SCIENTIFIC REPORTS
natureresearch

OPEN

The polygenic architecture of left ventricular mass mirrors the clinical epidemiology

Jonathan D. Mosley[1,2] ✉, Rebecca T. Levinson[1], Eric Farber-Eger[3], Todd L. Edwards[1,4], Jacklyn N. Hellwege[1,4,5], Adriana M. Hung[1,5], Ayush Giri[1,3,6,7], Megan M. Shuey[1], Christian M. Shaffer[1], Mingjian Shi[1], Evan L. Brittain[1], Wendy K. Chung[8,9], Iftikhar J. Kullo[10], Adelaide M. Arruda-Olson[10], Gail P. Jarvik[11], Eric B. Larson[12], David R. Crosslin[13], Marc S. Williams[14], Ken M. Borthwick[15], Hakon Hakonarson[16], Joshua C. Denny[1,7], Thomas J. Wang[1], Charles M. Stein[1,17], Dan M. Roden[1,2,17] & Quinn S. Wells[1,17]

Left ventricular (LV) mass is a prognostic biomarker for incident heart disease and all-cause mortality. Large-scale genome-wide association studies have identified few SNPs associated with LV mass. We hypothesized that a polygenic discovery approach using LV mass measurements made in a clinical population would identify risk factors and diseases associated with adverse LV remodeling. We developed a polygenic single nucleotide polymorphism-based predictor of LV mass in 7,601 individuals with LV mass measurements made during routine clinical care. We tested for associations between this predictor and 894 clinical diagnoses measured in 58,838 unrelated genotyped individuals. There were 29 clinical phenotypes associated with the LV mass genetic predictor at FDR $q < 0.05$. Genetically predicted higher LV mass was associated with modifiable cardiac risk factors, diagnoses related to organ dysfunction and conditions associated with abnormal cardiac structure including heart failure and atrial fibrillation. Secondary analyses using polygenic predictors confirmed a significant association between higher LV mass and body mass index and, in men, associations with coronary atherosclerosis and systolic blood pressure. In summary, these analyses show that LV mass-associated genetic variability associates with diagnoses of cardiac diseases and with modifiable risk factors which contribute to these diseases.

Left ventricular (LV) mass captures the cardiac response to the cumulative exposure of a diverse set of risk factors, and is strongly predictive of future cardiac events and all-cause mortality[1–3]. Determining the genetic factors associated with LV mass could identify important predisposing mechanisms that contribute to heart disease. Three genome-wide association studies (GWAS) have only identified a single nucleotide polymorphism (SNP) associated with LV mass[4–6]. Thus, the genetic architecture underlying LV mass remains uncharacterized.

[1]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. [2]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA. [3]Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University Medical Center, Nashville, TN, USA. [4]Vanderbilt Epidemiology Center, Vanderbilt University Medical Center, Nashville, TN, USA. [5]Tennessee Valley Healthcare System (626), Vanderbilt University, Nashville, TN, USA. [6]Department of Obstetrics and Gynecology, Vanderbilt University Medical Center, Nashville, TN, USA. [7]Institute for Medicine and Public Health, Vanderbilt University Medical Center, Nashville, TN, USA. [8]Office of Research & Development, Department of Veterans Affairs, Washington DC, DC, USA. [9]Departments of Pediatrics and Medicine, Columbia University Medical Center, New York, NY, USA. [10]Department of Cardiovascular Diseases, Mayo Clinic, Rochester, MN, USA. [11]Departments of Medicine (Medical Genetics) and Genome Sciences, University of Washington, Seattle, WA, USA. [12]Kaiser Permanente Washington Health Research Institute and Department of Medicine, University of Washington, Seattle, WA, USA. [13]Departments of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, USA. [14]Genomic Medicine Institute, Geisinger, Danville, PA, USA. [15]Biomedical and Translational Informatics, Geisinger, Danville, PA, USA. [16]Center for Applied Genomics, Division of Human Genetics, Department of Pediatrics, The Children's Hospital of Philadelphia, University of Pennsylvania, Philadelphia, PA, USA. [17]Department of Pharmacology, Vanderbilt University Medical Center, Nashville, TN, USA. ✉e-mail: jonathan.d.mosley@vumc.org

Data source	Independent variable	Additional covariate[a]	OR (95% CI)[b]	P-value
Derivation set[c]	LV Mass measured	None	2.1 (2.0–2.3)	$<2 \times 10^{-16}$
	LV Mass PRS	None	2.0 (1.9–2.1)	$<2 \times 10^{-16}$
	LV Mass PRS	LV Mass measured	1.1 (0.99–1.31)	0.06
	LV Mass PRS	Height	2.0 (1.9–2.2)	$<2 \times 10^{-16}$
	LV Mass PRS	BMI	1.9 (1.8–2.1)	$<2 \times 10^{-16}$
	LV Mass permuted PRS	None	0.99 (0.94–1.04)	0.66
Validation set 1: BioVU[d]	LV Mass PRS	None	1.08 (1.02–1.13)	0.002
Validation set 2: eMERGE[e]	LV Mass PRS	None	1.10 (1.06–1.14)	4.1×10^{-7}

Table 1. Validation of the LV mass polygenic predictor (PRS). [a]Additional covariate added to the logistic regression model. [b]Analyses are based on a logistic regression model using the PheWAS phenotype cardiomegaly as the dependent variable and specified independent variable. All independent variables were set to have a standard deviation of 1. All models were additionally adjusted for age, sex and 5 principal components. [c]The data set used to develop the PRS. There were 2929 cases and 3416 controls for cardiomegaly. [d]Cases (n = 2,624) and controls (n = 16,682) for the cardiomegaly phenotype that were not part of the derivation set. [e]Cases (n = 5,982) and controls (n = 21,198) for the cardiomegaly phenotype that were not part of the derivation set.

An individual's LV mass can change over time (referred to as LV remodeling). For instance, LV mass may increase due to prolonged exposure to risk factors such as high blood pressure, metabolic abnormalities and cardiac diseases[7]. In contrast, treating these risk factors using medications or other interventions attenuates or reverses these LV mass increases over time. These changes suggest that important processes contributing to adverse LV remodeling have extended latencies and are modulated by environmental influences.

We hypothesized that defining the polygenic architecture of LV mass in a clinical population comprised of healthy and diseased individuals, would identify genetic variation associated with variability in LV mass within that population. To test this hypothesis, we developed a SNP-based polygenic predictor of LV mass in a genotyped population of individuals who received transthoracic echocardiography (TTE) as part of routine clinical care at the Vanderbilt University Medical Center (VUMC). To identify the genetic diseases associated with this predictor, we interrogated it against a large collection of diseases ascertained through the Electronic Medical Records and Genomics (eMERGE) network, a consortium of medical centers with EHR-linked DNA biobanks[8,9]. We show that genetic diagnoses associated with LV mass include modifiable risk factors that suggest targets for directed treatment and prevention efforts within the clinical population.

Results

TTE population.

The TTE population was 53% male and had a mean age of 64 (standard deviation[s.d.], 12) years (Supplementary Table 1). The most prevalent diagnoses were hypertension (81%), respiratory symptoms (75%), arrhythmias (68%) and lipid disorders (67%). The mean LV mass was 212 (s.d. 72) grams in men and 37.0% had an LV mass greater than the upper limit of normal in men (224 grams). Among women, mean LV mass was 155 (s.d. 54) grams and 37.8% had a mass over 162 grams.

GWAS analysis.

We performed a GWAS to determine whether there were common SNP variants associated with LV mass. There were no SNPs associated with LV mass at the genome-wide significant threshold of $p < 5 \times 10^{-8}$ (Supplementary Figure 1). A prior GWAS reported an association between LV mass and the SNP rs2255167-T, located within the *TTN* gene[6]. There was a similar direction of effect for this SNP in these analyses, but the association was not significant ($\beta = 0.012$, standard error = 0.007, $p = 0.06$).

PheWAS analysis.

An alternative unbiased discovery strategy to identify genetic associations is to construct a polygenic predictor comprising common SNPs associated with the phenotype. We validated this predictor against the PheWAS diagnosis of cardiomegaly, which is a clinical diagnosis of an enlarged heart, and corresponds to an elevated LV mass. Within the training data set used to build the polygenic predictor, both measured LV mass and the genetic predictor were strongly positive associated ($p < 2 \times 10^{-16}$) with the risk of a cardiomegaly diagnosis (Table 1). While adjusting for measured LV mass eliminated the association between cardiomegaly and the predictor ($p = 0.06$), adjusting for either body mass index or height only minimally attenuated the association ($p < 2 \times 10^{-16}$). Thus, the phenotypic variation captured by the genetic predictor corresponds to LV mass. A genetic predictor derived from permuted LV mass measurements was not associated with the phenotype, indicating that a genetic predictor from a random phenotype does not associate with the cardiomegaly diagnosis. In two independent validation sets that did not include individuals used to build the genetic predictor, the LV mass genetic predictor was significantly positively associated with cardiomegaly (Table 1). Thus, the genetic predictor demonstrated the expected associations with the cardiomegaly positive control phenotype.

We used this predictor to perform a phenome scan to identify clinical diagnosis associated with genetic variability that also associates with LV mass (Fig. 1 and Supplementary Table 2). The estimated proportion of the LV mass variance accounted for by the SNPs used to construct the predictor was 12.4%. There were 29 diagnoses associated with the genetic predictor (FDR $q < 0.05$) (Fig. 2 and Supplementary Table 3). All significant

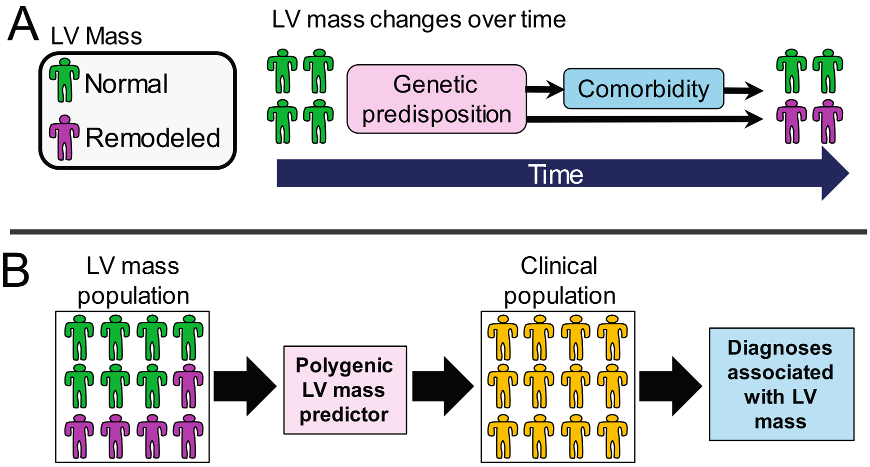

Figure 1. Overview of rationale and approach. (**A**) Over time, some individuals undergo changes in LV mass (remodeling). Among factors contributing to these changes is a genetic predisposition which includes disease variants which directly affect cardiac structure and variants which predispose to comorbidities which secondarily lead to remodeling. (**B**) LV mass measurements were derived from a clinical population who had undergone transthoracic echocardiography (TTE). LV mass variation in this population was modeled using a SNP-based polygenic predictor. This predictor was then tested for associations with clinical phenotypes ascertained in a larger clinical population. Clinical diagnoses associated with the predictor are associated with genetic factors that also associate with LV mass variation in the TTE population.

associations had positive odds-ratios, indicating that higher genetically predicted LV mass was associated with an increased risk of the clinical phenotype. Among the significant associations were modifiable risk factors including obesity ($p = 5.6 \times 10^{-8}$), hypertension ($p = 1.0 \times 10^{-5}$), coronary artery disease ($p = 2.0 \times 10^{-5}$), and type 2 diabetes (T2D) ($p = 7.6 \times 10^{-5}$) as well as cardiac diagnoses including cardiomyopathies ($p = 1.2 \times 10^{-5}$), cardiomegaly ($p = 1.8 \times 10^{-4}$), and atrial fibrillation/flutter ($p = 4.3 \times 10^{-4}$). There were also associations with renal and pulmonary vascular disease phenotypes.

Genetic risk score analyses.

Four of the associations were with modifiable risk factors that represent potential targets for intervention. To verify these associations, we employed a Mendelian Randomization approach to determine whether genetic risk scores (GRSs) derived from large GWAS studies for these risk factors associated with LV mass in the TTE population. A GRSs for each risk factor was strongly associated with a positive control phenotype measured in an independent population (Supplementary Table 4). There was a consistent positive association between a BMI GRS and LV Mass in both genders (Table 2). Among men, there were also significant positive associations with systolic blood pressure (SBP) ($p = 0.03$) and coronary artery disease ($p = 0.03$) with consistent directions of effects across all association methods evaluated, though the point estimates for the effect estimates were smaller for SBP when using the MR-Egger (change in log LV mass per unit change in SBP = 0.002) and Weighted Median (0.003) methods, which are less sensitive to the effects of pleiotropy, as compared to the Inverse-variance weighted average meta-analysis method (IVWM) (0.004).

Among women, increased type 2 diabetes (T2D) genetic risk was associated with higher LV mass (Table 2). However, the effect estimates were considerably weaker when using the MR-Egger and Weighted Median methods, suggesting pleiotropy. Also, among males, there was evidence of heterogeneity among the SNP associations (heterogeneity p-value = 0.02 and the MR-Egger intercept p-value = 0.01). Elevated BMI is a risk factor of T2D and when the analyses were repeated after excluding BMI-associated SNPs from the GRS, the T2D association was not significant in either men or women, suggesting the T2D association was likely due to the effects of elevated BMI on both T2D risk and LV mass. In sum, these analyses confirm the observed associations between LV Mass and obesity and, in males, associations with both hypertension and coronary artery disease.

Discussion

We used TTE measurements taken in a heterogeneous clinical population to identify diseases and risk factors associated with LV mass variability. We identified 29 clinical diagnoses associated with genetic variation that also associates with LV mass. We used GRSs to confirm associations between genetic predictors of adiposity, blood pressure and atherosclerotic disease, and LV mass, and found that a genetic predisposition toward these risk factors associates with higher LV mass measurements. In aggregate, these analyses are in agreement with the known clinical epidemiology of LV remodeling, and extend our understanding of the genetic epidemiology of this phenotype.

An individual's LV mass is not static and may increase over time due to the unmitigated effects of disease processes driven by gene-by-environment interactions. Thus, LV mass is a biomarker that measures the severity and duration of exposure to a broad range of pathological influences. Within a population, LV mass has been shown to be a prognostic measure of cardiac health[1–3]. Thus, the genetic architecture of LV mass in a clinical population could be expected to capture important genetic influences of disease.

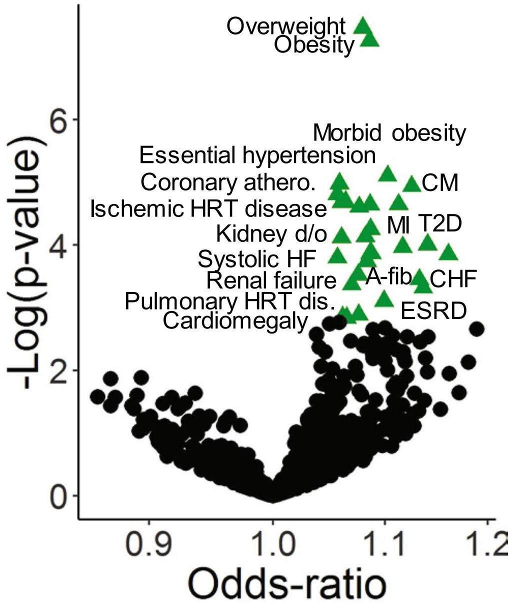

Figure 2. Diagnoses associated with genetically predicted LV mass. The funnel plot summarizes clinical associations between a PheWAS phenotype and predicted LV Mass. Each point represents a logistic regression association analysis, adjusting for age, sex and 5 principal components. Odds-ratios are the risk associated with a 1 standard deviation change in the value of the LV mass genetic predictor. Points highlighted by green triangles are significant at a false discover rate (FDR) $q < 0.05$, and selected points are labelled.

The genetic architecture of LV mass has remained largely elusive to SNP-based discovery approaches, and only one significant SNP association has been reported[4–6]. We did not identify individual SNPs associated with LV mass, and we did not replicate the previously reported association in the *TTN* gene. To gain further insights into the genetic architecture, we leveraged a polygenic approach whereby we constructed a genetic risk score that captured the additive contributions of a large number of SNPs. We then employed PheWAS to identify clinical phenotypes associated with the genetic variation captured by this predictor[9]. Among the phenotypes associated with the GRS were diagnoses related to elevations in risk factors such as obesity and hypertension, diagnoses of heart disease including diagnoses of end-stage disease such as heart failure as well cardiac diseases such as coronary artery disease.

A prior study in a Japanese population used genetic correlation analyses, another polygenic approach, to test for associations between LV mass and 30 candidate diseases and reported significant associations with T2D, stroke risk and atrial fibrillation risk[5]. Our study, which employed a discovery-based approach, refined and extends these findings by identifying a much broader range of clinical phenotypes associated with LV mass. Indeed, the extended range of pathologies associated with LV mass demonstrate that genetic variation underlying variability in LV mass also underlies diagnoses affecting multiple organ systems including the kidney, lungs and heart, which may account for the strong association between elevated LV mass and mortality[3].

Our findings have direct translational relevance to the target population from which the LV mass genetic predictor was derived. In the two sample approach used in these analyses, an association is observed when genetic variation associated with LV mass also associates with the disease or risk factor. The associations with obesity, hypertension and CAD indicate that genetic variation associated with increased risk for these diagnoses may also contribute to structural heart changes in our clinic population, an observation consistent with the Mendelian randomization analyses. It is important to note that the genetic architecture of a trait reflects gene-by-environmental interactions present in a population. These interactions can be exploited to develop approaches to attenuate genetic risk. For instance, the penetrance of obesity-associated genetic variation is modulated by environment[10,11]. Thus, one direct implication of our findings is that prevention efforts in the TTE population should be directed toward approaches with mitigate the genetic predisposition towards these risk factors including promoting health behaviors that reduce both obesity and CAD risk, and treating hypertension.

There were differences in the patterns of associations between men and women with respect to SBP and CAD. One explanation for the CAD difference is the fact that men at elevated CAD genetic risk are more likely to have CAD events, as compared to women with comparable genetic risk[12,13]. Thus, a CAD GRS is a poorer surrogate for CAD events among women, and this would lead to an attenuation of the association among women. For SBP, one reason that an association may not have been observed in women is that, for a given change in SBP, men manifest larger differences in LV mass[14]. Thus, the effect sizes in women may be beyond the power of this study to detect. An alternative explanation is that the women's blood pressures may be under better control than men and, thus, they have not undergone remodeling to the extent present among men.

There are limitations to these analyses. We used clinical phenotypes defined using diagnostic codes primarily used for reimbursement and which may be under-ascertained or inaccurate, and which can lead to false negative

	Males		Females	
Risk factor	Change in LV Mass[a]	P-value	Change in LV Mass[a]	P-value
Inverse-variance weighted average meta-analysis				
Body mass index	**0.079 (0.015)**	**1.2×10^{-7}**	**0.147 (0.016)**	**1.7×10^{-20}**
Systolic blood pressure	**0.004 (0.002)**	**0.03**	0.001 (0.001)	0.29
Coronary artery disease	**0.027 (0.013)**	**0.03**	−0.002 (0.014)	0.87
Type 2 diabetes (T2D)	−0.001 (0.008)	0.92[b]	**0.018 (0.008)**	**0.02**
T2D, excluding BMI SNPs[d]	0.017 (0.015)	0.25	**−0.008 (0.015)**	**0.62**
MR-Egger method				
Body mass index	**0.146 (0.047)**	**0.002**	**0.168 (0.050)**	**0.001**
Systolic blood pressure	0.002 (0.007)	0.7	0.000 (0.007)	0.95
Coronary artery disease	0.044 (0.030)	0.15	−0.040 (0.033)	0.23
Type 2 diabetes (T2D)	−0.042 (0.018)	0.02[c]	0.004 (0.019)	0.82
T2D, excluding BMI SNPs[d]	−0.026 (0.046)	0.58	0.049 (0.049)	0.31
Weighted median method				
Body mass index	**0.114 (0.025)**	**4.0×10^{-6}**	**0.145 (0.029)**	**6.8×10^{-7}**
Systolic blood pressure	0.003 (0.003)	0.24	−0.002 (0.003)	0.51
Coronary artery disease	0.037 (0.020)	0.06	−0.020 (0.022)	0.37
Type 2 diabetes (T2D)	−0.017 (0.014)	0.23	0.008 (0.015)	0.6
T2D, excluding BMI SNPs[d]	0.006 (0.023)	0.81	−0.003 (0.025)	0.9

Table 2. Associations between genetic predictors for selected risk factors and LV mass, by sex. [a]Estimates (standard error) represent the changes in the log(LV mass) per unit change of the risk factor. [b]Heterogeneity $p < 0.05$. [c]MR-Egger intercept $p < 0.05$. [d]Excludes SNPs associated with BMI at $p < 0.05$.

associations. Furthermore, the set of diagnoses that we interrogated were not inclusive of all clinical diagnoses. The TTE data set was a real-world clinical data set and the clinical protocols used may have varied over time. However, despite this potential heterogeneity, we were still able to recapitulate known LV mass associations. We did not confirm that the LV mass genetic predictor associated with measured LV mass in an independent genotyped sample. Thus, it is possible, though unlikely, that the predictor is measuring genetic variation that is unrelated to LV mass. These analyses were limited to individuals of European ancestry, and future analyses in other ancestries are needed to describe the local epidemiology of these populations. Treatments and other environmental factors may cause LV mass reverse remodeling which will attenuate or eliminate associations between LV mass and the mitigated diagnoses or risk factors. Finally, these findings are most relevant to the population from which the TTE measures were derived. We view this as a strength, as this approach identifies potentially untreated genetic risk mechanisms that directly impact the population that the biomarker was measured in.

In summary, we leveraged the polygenic architecture underlying LV mass variability in a clinical population to identify clinical diagnoses associated with structural heart disease. Consistent with the prognostic nature of this phenotype, a genetic predictor of LV mass was associated with end-stage organ disease such as heart failure and kidney failure. Importantly, we also identified and confirmed associations with modifiable risk factors including obesity, hypertension and coronary artery disease. These findings highlight the power of polygenic methods to elucidate the genetic architecture of disease, as compared to SNP-based analyses, and extend our understanding of genetic modulators of LV remodeling in a clinical population. Importantly, our results suggest that well-recognized modifiable risk factors of LV remodeling associate with LV mass increases, suggesting that they are incompletely treated among the patients we studied. Future studies should assess whether genetic risk factor associations are similar in diverse populations.

Methods

Study populations.

The echocardiography population was derived from the VUMC BioVU resource, a collection of individuals seen at VUMC whose EHR data was de-identified and linked to a DNA biobank constructed from discarded blood samples[15]. This IRB-approved resource includes individual-level clinical data and procedural reports (e.g., echocardiography). TTE measurements were extracted from VUMC's clinical echocardiography database for adults over 35 years old who had TTEs performed between 2008 and 2016 and who had DNA available for SNP genotyping. The majority of subjects were identified as white; thus, analyses were restricted to individuals of genetic European ancestry (EA)[16]. The final echocardiography population comprised 7,601 unrelated individuals (Supplementary Table 1).

The phenome-wide association study (PheWAS) included EA individuals born prior to 1990 from the eMERGE network (phases 1-3) (n = 31,773, excluding VUMC)[8] and additional BioVU subjects over 18 years old (n = 27,065) (Supplementary Table 2). The participating eMERGE sites were Columbia University, Geisinger, Marshfield Clinic, Northwestern University, Mayo Clinic, Harvard University, Mt. Sinai Health System, and Kaiser Permanente/University of Washington, Seattle.

Analyses were approved by each eMERGE institution's Institutional Review Board (IRB)[8,15].

Genetic data. BioVU subjects underwent SNP genotyping using the Illumina Infinium Multi-Ethnic Genotyping Array ($\mathrm{MEGA^{EX}}$) platform. eMERGE subjects were genotyped on multiple platforms and underwent QC analyses and imputation as previously described[17,18]. Quality control (QC) analyses used PLINK v 1.90β3[19] and included reconciling strand flips, verifying that allele frequencies were concordant among data sets, and identifying duplicate and related individuals (one of each pair of subjects with a pi-hat >0.05 was excluded)[17,20]. Data sets were standardized using the HRC-1000G-check tool v4.2.5 (http://www.well.ox.ac.uk/~wrayner/tools/) and pre-phased using SHAPEIT[21]. For the subjects with TTEs, data were imputed using IMPUTE2[22] in conjunction with the 10/2014 release of the 1000 Genomes cosmopolitan reference haplotypes. All other genetic data for were imputed using the Michigan Imputation Server (HRC v1.1)[23]. Imputed data were filtered for a sample missingness rate <2%, a SNP missingness rate <4% and a SNP deviation from Hardy-Weinberg $p < 10^{-6}$. There were 5,455,089 imputed SNPs with MAF > 1% that passed QC in all data sets. The LV mass genetic predictor was constructed using a LD-reduced (r-square<0.9) subset of 1,005,032 SNPs. Principal components were generated using the SNPRelate package[24].

Echocardiographic and phenotype data. LV mass was calculated using clinically acquired echocardiographic parameters according to the formula:[25]

$$\text{LV mass} = 0.8\{1.04[([\text{LVEDd} + \text{IVSd} + \text{PWd}]^3 - \text{LVEDd}^3)]\} + 0.6$$

where LVEDd = LV internal diameter at end diastole; IVSd = interventricular septal thickness at end diastole; and PWd = LV posterior wall thickness at end diastole. There were 7,601 individuals with LV mass measurements with a value between 50 and 500 grams (g). For individuals with multiple TTEs, only measurements from the first were used. LV mass was log-transformed for these analyses. While LV mass is often indexed to body surface area or height, we used unindexed values to avoid spurious genetic associations caused by adjusting a phenotype by another highly heritable phenotype (referred to as collider bias)[26].

PheWAS were conducted in the eMERGE and BioVU populations using clinical phecode phenotypes (https://phewas.mc.vanderbilt.edu/), which are collections of related ICD-9-CM (International Classification of Disease, Ninth revision) diagnosis codes[27,28]. Cases were individuals with two or more instances of a PheWAS diagnosis appearing in their medical record[29]. Phenotypes that affected a single sex (such as prostate cancer or uterine prolapse) were excluded. There were 894 clinical phenotypes with ≥300 cases (our minimum criteria for inclusion). Controls were subjects without the clinical phenotype or any closely related PheWAS code (using the standard phecode control groupings) and whose age (BioVU) or decade of birth (eMERGE) fell within the range of values observed among cases. The cardiomegaly PheWAS code (code 416) was used a positive control phenotype to validate the LV mass polygenic predictor. For these analyses, cases were individuals who had one or more instances of this diagnosis in their medical record, which is a more sensitive case definition.

GWAS summary statistics. To further explore the relationship between LV mass and candidate phenotype associations, summary statistics from prior large-scale GWAS were used to construct genetic risk scores representing these candidate phenotypes. Specifically, summary statistics were obtained for coronary artery disease (CAD) from the CARDIOGRAM C4D consortium GWAS[30], body mass index (BMI) from the GIANT Consortium[31,32], systolic blood pressure (SBP) from a GWAS from the Million Veterans Program[33], and type 2 diabetes (T2D) from the DIAGRAM Consortium[34]. Summary statistics were downloaded from the consortia websites.

Analysis. GWAS was performed assuming an additive model and employed a multivariable linear model adjusted for age, sex and 10 principal components. A $p < 5\text{x}10^{-8}$ was considered significant.

The vast majority of the heritability attributable to common SNPs is accounted for by SNPs that typically do not meet the criteria for genome-wide significance[35,36]. Thus, we used a modelling approach that assigns SNP weightings based on large numbers of common SNPs to construct a genetic risk score for LV mass. Genetically predicted LV mass was computed using a two-step approach, as previously described[9,16]. First, predictive weightings were assigned to SNPs using Bayesian sparse linear mixed modelling (BSLMM), as implemented in the GEMMA v0.95α package[37]. The BSLMM approach jointly models the contribution of all SNPs to the observed phenotypic variance by employing a hybrid of generalized linear mixed modelling and sparse regression models[38]. The models were adjusted for age on the date of the TTE, sex and 5 PCs; 100,000 sampling steps were run. The estimated proportion of additive genetic variance explained by the common SNPs used to model the SNP weightings is the median estimated value taken from the last 50,000 sampling steps[39]. SNP weightings are comprised of a small polygenic effect (α), a large effect (β) and a posterior probability that the SNP is in the large effect group (γ). The SNP weight (w) is computed from these estimates using the equation: $w=\alpha+\beta\gamma$. The SNP weightings were used to compute a predicted LV mass for each individual in the PheWAS analysis (i.e. subjects without a TTE measurement) using the following equation

$$\text{Predicted feature value} = \sum_{i=1}^{\#\text{SNPs}} (\mathrm{w_i} \times [\text{SNP genotype}]_\mathrm{i}) \tag{1}$$

where genotype is the number of alleles present for the SNP (coded as 0, 1 or 2).

To identify clinical phenotypes associated with genetically predicted LV mass, multivariable logistic regression, adjusting for 5 PCs, sex and either [birth decade and site for eMERGE sites] or [maximum age for BioVU], was used to test for an association with each PheWAS phenotype (dependent variable) and the predicted feature (independent variable) using the R PheWAS package[40]. Odds-ratios (ORs) are the risk of disease per standard deviation (s.d.) increase in the genetically predicted biomarker value. PheWAS analyses were run separately for the BioVU and eMERGE subjects and results were meta-analyzed using the METAL package[41]. To adjust for multiple testing, we employed a Benjamini-Hochberg (B-H) false discovery rate (FDR)[42] adjustment and a q-value < 0.05, which has previously been shown to perform well for these analyses[9], was considered significant.

To further characterize the candidate associations between the genetic risk score for LV mass and the PheWAS phenotypes, we generated weighted genetic risk scores for four candidate phenotypes and assessed their associations with LV mass within the TTE population. The genetic risk scores were based on summary statistics from GWAS of each phenotype. The SNPs comprising each GRS were selected using a clumping algorithm that identified an LD-reduced set (r-square < 0.05) of the most significantly-associated SNPs that had a minor allele frequency (MAF) > 5% and an association p-value $< 5 \times 10^{-6}$ in the original GWAS[43]. To validate the relevance of each GRS, its association was tested against the corresponding phenotype ascertained in BioVU subjects (n = 13,077). For continuous phenotypes (BMI and SBP), the phenotype represents the median for all available values for an individual. Binary phenotypes (coronary artery disease and type 2 diabetes) are based on PheWAS phenotypes.

Associations were tested using an inverse-variance weighted average meta-analysis (IVWA). Heterogeneity p-values are based on the Cochran's Q statistic, and a low p-value indicates that that one or more variants in the GRS may be pleiotropic. Though less powered than the IVWA, associations were also tested by the MR-Egger and Weighted Median methods, which provide more accurate estimates of effect sizes in the presence of horizontal pleiotropy, strong outliers or invalid instrumental variables. Associations were measured using the Mendelian Randomization R package[22]. Analyses were stratified by sex, and association estimates represent the change in log(LV mass) per unit change in the phenotype corresponding to the GRS. An IVWA association $p < 0.05$ was considered significant. For T2D, a GRS was also constructed that excluded SNPs associated with BMI at $p < 0.05$.

Data availability

G.W.A.S. summary statistics for LV Mass will be made available through dbGaP or can be obtained from the corresponding author. eMERGE data are available through dbGaP (phs000360.v3.p1).

Received: 15 September 2019; Accepted: 16 April 2020;
Published online: 05 May 2020

References

1. Levy, D. *et al*. Echocardiographically detected left ventricular hypertrophy: prevalence and risk factors. The Framingham Heart Study. *Ann. Intern. Med.* **108**, 7–13 (1988).
2. Levy, D., Garrison, R. J., Savage, D. D., Kannel, W. B. & Castelli, W. P. Left ventricular mass and incidence of coronary heart disease in an elderly cohort. The Framingham Heart Study. *Ann. Intern. Med.* **110**, 101–107 (1989).
3. Levy, D., Garrison, R. J., Savage, D. D., Kannel, W. B. & Castelli, W. P. Prognostic implications of echocardiographically determined left ventricular mass in the Framingham Heart Study. *N. Engl. J. Med.* **322**, 1561–1566 (1990).
4. Wild, P. S. *et al*. Large-scale genome-wide analysis identifies genetic variants associated with cardiac structure and function. *J. Clin. Invest.* **127**, 1798–1812 (2017).
5. Kanai, M. *et al*. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* **50**, 390–400 (2018).
6. Aung, N. *et al*. Genome-Wide Analysis of Left Ventricular Image-Derived Phenotypes Identifies Fourteen Loci Associated with Cardiac Morphogenesis and Heart Failure Development. *Circulation*, https://doi.org/10.1161/CIRCULATIONAHA.119.041161 (2019).
7. Kannel, W. B. Left ventricular hypertrophy as a risk factor: the Framingham experience. *J Hypertens Suppl* **9**, S3-8; discussion S8-9 (1991).
8. Gottesman, O. *et al*. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet. Med.* **15**, 761–771 (2013).
9. Mosley, J. D. *et al*. A study paradigm integrating prospective epidemiologic cohorts and electronic health records to identify disease biomarkers. *Nat Commun* **9**, 3522 (2018).
10. Rosenquist, J. N. *et al*. Cohort of birth modifies the association between FTO genotype and BMI. *Proc. Natl. Acad. Sci. USA* **112**, 354–359 (2015).
11. Abadi, A. *et al*. Penetrance of Polygenic Obesity Susceptibility Loci across the Body Mass Index Distribution. *Am. J. Hum. Genet.* **101**, 925–938 (2017).
12. Hajek, C. *et al*. Coronary Heart Disease Genetic Risk Score Predicts Cardiovascular Disease Risk in Men, Not Women. *Circ Genom Precis Med* **11**, e002324 (2018).
13. Inouye, M. *et al*. Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. *J. Am. Coll. Cardiol.* **72**, 1883–1893 (2018).
14. Savage, D. D., Levy, D., Dannenberg, A. L., Garrison, R. J. & Castelli, W. P. Association of echocardiographic left ventricular mass with body size, blood pressure and physical activity (the Framingham Study). *Am. J. Cardiol.* **65**, 371–376 (1990).
15. Roden, D. M. *et al*. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin. Pharmacol. Ther.* **84**, 362–369 (2008).
16. Mosley, J. D. *et al*. Investigating the Genetic Architecture of the PR Interval Using Clinical Phenotypes. *Circ Cardiovasc Genet* **10**, (2017).
17. Zuvich, R. L. *et al*. Pitfalls of merging GWAS data: lessons learned in the eMERGE network and quality control procedures to maintain high data quality. *Genet. Epidemiol.* **35**, 887–898 (2011).
18. Stanaway, I. B. *et al*. The eMERGE genotype set of 83,717 subjects imputed to ~40 million variants genome wide and association with the herpes zoster medical record phenotype. *Genet. Epidemiol.*, https://doi.org/10.1002/gepi.22167 (2018).
19. Purcell, S. *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
20. Mosley, J. D. *et al*. Defining a Contemporary Ischemic Heart Disease Genetic Risk Profile Using Historical Data. *Circ Cardiovasc Genet* **9**, 521–530 (2016).
21. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
22. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
23. Das, S. *et al*. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
24. Zheng, X. *et al*. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).

25. 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 Echocardiography, a branch of the European Society of Cardiology. *J Am Soc Echocardiogr* **18**, 1440–1463 (2005).
26. Day, F. R., Loh, P.-R., Scott, R. A., Ong, K. K. & Perry, J. R. B. A Robust Example of Collider Bias in a Genetic Association Study. *Am. J. Hum. Genet.* **98**, 392–393 (2016).
27. Denny, J. C. *et al*. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205–1210 (2010).
28. Denny, J. C. *et al*. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1110 (2013).
29. Wei, W.-Q. *et al*. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLoS ONE* **12**, e0175508 (2017).
30. Nikpay, M. *et al*. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47**, 1121–1130 (2015).
31. Yengo, L. *et al*. Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Hum. Mol. Genet.*, https://doi.org/10.1093/hmg/ddy271 (2018).
32. Shungin, D. *et al*. New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015).
33. Giri, A. *et al*. Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nat. Genet.* **51**, 51–62 (2019).
34. Mahajan, A. *et al*. Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat. Genet.* **50**, 559–571 (2018).
35. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).
36. Yang, J. *et al*. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
37. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
38. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet.* **9**, e1003264 (2013).
39. Wheeler, H. E. *et al*. Survey of the Heritability and Sparse Architecture of Gene Expression Traits across Human Tissues. *PLoS Genet.* **12**, e1006423 (2016).
40. Carroll, R. J., Bastarache, L. & Denny, J. C. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* **30**, 2375–2376 (2014).
41. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
42. Majumdar, A., Haldar, T. & Witte, J. S. Determining Which Phenotypes Underlie a Pleiotropic Signal. *Genet. Epidemiol.* **40**, 366–381 (2016).
43. International Schizophrenia, C. *et al*. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).

Acknowledgements

The authors wish to acknowledge the expert technical support of the VANTAGE and VANGARD core facilities, supported in part by the Vanderbilt-Ingram Cancer Center and Vanderbilt Vision Center. Data on coronary artery disease and myocardial infarction have been contributed by CARDIoGRAMplusC4D investigators and have been downloaded from www.CARDIOGRAMPLUSC4D.ORG. See supplementary attachment for additional Million Veterans Project (MVP) acknowledgments. This work was supported by a career development award from the American Heart Association (16FTF30130005) (JDM), AHA 17IBDG33780015, AHA 17SFRN33520017, R01HL140074 (QSW), NIH/NHLBI 1R01HL140074 (QSW), R01 LM010685 (JCD), RO1 GM109145 (CMS), K01HL124045 (AMA-O), MVP #BX003360-01 (AMH). JNH was supported by the Vanderbilt Molecular and Genetic Epidemiology of Cancer (MAGEC) training program, funded by T32CA160056. VUMC's BioVU resource is supported by numerous sources: institutional funding, private agencies, and federal grants and include the NIH funded Shared Instrumentation Grant S10RR025141; and CTSA grants UL1TR002243, UL1TR000445, and UL1RR024975. Genomic data are also supported by investigator-led projects that include U01HG004798, R01NS032830, RC2GM092618, P50GM115305, U01HG006378, U19HL065962, R01HD074711; additional funding sources listed at https://victr.vanderbilt.edu/pub/biovu/. This eMERGE Network (Phase III) was initiated and funded by the NHGRI through the following grants: U01HG8657 (Group Health Cooperative/University of Washington); U01HG8685 (Brigham and Women's Hospital); U01HG8672 (Vanderbilt University Medical Center); U01HG8666 (Cincinnati Children's Hospital Medical Center); U01HG6379 (Mayo Clinic); U01HG8679 (Geisinger Clinic); U01HG8680 (Columbia University Health Sciences); U01HG8684 (Children's Hospital of Philadelphia); U01HG8673 (Northwestern University); U01HG8701 (Vanderbilt University Medical Center serving as the Coordinating Center); U01HG8676 (Partners Healthcare/Broad Institute); and U01HG8664 (Baylor College of Medicine). This research is based on data from the Million Veteran Program, Office of Research and Development, Veterans Health Administration, and was supported by award # [I01BX003360]. This work was supported using resources and facilities of the VA Informatics and Computing Infrastructure (VINCI), VA HSR RES 13-457. This publication does not represent the views of the Department of Veterans Affairs or the United States Government.

Author contributions

J.D.M. and Q.S.W. developed the analyses. J.D.M., Q.S.W. and C.M.S. wrote the main manuscript. Analytic and content support was provided by R.T.L., E.F., C.M.S., T.J.W. and M.S. Data were provided by T.L.E., J.N.H., A.M.H., A.G., M.M.S., E.L.B., W.K.C., I.J.K., A.M.A., G.P.J., E.B.L., D.R.C., M.S.W., K.M.B., D.R.C., H.H., J.C.D., D.M.R. All authors reviewed the manuscript and provided input relevant to their expertise.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-64525-z.

Correspondence and requests for materials should be addressed to J.D.M.