SCIENTIFIC REPORTS

OPEN

Received: 27 November 2018
Accepted: 5 March 2019
Published online: 10 April 2019

A Multi-Trait Approach Identified Genetic Variants Including a Rare Mutation in RGS3 with Impact on Abnormalities of Cardiac Structure/Function

Akram Yazdani[1,2], Azam Yazdani[3], Raúl Méndez Giráldez[4], David Aguilar[5] & Luca Sartore[6]

Heart failure is a major cause for premature death. Given the heterogeneity of the heart failure syndrome, identifying genetic determinants of cardiac function and structure may provide greater insights into heart failure. Despite progress in understanding the genetic basis of heart failure through genome wide association studies, the heritability of heart failure is not well understood. Gaining further insights into mechanisms that contribute to heart failure requires systematic approaches that go beyond single trait analysis. We integrated a Bayesian multi-trait approach and a Bayesian networks for the analysis of 10 correlated traits of cardiac structure and function measured across 3387 individuals with whole exome sequence data. While using single-trait based approaches did not find any significant genetic variant, applying the integrative Bayesian multi-trait approach, we identified 3 novel variants located in genes, *RGS3*, *CHD3*, and *MRPL38* with significant impact on the cardiac traits such as left ventricular volume index, parasternal long axis interventricular septum thickness, and mean left ventricular wall thickness. Among these, the rare variant NC_000009.11:g.116346115C > A (rs144636307) in *RGS3* showed pleiotropic effect on left ventricular mass index, left ventricular volume index and maximal left atrial anterior-posterior diameter while *RGS3* can inhibit TGF-beta signaling associated with left ventricle dilation and systolic dysfunction.

Heart failure (HF) is a complex clinical syndrome characterized by abnormal cardiac structure and function that leads to reduced cardiac output and elevated filling pressures at rest or with exertion[1]. Although, there is increasing evidence that the risk and course of HF depend on genetic predispositions[2], genome wide association studies (GWAS) have identified only a handful of genetic variants associated with it. For instance, the chromosome region 9p21 includes several highly replicated genetic variants associated with HF risk factors (e.g. NC_000009.11:g.22096055A > G and NC_000009.11:g.22124477A > G)[3].

For better understanding heritability of HF, some studies combine results of multiple cohorts and involve more samples in the analysis through meta-analysis[4,5]. One of the largest studies on African-American population identified four variants associated with left ventricular mass and left ventricular internal diastolic diameter respectively using Echocardiography[6]. A meta-analysis on 5 cohorts of individuals with European ancestry identified five genetic loci harboring common variants associated with left ventricular diastolic dimensions and aortic root size[7]. More recently, a meta-analysis of a large set of samples including 73,518 individuals identified 32 novel loci associated with electrocardiographic markers of hypertrophy as an important and independent risk factor for the development of heart failure[8].

Taking advantage of dozens-to-hundreds of traits measured on each study participant creates opportunities to obtain insights into the biology of HF, and consequently reduces morbidity, and economic burden of HF.

[1]Department of Genetics and Genomic Science, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [2]Climax Data Pattern, Boston, MA, USA. [3]School of Medicine, Boston University, Boston, MA, USA. [4]Lineberger Comprehensive Cancer Center, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [5]Baylor College of medicine, Houston, TX, USA. [6]National Institute of Statistical Science, Washington, DC, USA. Correspondence and requests for materials should be addressed to A.Y. (email: akramyazdani16@gmail.com)

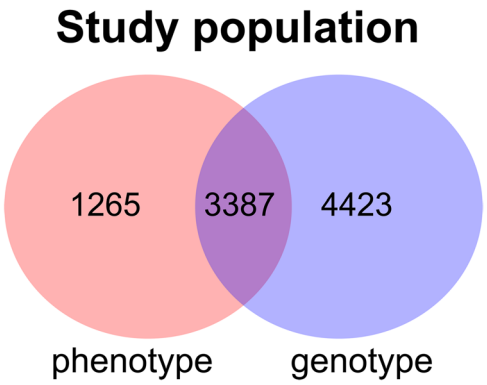

Figure 1. Venn diagram of study population with phenotype and genotype records.

Multi-trait analysis is toward this aim and increases the statistical power[9–12]. Although there are many studies on multi-trait approaches, applications of those methods have recently received increased attention e.g.[13–15]. A limitation of those methods is their complexity due to the large number of parameters in the model.

To reduce the complexity of the multi-trait approaches due to the number of parameters, we integrated a Bayesian network[16] with a Bayesian multi-trait polygenic mixed model while setting G-Wishart prior[17] on inverse of relatedness matrix and called it Integrative Bayesian Multi-Trait (IBMT) approach. Using IBMT approach, we conducted an analysis to identify genomic variants influencing 10 echocardiographic traits related to cardiac structure and function from Atherosclerosis Risk in Communities (ARIC) study[18]. We have genotype data of 7810 European American individuals from baseline measurement while 3387 of these individuals have phenotype records in Visit 5. The phenotype data are also recorded for 1265 new participants at visit 5 who do not have baseline genotype. Thus, we had three sets of data, including (i) individuals with only genotype data, (ii) individuals with only phenotype data, and (iii) individuals with both genotype and phenotype data. We incorporated information from all these three sets into the analysis to improve the statistical power, prevent overfitting, and avoid using data multiple times. The details are provided in the Methods section. These steps ultimately improve reliability and generalizability of the results.

After data preparation, we applied the IBMT method over the whole exome sequence data to investigate genomic and cardiac trait relationships. We identified 3 genetic variants (NC_000009.11:g.116346115C> A, NC_000017.10:g.7802658C > T, NC_000017.10: g.73897977C > T) in genes *RGS3*, *CHD3*, and *MRPL38* with significant impact on the cardiac traits. The variant in *RGS3* gene (NC_000009.11:g.116346115C> A) showed pleiotropic gene action on vertical mass index, left ventricular volume index and maximal left atrial anterior-posterior diameter). *RGS3* can inhibit TGF-beta signaling associated with left ventricle dilation and systolic dysfunction and codes for GTPase-activating protein that inhibits G-protein-mediated signal transduction[19,20]. *CHD3* encodes a protein with a chromatin organization modifier domain and a SNF2-related helicase/ATPase domain[21]. *MRPL38* produces a mitochondrial ribosomal protein, involved in the synthesis of proteins within the mitochondrion[22,23].

Materials and Methods

Study population. Echocardiographic and genomic data were collected on a subset of the ARIC study, a biracial longitudinal cohort of 15,792 middle-aged individuals who were randomly sampled from four US sites (Forsyth County, NC; Jackson, MS; suburbs of Minneapolis, MN; and Washington County, MD) and have been measured for risk factor traits related to health and chronic diseases. A detailed description of the ARIC study design and methods have been published elsewhere[18,24]. The data presented here includes 7810 European American individuals with baseline genotype available on dbGAP (https://www.ncbi.nlm.nih.gov/gap), accession number phs000090.v5.p1. A subset of individuals with genotype data, 3387 out of 7810 individuals, has phenotype records at visit 5. The phenotype data described in the following subsection are also recorded for 1265 new participants at visit 5 who do not have baseline genotype. Figure 1 visualized study population with phenotype and genotype records through Venn diagram.

Echocardiographic methods and measurements. Echocardiograms were obtained from participants at visit 5 using a standardized protocol as recommended by the American Society of Echocardiography. Images were digitally transferred to the Cardiovascular Imaging Core Laboratory at Brigham and Women's Hospital, Boston, MA, for offline analysis. The intra-observer variability (coefficient of variation and interclass correlation) for key echocardiographic measures has been previously published[25]. Images were obtained in the parasternal long- and short-axis and apical 2- and 4-chamber views. Primary measures of the traits such as left ventricular (LV) dimensions, volumes, and wall thickness; left atrial (LA) dimensions, volumes, and areas were made in triplicate from the 2-dimensional views in accordance with the recommendations of the American Society of Echocardiography[26]. This study includes 10 cardiac structure and function tabulated in Table 1. LV mass was calculated from LV linear dimension and indexed to body surface area. Relative wall thickness was calculated using the posterior wall thickness and LV end-diastolic dimension. LA volumes were measured by methods of disks using apical 4- and 3- chamber views. LV volumes were calculated from the apical 4- and 2- chamber views utilized the modified Simpson method.

Name	Measurement
Phenotypes of Interest	
Parasternal long axis interventricular septum thickness (PLAx-IST)	Cm
Parasternal long axis posterior wall thickness (PLAx-PWT)	Cm
End-diastolic volume (ED-V)	Ml
End-systolic volume (ES-V)	Ml
Ejection fraction (EF)	%
LV mass index (LV-MI)	G per m^2
LV relative wall thickness (LV-RWT)	Cm
Mean LV wall thickness (LV-WT)	Cm
Maximal left atrial anterior-posterior diameter (Max-LA-APD)	Ml per m^2
LA volume index (LA-VI)	

Table 1. Cardiac structural and functional traits under study.

Whole exome sequencing. Whole exome sequencing was performed on samples with the Illumina HiSeq platform. Mercury pipeline is applied for variant calling[27]. The reads are mapped into the Genome Reference Consortium Human Build 37 (GRCh37) sequence. Low-quality variants are filtered if they were outside the exon capture regions, belonged to multi-allelic sites, had missing rate >20% and had mean depth of coverage >500-fold. In addition, highly significant departures from race-specific Hardy-Weinberg equilibrium (P-value < 5e-6) are excluded from the data.

Statistical methods. The IBMT as an integrative Bayesian approach takes into account the underlying relationship among multiple traits and seeks for their significant association with genetic variants through multi-trait polygenic mixed model as the following

$$\boldsymbol{Y}_{\mathrm{nq}\times 1} = \boldsymbol{\mu}_{\mathrm{nq}\times 1} + (\boldsymbol{X}_{\mathrm{n}\times \mathrm{p}} \otimes \boldsymbol{I}_{\mathrm{q}})\boldsymbol{\beta}_{\mathrm{pq}\times 1} + \boldsymbol{U}_{\mathrm{nq}\times 1} + \epsilon_{\mathrm{qn}\times 1},\ \epsilon \sim N(\boldsymbol{0}, \Sigma) \tag{1}$$

Each entry of $\boldsymbol{Y} = \{y_{ij}\}_{\substack{i=1,\ldots,q \\ j=1,\ldots,n}}$ represents ith trait recorded for jth individual, β_{ik} is the effect of kth genomic variants on ith trait in coefficient vector $\boldsymbol{\beta} = \{\beta_{ik}\}_{\substack{i=1,\ldots,q \\ k=1,\ldots,p}}$, and vector $\boldsymbol{U} = \{u_{ij}\}_{\substack{i=1,\ldots,q \\ j=1,\ldots,n}}$ includes random effects corresponding to each sample where $\boldsymbol{U}\sim N(\boldsymbol{0}, \Psi)$ and Ψ is a relatedness matrix. The inverse of relatedness matrix (Ψ^{-1}) represents the conditional dependence of samples between and within traits.

In large scale settings, Ψ imposes numerous parameters into the model[1] and slows down the analysis. To overcome this limitation, we proposed to integrate the Bayesian network and the multi-trait polygenetic mixed model since Bayesian network represents the conditional dependence of the traits through graphical representation. Therefore, the IBMT first builds the Bayesian network among the traits of interest and then set the parameters in Ψ^{-1} to zero if their corresponding edges in Bayesian network are missing. To impose this sparsity to the model in posteriori, we set the G-Wishart prior on Ψ^{-1}. The G-Wishart places no probability mass on zero entries of Ψ^{-1} and accordingly it reduces the number of parameters of the model and optimizes the performance and computational time of Gibbs sampling scheme. Further details of the model and the Gibbs sampling scheme are provided in the Supplementary, statistical methods section, although in the following, we presented general guidline on the IBMT application.

Adjusting covariates. There is a broad consensus on analytic techniques for covariate adjustment to discover genomic variants associated with traits of interest independently of the correlated covariates, and to improve statistical power by gaining precision. The set of covariates typically include principal components of individual genotypes to account for population structure, and correlated environmental or demographic factors such as gender and age. However, some of those covariates may not have significant impact on the traits and adjusting for those covariates simply leads to a loss of power and cause variance inflation of the effects. Therefore, instead of adjusting each trait for all routine covariates, we first investigated the effect of covariates on the traits using 1265 individuals with only phenotypic record. This not only prevents decreasing the statistical precision or contaminating data due to adjusting for unrelated covariates but also avoids using data twice which prevents overfitting.

Applying the Integrative Bayesian Multi-Trait approach (IBMT). We first identified underlying relationships of the 10 cardiac functional and structural traits tabulated in Table 1 via application of Bayesian networks. We calculated hamming distance to measure structural similarities of networks based on different statistical significant levels (0.005, 0.01, 0.05) and found 0.05 as the best significant level to construct the network in our analysis[28–30]. The identified network, which is also supported by clinical background knowledge, revealed the sparsity level of the relationships. Figure 2 displays the network, where the nodes are the traits and the edges represent a significant relationship between the two corresponding traits after excluding the effect of the other traits. Clustering approaches[31–33] can be also applied for the same purpose, although they may estimate more connections among the traits. To avoid overfitting, we built the Bayesian network using the subset of individuals with only phenotype records.

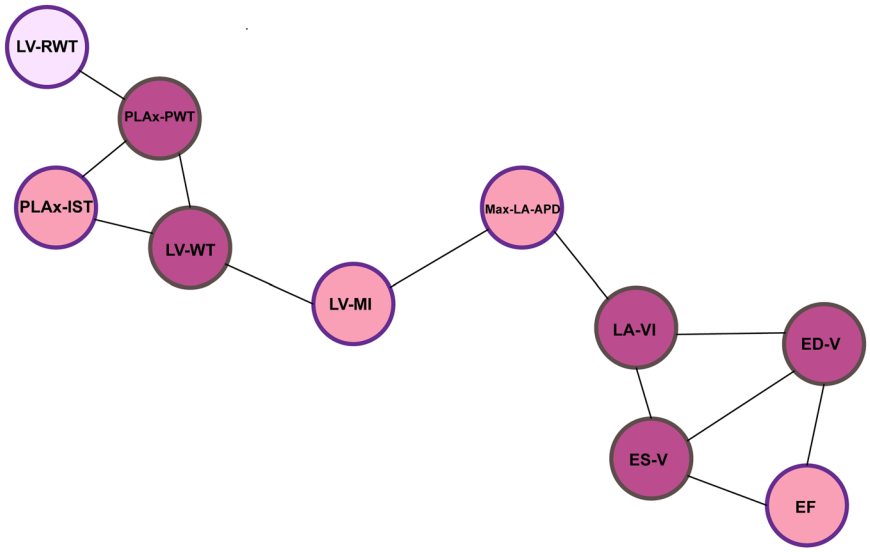

Figure 2. The Bayesian network over the cardiac traits. The colors correspond to the degree of connectivity of each trait; darker color means greater connectivity.

Using the identified Bayesian network over the traits, we implemented the IBMT method and analyzed the whole exome sequencing data including 260,688 variants based on sliding window. Each window included 100 variants with a step size of 25, such that each variant appears in 4 windows. If a variant is selected in all 4 windows, we reported the variant as the most promising variant. The parameters of the model were updated at each iteration of Markov Chain Monte Carlo (MCMC) algorithm and estimated with posterior means, which calculate optimal point estimators under square error loss, after 200 burn-in period.

To identify genomic variants significantly associated with the traits, we calculated 98% credible interval[34,35], (q^L, q^U), for the effects of all the variants to test a null hypothesis of no genomic effects. The endpoints of the intervals correspond to quantile (q) of the empirical distribution of the MCMC drawn from the marginal posterior distribution of genetic effects. A desired degree of precision for the endpoints of intervals is achieved by running a number of iterations until

$$P(|q_i^L - q_{i-1}^L|) < \zeta \ \& \ P(|q_i^U - q_{i-1}^U|) < \zeta,$$

where q^L and q^U are lower and upper quantile respectively and i represents the number of iterations. We set ζ to 0.01 as a small value.

Results

Since the IBMT method is based on a linear polygenic mixed model, we first tested the normality of the traits as the main assumption. Except the ejection fraction and the maximal left atrial anterior-posterior diameter that are normally distributed, the other traits were transformed to normal using log transformation. The histograms of the traits after winsorization, standardization, and log transformation are represented in Supplementary Figure 1.

We then investigated the effects of gender, age, ever-smoked, body mass index (BMI), hypertension, systolic and diastolic blood pressure on the traits. Among them gender and BMI showed highly significant relationships (P-value < 1e-8) with all traits except the mean LV Wall Thickness (LV-WT) and the ejection fraction, which they were relatively less significant with P-values 0.05 and 0.005 respectively. Hypertension also showed significant effects (P-value < 1e-6) with all traits except the ejection fraction. We obtained these results on the set of individuals without genotype data to avoid the use of data twice. To generalize the results to the set of interest (individuals with both genotype and phenotype records), we compared BMI distribution, gender ratio (Female/Male), and ratio of (with/without) hypertension in the two sets. We observed that in both sets, the distribution of BMI is similar (Figure 3); gender ratios 1.302 and 1.387 showed almost the same proportion of female to male; and hypertension ratios 2.502 and 2.579 also showed almost the same proportion of individuals with and without hypertension. Therefore, we adjusted the traits for BMI, gender, and hypertension, in addition to the first 10 PCs from population stratification analysis.

Applying IBMT, we first identified Bayesian network among 10 traits listed in Table 1. As shown in Figure 2, underlying relationships among the traits are sparse. Therefore, we do not need to consider all pairwise connectivity in the analysis. Incorporating this result into multi-trait mixed model, we reduced the number of parameters in the model and consequently increased the power of genotype-phenotype association identification (Supplementary Statistical methods). We could identify 3 genetic variants with significant impact on 4 cardiac traits using a 98% credible interval (Table 2 and Figure 4). Minor allele frequencies (MAF) in Table 2 show the identified variants are rare. The estimated effects of the identified rare variants are reported in Table 3.

Figure 5 shows empirical distributions of LV-MI (red/blue) for individuals with reference/alternative allele of one of the identified variants (NC_000017.10:g.73897977C > T). The noticeable shift of the distribution for

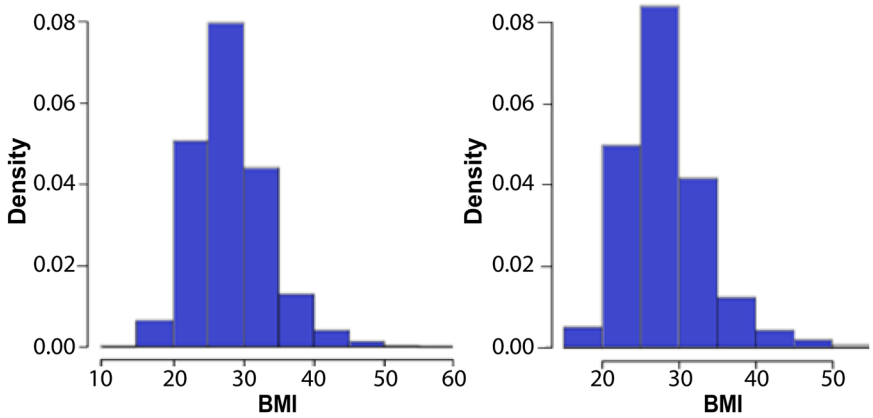

Figure 3. Histogram of BMI. Right: individuals without genotype data. Left: individuals with genotype data.

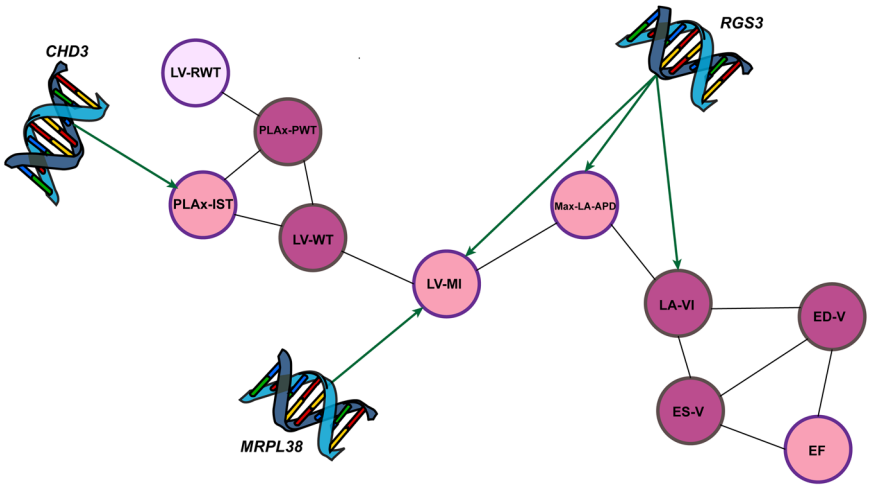

Figure 4. Identified genetic pathway to cardiac and structure and function using IBMT.

HGVS name	refSNP ID	CHR	MAF%	Gene name	Related trait
Identified variants					
NC_000009.11:g.116346115C > A	rs144636307	9	0.38	*RGS3*	LV-MI
					LA-VI
					Max-LA-APD
NC_000017.10:g.7802658C > T	rs200287864	17	0.25	*CHD3*	PLAx-IST
NC_000017.10:g.73897977C > T	rs76054219	17	0.32	*MRPL38*	LV-MI

Table 2. Selected genomic variants related to the traits, using a 98% Bayesian credible interval. HGVS name is description of sequence variation in genomic established by The Human Genome Variation Society; refSNP ID is a unique identifier provided by NCBI; CHR is the chromosome number; and MAF stands for minor allele frequency.

HGVS name	Trait	Est-Eff	SD-Eff
NC_000009.11:g.116346115C > A	LV-MI	1.13	0.104
NC_000009.11:g.116346115C > A	LA-VI	1.27	0.11
NC_000009.11:g.116346115C > A	Max-LA-APD	1.29	0.102
NC_000017.10:g.7802658C > T	PLAx-IST	1.48	0.186
NC_000017.10:g.73897977C > T	LV-MI	1.28	0.153

Table 3. Estimated effect (Est-Eff) and standard deviation (SD-Eff) of the identified genes with significant effect on the traits.

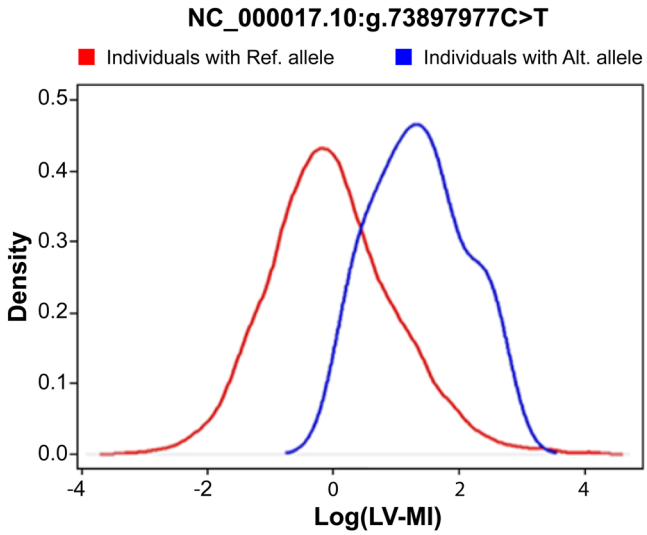

Figure 5. Empirical Distributions of log (LV-MI) for individuals with reference/alternate alleles in NC_000017.10:g.73897977C > T variant.

individuals with mutation is observable. Supplementary Figure 2 shows levels of the traits for individuals with identified rare mutation over distribution of the traits.

One of the identified variants located in chromosome 9 (NC_000009.11:g.116346115C > A) showed pleotropic effect on LV-MI, LA-VI and Max-LA-APD. Among 14 individuals (5 Females and 9 Males) having rare mutation in *RGS3* (NC_000009.11:g.116346115C > A) with pleiotropic action, two of the females and one male showed different patterns with other individuals, such that their level of LV-MI, LA-VI and Max-LA-APD are not greater than third quartiles (Supplementary Figure 2). This suggests that the rare mutation identified in *RGS3* may have higher impact on males. In addition, pleiotropic effect of this variant may represent different functions of the *RGS3* gene.

According to the Variant Effect Predictor (VEP) analysis[36], this variant is most likely a missense mutation in *RGS3* (regulator of G protein signaling 3) gene that could yield a codon change from ACC to AAC replacing the amino acid Threonine (Thr) by an Asparagine (Asn) in the protein sequence. *RGS* genes encode GTPase activating proteins (GAPs) and down-regulate G protein signaling. More details about the results from the VEP analysis is provided in Supplementary Table 1.

Another identified variant with impact on PLAx-IST is in chromosome 17 (NC_000017.10:g.7802658C > T). Individuals with T allele of this rare variant, including 2 Females and 8 Males, all have PLAx-IST level greater than third quartile of the trait (Supplementary Figure 2).

The variant NC_000017.10:g.7802658C > T is intronic to Chromatin Helicase DNA Binding Protein 3 (*CHD3*) gene based on VEP analysis (Supplementary Table 2). *CHD3* belongs to a family of genes coding for proteins that bear CHROMO (chromatin organization modifier) and SNF2-related helicase/ATPase domains. CHD3 protein is one of the components of the Mi-2/NuRD (histone deacetylase) complex that participates in the remodeling of chromatin structure via histone deacetylation.

The third identified genetic variant influencing LV-MI is located in chromosome 17, gene *MRPL38* (NC_000017.10:g.73897977C > T). All 8 male individuals with rare mutation in *MRPL38* have LV-MI level greater than third quartiles of distribution (Supplementary Figure 2). However, the 2 female individuals do not show any different pattern.

The variant NC_000017.10:g.73897977C > T can be in a non-coding exon or be a missense mutation of *MRPL38* gene depending on splicing. The missense mutation of CGG into CAG codon causes the substitution of amino acid Arginine (Arg) by a Glutamine (Gln) concluded. Gln is a non-charged amino acid and smaller than Arg with putatively less capacity to create hydrogen bonds and favorable electrostatic interactions than Arg (see Supplementary Table 3 for VEP results). *MRPL38* encodes mitochondrial ribosomal proteins (MRP). The MRP family stabilizes mitochondrial ribosome (mitoribosome) and are responsible for the mitochondrial translation of 13 protein components of the Oxidative Phosphorylation (OXPHOS) gene complex in the mitochondrial DNA[23].

Discussion

Single trait analysis did not identify any genetic variants with significant impact on the 10 considered traits of cardiac structure and function in European American individuals from ARIC study. Therefore, we integrated Bayesian network and Bayesian multi-trait approach to improve the performance of the analysis. This Integrative Bayesian Multi-Trait (IBMT) approach provides a sparse relatedness matrix and, eventually, more precise estimates of parameters in the model. Utilizing the IBMT method to increase the power of identification, in addition to carefully adjusting for covariates to avoid data contamination, and choosing the appropriate transformation function for each trait, we identified three significant genetic variants. These variants located in *RGS3*, *CHD3*, and *MRLP38* genes are rare, hence a high statistical power is required to detect the association with the trait(s)[37–40].

Among those, the variant NC_000009.11:g.116346115C > A in *RGS3* showed pleiotropic action on vertical mass index, left vertical volume index, maximal left atrial anterior-posterior diameter (Figure 4).

The variant NC_000009.11:g.116346115C > A is in the exon of gene *RGS3* (Regulator of G protein signaling 3) which belongs to *RGS* family. *RGS* family codes for proteins that act as GAPs and down-regulate G protein signaling. Many studies have proven that *RGS* gene expression is highly regulated in myocardium[41–43]. Quantitative messenger RNA (mRNA) analysis revealed that *RGS3* is most highly expressed in human heart[44–46]. The N-terminus of *RGS3* can inhibit TGFβ induced differentiation of pulmonary fibroblasts which is associated with left ventricular dilation and systolic dysfunction[19,20,47]. The identified variant in *RGS3* was associated with pleotropic effect on LV mass (LV-MI) and measures of left atrial size (LA-VI and Max-LA-APD). The left atrial size may reflect the cumulative effects of increased LV filling pressure and diastolic function[48] and is a predictor of heart failure, ischemic stroke, and death. Thus, genetic variants contributing to abnormalities of LV mass and worsened diastolic function would be expected to potentially be associated with LA size.

As missense mutation, NC_000009.11:g.116346115C > A could yield a codon change from ACC to AAC, which replaces a Threonine (Thr) to an Asparagine (Asn) in the protein amino acid sequence. The change from Thr to Asn does not alter side chain electrostatic charge because of both amino acids being electrostatically neutral. This could eventually affect hydrogen bond pattern since Asn has an extra hydrogen bond donor group (NH2). However, it is difficult to evaluate the final effect in the protein three-dimensional structure since there may be alternative spliced transcripts. If this mutation happened at the interaction interface between *RGS3* and Gα subunit, it could eventually affect GTPase activity.

The variant NC_000017.10:g.7802658C > T that is significantly associated with parasternal long axis interventricular septum thickness (PLAx-IST) is in an intron of the *CHD3* gene, which codes for the Chromatin Helicase DNA Binding Protein 3 as a part of the chromatin structure remodeling complex. This is a potential splicing variant that could affect the rate of mature mRNA synthesis, and ultimately impact gene transcription.

The other identified variant NC_000017.10:g.73897977C > T is within *MRPL38* gene, a member of Mitochondrial ribosomal proteins (MRP) family that are part of the large subunit of the mitochondrial ribosome. As missense mutation, the change of CGG into CAG codon causes the substitution of amino acid Arginine (Arg) by a Glutamine (Gln). Gln is a non-charged amino acid and smaller than Arg with putatively less capacity to create hydrogen bonds and favorable electrostatic interactions than Arg. If the amino acid mutation takes place at the interface between *MRPL38* and mitochondrial ribosome, it could decrease binding affinity and destabilize mitochondrial ribosome. This in turn may reduce ribosomal protein synthesis levels and affect the oxidative phosphorylation pathway.

Overall, this study suggests that rare mutation might provide a better understanding of genetic impact on cardiovascular structure and resulting in remodeling cardiovascular disease and/or heart failure, although further studies are required.

Data Availability

The genotype data is available on dbGAP (https://www.ncbi.nlm.nih.gov/gap) with accession number phs000090.v5.p1. The phenotype data is available upon request through https://www2.cscc.unc.edu/aric/.

References

1. McMurray, J. J. & Pfeffer, M. A. Heart failure. *Lancet.* **365**(9474), 1877–89 (2005).
2. MacRae, C. A. The Genetics of Congestive Heart Failure. *Heart Failure. Clinics* **6**(2), 223–30 (2010).
3. Yamagishi, K., Folsom, A. R., Rosamond, W. D. & Boerwinkle, E. A genetic variant on chromosome 9p21 and incident heart failure in the ARIC study. *Eur Heart J* **30**(10), 1222–8 (2009).
4. Harismendy, O. *et al.* 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. *Nature.* **470**, 264 (2011).
5. Méndez-Giráldez, R. *et al.* GWAS of the electrocardiographic QT interval in Hispanics/Latinos generalizes previously identified loci and identifies population-specific signals. *Sci Rep.* **7**(1), 17075 (2017).
6. Lieb, W. *et al.* Genome-wide meta-analyses of plasma renin activity and concentration reveal association with the kininogen 1 and prekallikrein genes. *Circ Cardiovasc Genet* **8**(1), 131–40 (2015).
7. Vasan, R. S. *et al.* Genetic variants associated with cardiac structure and function. *JAMA J Am Med Assoc* **302**(2), 168 (2009).
8. van der Harst, P. *et al.* 52 Genetic Loci Influencing Myocardial Mass. *J Am Coll Cardiol* **68**(13), 1435–48 (2016).
9. Korte, A. *et al.* A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat Genet.* **44**(9), 1066–71 (2012).
10. Galesloot, T. E., Van Steen, K., Kiemeney, L. A., Janss, L. L. & Vermeulen, S. H. A comparison of multivariate genome-wide association methods. *PLoS One.* **9**(4) (2014).
11. Kwak, I. Y. & Pan, W. Gene- and pathway-based association tests for multiple traits with GWAS summary statistics. *Bioinformatics.* **33**(1), 64–71 (2017).
12. Yazdani, A., Yazdani, A., Samiei, A. & Boerwinkle, E. Generating a robust statistical causal structure over 13 cardiovascular disease risk factors using genomics data. *J Biomed Inform* **60**, 114–9 (2016).
13. Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M. & Wray, N. R. Estimation of pleiotropy between complex diseases using singlenucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics.* **28**(19), 2540–2 (2012).
14. Schaid, D. J. *et al.* Statistical Methods for Testing Genetic Pleiotropy. *Genetics.* **204**(2), 483–97 (2016).
15. Yazdani, A., Yazdani, A. & Boerwinkle, E. A Causal Network Analysis of the Fatty Acid Metabolome in African-Americans Reveals a Critical Role for Palmitoleate and Margarate. Omi A. *J Integr Biol* **20**(8), 480–4 (2016).
16. Pearl, J. Probabalistic Reasoning in Intelligent Systems. *Probabalistic Reason Intell Syst*. 552 (1988).
17. Roverato, A. Hyper Inverse Wishart Distribution for Non-decomposable Graphs and its Application to Bayesian Inference for Gaussian Graphical Models. *Scand J Stat* **29**(1993), 391–411 (2002).
18. The ARIC Investigators. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. *Am J Epidemiol* **129**(4), 687–702 (1989).

19. Lucas, J. A. *et al*. Inhibition of transforming growth factor- signaling induces left ventricular dilation and dysfunction in the pressure-overloaded heart. *AJP Hear Circ Physiol* **298**(2), 424–32; Available from: http://ajpheart.physiology.org/cgi/doi/10.1152/ajpheart.00529.2009 (2010).
20. Dobaczewski, M., Chen, W. & Frangogiannis, N. G. Transforming growth factor (TGF)-β signaling in cardiac remodeling. *Journal of Molecular and Cellular Cardiology* **51**, 600–6 (2011).
21. [provided by RefSeq, Jul 2008].
22. Owen, V. J. *et al*. Expression of RGS3, RGS4 and Gi alpha 2 in acutely failing donor hearts and end-stage heart failure. *Eur Heart J* **22**(12), 1015–20 (2001).
23. Galmiche, L. *et al*. Exome sequencing identifies MRPL3 mutation in mitochondrial cardiomyopathy. *Hum Mutat.* **32**(11), 1225–31 (2011).
24. Bello, N. A. *et al*. Association of weight and body composition on cardiac structure and function in the ARIC study (Atherosclerosis Risk in Communities). *Circ Hear Fail.* **9**(8) (2016).
25. Shah, A. M. *et al*. Rationale and design of a multicenter echocardiographic study to assess the relationship between cardiac structure and function and heart failure risk in a biracial cohort of community-dwelling elderly persons: The atherosclerosis risk in communities stud. *Circ Cardiovasc Imaging* **7**(1), 173–81 (2017).
26. Lang, R. M. *et al*. Recommendations for chamber quantification: A report from the American Society of Echocardiography's guidelines and standards committee and the Chamber Quantification Writing Group, developed in conjunction with the European Association of Echocardiograph. *Journal of the American Society of Echocardiography* **18**, 1440–63 (2005).
27. Reid, J. G. *et al*. Launching genomics into the cloud: Deployment of Mercury, a next generation sequence analysis pipeline. *BMC Bioinformatics*. **15**(1) (2014).
28. Tsamardinos, I., Brown, L. E. & Aliferis, C. F. The max-min hill-climbing Bayesian network structure learning algorithm. *Mach Learn.* **65**(1), 31–78 (2006).
29. Norouzi, M., Fleet, D. J., Salakhutdinov, R. & Blei, D. M. Hamming distance metric learning. *Adv Neural Inf Process Syst*. 1–9 (2012).
30. Schuster, P., Fontana, W., Stadler, P. F. & Hofacker, I. L. From Sequences to Shapes and Back: A Case Study in RNA Secondary Structures. *Proc R Soc B Biol Sci* **255**(1344), 279–84 (1994).
31. Yazdani, H., Ortiz-Arroyo, D., Choroś, K. & Kwasnicka, H. On High Dimensional Searching Spaces and Learning Methods. *Data Sci big data An Environ Comput Intell*. 29–48 (2017).
32. Broumand, A., Esfahani, M. S., Yoon, B. J. & Dougherty, E. R. Discrete optimal Bayesian classification with error-conditioned sequential sampling. *Pattern Recognit.* **48**(11), 3766–82 (2015).
33. Knight, J. M., Ivanov, I. & Dougherty, E. R. MCMC implementation of the optimal Bayesian classifier for non-Gaussian models: model-based RNA-Seq classification. *BMC bioinformatics.* **15**(1), 401 (2014).
34. Tseng, G. C. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res* **29**(12), 2549–57 (2001).
35. Chiara, S. Empirical Bayes Estimation of a Sparse Vector of Gene Expression Changes. *Statistical Applications in Genetics and Molecular Biology*. **4**(1) (2005).
36. McLaren, W. *et al*. The Ensembl Variant Effect Predictor. *Genome Biol*. **17**(1) (2016).
37. Yousri, N. A. *et al*. Whole-exome sequencing identifies common and rare variant metabolic QTLs in a Middle Eastern population. *Nat Commun*. **9**(1) (2018).
38. Yazdani, A., Yazdani, A., Liu, X. & Boerwinkle, E. Identification of Rare Variants in Metabolites of the Carnitine Pathway by Whole Genome Sequencing Analysis. *Genet Epidemiol.* **40**(6), 486–91 (2016).
39. Yazdani, A., Yazdani, A. & Boerwinkle, E. Rare variants analysis using penalization methods for whole genome sequence data. *BMC Bioinformatics.* **16**(1), 405 (2015).
40. Graham, E. *et al*. Integration of genomics and metabolomics for prioritization of rare disease variants: a 2018 literature review. *Journal of Inherited Metabolic Disease* **14**, 435–45 (2018).
41. Scheschonka, A. *et al*. RGS3 is a GTPase-activating protein for g(ialpha) and g(qalpha) and a potent inhibitor of signaling by GTPase-deficient forms of g(qalpha) and g(11alpha). *Mol Pharmacol.* **58**(4), 719–28 (2000).
42. Liu, Y. *et al*. Regulator of G protein signaling 3 protects against cardiac hypertrophy in mice. *J Cell Biochem* **115**(5), 977–86 (2014).
43. Zhang, S. *et al*. RGS3 and RGS4 are GTPase activating proteins in the heart. *J Mol Cell Cardiol* **30**(2), 269–76 (1998).
44. Zhang, P. & Mende, U. Regulators of g-protein signaling in the heart and their potential as therapeutic targets. *Circulation Research.* **109**, 320–33 (2001).
45. Larminie, C. *et al*. Selective expression of regulators of G-protein signaling (RGS) in the human central nervous system. *Mol Brain Res* **122**(1), 24–34 (2004).
46. Wieland, T. & Mittmann, C. Regulators of G-protein signalling: Multifunctional proteins with impact on signalling in the cardiovascular system. *Pharmacology and Therapeutics* **97**(2), 95–115 (2003).
47. Talasaz, A. H. *et al*. N-Acetylcysteine Effects on Transforming Growth Factor-β and Tumor Necrosis Factor-α Serum Levels as Pro-Fibrotic and Inflammatory Biomarkers in Patients Following ST-Segment Elevation Myocardial Infarction. *Drugs R D* **13**(3), 199–205 (2013).
48. Douglas, P. S. The left atrium: A biomarker of chronic diastolic dysfunction and cardiovascular disease risk. *Journal of the American College of Cardiology* **42**, 1206–7 (2003).

Acknowledgements

Thanks go to Dr. Boerwinkle for providing the data. Thanks also go to the staff and participants of the Atherosclerosis Risk in Communities (ARIC) Study for gathering the data. The ARIC Study is a collaborative study supported by the National Heart, Lung, and Blood Institute, National Institutes of Health, Contracts HHSN268201100005C, HHSN268201100006C and HHSN26-82011-00008C.

Author Contributions

Akram Yazdani defined the project, implemented the method, carried out the analysis and wrote the manuscript. Azam Yazdani collaborated on defining the approach and writing the manuscript. Raúl Méndez Giráldez collaborated on interpretating the results and writing the manuscript. David Aguilar contributed on interpretating the results and commented on the mauscript. Luca Sartore collaborated on implementating the algorithm.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-41362-3.

Competing Interests: The authors declare no competing interests.