultiNet) but only 8769 interactions (23.8% of HPRDNet and 8% of MultiNet). These two networks were both partitioned into gene clusters using the Markov Cluster (MCL)

Algorithm (van Dongen 2000). Clustering was done with the inflation parameter I ranging from 1.1 to 2.0 with a step of 0.1. Descriptive statistics of the two networks and their clustering results are summarized in Supporting Table S1.

Mapping of CNVs and SNPs

CNV coordinates were obtained from a CNV map in African Americans and Caucasians (McElroy et al. 2009). There are three types of CNVs in this map: (1) CNVs only occur in African Americans, (2) CNVs only occur in Caucasians, and (3) CNVs occurred in both African Americans and Caucasians. To simplify the analysis, we further partitioned the last type: CNVs that occurred more than 50% in African Americans or in Caucasians were combined with the first and second types of CNVs, respectively. This repartition resulted in two modified CNV sets with differential population frequencies. The coordinates of these CNVs were then searched in the UCSC Genome Database (Karolchik et al. 2014) through its MySQL API to obtain the corresponding gene sets. For simplicity, CNVs that occur more frequently in African Americans were called African-American CNVs or CNV_AA; CNVs that occur more frequently in Caucasians were called Caucasian CNVs or CNV_CA.

Disease-associated SNPs were retrieved from a file, OmimVarLocusIdSNP.bcp, from the FTP site of Single Nucleotide Polymorphism Database (dbSNP) (Sherry et al. 2001). Coordinates of these SNPs were then queried against the

MySQL API of the UCSC Genome Database to identify genes in which those SNPs are located. This identified gene set was termed as pathogenic genes. Details of gene mapping results are shown in Supporting Table S2.

Cluster Analyses

Clusters were obtained from both HPRDNet and MultiNet using MCL with a range of ten inflation parameters. For each cluster, contingency tables were constructed using the numbers of pathogenic genes and CNVs related genes (Table 1A and 1B). Right-tailed Fisher's exact tests were applied to these contingency tables to calculate enrichment significance of pathogenic genes, and CNV_AA or CNV_CA genes, respectively. Based on obtained *p*-values, false discovery rates (FDRs) were calculated using the Robust FDR Routine (Pounds & Cheng 2006). Fisher's exact tests and Robust FDR Routine were both performed in the R statistical environment (R Development Core Team 2013). Ranking were applied to clusters with *p*-value<0.10 and FDR<0.20 in both enrichment tests for pathogenic genes and population-preferred CNVs genes. Assuming both enrichment tests are independent, the FDR values were multiplied to jointly rank the network clusters. The same cluster analysis procedure was applied to clustering results with different MCL inflation parameters.

For clarity, we focused our functional analyses on clusters that were consistently ranked at the first place with different MCL inflation parameter values.

Biological Significance Analyses

Biological relevance of selected network clusters were analyzed by GOrilla (Eden et al. 2009) to search for enriched gene ontology (GO) terms. In GOrilla search, genes in the selected clusters were target genes, and all genes in the network were treated as background genes. To investigate the possible links of population-specific CNVs to heath disparities, we first identified significantly enriched GO terms that are associated with CNV_AA or CNV_CA genes. We then focused on the pathogenic genes with the enriched GO terms, and examined their associated disease phenotypes in OMIM database (Online Mendelian Inheritance in Man 2014).

Results and Discussions

Top-ranked network clusters

We performed cluster analyses with ten MCL inflation parameters values for both HPRDNet and MultiNet (Table S1), and scored the resulted clusters for their potential roles in CNV related health disparities (Table S3). For clarity, we focused on clusters that are consistently top-ranked with different MCL inflation parameters. The graph representations of selected clusters are shown in Figure 2.

We found four similar clusters, (AA1, AA2, and AA3 in HPRDNet and AA4 in Multinet), that are enriched both for pathogenic genes and for genes located in African-American CNVs (Table 2). In HPRDNet, cluster AA1, AA2 and AA3 together were ranked at first place five times; and cluster AA4 were ranked five times in Multinet (Table S3). Cluster AA1 contains 11 genes, within which eight are pathogenic genes (Figure 2A). Cluster AA2 and AA3 contain one and two more genes than cluster AA1, respectively (Figure S1). In MultiNet, cluster AA4 contains five genes and can be considered as a sub-cluster of cluster AA1, AA2 and AA3 (Figure 2B). In these four clusters, gene *HSPB1* is mainly duplicated in African Americans (Table 2 and Table 3). Since cluster AA1, AA2 and AA3 were selected from the same network and are highly similar to each other, only cluster AA1 and AA4 were studied in biological significance analyses.

In both HPRDNet and MultiNet, the same cluster, named as CA1, was identified to be enriched with both pathogenic genes and genes located in Caucasian CNVs (Table 2). Cluster CA1 was ranked at first place four times in HPRDNet and seven times in MultiNet (Table S3). This cluster contains five genes, and four of them are associated with diseases (Figure 2C). Cluster CA1 contains gene *ATP2A1* that is duplicated only in Caucasians (Table 3).

Duplication of *HSPB1* and health disparities in African Americans.

Gene *HSPB1* is located in genomic duplication regions occurring more frequently in African Americans (Table 3), and is found in the cluster family of AA1, AA2, AA3, and AA4 (Table 2). For cluster AA1, only one GO molecular function term related to gene *HSPB1* is significantly enriched (Cluster AA1 in Table 4). For cluster AA4, in addition to the same enriched GO molecular functions term, three GO biological process terms and one GO cellular component term are found significantly enriched (Cluster AA4 in Table 4). In the genes with the enriched GO terms, four of them are known to be associated with diseases (Cluster AA1/AA4 in Table 5). Among these four genes, three of them are implicated in health disparities of African Americans. Specifically, gene *CRYAB* is related to dilated cardiomyopathy and myofibrillar myopathy. African Americans were found at higher risk for idiopathic dilated cardiomyopathy compared with Caucasian, and this could not be explained by income, education, alcohol use, smoking, or history of some other diseases (Coughlin et al. 1993). Moreover, gene *CRYAA*, *CRYAB* and *CRYBB2* are all related to various types of cataract. It was reported that age-specific blindness prevalence was higher for African Americans compared with Caucasian, and cataract accounts for 36.8% of all blindness in African American, but for only 8.7% in Caucasian (Congdon et al. 2004).

How could *HSPB1* duplication contribute to health disparities? Based on the direct interaction between *HSPB1* and *CRYAB* and the fact that both genes are expressed in Z-disc (Table 4), it is plausible that *HSPB1* may play an unknown role in cardiomyopathy. Alternatively, *HSPB1* might be involved in cataract, because *HSPB1*, *CRYAA* and *CRYAB* interact with each other and all can negatively regulate apoptotic process (Table 4). Studies suggested that lens epithelial cell apoptosis may be a common cellular basis for initiation of non-congenital cataract formation (Li et al. 1995), and inhibition of epithelial cell apoptosis may be one possible mechanism that inhibits cataract development (Nahomi et al. 2013). Our results here argue for further experimental studies to test the possible role of *HSPB1* CNVs in cardiomyopathy or cataract/blindness in African Americans.

Duplication of *ATP2A1* and cardiomyopathy.

Gene *ATP2A1* in cluster CA1 is located in a genomic duplication region that occurs only in Caucasians (Table 3). We found that three genes in cluster CA1 are enriched with various GO biological process terms that involve *ATP2A1* (Cluster CA1 in Table 4). All of the three genes are related to diseases when they are mutated (Cluster CA1 in Table 5).

How would *ATP2A1* influence health disparities? Among the diseases related to the pathogenic genes in cluster CA1, idiopathic dilated cardiomyopathy occurs less often in Caucasians than in African Americans (Coughlin et al. 1993). One possibility is that higher copies of *ATP2A1* may offer some benefits to Caucasians. Studies have shown that increased activity of sarco/endoplasmic reticulum Ca^{2+}-ATPase 1 (SERCA1), which is encoded by *ATP2A1*, can partially rescue the heart from ·OH-induced injury (Hiranandani et al. 2006), and protect the heart from ischemia-reperfusion (I/R) injury (Talukder et al. 2007). Another possibility is that higher copies of *ATP2A1* only lead to moderate risk of cardiomyopathy in Caucasians, and this moderate effect is overshadowed by other genetics factors not covered by our CNV dataset.

Remarks and future directions

Although genetic factors play a crucial role in health disparities, only a few association studies have been reported in health disparities in common complex diseases, such as breast cancer (Long et al. 2013), prostate cancer (Bensen et al. 2014; Bensen et al. 2013; Xu et al. 2011), type 2 diabetes (Ng et al. 2014) and vascular diseases (Wei et al. 2011).

Our study here is closely related to network-based meta-analyses of GWAS results (Atias et al. 2013; Leiserson et al. 2013). One important aim of network-based meta-analysis of GWAS data is to distinguish the bona fide causal gene from others in the same haplotype block associated with the significant SNP. Likewise, our network approach aims to predict a potential causal gene from a population-specific CNV that can be associated with pathogenic genes.

Noticeably, our method does not require network permutations, whereas many existing methods of network/pathway based meta-analyses of GWAS data do. This difference is because we first partitioned the network into clusters and then perform association tests. In comparison, many network based GWAS meta-analysis methods use traversal distances to seed genes to evaluate candidate genes. This kind of traversal distance based method generally prohibits pre-partition of network into clusters and require network permutations for estimation of p-values. It can be seen that our cluster-based method naturally accommodate multiple candidate genes in the association analysis, whereas traversal distance in a network is by definition often limited to single candidate gene evaluation.

In future studies, we plan to improve network clustering results by integrating functional genomics data sets, such as gene expressions, into gene networks to generate weighted interactions.

Conclusions

In this study, gene clusters were inferred from two human gene/protein networks, HPRDNet and MultiNet, by MCL clustering algorithm with different parameters. Each cluster was ranked based on products of FDR values based on the right-tailed Fisher's exact tests for enrichment of pathogenic or CNV-genes. Five clusters were consistently found to be enriched with both pathogenic genes and genes located in African-American or Caucasian CNVs. In cluster AA1, AA2, AA3 and AA4, gene *HSPB1* is duplicated more frequently in African-Americans. In clusters CA1, gene *ATP2A1* is duplicated only in Caucasians. All gene clusters are associated with certain diseases that occur more often in one population than in the other. Although we only studied population-preferred CNVs and did not consider the roles of other genetic factors, our computational studies have generated some interesting hypotheses for further experimental studies to understand health disparities in these diseases.

Author contributions

HQ initiated this study. HQ and LY designed the overall project. YJ implemented the methods and performed data analyses. All authors participated in writing.

Acknowledgements

The authors thank three reviewers for constructive comments that have greatly improved the quality and presentation of this work. YJ and LY were partially supported by Tennessee Higher Education Commission's Center of Excellence in Applied Computational Science and Engineering.

Reference:

Atias N, Istrail S, and Sharan R. 2013. Pathway-based analysis of genomic variation data. *Curr Opin Genet Dev* 23:622-626.

Barabasi AL, Gulbahce N, and Loscalzo J. 2011. Network medicine: a network-based approach to human disease. *Nat Rev Genet* 12:56-68.

Baranzini SE, Galwey NW, Wang J, Khankhanian P, Lindberg R, Pelletier D, Wu W, Uitdehaag BM, Kappos L, Gene MSAC, Polman CH, Matthews PM, Hauser SL, Gibson RA, Oksenberg JR, and Barnes MR. 2009. Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum Mol Genet* 18:2078-2090.

Bensen JT, Xu Z, McKeigue PM, Smith GJ, Fontham ET, Mohler JL, and Taylor JA. 2014. Admixture mapping of prostate cancer in African Americans participating in the North Carolina-Louisiana Prostate Cancer Project (PCaP). *Prostate* 74:1-9.

Bensen JT, Xu Z, Smith GJ, Mohler JL, Fontham ET, and Taylor JA. 2013. Genetic polymorphism and prostate cancer aggressiveness: a case-only study of 1,536 GWAS and candidate SNPs in African-Americans and European-Americans. *Prostate* 73:11-22.

Congdon N, O'Colmain B, Klaver CC, Klein R, Munoz B, Friedman DS, Kempen J, Taylor HR, and Mitchell P. 2004. Causes and prevalence of visual impairment among adults in the United States. *Arch Ophthalmol* 122:477-485.

Coughlin SS, Labenberg JR, and Tefft MC. 1993. Black-white differences in idiopathic dilated cardiomyopathy: the Washington DC dilated Cardiomyopathy Study. *Epidemiology* 4:165-172.

Eden E, Navon R, Steinfeld I, Lipson D, and Yakhini Z. 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10:48.

Gilman SR, Iossifov I, Levy D, Ronemus M, Wigler M, and Vitkup D. 2011. Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron* 70:898-907.

Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S, Zhang H, Estes A, Brune CW, Bradfield JP, Imielinski M, Frackelton EC, Reichert J, Crawford EL, Munson J, Sleiman PM, Chiavacci R, Annaiah K, Thomas K, Hou C, Glaberson W, Flory J, Otieno F, Garris M, Soorya L, Klei L, Piven J, Meyer KJ, Anagnostou E, Sakurai T, Game RM, Rudd DS, Zurawiecki D, McDougle CJ, Davis LK, Miller J, Posey DJ, Michaels S, Kolevzon A, Silverman JM, Bernier R, Levy SE, Schultz RT, Dawson G, Owley T, McMahon WM, Wassink TH, Sweeney JA, Nurnberger JI, Coon H, Sutcliffe JS, Minshew NJ, Grant SF, Bucan M, Cook EH, Buxbaum JD, Devlin B, Schellenberg GD, and Hakonarson H. 2009. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* 459:569-573.

Hiranandani N, Bupha-Intr T, and Janssen PM. 2006. SERCA overexpression reduces hydroxyl radical injury in murine myocardium. *Am J Physiol Heart Circ Physiol* 291:H3130-3135.

Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, Harte RA, Heitner S, Hinrichs AS, Learned K, Lee BT, Li CH, Raney BJ, Rhead B, Rosenbloom KR, Sloan CA, Speir ML, Zweig AS, Haussler D, Kuhn RM, and Kent WJ. 2014. The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* 42:D764-770.

Khurana E, Fu Y, Chen J, and Gerstein M. 2013. Interpretation of genomic variants using a unified biological network approach. *PLoS Comput Biol* 9:e1002886.

Lee I, Blom UM, Wang PI, Shim JE, and Marcotte EM. 2011. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res* 21:1109-1121.

Leiserson MD, Eldridge JV, Ramachandran S, and Raphael BJ. 2013. Network analysis of GWAS data. *Curr Opin Genet Dev* 23:602-610.

Li WC, Kuszak JR, Dunn K, Wang RR, Ma W, Wang GM, Spector A, Leib M, Cotliar AM, Weiss M, and et al. 1995. Lens epithelial cell apoptosis appears to be a common cellular basis for non-congenital cataract development in humans and animals. *J Cell Biol* 130:169-181.

Liu ZP, Wang Y, Zhang XS, and Chen L. 2012. Network-based analysis of complex diseases. *IET Syst Biol* 6:22-33.

Long J, Zhang B, Signorello LB, Cai Q, Deming-Halverson S, Shrubsole MJ, Sanderson M, Dennis J, Michailiou K, Easton DF, Shu XO, Blot WJ, and Zheng W. 2013. Evaluating genome-wide association study-identified breast cancer risk variants in African-American women. *PLoS One* 8:e58350.

McCarroll SA, and Altshuler DM. 2007. Copy-number variation and association studies of human disease. *Nat Genet* 39:S37-42.

McElroy JP, Nelson MR, Caillier SJ, and Oksenberg JR. 2009. Copy number variation in African Americans. *BMC Genet* 10:15.

Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, Menon S, Hanumanthu G, Gupta M, Upendran S, Gupta S, Mahesh M, Jacob B, Mathew P, Chatterjee P, Arun KS, Sharma S, Chandrika KN, Deshpande N, Palvankar K, Raghavnath R, Krishnakanth R, Karathia H, Rekha B, Nayak R, Vishnupriya G, Kumar HG, Nagini M, Kumar GS, Jose R, Deepthi P, Mohan SS, Gandhi TK, Harsha HC, Deshpande KS, Sarker M, Prasad TS, and Pandey A. 2006. Human protein reference database--2006 update. *Nucleic Acids Res* 34:D411-414.

Nahomi RB, Wang B, Raghavan CT, Voss O, Doseff AI, Santhoshkumar P, and Nagaraj RH. 2013. Chaperone peptides of alpha-crystallin inhibit epithelial cell apoptosis, protein insolubilization, and opacification in experimental cataracts. *J Biol Chem* 288:13022-13035.

National Center for Health Statistics. 2013. Health, United States, 2012: With Special Feature on Emergency Care. Hyattsville, MD.

Ng MC, Shriner D, Chen BH, Li J, Chen WM, Guo X, Liu J, Bielinski SJ, Yanek LR, Nalls MA, Comeau ME, Rasmussen-Torvik LJ, Jensen RA, Evans DS, Sun YV, An P, Patel SR, Lu Y, Long J, Armstrong LL, Wagenknecht L, Yang L, Snively BM, Palmer ND, Mudgal P, Langefeld CD, Keene KL, Freedman BI, Mychaleckyj JC, Nayak U, Raffel LJ, Goodarzi MO, Chen YD, Taylor HA, Jr., Correa A, Sims M, Couper D, Pankow JS, Boerwinkle E, Adeyemo A, Doumatey A, Chen G, Mathias RA, Vaidya D, Singleton AB, Zonderman AB, Igo RP, Jr., Sedor JR, Kabagambe EK, Siscovick DS, McKnight B, Rice K, Liu Y, Hsueh WC, Zhao W, Bielak LF, Kraja A, Province MA, Bottinger EP, Gottesman O, Cai Q, Zheng W, Blot WJ, Lowe WL, Pacheco JA, Crawford DC, Grundberg E, Rich SS, Hayes MG, Shu XO, Loos RJ, Borecki IB, Peyser PA, Cummings SR, Psaty BM, Fornage M, Iyengar SK, Evans MK, Becker DM, Kao WH, Wilson JG, Rotter JI, Sale MM, Liu S, Rotimi CN, and Bowden DW. 2014. Meta-analysis of genome-wide association studies in African Americans provides insights into the genetic architecture of type 2 diabetes. *PLoS Genet* 10:e1004517.

Online Mendelian Inheritance in Man O. 2014. Baltimore, MD: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University.

Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JG, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, and Pandey A. 2003. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* 13:2363-2371.

Pizzuti C, Rombo S, and Marchiori E. 2012. Complex Detection in Protein-Protein Interaction Networks: A Compact Overview for Researchers and Practitioners. In: Giacobini M, Vanneschi L, and Bush W, eds. *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*: Springer Berlin Heidelberg, 211-223.

Pounds S, and Cheng C. 2006. Robust estimation of the false discovery rate. *Bioinformatics* 22:1979-1987.

Prasad TSK, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, and Pandey A. 2009. Human Protein Reference Database--2009 update. *Nucleic Acids Res* 37:D767-772.

R Development Core Team. 2013. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Ramos E, and Rotimi C. 2009. The A's, G's, C's, and T's of health disparities. *BMC Med Genomics* 2:29.

Sharma A, Gulbahce N, Pevzner SJ, Menche J, Ladenvall C, Folkersen L, Eriksson P, Orho-Melander M, and Barabasi AL. 2013. Network-based analysis of genome wide association data provides novel candidate genes for lipid and lipoprotein traits. *Mol Cell Proteomics* 12:3398-3408.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, and Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 29:4.

Talukder MA, Kalyanasundaram A, Zhao X, Zuo L, Bhupathy P, Babu GJ, Cardounel AJ, Periasamy M, and Zweier JL. 2007. Expression of SERCA isoform with faster Ca2+ transport properties improves postischemic cardiac function and Ca2+ handling and decreases myocardial infarction. *Am J Physiol Heart Circ Physiol* 293:H2418-2428.

van Dongen S. 2000. Graph Clustering by Flow Simulation PhD. University of Utrecht.

Vidal M, Cusick ME, and Barabasi AL. 2011. Interactome networks and human disease. *Cell* 144:986-998.

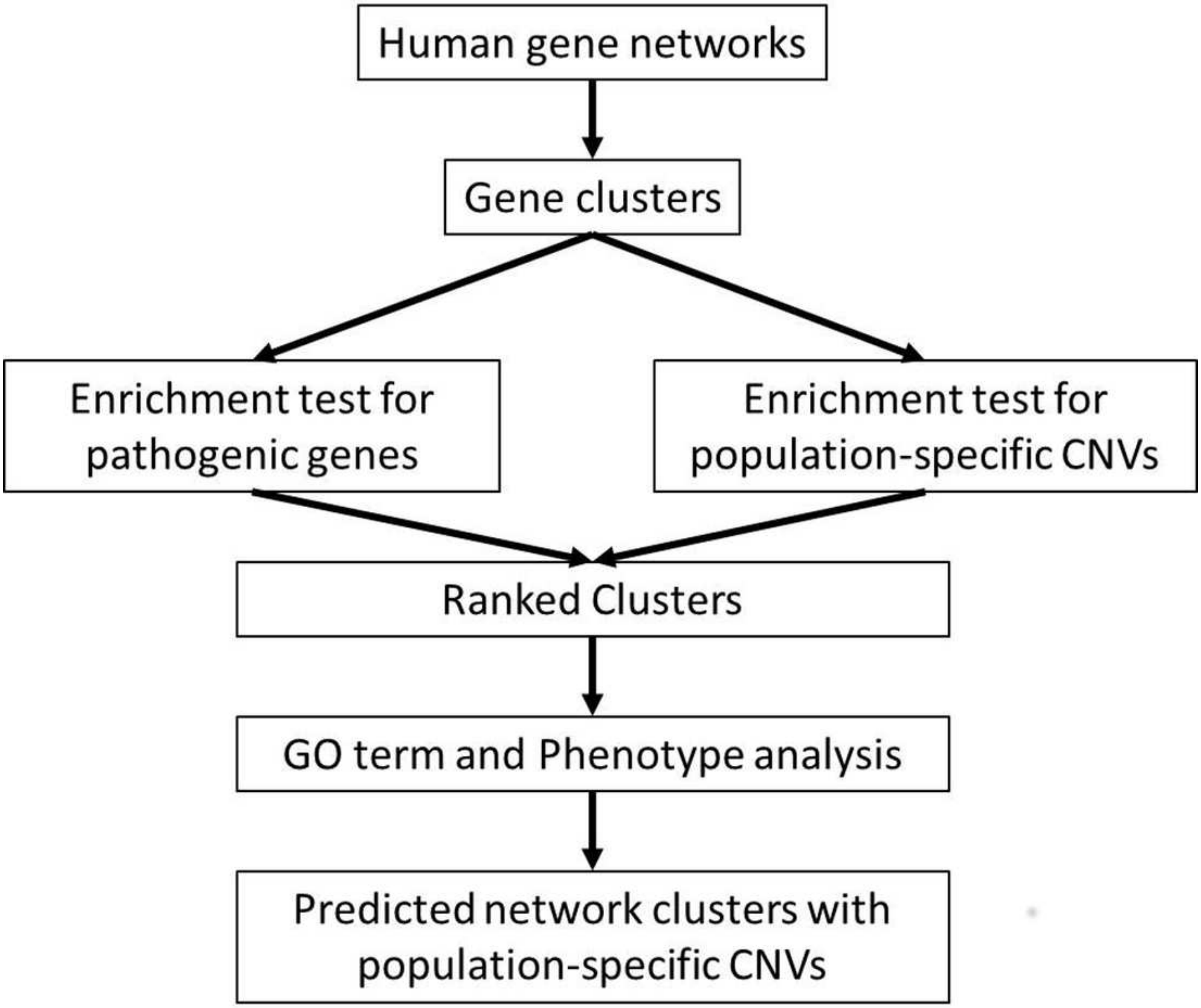
Wei P, Milbauer LC, Enenstein J, Nguyen J, Pan W, and Hebbel RP. 2011. Differential endothelial cell gene expression by African Americans versus Caucasian Americans: a possible contribution to health disparity in vascular disease and cancer. *BMC Med* 9:2.

Xu Z, Bensen JT, Smith GJ, Mohler JL, and Taylor JA. 2011. GWAS SNP Replication among African American and European American men in the North Carolina-Louisiana prostate cancer project (PCaP). *Prostate* 71:881-891.

Zhao M, and Zhao Z. 2013. CNVannotator: a comprehensive annotation server for copy number variation in the human genome. *PLoS One* 8:e80170.

Figure 1

Overview of our approach to identify CNVs associated with health disparities

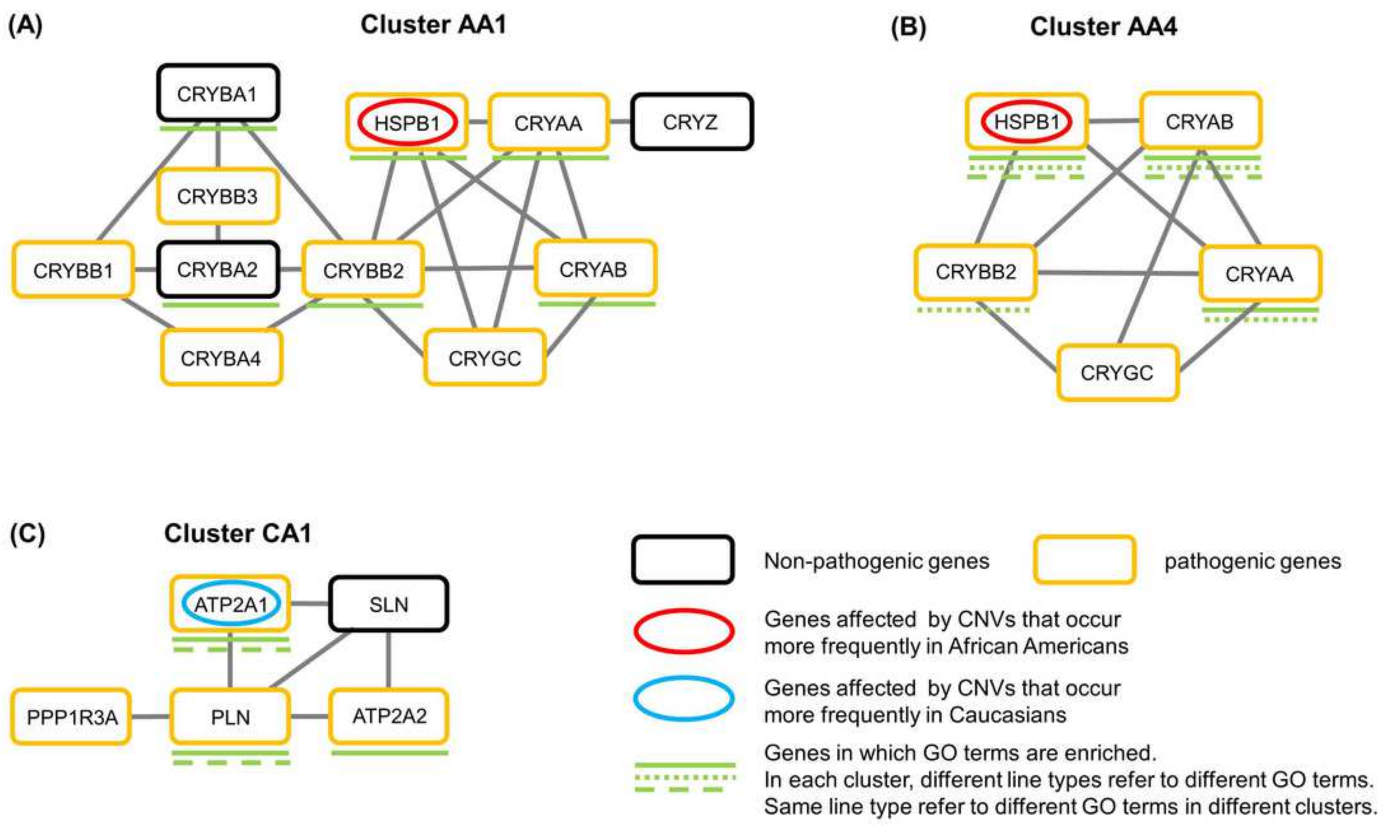

Human gene networks
Gene clusters
Enrichment test for pathogenic genes
Enrichment test for population-specific CNVs
Ranked Clusters
GO term and Phenotype analysis
Predicted network clusters with population-specific CNVs

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.584v1 | CC-BY 4.0 Open Access | rec: 3 Nov 2014, publ: 3 Nov 2014

Figure 2

Graph representations of selected clusters for biological significance analysis.

Each rounded rectangle represents a gene and each gray line represents a gene-gene interaction. Black rounded rectangles represent non-pathogenic genes and orange rounded rectangles represent pathogenic genes. Genes labeled with red or blue ovals are located in African American CNVs or in Caucasian CNVs. Genes with Green lines share the same GO terms. In each cluster, different line types represent the enrichment of different GO terms. Line types shown in different clusters refer to the enrichment of different GO terms.


(A)
Cluster AA1
CRYBA1
HSPB1
CRYAA
CRYZ
CRYBB3
CRYBB1
CRYBA2
CRYBB2
CRYAB
CRYBA4
CRYGC
(B)
Cluster AA4
HSPB1
CRYAB
CRYBB2
CRYAA
CRYGC
(C)
Cluster CA1
ATP2A1
SLN
PPP1R3A
PLN
ATP2A2
Non-pathogenic genes
pathogenic genes
Genes affected by CNVs that occur more frequently in African Americans
Genes affected by CNVs that occur more frequently in Caucasians
Genes in which GO terms are enriched.
In each cluster, different line types refer to different GO terms.
Same line type refer to different GO terms in different clusters.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.584v1 | CC-BY 4.0 Open Access | rec: 3 Nov 2014, publ: 3 Nov 2014

Table 1(on next page)

Contingency tables

Table 1A. Contingency Table for Fisher’s exact Test on Pathogenic Genes. Table 1B. Contingency Table for Fisher’s exact Test on CNV genes. For each cluster, contingency tables were constructed for right-tailed Fisher’s exact Tests. Table 1A is for pathogenic significance test, and Table 1B is for tests of enrichment significance of CNV genes (CNV_AA or CNV_CA genes). Q and q are the number of pathogenic genes in the whole networks and that in current cluster, respectively. N and m are the number of genes in whole networks and that in current cluster, respectively. S and s are the number of CNV_AA or CNV_CA genes in the whole networks and that in current cluster, respectively.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.584v1 | CC-BY 4.0 Open Access | rec: 3 Nov 2014, publ: 3 Nov 2014

Table 1A. Contingency Table for Fisher's exact Test on Pathogenic Genes

	Pathogenic Genes	**Non-pathogenic Genes**	**Total**
Genes in this cluster	q	m-q	m
Genes in other clusters	Q-q	N-Q-m+q	N-m
Total	Q	N-Q	N

Table 1B. Contingency Table for Fisher's exact Test on CNV genes

	CNV Genes	**Non-CNV Genes**	**Total**
Genes in this cluster	s	m-s	m
Genes in other clusters	S-s	N-S-m+s	N-m
Total	S	N-S	N

For each cluster, contingency tables were constructed for right-tailed Fisher's exact Tests. Table 1A is for pathogenic significance test, and Table 1B is for tests of enrichment significance of CNV genes (CNV_AA or CNV_CA genes). Q and q are the number of pathogenic genes in the whole networks and that in current cluster, respectively. N and m are the number of genes in whole networks and that in current cluster, respectively. S and s are the number of CNV_AA or CNV_CA genes in the whole networks and that in current cluster, respectively.

Table 2(on next page)

Cluster analysis results for HPRDNet and MultiNet

Table 2. Cluster analysis results for HPRDNet and MultiNet

Network	Cluster Name	CNV_AA	CNV_CA	Pathogenic gene number	Cluster Size
HPRDNet	AA1	*HSPB1*	-	8	11
	AA2	*HSPB1*	-	8	12
	AA3	*HSPB1*	-	8	13
	CA1	-	*ATP2A1*	4	5
MultiNet	AA4	*HSPB1*	-	5	5
	CA1	-	*ATP2A1*	4	5

Selected clusters were listed. CNV_AA and CNV_CA are CNV-related genes.

Table 3(on next page)

Detected genes with potential roles in health disparity and their located CNVs

Table 3. Detected genes with potential roles in health disparity and their located CNVs

Gene	Chr	Gene Coordinates	CNV Region	CNV Type	CNV Occurrence preference
HSPB1	7	75,931,861-75,933,614	75,867,431-76,481,102	Duplication	Only in African American
			75,929,740-76,481,102	Duplication	Only in African American
			75,929,740-76,568,388	Duplication	More in African American than in Caucasian
ATP2A1	16	28,889,726-28,915,830	28,306,730-28,936,772	Duplication	Only in Caucasian

Chr represents chromosomes. CNV Regions are regions of CNVs identified in more than a single individual; all CNVs listed have a type of Duplication, referring to one copy increase. CNV Regions and Types are from the CNV map (McElroy et al. 2009). CNV Occurrence preference describes in which population those CNVs have higher occurrence frequency.

Table 4(on next page)

Enriched GO terms with CNV-genes in the identified network clusters

Table 4. Enriched GO terms with CNV-genes in the identified network clusters

Clusters	Involved Genes	GO Domain	GO ID	GO term
AA1	***HSPB1, CRYAA, CRYAB, CRYBB2, CRYBA1, CRYBA2***	Molecular Function	GO:0042802	Identical protein binding
AA4	***HSPB1, CRYAA, CRYAB***	Biological Process	GO:0043086 GO:0043066 GO:0043069	negative regulation of catalytic activity negative regulation of apoptotic process negative regulation of programmed cell death
	HSPB1, CRYAA, CRYAB, CRYBB2	Molecular Function	GO:0042802	Identical protein binding
	HSPB1, CRYAB	Cellular Component	GO:0030018	Z disc
CA1	***ATP2A1, ATP2A2, PLN***	Biological Process	GO:0048878	chemical homeostasis
	ATP2A1, PLN	Biological Process	GO:0006937 GO:0008016	regulation of muscle contraction regulation of heart contraction

Biological relevance of network clusters was analyzed by GOrilla (Eden et al. 2009) to search for enriched gene ontology (GO) terms. Genes in the selected clusters were used as target genes, and all genes in the networks were treated as background genes. Three types of GO terms were analyzed: biological process, molecular function and cellular component. The default *p*-value threshold (1×10^{-3}) was used. In the results, enriched GO terms that are associated with CNV_AA gene HSPB1 and CNV_CA gene ATP2A1 were selected and listed in the table.

Table 5(on next page)

Associated diseases of genes with enriched GO terms.

Table 5. Associated diseases of genes with enriched GO terms.

Cluster	Gene	Associated Disease
AA1 and AA4	*HSPB1*	Axonal Charcot-Marie-Tooth disease type 2F
		Distal hereditary motor neuronopathy type 2B
	CRYAA	Multiple types of cataract 9
	CRYAB	Multiple types of cataract 16
		Dilated cardiomyopathy-1II
		Myofibrillar myopathy-2
		CRYAB-related fatal infantile hypertonic myofibrillar myopathy
	CRYBB2	Multiple types of Cataract 3
CA1	*ATP2A1*	Brody myopathy
	ATP2A2	Acrokeratosis verruciformis
		Darier disease
	PLN	Dilated cardiomyopathy-1P
		Familial hypertrophic cardiomyopathy-18

Only GO terms that contain CNV-genes are studied due to our focus on the role of CNV-genes in health disparity.