






Unsupervised ensemble-based phenotyping enhances discoverability of genes related to left-ventricular morphology

Received: 7 January 2023

Accepted: 25 January 2024

Published online: 11 March 2024


 Check for updates

Rodrigo Bonazzola ^{1,2}, Enzo Ferrante ³, Nishant Ravikumar^{1,2}, Yan Xia^{1,2}, Bernard Keavney ^{4,5,6}, Sven Plein², Tanveer Syeda-Mahmood⁷ & Alejandro F. Frangi ^{6,8,9,10,11} 

Recent genome-wide association studies have successfully identified associations between genetic variants and simple cardiac morphological parameters derived from cardiac magnetic resonance images. However, the emergence of large databases, including genetic data linked to cardiac magnetic resonance facilitates the investigation of more nuanced patterns of cardiac shape variability than those studied so far. Here we propose a framework for gene discovery coined unsupervised phenotype ensembles. The unsupervised phenotype ensemble builds a redundant yet highly expressive representation by pooling a set of phenotypes learnt in an unsupervised manner, using deep learning models trained with different hyper-parameters. These phenotypes are then analysed via genome-wide association studies, retaining only highly confident and stable associations across the ensemble. We applied our approach to the UK Biobank database to extract geometric features of the left ventricle from image-derived three-dimensional meshes. We demonstrate that our approach greatly improves the discoverability of genes that influence left ventricle shape, identifying 49 loci with study-wide significance and 25 with suggestive significance. We argue that our approach would enable more extensive discovery of gene associations with image-derived phenotypes for other organs or image modalities.

Genome-wide association studies (GWAS) have accelerated the discovery of associations between genomic and complex traits¹. In general, they analyse genetic variants (that is, the genotype) in a sample of individuals to test their possible association with the presence of disease

or with systematic changes in measurable traits, known broadly as phenotypes in this context. GWAS have already successfully identified genetic variants associated with a broad range of diseases and other complex traits, such as metabolic, anthropometric or behavioural ones.

¹Centre for Computational Imaging and Simulation Technologies in Biomedicine, School of Computing and School of Medicine, University of Leeds, Leeds, UK. ²Leeds Institute of Cardiovascular and Metabolic Medicine, School of Medicine, University of Leeds, Leeds, UK. ³Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH-UNL/CONICET, Santa Fe, Argentina. ⁴Division of Cardiovascular Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK. ⁵Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK. ⁶NIHR Manchester Biomedical Research Centre, Manchester Academic Health Science Centre, Manchester, UK. ⁷IBM Almaden Research Center, San Jose, CA, USA. ⁸Division of Informatics, Imaging and Data Sciences, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK. ⁹Department of Computer Science, School of Engineering, Faculty of Science and Engineering, University of Manchester, Manchester, UK. ¹⁰Medical Imaging Research Center (MIRC), University Hospital Gasthuisberg, Cardiovascular Sciences and Electrical Engineering Departments, KU Leuven, Leuven, Belgium. ¹¹Alan Turing Institute, London, UK.  e-mail: alejandro.frangi@manchester.ac.uk

These findings have improved our understanding of disease pathogenesis, facilitating the development of better treatments, supporting drug discovery and assisting advances towards precision medicine.

Large-scale epidemiological imaging studies have correlated image-derived phenotypes (IDPs) with genetic data to identify the genetic basis of organ structure and function in health and disease. In cardiology, GWASs have been performed on clinically relevant quantitative left-ventricular (LV) indices, such as LV volumes, LV mass and LV ejection fraction. Diagnosis of patients with heart disease usually involves a quantitative analysis of the LV as a key component^{2,3}. Although there are discrepancies in the number of genetic loci associated with changes in LV IDPs from recently reported GWAS^{2,4}, some consistent genetic factors have been established.

These cardiac imaging genetics studies were based on traditional approaches, where handcrafted features characterizing LV IDPs were first determined, before running GWAS to find the associated genetic loci. Although these IDPs have been clinically used to diagnose heart disease, they do not provide detailed representations of the chamber morphology and its variation across the population. In this paper, we advance the view that shape features encoded in a learnt latent space can provide a more refined imaging phenotype, which is more informative than traditional measurements. When associated with genetic variation, this can provide novel insights into the genetic basis of cardiac structure and function.

The unprecedented amount of linked genetic and cardiac imaging data available within the UK Biobank (UKBB)⁵ facilitates using unsupervised machine learning techniques to automatically learn a set of characteristics that best describe the morphology of the heart. At the same time, atlas-based methods have been proposed to generate three-dimensional (3D) meshes that represent cardiac anatomy from volumetric images^{6,7}. On top of this work, we use the latest advances in graph-convolutional neural networks⁸ to learn low-dimensional representations that consider mesh topology. While standard convolutional neural networks operate on domains with an underlying Euclidean or grid-like structure (for example, images), geometric deep learning generalizes convolutions to non-Euclidean domains such as graphs, meshes and manifolds, taking into account their topology and irregular structure. Previous studies used mesh autoencoders to model the expression space of human face surfaces⁹. Here, we show that such models can enable anatomical variation in cardiac structures to be learnt and correlated with genetic data.

In this work, we learn compact and nonlinear representations of cardiac anatomy in an unsupervised setting via convolutional-mesh autoencoders (CoMA). We propose that the learnt features can identify genetic loci that affect cardiac morphology due to their ability to explain shape variability across the population. We show that such representations can indeed be used to discover novel genetic associations via GWAS, which was not previously possible with traditional handcrafted IDPs such as volume, mass and function indices.

In a previous conference communication¹⁰, we reported on a much simpler exploratory methodology and analysis, wherein we demonstrated that latent representations learnt from LV surface meshes can find significant genetic associations. In contrast, using latent representations of anatomical meshes of the entire surface, and not just LV functional parameters^{2,4,11} or individual mesh nodes independently³ as in previous genetic studies, could reproduce but only marginally expand the knowledge about previously discovered loci. We proposed that this was partially due to the high dimensionality and insufficient expressiveness of the image-derived anatomical phenotypes. In this study, we address these two concerns. First, a new framework, namely, the unsupervised phenotype ensemble (UPE), adds robustness and discoverability: we replicate recently reported genes and discover several novel genetic associations, not yet reported in the literature. Furthermore, this paper expands the size of our cardiac magnetic resonance (CMR) dataset, as well as the accuracy of the derived meshes.

We analysed 48,651 participants from the UKBB, deriving high-quality phenotypes and robust latent representations from cardiac segmentations and meshes with a state-of-the-art high-throughput and validated CMR analytic pipeline¹². We conducted an extensive analysis of the stability of the results. This article underlines the crucial role of high-quality latent representations in imaging genetics to greatly improve gene discoverability associated with LV morphology.

A schematic overview of the proposed methodology is presented in Fig. 1. The details of each step are outlined in the Methods section. First, we extracted a surface mesh representation of the anatomical structures. In particular, we studied 3D meshes representing LV at the end of diastole from CMR images of the UKBB database using an automatic deep learning-based segmentation method¹². We then learn a low-dimensional representation of the 3D meshes, which captures anatomical variations using an encoder–decoder model. All meshes were projected onto this latent space to derive a few shape descriptors (or latent variables) for each of them. GWAS used these features to discover genetic variants associated with shape patterns. Furthermore, to enhance discoverability, we adopt an ensemble-based approach: a set of phenotypes obtained through different models trained and configured with varying network metaparameters and weight initializations (which induce diversity in the learnt representations) are pooled together in one ensemble, yielding redundant yet more expressive representations than the individual latent vectors. The expected improvement of UPE is based on previous work providing evidence that the use of deep ensembles can lead to diverse data representations that are linked in non-trivial ways, even when only the random initialization differs¹³. GWAS is performed against each phenotype of the ensemble, one at a time. A corrected Bonferroni threshold is then calculated to keep the false discovery rate below 5%, by dividing the usual genome-wide threshold by the number of phenotypes of the ensemble being tested.

We demonstrate that this approach effectively discovers additional biologically relevant genetic associations. It expands on previous knowledge by identifying 49 loci with study-wide significance. From this, only nine loci had been reported in previous GWAS of LV phenotypes. This leaves a total of 40 novel LV associations, with eight loci that were reported here in association with handcrafted LV phenotypes, 12 additional associations obtained through shape principal component analysis (PCA) and 20 that are exclusively attributed to our UPE framework with CoMAs. Furthermore, we report 24 suggestive associations, with some highly plausible causative genes according to pre-existing knowledge.

Results

In the following, we present our GWAS results. First, we investigate handcrafted phenotypes. Second, we examine unsupervised phenotypes obtained via shape PCA. Finally, we examine the results of our proposed UPE approach.

The loci were annotated with gene names on the basis of proximity to the lead single nucleotide polymorphism (SNP) if there was no additional causal evidence in the literature, or with nearby genes likely to mediate the association. For this, we used a diverse array of tools: the functional mapping and annotation (FUMA) web tool¹⁴, g:Profiler¹⁵, S-PrediXcan¹⁶ and the Ensembl Biomart database¹⁷. Among the candidate genes provided by these tools, a literature review was conducted to find evidence of an association with cardiovascular phenotypes, or experimental. Genes with asterisks were annotated solely on the basis of proximity and hence constitute totally novel findings.

Genetic findings

Handcrafted phenotypes. We performed GWAS on traditional cardiac indices obtained using our segmentation approach. These indices were LVEDV, LV sphericity index at end diastole (LVEDSph), LV myocardial mass (LVM) and LV mass-to-volume ratio (LVMVR = LVM/LVEDV). Note

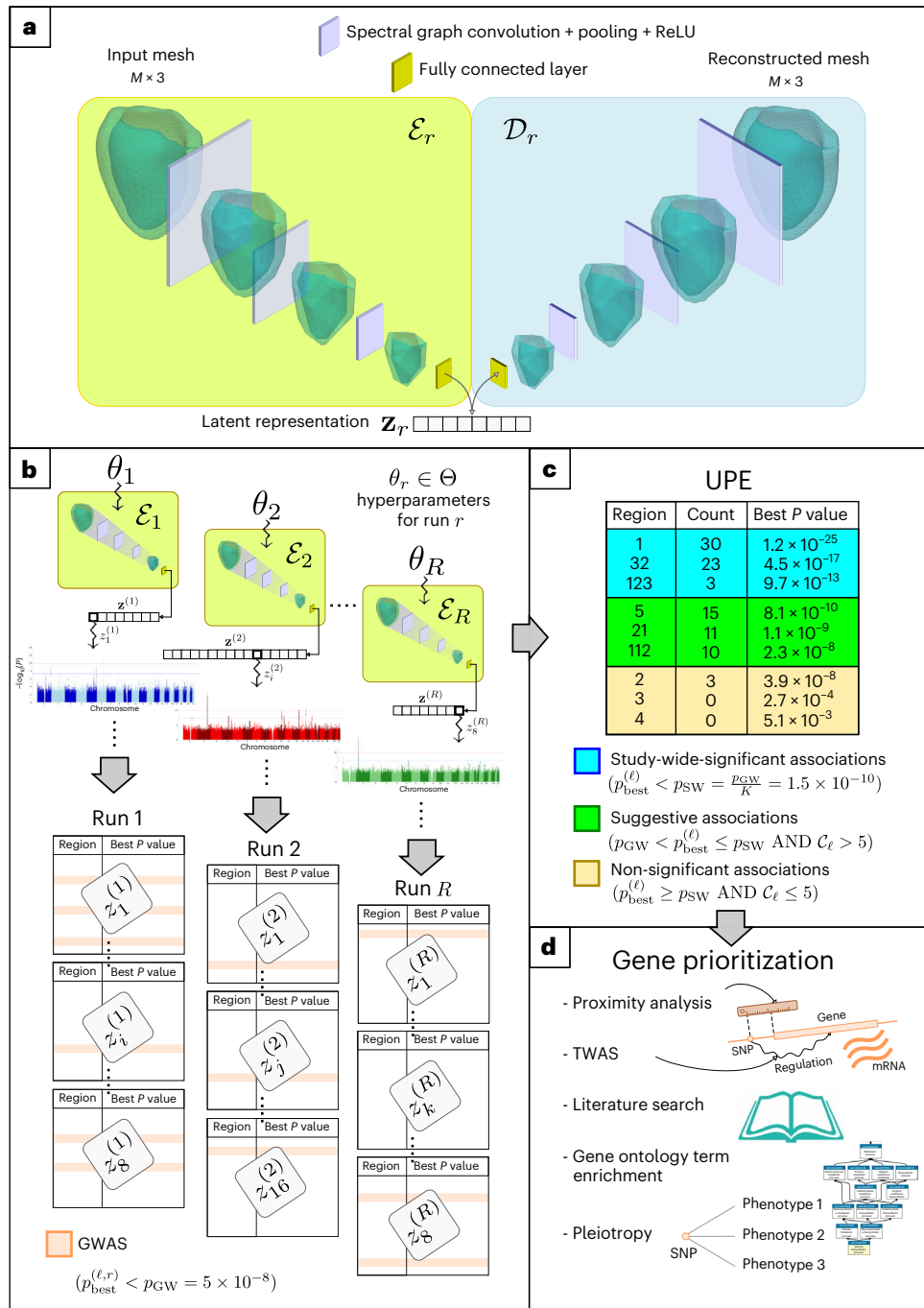


Fig. 1 | Flowchart of the proposed UPE framework. a, A graph-convolutional autoencoder is trained and applied to our set of CMR-derived LV meshes (number of vertices $M = 5,220$) to produce low-dimensional representations of these shapes. In each layer, a representation with fewer vertices is obtained. The bottleneck \mathbf{z} , of the autoencoder with hyperparameters θ , is a n_z^r -dimensional vector for each run r ($n_z^r \in \{8, 16\}$). ReLU, rectified linear unit. **b**, The different latent vectors obtained for each run, $\{\mathbf{z}_r\}_{r=1}^R$ are then tested in a GWAS component

by component, for association with genetic variants. The best association, with P value $p_{\text{best}}^{(\ell,r)}$, is found for each run r and region ℓ (each region is of around 2 Mb in length). **c**, Associations are then aggregated across the ensemble and classified as significant, suggestive and non-significant according to $p_{\text{best}}^{(\ell)}$ and the count C_{ℓ} (as indicated by the rules). **d**, For each significant association, a downstream analysis is conducted to identify potentially causative genes. mRNA, messenger RNA.

that the LVEDSph as calculated here has not been investigated in previous GWAS (although a related phenotype, named ‘LV internal dimensions’ was studied in an early GWAS of echocardiography-derived LV traits¹⁸). Details on how to compute this phenotype can be found in the Supplementary Information.

In the following, we discuss the associations found for each of these phenotypes. The Manhattan plots are shown in Extended Data Figs. 1–4.

For LVEDV, we discover nine independent associations. The association at intergenic SNP rs11153730 is probably related to *PLN*. This gene plays a crucial role in cardiomyocyte calcium handling by acting as a primary regulator of the SERCA protein (sarco- or endoplasmic reticulum Ca^{2+} -ATPase), which transports calcium from the cytosol into the SRI (ref. 19). Mutations in *PLN* have a well-established relationship with dilated cardiomyopathy (DCM)²⁰. In ref. 4, *PLN* was found to be associated with LVEDV and LVESV. However, ref. 2 does not report

this locus for the same phenotypes. The locus on chromosome 2 (with lead SNP rs2042995) is widely known to be associated with *TTN*. This gene encodes the protein titin, which is responsible for assembling myocyte sarcomere, and determines the stretching, contraction and passive stiffness of the myocardium²¹. This gene has been reported by refs. 2,4,11. rs375034445 lies within the body of *BAG3*; this is a well-known cardiac gene coding for a cellular protein that is predominantly expressed in skeletal and cardiac muscle, which plays a role in myocyte homeostasis and in the development of heart failure²²; also, it shows a stronger association with LVESV and LV ejection fraction (LVEF), as found in previous studies^{2,4}. The locus near the *ATXN2* gene has previously been reported for LVEDV and stroke volume (SV)⁴. A candidate causal gene for this association is gene *MYL2*, the lead SNP (rs35350651) lies 558808 base pairs away from this gene's transcription start site (TSS)²³. The gene *TMEM43* has been found in ref. 4 in association with LVESV and LVEF. Finally, gene *MYH6* harbours SNP rs365990. This gene provides instructions for making a protein known as the cardiac α -myosin heavy chain, which is expressed throughout the myocardium during early cardiac development²⁴. Mutations in this gene, as well as the neighbouring *MYH7* responsible for the β -myosin heavy chain, have been linked to several pathologies: cardiomyopathies, arrhythmias and congenital heart disease (CHD). Two additional associations are located close to genes *RRAS2* and *ATG4D*, respectively.

For LVEDSph, we find nine additional independent associations, apart from the *PLN* locus. rs35564079 is located 8,250 bp upstream of the TSS of *NKX2-5*, in chromosome 5. This gene plays a crucial role in heart development; in particular, in the formation of the heart tube, which is a structure that will eventually give rise to the heart and great vessels. *NKX2-5* helps determine the heart's position in the chest and also develops the heart valves and septa. Mutations in the *NKX2-5* gene have been associated with several types of congenital heart defect, including atrial septal defects and atrioventricular block²⁵. It has not been reported in refs. 2 or 4, but shows borderline significance with the fractal dimension of the LV trabeculae¹¹. rs72007904 is located 300 kb upstream of the TSS of the gene *ABRA*. *ABRA* codes for a cardiac and skeletal muscle-specific actin-binding protein located in the Z disc and M-line and binds with actin. Consistent with this, it is differentially expressed in cardiac tissues and skeletal muscle in the genotype-tissue expression (GTEx) data. *ABRA* has been associated with DCM in mice²⁶. rs35001652 is close to *KDM1A*, a gene that codes for a histone demethylase involved in cardiac development, according to studies in mice²⁷. rs463106 lies in the body of gene *PRDM6*. The mouse homologue of this gene, *Prdm6*, has been found to be important in early cardiac development²⁸. An interesting association, with SNP rs162746, is close to gene *ENI*, however, we were not able to find a strong candidate gene in this region. Finally, rs573709385 lies in a gene desert in chromosome 2, the closest protein-coding genes are *ACVR2A* and *ZEB2* (both at around 1.6 Mb).

For LVM, four associations are found: rs4767239 is probably related to developmental gene *TBX5* (T-box transcription factor 5), which has a known role in developing the heart and the limbs²⁹. Through familial studies, mutations in this gene have been associated with Holt-Oram syndrome, a developmental disorder affecting the heart and upper limbs. In particular, there have been no recent reports on GWAS on LV phenotypes. The locus near the *CENPW* gene has a cardiac gene, *HEY2*, possibly causal for this association. *HEY2* has been shown to suppress cardiac hypertrophy through an inhibitory interaction with *GATA4*, a transcription factor that plays a key role in cardiac development and hypertrophy³⁰. *HEY* proteins are direct targets of Notch signalling and have been shown to regulate multiple key steps in cardiovascular development. Studies have found that the loss of *HEY2* in mice leads to cardiac defects with high postnatal lethality³¹. This locus has also been reported as associated to right-ventricular phenotypes³². rs3740293 overlaps gene *SYNPO2L*, which is highly expressed in cardiac tissues (LV and atrial appendage) and skeletal muscle, making it a strong candidate gene. This SNP is also close to

gene *MYOZ1*, which is also supported by our GWAS study (section on transcriptome-wide association studies, 'TWAS'). Both genes have been previously proposed as candidates for cardiac phenotypes, in particular atrial fibrillation^{33,34}. However, *MYOZ1* shows very high expression only in the latter. Loss-of-function variants in this *SYNPO2L* have also been found causative of atrial fibrillation³⁵, supporting this gene as a more likely candidate. rs73243622 is close to the candidate gene *PPARGCIA*. Finally, gene *CDKN1A* has been found in ref. 4 in association with LVESV and LVEF. Finally, for LVMVR, three new loci were found, apart from the *PLN* locus: rs2070458 close to *SMARCB1* (in chromosome 22), rs17460016 in the *FNDC3B* locus (in chromosome 3) and rs12542527 (in chromosome 8). The last is an eQTL for the *MTSS1* gene also linked to LV fractal dimension¹¹.

The detailed summary statistics for the significant associations with handcrafted phenotypes are provided as Supplementary Data.

Shape PCA. A shape PCA model was fit to our set of meshes (Methods). The effect on LV shape for the first 16 modes is shown in the Supplementary Fig. 7. GWAS was performed for these 16 modes and 18 independent loci were found with study-wide significance ($P < 3.1 \times 10^{-9}$). PC1, which is highly correlated with LVEDV, reconfirms the associations with *TTN*, *MYL2* and *MYH6*. A new association, in chromosome 4, is an indel (chr4:120304290_GC_G) located 200 kb downstream of *MYOZ2*. This gene codes for protein that functions by tethering calcineurin to alpha-actinin at Z-discs in muscle cells and inhibits the pathological cardiac hypertrophic response³⁶. Another candidate gene in this locus is *PDE5A*. Indeed, some of the strongest associations overlap the body of this gene (although not the lead variant, which is the indel mentioned above). It has been shown that *PDE5A* is expressed in cardiac myocytes and may have pro-hypertrophic effects³⁷.

PC2 is strongly linked with a new locus in chromosome 17, *GOSR2*. This component seems to be linked to LV conicity. Ref. 11 reports the *GOSR2* locus as significantly associated with trabecular fractal dimension in slices 3 and 4, however, previous GWAS in global LV indices have not reported this locus. More broadly in the literature on genetics of cardiovascular phenotypes, it has been reported as associated to ascending aorta distensibility³⁸, mitral valve geometry³⁹ and CHD⁴⁰.

PC3, highly correlated with LVEDSph, re-discovers the *PLN* and *NKX2-5* loci. It also adds an association in chromosome 1, the SNP rs12142143, which lies within the *ACTN2* gene. This gene codes for the Z disc protein α -actinin-2. This locus has been reported for SV in ref. 4.

PC6 has hits in the *TBX5* and *NKX2-5* loci, with a new association near the *NAV3* gene, that has been found to play a role in heart development in zebrafish⁴¹. PC7 is associated to a SNP near the TSS of *PITX2* gene. It encodes for a transcription factor required for mammalian development, and disruption in its expression in humans causes CHD and is associated with atrial fibrillation. PC10 is linked to the *PRDM6* locus (discussed before in connection with LVEDSph). PC11 is associated to SNPs rs59894072 (close to *TBX3*, a known cardiac gene⁴²) and rs56229089. The second, in turn, is close (1 Mb) to two possible candidate genes: *KCNJ2*, a potassium channel gene that is active in skeletal muscles and cardiac muscles⁴³ and *SOX9*, a gene implicated in cardiac development⁴⁴. The detailed summary statistics for the significant associations with shape PCs are provided in Supplementary Data.

UPE. CoMAs were trained on LV meshes at end diastole, using a range of network hyperparameters. The reconstruction performance for these models is shown in Supplementary Fig. 1.

GWAS was performed on all latent variables, for all training runs achieving a good reconstruction performance (Methods). A run is an instance of model training, defined by the choice of hyperparameters: in particular, random seeds controlling training and validation samples, weight initialization, network architecture and Kullback–Leibler divergence weight. The number of such runs was $R = 36$. The results obtained with $n_z = 8$ and $n_z = 16$ (8 and 16 latent variables, respectively)

are reported, with a total number of 384 latent variables in the pooled representation. First, we examine the prevalence of significant GWAS loci found in all runs of our ensemble. To count the loci, we split the genome into approximately linkage disequilibrium-independent genomic regions⁴⁵ and computed the number of loci below the usual genome-wide significance threshold of 5×10^{-8} (see details in the Methods section); Table 1 shows the results.

We found 49 independent associations with study-wide significance. All of the previously discussed findings are recovered by UPE with study-wide significance, except the following loci: *MTSS1*, *TBX3*, *PPARGC1A* and *FNDC3B* (the last two show with suggestive significance in UPE). The summary statistics of the GWAS for the best latent variable of each of these 49 loci are displayed in Table 1. When a gene name is displayed in bold letters, it means that this locus was found only via the ensemble approach. Most loci have previous evidence supporting their plausible role in cardiac pathways. In addition, many of them are totally novel and represent interesting avenues for further research.

In what follows, we perform an in-depth analysis of our novel genetic findings in the light of recent literature.

Loci with previous evidence. We now describe loci that have not been linked to structural LV phenotypes in recent GWAS, but count with other types of evidence.

rs11706187 is probably linked to developmental gene *SHOX2*. The mouse homologue of *SHOX2*, *Shox2*, is essential to differentiate cardiac pacemaker cells by repressing *Nkx2-5* (ref. 46). Whereas both *TBX5* and *NKX2-5* are highly expressed in adult cardiac tissues according to GTEx data, *SHOX2* is not highly expressed in these tissues. A possible hypothesis is that rs11706187 regulates the expression of *SHOX2* in developmental or pre-adult stages.

A particularly interesting association, with the SNP rs2245109, is located within the body of the *STRN* gene on chromosome 2 and is probably causally related to it: this gene encodes the protein striatin, which is expressed in cardiomyocytes and has been shown to interact with other proteins involved in the mechanism of myocardial function⁴⁷. Mutations in this gene have been shown to lead to DCM in dogs⁴⁸. In humans, there has been a recent GWAS on heart failure that reported this locus, but our study links it with cardiac morphology. Moreover, our estimated effect size is substantially higher; suggesting that this latent variable is an endophenotype closer to the underlying biology. This could provide insight to unravel the aetiology of a heterogeneous condition such as heart failure. The lead SNP has a high minor allele frequency (MAF) of 47.4%. This locus also contains eQTLs for this gene, as evidenced by TWAS (section 'TWAS'). Something similar occurs with the *RNF11* locus, although this does not reach genome-wide significance for heart failure ($P = 3.2 \times 10^{-6}$). The lead variant for this locus is an indel with low frequency (MAF 1.4%) and large estimated standardized effect size ($\beta = 0.138$). This locus has also been linked to the QRS (a combination of the Q, R and S waves) interval, although the causative gene is not clear⁴⁹, some candidates being *RNF11* itself, *CDKN2C*, *C1orf185* and *FAF1*.

The *SRL* gene, which encodes the sarcalumenin protein, harbours the SNP rs889807. Sarcalumenin is a protein that binds Ca^{2+} located in the longitudinal sarcoplasmic reticulum of the heart. Its main function is to regulate Ca^{2+} reuptake in the sarcoplasmic reticulum by interacting with the cardiac sarco (endo)plasmic reticulum Ca^{2+} -ATPase 2a (SERCA2a). According to GTEx data, this gene is highly expressed in adult cardiac tissue (both in the LV and atrial appendage tissues) and skeletal muscle.

Several associations lie near genes of the *ADATMS* (a disintegrin and metalloproteinase with thrombospondin motifs) family⁵⁰: *ADAMTS1* and *ADAMTS5* (near rs2830977 on chromosome 21, with $P = 1.4 \times 10^{-10}$), *ADAMTS6* (rs753963943 on chromosome 5, $P = 5.6 \times 10^{-11}$) and *ADAMTS18* (chromosome 16, $P = 5.2 \times 10^{-13}$).

An association lies 260 kb upstream of *GATA6*, a transcription factor that plays a critical role in the development of the heart. It has been found to regulate the hypertrophic response⁵¹. Sequence variants in this gene have been discovered to predispose for CHD phenotypes^{52,53}.

rs12889267 lies 3,700 kb upstream of the TSS of *NDRG2*. This gene has been demonstrated to play a role in protection against ischaemia and/or reperfusion injury, in a study in rats⁵⁴.

One SNP overlaps *KDM2A*. As *KDM1A*, it is a histone demethylase gene. Although its link to the heart is less clear, there exists evidence from knockout studies in mice that supports its importance in embryonic development, including heart development⁵⁵.

rs206524 is located within a gene for long non-coding RNA, *LINCO1254*. A possible candidate protein-coding gene is *NDUVF2*, located 1.3 Mb upstream of the SNP. According to the GTEx dataset, *NDUVF2* is highly expressed in cardiac and skeletal muscle tissue.

rs12046416 is located 8,268 bp upstream of the TSS of *GJA5*, a gene that is expressed in atrial myocytes and mediates the coordinated electrical activation of the atria⁵⁶.

Novel loci. In addition to the loci with previous evidence discussed above, we report a number of novel genetic loci with $P < P_{sw}$, which have not been previously reported in connection with cardiac phenotypes or pathways. These loci were annotated on the basis of the closest gene: *CCDC91*, *FILIP1L*, *EN1*, *AFAP1*, *IGFBP3*, *CCDC34*, *WASF3*, *DOCK9* and *MAF*. Of particular interest are those loci with a small number of counts, for example $c_{\ell} \leq 15$. These are the loci for which the ensemble approach seems most relevant, since they are unlikely to be pinpointed by one particular run. Furthermore, they are typically not found by testing the shape PCs, as evidenced by the higher frequency of bold letters towards the bottom of Table 1.

Loci with suggestive significance. In addition to genetic loci with $P < P_{sw}$, several SNPs show $P_{sw} < P < P_{cw}$ in five or more independent runs. We consider these associations suggestive and briefly discuss some of them here. The summary statistics for these associations are shown in Supplementary Table 3. Some of these loci have been found in previous studies: GWAS studies, familial studies or studies with model organisms. For example, variants in gene *RBM20* are associated to DCM⁵⁷. We observe that the lead SNP in this region has a low MAF (1.4%), and the effect size estimate is high (standardized $\beta = 0.20$).

A cluster of associations in chromosome 1 is located in a region that includes the *S100* family of genes. In particular, the lead SNP in this region, rs985242, is located within the genes *S100A1* and *S100A13*. The *S100* is a family of low-weight Ca^{2+} -binding EF-hand proteins, with 25 human genes identified.

The SNP rs28681517 lies within gene *ADAMTSL3*, whose associated protein has been shown to play a crucial role in maintaining cardiac structure and function in mice⁵⁸.

SNP rs569550 lies 578,846 base pairs away from *KCNQ1*, which belongs to a large family of genes that provide instructions for making potassium channels. *KCNQ1* encodes the alpha subunit of the potassium channel KvLQT1. Mutations in *KCNQ1* are responsible for the long QT syndrome⁵⁹.

Deletion 15:48690566_TC_T is a relatively common variant (MAF 14.4%), and is located 10 kb downstream of the transcription end site of *FBN1*. Mutations in this gene are associated with Marfan syndrome, a genetic disorder that affects connective tissues in the body. It can have various manifestations, including cardiovascular complications.

rs9814240 is a coding variant in the *LMCD1* gene. Mutations in this gene are causative of hypertrophic cardiomyopathy in mice⁶⁰, however, no association had been found between variants in this gene and human cardiac phenotypes. Moreover, this gene has been found to interact with (the homologous of) *GATA6* in mice⁶¹. *GATA6* is located near one of the loci discovered with study-wide significance.

Table 1 | Counts of GWAS hits across runs in the UPE framework, C_ℓ for each locus ℓ , which represents the number of runs for which the corresponding locus shows at least one association with $P < P_{GW} = 5 \times 10^{-8}$ (see details in the Methods section)

Chromosome	Region	Candidate gene	Count	Minimum P value	Lead variant	NEA/EA	EAF (%)	$ \beta \pm se(\beta) (\times 10^{-2})$
10	120591353-122407323	BAG3	35	4.1×10^{-18}	rs375034445	A/AT	21.2	5.29 ± 0.79
2	178553183-181312739	TTN	35	1.4×10^{-17}	rs2042995	T/C	23.2	5.70 ± 0.77
6	117672972-118963115	PLN	35	2.0×10^{-29}	rs11153730	T/C	48.6	4.29 ± 0.64
14	23018665-24905123	MYH6	34	2.7×10^{-14}	rs365990	A/G	36.9	4.04 ± 0.66
12	113986709-115036602	TBX5	34	2.3×10^{-11}	rs4767239	G/C	82.5	4.12 ± 0.85
12	110336719-113263518	MYL2	34	2.8×10^{-15}	rs35350651	A/AC	51.4	4.36 ± 0.64
4	119933512-120392684	MYOZ2	33	2.4×10^{-13}	4:120304290_GC_G	GC/G	29.0	3.96 ± 0.71
10	73508512-75422550	SYNPO2L	31	2.5×10^{-15}	rs3740293	A/C	14.3	5.07 ± 0.92
3	157312028-159477890	SHOX2	30	7.0×10^{-15}	rs11706187	A/G	50.1	3.23 ± 0.64
1	154770403-156336133	FDPS	29	9.7×10^{-13}	rs41314549	T/C	2.81	13.7 ± 1.9
17	43056905-45876022	GOSR2	26	8.3×10^{-22}	rs17608766	T/C	14.3	5.73 ± 0.90
3	99373762-100592217	FILIP1L^a	25	8.4×10^{-14}	rs9811920	G/A	40.8	3.23 ± 0.65
16	4001196-5118345	SRL	24	9.4×10^{-12}	rs889807	T/C	50.8	3.24 ± 0.64
1	21736588-23086883	KDM1A	22	2.0×10^{-12}	rs35001652	G/A	37.3	2.59 ± 0.67
7	45952922-46986720	IGFBP3	22	2.0×10^{-11}	rs143741275	A/AGTGTGT	42.4	2.59 ± 0.66
5	171074292-172678327	NKX2-5	21	9.0×10^{-14}	rs35564079	C/CT	28.5	2.76 ± 0.72
3	13070799-14816900	TMEM43	21	2.7×10^{-11}	rs900173	T/C	34.0	3.48 ± 0.68
1	144977494-148361253	GJA5	20	1.1×10^{-10}	rs12046416	A/G	33.5	2.81 ± 0.68
1	235819436-237555628	ACTN2	20	2.0×10^{-12}	rs12142143	T/C	53.1	4.11 ± 0.65
13	20686720-22242174	FGF9	20	8.8×10^{-13}	rs10628955	G/GAA	47.7	3.95 ± 0.67
4	111256567-113870102	PITX2	19	4.4×10^{-14}	rs2723294	C/T	69.5	4.65 ± 0.70
1	14891511-16897730	HSPB7	17	2.3×10^{-11}	rs1763605	T/G	67.5	2.80 ± 0.68
2	146445570-147277162	ACVR2A/ZEB2	16	2.9×10^{-11}	rs573709385	A/AT	44.9	1.93 ± 0.65
6	125424383-127540461	HEY2	16	2.9×10^{-11}	rs11423823	C/CT	50.2	3.23 ± 0.66
2	36122006-38132712	STRN	15	9.9×10^{-16}	rs2110944	T/C	52.6	3.71 ± 0.64
16	79134815-80297374	MAF^a	15	5.2×10^{-11}	rs558328129	A/AT	45.4	2.97 ± 0.70
8	107410754-108648177	ABRA	15	8.1×10^{-11}	rs72007904	A/AACTATTC	50.0	3.87 ± 0.64
21	27271019-29125226	ADAMTS1	15	1.4×10^{-10}	rs2830977	G/A	22.0	4.29 ± 0.78
1	49894177-51713726	RNF11^a	13	3.1×10^{-11}	rs7555411	C/T	1.42	13.8 ± 2.7
13	98938919-100574095	DOCK9^a	13	1.2×10^{-10}	rs34138434	C/A	29.7	4.62 ± 0.72
2	118367466-121303783	EN1 ^a	13	6.7×10^{-14}	rs162746	A/G	67.5	4.01 ± 0.68
12	27799773-29651255	CCDC91^a	12	2.0×10^{-14}	rs5797270	G/GT	20.3	3.35 ± 0.81
13	25784362-27284362	WASF3 ^a	11	9.3×10^{-11}	rs61944841	G/A	41.3	3.07 ± 0.67
18	19485844-20649472	GATA6	11	6.6×10^{-11}	rs62094198	T/A	39.6	2.21 ± 0.66
14	19002084-21589402	NDRG2	10	2.5×10^{-11}	rs12889267	A/G	16.7	3.43 ± 0.85
6	35455756-37572596	CDKN1A	10	3.2×10^{-13}	rs3176326	G/A	19.8	3.47 ± 0.81
7	118351581-121045273	WNT16	10	2.4×10^{-11}	rs3801387	A/G	28.1	2.79 ± 0.72
16	75977954-77523678	ADAMTS18	10	5.2×10^{-13}	rs62046468	C/T	37.6	4.80 ± 0.66
17	41772087-43056905	SOST	10	5.5×10^{-11}	rs17881550	G/GC	43.4	3.65 ± 0.65
11	27020461-28481593	CCDC34^a	8	5.7×10^{-12}	rs10835164	C/T	25.6	3.84 ± 0.74
5	120452166-122556905	PRDM6	8	6.2×10^{-11}	rs463106	T/C	47.2	3.68 ± 0.65
17	67858770-69387817	KCNJ2/SOX9	7	3.6×10^{-12}	rs56229089	G/C	55.7	2.89 ± 0.65
5	63968304-65911286	ADAMTS6	7	1.6×10^{-11}	rs753963943	ATT/A	42.5	2.88 ± 0.66
18	8498931-11075913	NDUFV2	6	6.1×10^{-12}	rs206524	T/C	70.7	3.01 ± 0.71
11	65898631-68005825	KDM2A	4	2.1×10^{-11}	rs12785906	G/C	5.84	7.40 ± 1.38
1	1892607-3582736	PRDM16	3	1.7×10^{-11}	rs781212641	G/GC	9.33	5.68 ± 1.12
4	7539692-8152235	AFAP1	3	1.4×10^{-11}	rs28542374	G/A	63.5	3.10 ± 0.67
12	76511314-78570570	NAV3	3	2.7×10^{-12}	rs7965680	T/C	55.7	2.43 ± 0.65
17	15019097-16412342	CENPV^a	3	3.4×10^{-11}	rs7477	A/C	49.8	3.18 ± 0.64

The total number of runs was 36. The lead variant is the one for which the minimum P value occurs. P values are two-sided and derived from a linear association t-statistic (no adjustments were made for multiple comparisons). NEA and EA stand for non-effect and effect allele, respectively, whereas β is the standardized effect size estimate. EAF is the frequency of EA. Note that these values correspond to different phenotypes of the ensemble, therefore they are not comparable in terms of the magnitude of the morphological change that they produce. The directions of effect can be understood with the help of Supplementary Table 5. ^aThe genes were annotated based purely on closest proximity to the lead variant in that region (the rest of the genes have additional previous evidence of a link to cardiac physiology as discussed in the text). Gene names that are in bold mean that the corresponding locus is only discovered with UPE, that is the remaining loci were already found by testing the handcrafted phenotypes or the shape PCs.

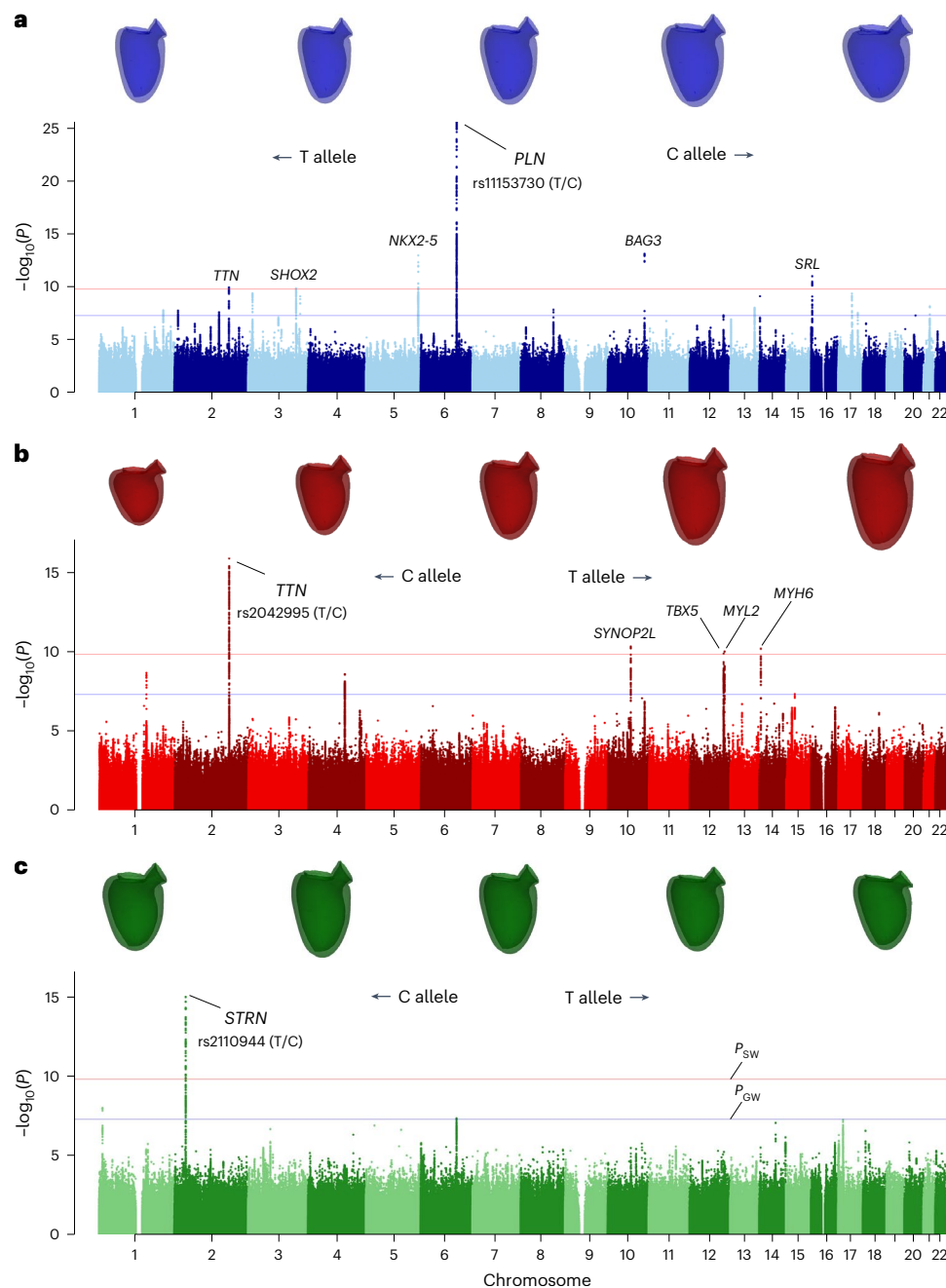


Fig. 2 | Variants in the *TTN*, *PLN* and *STRN* loci show distinct effects on LV morphology. **a–c**, Manhattan plots for LV latent variables with best association for SNPs at the *PLN* (**a**), *TTN* (**b**) and *STRN* (**c**) loci. On top are shown the average

meshes corresponding to the following range of quantiles, for each latent variable (from left to right): [0, 0.01], [0.095, 0.105], [0.495, 0.505], [0.895, 0.905] and [0.99, 1].

Effect on LV morphology

The effect of these loci on the LV morphology was evaluated by selecting the single phenotype with the strongest P value for the associated locus. To help characterize these latent variables, the Spearman correlation coefficient between the latter and the handcrafted LV indices were calculated and shown in Supplementary Table 4. We also examine the shapes of the average mesh within different ranges of quantiles for this latent variable, from 0 through 1. This is shown in Fig. 2, along with the associated Manhattan plots, for the loci *PLN*, *TTN* and *STRN*. The direction of effect is shown by indicating with arrows which allele favours which shape. We observe a very distinct effect on the morphology of each of these SNPs. While the *PLN* variant influences a latent variable that has a smaller effect on LVEDV (Spearman $r = 0.722$)

and a strong link to LVEDSph ($r = 0.532$), the best latent variable for *TTN* gene shows a greater correlation with LVEDV ($r = 0.910$). Consistent with this, the GWAS on LVEDSph shows no significant signal for *TTN*, but a strong one for *PLN* ($P = 10^{-20}$, Extended Data Fig. 2), which is also in line with a previous finding of ours¹⁰. Furthermore, these findings are in line with the effects of PC1 and PC3, where *TTN* and *PLN* loci are found, respectively.

The SNP in the *STRN* gene is associated with a subtle phenotype that controls mitral orientation without a concomitant change in LV size (Fig. 2). This is consistent with the fact that it was not discovered in previous studies of structural LV phenotypes. Notably, this effect is consistent with the observed effect of PC4, for which this locus reaches genome-wide significance (see Supplementary Fig. 2 for the effect of PC4).

TWAS

We performed TWAS using the S-PrediXcan tool¹⁶, to test the possibility of a mediating effect of gene expression and intron excision events on structural phenotypes. This tool is fed with models that impute gene expression and intron excision data on the basis of the genotype, which in turn were trained using data from the GTEx project, v.8 (ref. 62).

Our focus was on cardiovascular tissues, specifically the LV, atrial appendage and coronary, aortic and tibial arteries. To maintain statistical rigour, we applied a significance threshold of $P_{\text{GEx}} = 2.2 \times 10^{-9}$, which adjusts for multiple comparisons (324 phenotypes and 68,919 tissue–gene pairs). Similarly, for alternative splicing, the threshold was set at $P_{\text{AS}} = 8.2 \times 10^{-10}$, considering the same multiple testing correction (187,535 being the number of intron–tissue pairs tested).

In the cardiac tissues (LV and atrial appendage), we identified genes located within loci of previously reported genes. In the LV, these included *NKX2-5*, *STRN*, *SYNPO2L* (*FUT11*, *SEC24C* and *SYNPO2L* itself), *PLN*, *HEY2* (*CENPW* gene), *TTN* (*FKBP7* gene), *CENPV*, *GOSR2* (*MAPT* and *GOSR2* itself) and *FDPS* (*SCAMP3*, *ARHGEF2*, *RIT1*, *GOSR2*, *MAPT*, *HCN3*, *GBA*, *MSTO1*, *RUSC1*, *FUT11*, *SYT11*, *ADAM15* and *FDPS* itself). For the atrial appendage, the genes included *PLN*, *STRN*, *NKX2-5*, *SYNPO2L* and *MYOZ1* within the *SYNPO2L* locus, as well as *FKBP7* and *SCAMP3*. Many of these genes had been previously implicated on the basis of independent knowledge, bolstering the evidence for their potential causal roles. Notably, our analysis also revealed the direction of the effect on gene expression: higher *PLN* expression was associated with a more spherical LV morphology, while lower *NKX2-5* expression was linked to the same phenotype (refer to Fig. 2b). Furthermore, an elevated *STRN* expression (in both cardiac tissues) was associated with a more horizontal mitral orientation (Fig. 2c). Detailed results for significant gene expression associations are provided as Supplementary Data.

In the case of arterial tissues, we found significant associations within various loci, such as the *SYNPO2L* locus (with the genes *AGAP5*, *FUT11*, *SEC24C* and *ARHGAP27*), *FDPS* (*ARHGEF2*, *CLK2*, *FAM189B*, *GBA*, *GON4L*, *HCN3*, *NPR1* and *SYT11*), *CENPW*, *TTN* (*PRKRA* and *FKBP7* genes), *PLN* (*CEP85L* and *PLN*), *GOSR2* (*WNT3*, *CRHR1*, *LRRC37A* and *MAPT*), *KDM2A*, *LINC01562*, *MYH6* (*MYH6* and *MYH7*), *RP11-383123.2*, *RP11-574K11.29*, *SCAMP3*, *MYL2* (*SH2B3* gene), *SOST* and *TCF21*.

Detailed results for intron excision events are provided in Supplementary Data.

Gene ontology enrichment analysis

We use the tool g:Profiler to find pathways for which our sets of genes were enriched. To define the gene sets, we selected a region of 100 kb around each lead variant and chose the genes whose TSS was located within that window. Gene ontology terms belong to one of three different categories: molecular functions, cellular components and biological processes. Within the cellular component category, we have found a relevant enriched term, ‘Sarcomere’, comprising the following nine genes from our query: *ACTN2*, *MYOZ1*, *SYNPO2L*, *BAG3*, *TNNT3*, *TNNI2*, *MYH6*, *MYH7*, *KY* ($P = 9.2 \times 10^{-3}$). Within the biological process category, the terms ‘Myofibril assembly’, ‘striated muscle cell development’ and ‘sarcomere organization’ result enriched ($P = 1.2 \times 10^{-3}$, $P = 1.4 \times 10^{-3}$ and $P = 1.5 \times 10^{-3}$, respectively). Within the molecular function category, the term ‘calcium-dependent protein binding’ is enriched ($P = 2.9 \times 10^{-8}$), although it is composed of nine members of the S100A family (which encompass a single locus), apart from *SYT8* and *TNNT3*.

Phenome-wide association studies

To detect pleiotropic effects, we performed a phenome-wide association study of the lead SNPs from Table 1. For this, we queried the Integrative Epidemiology Unit OpenGWAS Project’s database. The results are included in the Supplementary Data File. We discuss briefly here some associations with cardiovascular phenotypes. A number of loci were associated to cardiac electrical phenotypes: *CDKN1A*, *NDRG2*, *PLN*, *TBX5* and *MYH6*. The following loci were associated to pulse rate: *SYNPO2L*,

NDRG2, *MYH6*, *SRL*, *GOSR2*, *GATA6*, *ACTN2*, *KIAA1755*, *TMEM43*, *SLC27A6* and *FNDC3B*. The lead SNP at the *PRDM6* locus was associated to heart rate recovery post exercise. The following loci were associated to blood pressure phenotypes (diastolic, systolic or hypertension): *SYNPO2L*, *KCNQ1*, *MYL2*, *NDRG2*, *MYH6*, *SRL*, *GOSR2*, *GATA6*, *HSPB7*, *RNF11*, *EFEMP1*, *FNDC3B*, *NME9*, *PRDM6* and *PLN*. Finally, *SYNPO2L*, *TBX5*, *MYH6*, *GOSR2*, *PITX2* and *CDKN1A* were associated to cardiac arrhythmias.

Replication study

We set apart a subset of 5,470 UKBB participants of British ancestry for which the whole pipeline was run identically to the individuals from the discovery set. We report the detailed results in the Supplementary Material, including the estimated statistical power for each SNP on the basis of the effect size estimate $\hat{\beta}$ from the discovery phase. Among the 49 study-wide significant loci, we report 28 that replicate with $P < 0.05$ (whereas seven replicate with the more stringent Bonferroni threshold of $P < 0.05/49$), as well as 47 loci for which the estimated direction of effect is consistent with that found in the discovery phase. For the suggestive associations, 11 loci replicated (out of 25) with the threshold of $P < 0.05$, whereas 22 have a concordant direction of effect between the discovery and replication phases.

Comparison with GWAS on traditional LV indices

For comparison, we collected the GWAS summary statistics from previous studies on LV phenotypes, derived also from UKBB CMR images, namely refs. 2,4 and 11. We also include the results for LVESV, SV and LVEF from these studies. However, note that the unsupervised features studied in this work are static and were extracted using only the end-diastolic phase.

The comparison can be seen in Fig. 3. For each locus in Table 1 (which all pass the Bonferroni threshold), this figure displays the association P value found in previous GWAS and on our own GWAS of handcrafted phenotypes. Shades of red represent non-genome-wide significant associations, whereas shades of blue represent genome-wide significant ones and white corresponds to the P_{GW} threshold. The second column represents the best P value across all traditional phenotypes for the loci given in the columns. Therefore, a shade of red in this column means that the locus is novel in the context of LV structural phenotypes.

Discussion

As shown in ‘Results’, we were able to retrieve study-wide significant loci that had been found in previous GWAS on handcrafted phenotypes (*PLN*, *TTN*, *MYL2*, *GOSR2*, *BAG3*, *TMEM43*, *HSPB7*, *CPKN1A*, *NKX2-5*). Furthermore, genes with a known role in cardiac physiology (for example, *TBX5*, *SHOX2* and *STRN*) were identified, but no previous association with GWAS of LV phenotypes had been found in previous studies. Thirteen additional loci constitute potential avenues for future research. Finally, 24 additional independent loci of suggestive significance ($P_{\text{SW}} < P < P_{\text{GW}}$ and $C_e > 5$). Several of these have previous evidence of a link to cardiac pathways, for example *RBM20* and genes from the S100A family.

For some loci, a relatively small number of runs produced a latent variable with a genome-wide significant association to the locus: the UPE approach seems crucial for pinpointing this association, as it is likely to be missed in one individual autoencoder run. Also, they are typically missed by shape PCA or handcrafted phenotypes (Table 1). Our approach allows us to detect the milder effect on morphology of common variants near genes whose mutations are known to have highly deleterious effects, either by study of Mendelian diseases in humans or by studies on model organisms. One example of the first is the suggestive association near *FBNI*. It is likely that these variants and the associated unsupervised LV features hold prognostic value; however, this is uncertain at this point, and it should be possible to assess it once UKBB releases more longitudinal data on the same participants studied here.

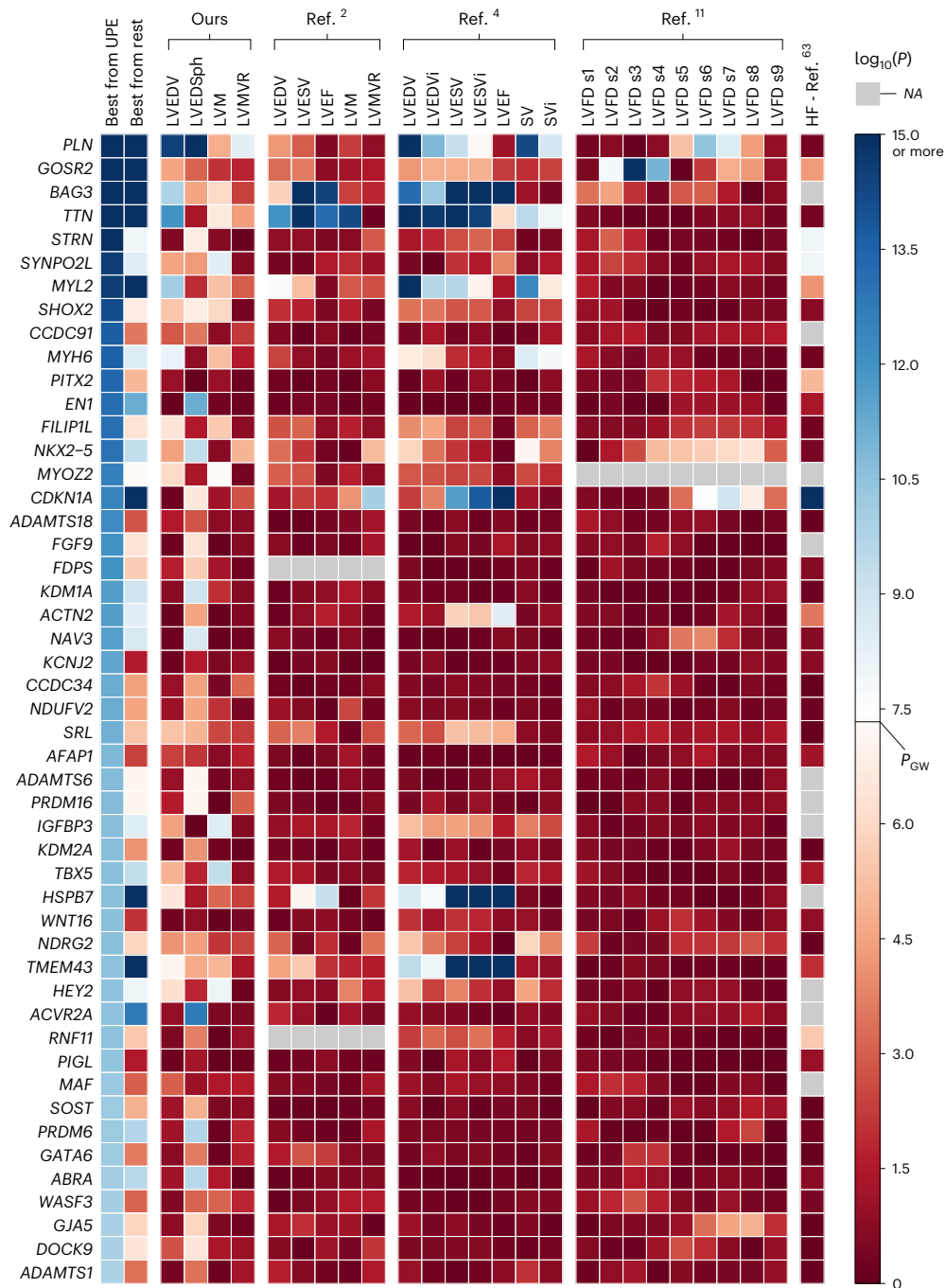


Fig. 3 | Comparison of the $-\log_{10}(P)$ values for the lead variants of the 49 study-wide significant genetic loci found in this work, with GWAS on handcrafted cardiac indices and a GWAS on heart failure. The leftmost column corresponds to the best association found for that locus across the ensemble of phenotypes, whereas the second column corresponds to the best P value for that locus across the previous GWAS, where the P values are two-sided and derived from a linear association t -statistic (no adjustments were made for multiple comparisons). The white colour corresponds to the genome-wide significance

threshold of 5×10^{-8} , whereas the shades of red and blue correspond to weaker and stronger associations, respectively. SV denotes stroke volume. LVEDVi, LVESVi and SVi denote the indexed versions of the phenotypes, that is, the phenotype divided by the participant’s body surface area. Finally, LVFD s_n stands for LV trabecular fractal dimension measured at the n th slice of the LV longitudinal axis (for details, refer to the original publication in ref. 11). Grey squares (NA, not applicable) mean that the genetic variant was not tested in the corresponding study.

The SNP rs2245109 is located within the body of the *STRN* gene, on chromosome 2, and is most probably causally related to it. This gene codes for the protein striatin, which is expressed in cardiomyocytes and has been shown to interact with other proteins that affect the mechanism of myocardial function⁴⁷. Mutations in this gene have been shown to lead to DCM in dogs⁴⁸. A recent GWAS on heart failure reported this locus in humans, and our study links it to cardiac morphology. Furthermore, the estimated effect size that we find is notably higher than that

for heart failure; this suggests that this latent variable is an endophenotype closer to the underlying biology. This could provide insight to unravel the aetiology of a heterogeneous condition such as heart failure. Furthermore, it makes *STRN* a promising therapeutic target.

As an interesting observation, we note that the phenotypes extracted by UPE and shape PCA show a remarkable oligogenicity, that is, they are controlled by few genes (Extended Data Fig. 5) for shape PCA and Supplementary Figures through for UPE). This is in

contrast to what is observed for heterogeneous conditions such as heart failure. For example, heart failure (a single phenotypic score) is linked to 47 loci with genome-wide level of significance⁶³. However, a much larger sample size is needed to detect them; indeed, note that this GWAS involves more than 110,000 cases and 1.5 million controls (compared to almost 49,000 participants in our study). Our results confirm (1) the view that endophenotypes are better suited for detecting risk genes for higher-level phenotypes (such as heart failure), due to their higher oligogenicity and stronger link to causal genes (that is, higher effect size) and (2) that the use of unsupervised phenotypes, and in particular the UPE approach, allows one to identify more optimal endophenotypes for each genetic locus, as compared to traditional handcrafted phenotyping approaches, thus boosting discoverability.

In terms of gene discovery, the advantages of an unsupervised phenotyping approach are best conveyed by examining the associated *P* values of the loci found in GWAS performed against traditional handcrafted phenotypes, shown in Fig. 3. For example, when examining the *GOSR2* locus, we found no genome-wide significant association when performing GWAS on traditional LV indices derived from the same meshes; neither have previous studies, except for ref. 11 that investigated the trabecular fractal dimension of LV. However, we were able to find it linked to shape PC2, which seems to model LV conicity. Similarly, the UPE approach finds it in 26 (out of 36) runs, where the best latent variable models a similar phenotype (Supplementary Fig. 6). Other examples of novel associations found via shape PCA and UPE are *ACTN2*, *PITX2*, *NAV3* and *PRDM6*.

Likewise, other genes, such as *STRN*, which have previous knowledge of being implicated in cardiac pathways, have not been reported to date in mostly healthy cohorts such as UKBB. It reaches a strong *P* value ($P = 9.9 \times 10^{-16}$) in our UPE approach, but with shape PCA it only reaches genome-wide significance for PC4, whereas no significant signal is detected for traditional phenotypes. Other examples of highly plausible genes that are found only via UPE are *SHOX2*, *SRL*, *KDM2A*, *NRDG2* and four genes from the *ADAMTS* family.

Some other loci have little evidence to the best of our knowledge, and represent interesting avenues for further research. Examples are the loci near genes *CCDC91*, *FILIP1L* and *CCDC34*, which are of study-wide significance in our approach; however, they have not been reported in previous GWAS on LV phenotypes (that is, all remaining squares are coloured in red shades). Similarly, they are not captured by shape PCA. This highlights the shortcomings of traditional image-derived phenotyping techniques when it comes to the discoverability of relevant genes.

In addition to improved discoverability, the UPE framework enables a more refined understanding of the genetic architecture of cardiac phenotypes, even for genetic loci that were known from previous studies. Most notably, the top SNP in the *TTN* locus was shown to be distinctly related to the size of the LV, while the *PLN* variant (which has been previously found in GWAS of LVEDV) controls a feature that jointly models changes in the size and sphericity of the LV. The *STRN* locus is most strongly associated with a subtle feature that controls mitral orientation and was therefore not discovered in previous studies, which investigated more global phenotypes.

On the basis of our findings, we argue that, in large-scale imaging studies, it is crucial, along with increasing sample size, to count with good techniques to perform deep phenotyping that allow to boost gene discoverability in GWAS.

Conclusions

In this work, we proposed a framework for LV phenotyping based on unsupervised geometric deep learning techniques in image-derived 3D meshes to discover genetic variations that affect the shape of the LV through GWAS. The proposed methodology is based on finding a latent low-dimensional representation of the CMR-derived LV 3D meshes using CoMAs and then performing GWAS on the learnt latent

features. As proposed, this dimensionality reduction method, using Kullback–Leibler regularization, yielded phenotypes with statistically significant genetic associations.

The methodology of ensembling SNP associations across representations obtained through different network metaparameters, followed by the correction in the Bonferroni threshold necessary to control for false discovery rate, has proven effective in identifying novel associations of mesh-derived phenotypes with genetic loci. In addition to previously identified loci, namely *TTN*, *PLN*, *GOSR2* and *ATXN2*, we report 40 additional genetic loci that have not been discovered in recent GWAS of LV phenotypes. Moreover, we report 24 independent associations that do not exceed our corrected Bonferroni threshold; however, their association remains suggestive by virtue of exceeding the usual genome-wide significance threshold of $P_{\text{GW}} = 5 \times 10^{-8}$ in more than five unsupervised phenotypes, obtained from independently trained autoencoder networks. Some of the last genes, such as *SIOOAI*, *LMCD1*, *RBM2O* and *FBN1*, have been previously linked to cardiac pathways.

We argue that the proposed assembly approach is not only useful for discovering novel associations but also enables a deeper understanding of the effect of previously known genes: in fact, the effect of the latent variables with the strongest associations *P* values for each locus can be used as suggestive evidence of the role of that locus in LV shape. For example, we found that the *TTN* and *PLN* variants, which had been previously found to correlate with LV volume, actually have a distinct effect on the shape of the LV. Whereas the *TTN* variant shows in fact a clear effect on LV size, the *PLN* variant is linked to a more complex phenotype that involves a concomitant change in LV volume and sphericity.

More generally, these results validate our methodology to extract knowledge about the genetics driving the morphology of organs, leveraging databases that provide linked genetic and imaging data, such as the UKBB. This methodology can be used seamlessly to study surface meshes of other organs, such as the brain or the skull^{64,65}. Additionally, the algorithm proposed here can be extended to process 3D cardiac meshes throughout the cardiac cycle to capture anatomy and quantitative features related to contraction and relaxation patterns. Future studies will explore these directions.

Methods

The proposed method is outlined in Fig. 1. It starts with extracting 3D meshes representing LV from CMR images using an automatic segmentation method¹². We then train several models with different metaparameters (network architecture, random seeds controlling weight initialization and dataset partitioning, and relative weight of the variational loss) to learn low-dimensional representations of the 3D meshes that capture anatomical variations using an encoder–decoder model. All meshes are then projected to this latent space to derive a few shape descriptors (or latent variables) for each mesh. To take advantage of the variability induced in the representation obtained by the metaparameters, we pooled the different latent vectors together to obtain a richer representation. The features that make up this pooled representation are finally used in GWAS to discover genetic variants associated with shape patterns.

Description of the data

The proposed framework can discover novel associations between genetic variations and morphological changes in anatomical structures. We present its potential in the context of cardiac images acquired within the UKBB project (data accession number 11350). The UKBB is a prospective cohort study that between 2006 and 2010 recruited around half a million volunteers in the United Kingdom, aged 40 to 69 years at the time of recruitment⁵. The project collected vast phenotypic information about its participants and linked them to their electronic health records. The collected data includes, among

others, genetic data from SNP microarrays for all the individuals and also CMR data for a subset of them (which comprises more than 50,000 individuals at the moment of this writing, but is planned to reach 100,000). These datasets are described in refs. 66 and 67, respectively.

CMR data. The CMR imaging protocol used to obtain the raw imaging data is described in ref. 67. We used an automatic segmentation method¹² to segment the LV in the CMR images. This method generates a set of registered 3D meshes: that is, meshes with the same number of vertices with consistent identical connectivity between them. There is one mesh per participant and per time point. In this work, we only use the LV mesh at end diastole. The LV mesh for the participant i , $i = 1, \dots, N$, can then be represented as pairs (S_i, A) , where $S_i = [x_{i1}, y_{i1}, z_{i1} | \dots | x_{iM}, y_{iM}, z_{iM}] \in \mathbb{R}^{M \times 3}$ is the shape and A is the mesh adjacency matrix. The adjacency matrix is such that $A_{jk} = 1$ if and only if there is an edge between vertices j and k and $A_{jk} = 0$ otherwise. The cardiac meshes also have the property of being triangular and closed, so $A_{jk} = A_{kl} = 1 \Rightarrow A_{jl} = 1$ for all vertices j, k and l .

Genotype data. SNP microarray data are available for all individuals in the UKBB cohort. This microarray covers 801,526, genetic variants that include SNPs and short insertions and deletions. The SNP microarrays used in UKBB have been described in ref. 66. An augmented set of more than 90 million variants was imputed from these genotyped markers. GWAS was performed on the latter dataset, particularly on autosomes (chromosomes 1 to 22). The usual quality control steps on the genetic data were performed. This included filtering out rare variants using a threshold for MAF of 1% (within the subcohort of 48,651 participants), a Hardy–Weinberg equilibrium value $P < 10^{-5}$ and a low imputation information score (less than 0.3). This results in a set of 9,472,708 genetic variants.

Unsupervised representation learning for genetic discovery

Given the set of meshes representing the anatomical structure of interest (LV meshes), the pose-sensitive parameters (translation and rotation) were removed using generalized Procrustes analysis. Here we propose to learn a reduced set of features that best describe cardiac shape using CoMA. We will compare the proposed approach with the well-known PCA method. While in PCA only vectorized 3D point clouds s_i will be provided as input (therefore ignoring the data structure and topology), CoMAs leverage topological information about the connectivity between the vertices for learning more powerful nonlinear representations. However, both approaches can be thought of as particular cases of the encoder–decoder paradigm.

In such a model, there is a pair of encoding and decoding functions, $E_\theta : \mathbb{R}^{3M} \rightarrow \mathbb{R}^{n_z}$ and $D_\phi : \mathbb{R}^{n_z} \rightarrow \mathbb{R}^{3M}$ that are parameterized by a set of learnable coefficients θ and ϕ , respectively. $n_z \in \mathbb{N}$ is the size of the latent space, and is usually chosen so that $n_z \ll M$ (hence the reduction in dimension).

Optimal parameters θ^* and ϕ^* for reconstruction can be estimated by making the composite function $D_{\phi^*} \circ E_{\theta^*}$ as close to the identity function I as possible over the training set $\mathcal{S}_{\text{train}} \subset \mathcal{S}$, using some reasonable measure of reconstruction error L_{rec} (examples of which are the norm L_1 norm, the norm L_2 or the mean squared error) along with a regularization term Ω , which will account for additional constraints we want to impose on the model. We want to minimize the following function with respect to ϕ and θ :

$$L(\mathcal{S}_{\text{train}} | \theta, \phi) = L_{\text{rec}}(\mathcal{S}_{\text{train}} | \theta, \phi) + \beta \Omega(\mathcal{S}_{\text{train}} | \theta, \phi). \tag{1}$$

where $\beta \in \mathbb{R}_{\geq 0}$ is a weighting coefficient for the regularization term. $\mathbf{z}_i := E_{\theta^*}(S_i) \in \mathbb{R}^{n_z}$ would then be a low-dimensional representation of the shape S_i , while $\hat{S}_i := (D_{\phi^*} \circ E_{\theta^*})(S_i)$ is the associated reconstructed shape.

PCA. PCA is a standard linear technique for reducing the dimensionality⁶⁸. In terms of the encoder–decoder framework detailed above, it can be obtained by requiring D and E to be linear transformations and using the norm L_2 , in addition to imposing an orthogonality constraint on the latent vectors⁶⁹.

The idea is to find a basis of vectors $\mathcal{B}_{n_z} = \{\mathbf{v}_i\}_{i=1}^{n_z} \subset \mathbb{R}^M$ for a fixed $n_z < M$, such that the linear subspace generated by \mathcal{B}_{n_z} captures as much variability in the data as possible. It can be shown that this basis corresponds to the n_z eigenvectors of the covariance matrix of the data, C , with the largest n_z eigenvalues; that is, if $C = U^t \Lambda U$ where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots)$, that is it is a diagonal matrix composed of the eigenvalues ordered such that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{n_z}$ (all of which are necessarily non-negative). This technique can be used to reduce the dimensionality of shapes or, more generally, point clouds where the vertices are in correspondences. We define, for convenience, the vectorized form of the shapes, $\mathbf{s}_i = (x_{i1}, y_{i1}, z_{i1}, \dots, x_{iM}, y_{iM}, z_{iM}) \in \mathbb{R}^{3M}$. We refer to this approach as shape PCA throughout the text. Given a set of 3D shapes $\mathcal{S} = \{s_i\}_{i=1}^N$, we derive the mean shape $\bar{\mathbf{s}}$ and the shape covariance matrix C :

$$\bar{\mathbf{s}} = \frac{1}{N} \sum_{i=1}^N \mathbf{s}_i, \tag{2}$$

$$C = \frac{1}{N-1} \sum_{i=1}^N (\mathbf{s}_i - \bar{\mathbf{s}})(\mathbf{s}_i - \bar{\mathbf{s}})^t. \tag{3}$$

In this work, we implemented shape PCA by singular value decomposition of the data matrix (composed of the \mathbf{s}_i vectors), using the Python scikit-learn package.

CoMA. In an autoencoder, both the encoding and decoding functions are feedforward neural networks. Inspired by recent works on unsupervised geometric deep learning⁹ for facial meshes, we propose the construction of a CoMA that uses spectral convolutions⁷⁰ to learn low-dimensional and nonlinear representations of cardiac mesh structures. Here each layer of the encoder and decoder implements convolution operations parameterized by the graph Laplacian, to leverage information about the local context of each vertex. A hierarchical approach is used to learn global features where each layer of the encoder and decoder implements downsampling and upsampling operations, respectively. Since the vertices are not in a rectangular grid, the usual convolution, pooling and unpooling operations defined for such a topology (usually used in image analysis) are inadequate for this task and must be suitably adapted. Several methods have been proposed to do this⁸, which can be mainly classified into two broad groups: spatial or spectral. The approach proposed in this work belongs to the latter category, which relies on expressing the features in the Fourier basis of the graph, as explained below.

Spectral convolutions. The Laplace–Beltrami operator \mathcal{L} (or, more simply, the Laplacian) of a graph with adjacency matrix A is defined as $\mathcal{L} := D - A$, where D is the degree matrix, that is, a diagonal matrix with $D_{ii} := \sum_j A_{ij}$ being the number of edges that connect to the vertex i . The Fourier basis of the graph can be obtained by diagonalizing the Laplace operator, $\mathcal{L} = U^t \Lambda U$. The columns of U constitute the Fourier basis and the operation of convolution \star for a graph can be defined as follows:

$$x \star y := U(U^t x \odot U^t y), \tag{4}$$

where \odot is the element-wise product (also known as the Hadamard product), and x and y , are arbitrary functions defined on the graph’s vertices. Spectral methods rely on this definition of convolution and differ from one another in the specific filter used. This work will use a parameterization proposed in ref. 70. This method is

based on the Chebyshev family of polynomials $\{T_i\}$. The kernel g_ξ is defined as:

$$g_\xi(\mathcal{L}) = \sum_{i=1}^K \xi_i T_i(\mathcal{L}). \tag{5}$$

K is the highest degree of polynomials considered (in this work, $K = 6$). Chebyshev polynomials have the advantage of being computable recursively through the relation $T_i(x) = xT_{i-1}(x) - T_{i-2}(x)$ and the base cases $T_1(x) = 1$ and $T_2(x) = x$. It is also worth mentioning that the filter described by equation (5), despite its spectral formulation, has the characteristic of being local.

Autoencoder. The downsampling and upsampling operations used in this study were proposed in ref. 9 based on a surface simplification algorithm proposed in ref. 71. These operations are defined before training each layer using a single template shape. Here we use the mean shape \bar{s} as a template.

In each encoder layer, the downsampling operation generates a new triangular mesh (with its corresponding new Laplacian) to minimize the quadric error. Upsampling operations are created while downsampling: the coordinates of the decimated vertices with respect to the remaining vertices are stored for each layer.

Variational autoencoder. For some runs, a Kullback–Leibler (KL) divergence term was added to encourage the statistical independence of the different components of the latent representation, which is expected to improve its interpretability⁷². We propose that it will also contribute to producing features with higher heritability, that is, suitable candidate phenotypes on which to perform GWAS.

To train a model with such a loss function, the framework of variational autoencoder is used. In this framework, during the training phase the encoder maps the input into a probability distribution instead of a fixed vector. To emphasize this, we will replace the notation $E_\theta(\mathbf{S})$ for the encoder network with $q_\theta(\mathbf{Z}|\mathbf{S})$, the conditional probability of the (now random) latent variable \mathbf{Z} given the shape \mathbf{S} , also a random variable that represents the shapes. During training, for the j th latent variable (with $1 \leq z_j \leq n_z$) two quantities are learnt, μ_j and σ_j , and a realization z_j of the random variable. $Z_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$ is produced and passed through the decoder to generate the output mesh. The aforementioned Kullback–Leibler-divergence term is then used to encourage the variational approximate posterior to be a multivariate Gaussian with a diagonal covariance structure. The regularization term is computed as:

$$\begin{aligned} \Omega(\mathbb{S}_{\text{train}}|\theta, \phi) &= \mathbb{E}_{\mathbf{s} \sim \hat{p}_{\text{train}}} D_{\text{KL}}(q_\theta(\mathbf{Z}|\mathbf{S})||\mathcal{N}(\mathbf{Z}; \mathbf{0}, \mathbb{1}_{n_z})) \\ &= \mathbb{E}_{\mathbf{s} \sim \hat{p}_{\text{train}}} \frac{-1}{2n_z} \sum_{j=1}^{n_z} (\log \sigma_j^2 - \sigma_j^2 - \mu_j^2 + 1), \end{aligned} \tag{6}$$

where $\mathbb{1}_n$ is the identity matrix $n \times n$, $D_{\text{KL}}(p||q)$ is the Kullback–Leibler divergence between the probability distributions p and q , and \hat{p}_{train} is the empirical probability distribution associated with $\mathbb{S}_{\text{train}}$. $D_{\text{KL}}(p||q) := \int p(x) \ln \frac{p(x)}{q(x)} dp(x)$. The last equality in equation (6) arises from the formula for the Kullback–Leibler divergence between two normal distributions, where the second is also standardized. During testing, the mode of the latent distribution, $\mu(\mathbf{S})$, is the latent representation of the shape \mathbf{s} . In the following, we will rename the weighting coefficient β of equation (1) as w_{KL} to make it more memorable.

GWAS

According to the traditional GWAS scheme, we tested each genetic variant, $X_i \in \{0, 1, 2\}$, for association with each latent variable z_k through a univariate linear additive model of genetic effects:

$$z_k = \beta_{ik} X_i + \epsilon_{ik} \tag{7}$$

where ϵ_{ik} is the component not explained by the genotype, assumed to be normally distributed. The null hypothesis tested is that $\beta_{ik} = 0$.

Only unrelated individuals with self-reported British ancestry were included in the study to avoid problems related to population stratification. This produced a sample size of 48,651 individuals. Summary statistics of demographic data from these subsamples can be found in Supplementary Table 1. For the results presented in the main text, no individuals were excluded according to previous diagnoses or parameters of cardiac function derived from images (such as ejection fraction). Before GWAS, the phenotypes (that is, latent variables) were adjusted for a set of covariates: sex, age, height, weight, body mass index, body surface area, systolic and diastolic blood pressure, alcohol consumption, smoking status and the top ten genomic principal components (computed within the British population). Details on how to compute the genomic principal component loadings and the preprocessing of demographic data are provided in the Supplementary Information (section 1). To make this adjustment, a multivariate linear regression was performed on these covariates and then the residues of this regression were rank-inverse normalized. These inverse normalized residues are the phenotypic scores to be tested in the GWAS.

It is worth mentioning that the GWAS is performed on all individuals, including those on which the dimensionality reduction algorithm was trained. This is correct because the algorithm does not optimize association with genetic variants, and therefore a uniform distribution of P values under the null distribution can be safely assumed even when including these participants in the sample.

UPE

Given that the evaluation metric that guides training, that is, the reconstruction error with variational loss, is not necessarily aligned with the final objective of discovering genes that influence the shape of the LV, there is no reason to adopt the single run with the best value for this metric. This approach was followed in our previous work¹⁰. Indeed, the observation that several loci are detected in only a small subset of runs indicates that following such a procedure would lead to failure to discover some relevant genetic loci. For this reason, here we propose to adopt an ensemble-based approach, in which we pool the different phenotypes together in a redundant yet more expressive representation. On the basis of the observation that different network metaparameters, dataset partitioning and weight initializations yielded latent representations with different genome-wide significant loci, we proposed building an ensemble of phenotypes by concatenating the latent vectors for each run. This composite representation provides a redundant, yet more expressive representation of the LV shape at the end of the diastole. These runs covered a wide range of w_{KL} , and variations in network architectures, most importantly in the latent dimension n_z . Also, for a given combination of metaparameters (including architecture), an optimal learning rate was found and then five different random seeds were used to initialize the network’s weights and to partition the full dataset into training, validation and test sets (each seed constitutes a different run). Details on the architectural parameters are given in Supplementary Table 2.

Run selection. From the complete set of runs, we selected 36 training runs that achieved good reconstruction performance: a root mean squared deviation (r.m.s.d.) of less than 1 mm (averaged over participants from the test set). The deviation is taken to be the vertex-wise Euclidean distance, and the mean is taken over the $M = 5,220$ vertices of the LV mesh. In other words, the r.m.s.d. for participant i in run r is:

$$\text{r.m.s.d.}_{i,r} = \sqrt{\frac{1}{M} \sum_{j=1}^M \|\mathbf{x}_{i,j} - \hat{\mathbf{x}}_{i,j}^{(r)}\|_2^2}, \tag{8}$$

where $\mathbf{x}_{i,j}$ denotes the triad of spatial coordinates for vertex j in the mesh of the participant i , and $\hat{\mathbf{x}}_{i,j}^{(r)}$ is the same for the mesh reconstructed

in run r of the autoencoder. $\|\cdot\|_2^2$ denotes the squared Euclidean norm. The runs were selected based only on mesh reconstruction error and not in the presence or absence of GWAS hits. This allows us to assume a uniform distribution of P values over the $[0, 1]$ interval under the null distribution.

P value threshold correction. These 36 autoencoder runs produced a total of 384 phenotypes (where the latent dimension was eight for some runs and 16 for others). To control for the false discovery rate, this procedure requires correcting the usual genome-wide Bonferroni P value threshold, $P_{\text{GW}} = 5 \times 10^{-8}$, since the number of statistical tests that are performed increases with the size of the (pooled) representation. To avoid overcorrecting this threshold, one was dropped at random whenever a pair of latent variables (within the same run or not) had a Spearman correlation coefficient greater than 0.95 in absolute value. This procedure resulted in $K = 324$ phenotypes to be tested in GWAS. The new study-wide threshold P_{SW} is then Bonferroni-corrected dividing the standard genome-wide threshold P_{GW} by K . Thus, the final threshold is defined as $P_{\text{SW}} = \frac{P_{\text{GW}}}{K} = \frac{5 \times 10^{-8}}{324} = 1.5 \times 10^{-10}$. We note that, given the correlation present between the latent variables, this is a conservative threshold.

Genome partitioning and GWAS hit counting. Given that for each genomic locus, the lead variant might vary across different phenotypes by virtue of high linkage disequilibrium with close genetic variants, we adopted the following approach for locus counting: the genome is partitioned into 1,703 approximately LD-independent regions, where each is region is nearly 2 megabases (Mb) in length⁴⁵. We compute the number of autoencoder runs in which each region ℓ was genome-wide significant, denoting this quantity c_ℓ : for each run r and region ℓ , we retrieve the minimum value p , across the different latent variables $z_k^{(r)}$ (recall that $1 \leq k \leq 8$ or $1 \leq k \leq 16$, depending on the run r) that we call $p_{\ell,r}$. We then count the number of runs for which $p_{\ell,r} < P_{\text{GW}}$: $c_\ell = \sum_{r=1}^R \mathbb{1}_{p_{\ell,r} < P_{\text{GW}}}$ where $\mathbb{1}$ denotes the indicator function and $R = 36$. This c_ℓ is the quantity labelled 'count' in Table 1.

Downstream analysis of GWAS findings

Proximity analysis. We used the Ensembl Biomart database to query the positions of genes surrounding the lead SNPs in each region. We computed the distance between the genetic variant and the TSS and transcription end site (considering the information of the strands present in this database).

Transcriptome-wide associations studies. We used the S-PrediXcan tool to assess the correlation between imputed gene expression and intron excision occurrences with the extracted phenotypic data. The primary objective of this analysis is to identify potential candidate genes and the underlying mechanisms that may be involved in the observed genetic associations. The S-PrediXcan tool was supplied with summary statistics from GWAS as well as SNP dosage covariance matrices and gene expression (or intron excision) imputation models that were developed using GTEx data (v.8). These imputation models we used were constructed using the MASHR statistical methodology, which leverages on coexpression patterns across tissues to enhance the precision of estimated effect sizes for expression quantitative trait loci (eQTLs).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data for performing the GWAS in this work comes in its integrity from the UKBB. The UKBB Accession code for this application was 11350. Individual-level data are protected and therefore need to be

downloaded from the UKBB. 3D mesh data have been produced by ourselves via segmentation of the UKBB CMR imaging data. Interested researchers authorized by UKBB can be advised on how to reproduce these mesh data upon request. Publicly available datasets used for GWAS downstream analyses have been queried for this work: the Ensembl Biomart database (www.ensembl.org), the Integrative Epidemiology Unit OpenGWAS Project (gwas.mrcieu.ac.uk) for GWAS summary statistics, g:Profiler (biit.cs.ut.ee/gprofiler) for gene ontology terms and predictdb.org for GTEx-based prediction models and SNP covariance matrices needed to run S-PrediXcan. In all cases, the date of last access was 12 August 2023. For comparison, GWAS summary statistics were downloaded from <http://ftp.ebi.ac.uk> using the following study accession codes: GCST009393 through GCST009397 for ref. 2, GCST010125 through GCST010131 for ref. 4, GCST90000287 through GCST90000295 for ref. 11 and GCST90162626 for ref. 63. Relevant data for this study has been uploaded to Zenodo: network weights for the ensemble of 36 autoencoders⁷³ and the GWAS summary statistics for the traditional indices (LVEDV, LVEDSph, LVM and LVMVR) and for the first 16 shape PCs⁷⁴. A web application has been developed on which researchers can access detailed results derived from this work. Instructions on how to connect to this can be found at www.github.com/cistib/CardiacUPE. Source data are provided with this paper.

Code availability

The code for this work is split into several repositories publicly available on GitHub. All of them are accessible through a main repository: www.github.com/cistib/CardiacUPE (ref. 75). The www.github.com/cistib/CardiacCOMA repository, to which the previous points, is included as a Git submodule and contains the code for an implementation of the Chebyshev-based CoMA, using PyTorch and PyTorch Lightning. Hyperparameters and metrics are logged using the MLflow Python API. This repository also contains code to perform shape PCA on the cardiac meshes, using the scikit-learn Python package. Finally, it contains instructions on how to reproduce the software environment necessary to train the network and produce the latent variables that act as quantitative phenotypes in this work. The second submodule, www.github.com/cistib/GWAS_pipeline, contains the code to carry out preprocessing of genetic data for GWAS, GWAS execution, results visualization and downstream analysis. This repository is written in R and Python, and also contains bash scripts invoking standard GNU command-line tools. Additional tools required for this work are: bgenie, qctool, flashpca, plink and S-PrediXcan.

References

- Visscher, P. M. et al. 10 Years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
- 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* **140**, 1318–1330 (2019).
- Biffi, C. et al. Three-dimensional cardiovascular imaging-genetics: a mass univariate framework. *Bioinformatics* **34**, 97–103 (2018).
- Pirruccello, J. P. et al. Analysis of cardiac magnetic resonance imaging in 36,000 individuals yields genetic insights into dilated cardiomyopathy. *Nat. Commun.* **11**, 2254 (2020).
- Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
- Attar, R. et al. Quantitative CMR population imaging on 20,000 subjects of the UK Biobank imaging study: LV/RV quantification pipeline and its evaluation. *Med. Image Anal.* **56**, 26–42 (2019).
- Zhuang, X., Rhode, K. S., Razavi, R., Hawkes, D. J. & Ourselin, S. A registration-based propagation framework for automatic whole heart segmentation of cardiac MRI. *IEEE Trans. Med. Imaging* **29**, 1612–1625 (2010).

8. Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A. & Vandergheynst, P. Geometric deep learning: going beyond Euclidean data. *IEEE Signal Process. Magazine* **34**, 18–42 (2017).
9. Ranjan, A., Bolkart, T., Sanyal, S. & Black, M. J. Generating 3D faces using convolutional mesh autoencoders. In *Proc. Computer Vision - ECCV 2018*, Vol. 11207 (eds Ferrari, V. et al.) 725–741 (Springer International Publishing, 2018).
10. Bonazzola, R. et al. Image-derived phenotype extraction for genetic discovery via unsupervised deep learning in CMR images. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention* (eds de Bruijne, M. et al.) 699–708 (Springer, 2021).
11. Meyer, H. V. et al. Genetic and functional insights into the fractal structure of the heart. *Nature* **584**, 589–594 (2020).
12. Xia, Y. et al. Automatic 3D+t four-chamber CMR Quantification of the UK Biobank: integrating imaging and non-imaging data priors at scale. *Med. Image Anal.* **80**, 102498 (2022).
13. Fort, S., Hu, H. & Lakshminarayanan, B. Deep ensembles: a loss landscape perspective. Preprint at *arXiv* arXiv:1912.02757 (2019).
14. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
15. Kolberg, L. et al. g:profiler-interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Res.* **51**, W207–W212 (2023).
16. Barbeira, A. N. et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1825 (2018).
17. Smedley, D. et al. Biomart—biological queries made easy. *BMC Genomics* **10**, 22 (2009).
18. Vasani, R. S. et al. Genetic variants associated with cardiac structure and function: a meta-analysis and replication of genome-wide association data. *J. Am. Med. Assoc.* **302**, 168–178 (2009).
19. MacLennan, D. H., Asahi, M. & Tupling, A. R. The regulation of SERCA-type pumps by phospholamban and sarcolipin. *Ann. N.Y. Acad. Sci.* **986**, 472–480 (2003).
20. Eijgenraam, T. R., Silljé, H. H. & de Boer, R. A. Current understanding of fibrosis in genetic cardiomyopathies. *Trends Cardiovasc. Med.* **30**, 353–361 (2020).
21. Granzier, H. L. & Labeit, S. The giant protein titin: a major player in myocardial mechanics, signaling, and disease. *Circ. Res.* **94**, 284–295 (2004).
22. Knezevic, T. et al. BAG3: a new player in the heart failure paradigm. *Heart Fail. Rev.* **20**, 423–434 (2015).
23. Sheikh, F., Lyon, R. C. & Chen, J. Functions of myosin light chain-2 (MYL2) in cardiac muscle and disease. *Gene* **569**, 14–20 (2015).
24. Anfinson, M. et al. Significance of α -myosin heavy chain (MYH6) variants in hypoplastic left heart syndrome and related cardiovascular diseases. *J. Cardiovasc. Dev. Dis.* **9**, 144 (2022).
25. Xu, Y.-J. et al. Prevalence and spectrum of NKX2.5 mutations in patients with congenital atrial septal defect and atrioventricular block. *Mol. Med. Rep.* **15**, 2247–2254 (2017).
26. Li, B. et al. Isogenic human pluripotent stem cell disease models reveal ABRA deficiency underlies cTnT mutation-induced familial dilated cardiomyopathy. *Protein Cell* **13**, 65–71 (2022).
27. Astro, V. et al. Fine-tuned KDM1A alternative splicing regulates human cardiomyogenesis through an enzymatic-independent mechanism. *iScience* **25**, 104665 (2022).
28. Hong, L. et al. Prdm6 controls heart development by regulating neural crest cell differentiation and migration. *JCI Insight* **7**, e156046 (2022).
29. Steimle, J. & Moskowitz, I. TBX5: a key regulator of heart development. *Curr. Top. Dev. Biol.* **122**, 195–221 (2017).
30. Xiang, F. et al. Transcription factor CHF1/Hey2 suppresses cardiac hypertrophy through an inhibitory interaction with GATA4. *Am. J. Physiol. Heart Circ. Physiol.* **290**, H1997–H2006 (2006).
31. Fischer, A., Schumacher, N., Maier, M., Sendtner, M. & Gessler, M. The Notch target genes Hey1 and Hey2 are required for embryonic vascular development. *Genes Dev.* **18**, 901–911 (2004).
32. Pirruccello, J. P. et al. Genetic analysis of right heart structure and function in 40,000 people. *Nat. Genet.* **54**, 792–803 (2022).
33. Martin, R. I. et al. Genetic variants associated with risk of atrial fibrillation regulate expression of PITX2, CAV1, MYOZ1, C9orf3 and FANCC. *J. Mol. Cell. Cardiol.* **85**, 207–214 (2015).
34. Nielsen, J. B. et al. Genome-wide study of atrial fibrillation identifies seven risk loci and highlights biological pathways and regulatory elements involved in cardiac development. *Am. J. Hum. Genet.* **102**, 103–115 (2018).
35. Clausen, A. G., Vad, O. B., Andersen, J. H. & Olesen, M. S. Loss-of-function variants in the SYNPO2L gene are associated with atrial fibrillation. *Front. Cardiovasc. Med.* **8**, 650667 (2021).
36. Ruggiero, A., Chen, S. N., Lombardi, R., Rodriguez, G. & Marian, A. J. Pathogenesis of hypertrophic cardiomyopathy caused by myozenin 2 mutations is independent of calcineurin activity. *Cardiovasc. Res.* **97**, 44–54 (2013).
37. Zhang, M. et al. Expression, activity, and pro-hypertrophic effects of PDE5A in cardiac myocytes. *Cell. Signal.* **20**, 2231–2236 (2008).
38. Pirruccello, J. P. et al. Deep learning enables genetic analysis of the human thoracic aorta. *Nat. Genet.* **54**, 40–51 (2022).
39. Yu, M. et al. Computational estimates of annular diameter reveal genetic determinants of mitral valve function and disease. *JCI Insight* **7**, e146580 (2022).
40. Lahm, H. et al. Congenital heart disease risk loci identified by genome-wide association study in European patients. *J. Clin. Invest.* **131**, e141837 (2021).
41. Lv, F. et al. Neuron navigator 3 (NAV3) is required for heart development in zebrafish. *Fish Physiol. Biochem.* **48**, 173–183 (2022).
42. Bakker, M. L. et al. Transcription factor tbx3 is required for the specification of the atrioventricular conduction system. *Circ. Res.* **102**, 1340–1349 (2008).
43. Reilly, L. & Eckhardt, L. L. Cardiac potassium inward rectifier kir2: review of structure, regulation, pharmacology, and arrhythmogenesis. *Heart Rhythm* **18**, 1423–1434 (2021).
44. Deepe, R. N. et al. Sox9 expression in the second heart field; a morphological assessment of the importance to cardiac development with emphasis on atrioventricular septation. *J. Cardiovasc. Dev. Dis.* **9**, 376 (2022).
45. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283 (2016).
46. Espinoza-Lewis, R. A. et al. Shox2 is essential for the differentiation of cardiac pacemaker cells by repressing Nkx2-5. *Dev. Biol.* **327**, 376–385 (2009).
47. Nader, M. et al. Cardiac striatin interacts with caveolin-3 and calmodulin in a calcium sensitive manner and regulates cardiomyocyte spontaneous contraction rate. *Can. J. Physiol. Pharmacol.* **95**, 1306–1312 (2017).
48. Meurs, K. M. et al. Association of dilated cardiomyopathy with the striatin mutation genotype in boxer dogs. *J. Vet. Intern. Med.* **27**, 1437–1440 (2013).
49. Sotoodehnia, N. et al. Common variants in 22 loci are associated with QRS duration and cardiac ventricular conduction. *Nat. Genet.* **42**, 1068–1076 (2010).
50. Santamaria, S. & de Groot, R. ADAMTS proteases in cardiovascular physiology and disease. *Open Biology* **10**, 200333 (2020).
51. Van Berlo, J. H. et al. The transcription factor GATA-6 regulates pathological cardiac hypertrophy. *Circ. Res.* **107**, 1032–1040 (2010).

52. Maitra, M., Koenig, S. N., Srivastava, D. & Garg, V. Identification of gata6 sequence variants in patients with congenital heart defects. *Pediatr. Res.* **68**, 281–285 (2010).
53. Williams, S. G., Byrne, D. J. & Keavney, B. D. Rare gata6 variants associated with risk of congenital heart disease phenotypes in 200,000 UK Biobank exomes. *J. Hum. Genet.* **67**, 123–125 (2022).
54. Sun, Z. et al. NDRG2: a newly identified mediator of insulin cardioprotection against myocardial ischemia-reperfusion injury. *Basic Res. Cardiol.* **108**, 341 (2013).
55. Kawakami, E., Tokunaga, A., Ozawa, M., Sakamoto, R. & Yoshida, N. The histone demethylase Fbxl11/Kdm2a plays an essential role in embryonic development by repressing cell-cycle regulators. *Mech. Dev.* **135**, 31–42 (2015).
56. Gollob, M. H. et al. Somatic mutations in the connexin 40 gene (gja5) in atrial fibrillation. *New Engl. J. Med.* **354**, 2677–2688 (2006).
57. Koelemen, J., Gotthardt, M., Steinmetz, L. M. & Meder, B. RBM20-related cardiomyopathy: current understanding and future options. *J. Clin. Med.* **10**, 4101 (2021).
58. Rypdal, K. B. et al. ADAMTSL3 knock-out mice develop cardiac dysfunction and dilatation with increased TGF β signalling after pressure overload. *Commun. Biol.* **5**, 1392 (2022).
59. Boulet, I. R., Raes, A. L., Ottschytch, N. & Snyders, D. J. Functional effects of a KCNQ1 mutation associated with the long QT syndrome. *Cardiovasc. Res.* **70**, 466–474 (2006).
60. Frank, D. et al. Lmcd1/Dyxin, a novel Z-disc associated LIM protein, mediates cardiac hypertrophy in vitro and in vivo. *J. Mol. Cell. Cardiol.* **49**, 673–682 (2010).
61. Rath, N., Wang, Z., Lu, M. M. & Morrisey, E. E. LMCD1/Dyxin is a novel transcriptional cofactor that restricts GATA6 function by inhibiting dna binding. *Mol. Cell. Biol.* **25**, 8864–8873 (2005).
62. GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
63. Levin, M. G. et al. Genome-wide association and multi-trait analyses characterize the common genetic architecture of heart failure. *Nat. Commun.* **13**, 6914 (2022).
64. Roosenboom, J. et al. Mapping genetic variants for cranial vault shape in humans. *PLoS ONE* **13**, e0196148 (2018).
65. Fan, C. C. et al. Multivariate genome-wide association study on tissue-sensitive diffusion metrics highlights pathways that shape the human brain. *Nat. Commun.* **13**, 2423 (2022).
66. Bycroft, C. et al. Genome-wide genetic data on 500,000 UK Biobank participants. Preprint at *bioRxiv* <https://doi.org/10.1101/166298> (2017).
67. Petersen, S. E. et al. Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of UK Biobank - rationale, challenges and approaches. *J. Cardiovasc. Magn. Reson.* **15**, 46 (2013).
68. Pearson, K. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **2**, 559–572 (1901).
69. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT, 2016).
70. Defferrard, M., Bresson, X. & Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. In *Proc. Advances in Neural Information Processing Systems (NIPS)* (eds Lee, D.D. et al.) 3844–3852 (Curran Associates, 2016).
71. Garland, M. & Heckbert, P. S. Surface simplification using quadric error metrics. In *Proc. 24th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '97* (eds Pocock, L. et al.) 209–216 (ACM, 1997).
72. Higgins, I. et al. beta-VAE: learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations (ICLR)* 60–81 (Curran Associates, 2017).
73. Bonazzola, R. et al. Ensemble of 36 convolutional mesh autoencoders for left-ventricular meshes at end-diastole. *Zenodo* <https://doi.org/10.5281/zenodo.10536836> (2024).
74. Bonazzola, R. et al. GWAS summary statistics for left-ventricular phenotypes at end-diastole. *Zenodo* <https://doi.org/10.5281/zenodo.10537202> (2024).
75. Bonazzola, R. et al. Codebase for unsupervised phenotype ensembles. *Zenodo* <https://doi.org/10.5281/zenodo.10537131> (2024).

Acknowledgements

This project was funded by the following institutions: The Royal Academy of Engineering INSILEX (grant no. CIET1819\19), UKRI Frontier Research Guarantee INSILICO (grant no. EP/Y030494/1) (R.B., N.R. and A.F.F.), The Royal Society, through the International Exchanges scheme (grant no. IES\R2\202165) (R.B., E.F. and A.F.F.). E.F. was also funded by the Agencia Nacional de Promoción Científica y Tecnológica (grant no. PICT2018-3907) and UNL (grant nos. CAI+D 50220140100-084LI and 50620190100-145LI) (E.F.). B.K. and S.P. were supported by a British Heart Foundation Personal Chair. We thank A. Diaz-Pinto, P. Claes, R. Attar, F. Ibarrola and S. Raza for useful discussions as well as editorial reviews on the manuscript. This research has been conducted using data from UKBB, a major biomedical database. We thank the participants and members of the staff for enabling this research. This work was partly undertaken on ARC3 and ARC4, part of the High Performance Computing facilities at the University of Leeds, UK. The NIHR Manchester Biomedical Research Centre also funds the work of A.F.F. and B.K. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Author contributions

R.B. was responsible for the design of the work, implementation, data analysis and interpretation, and article writing. E.F. carried out the design of the work, drafting the article, critical revision of the article and project supervision. N.R. drafted the article and supervised the project. Y.X. was responsible for provision of input data and critical revision of the article. B.K. carried out critical revision of the article and interpretation of data. S.P. carried out critical revision of the article and interpretation of data. T.S.-M. conducted critical review of the article and supervision of the project. A.F.F. carried out conception of the work, drafting of the article, critical revision of the article and supervision of the project. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s42256-024-00801-1>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42256-024-00801-1>.

Correspondence and requests for materials should be addressed to Alejandro F. Frangi.

Peer review information *Nature Machine Intelligence* thanks Christoph Lippert, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

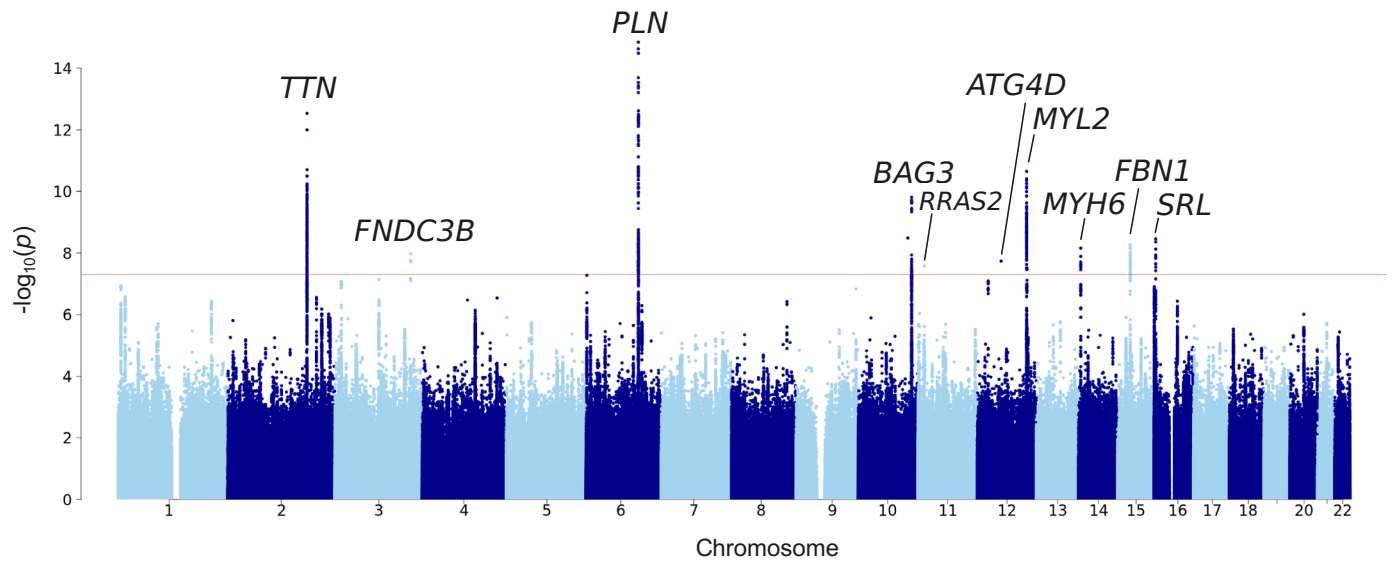
Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

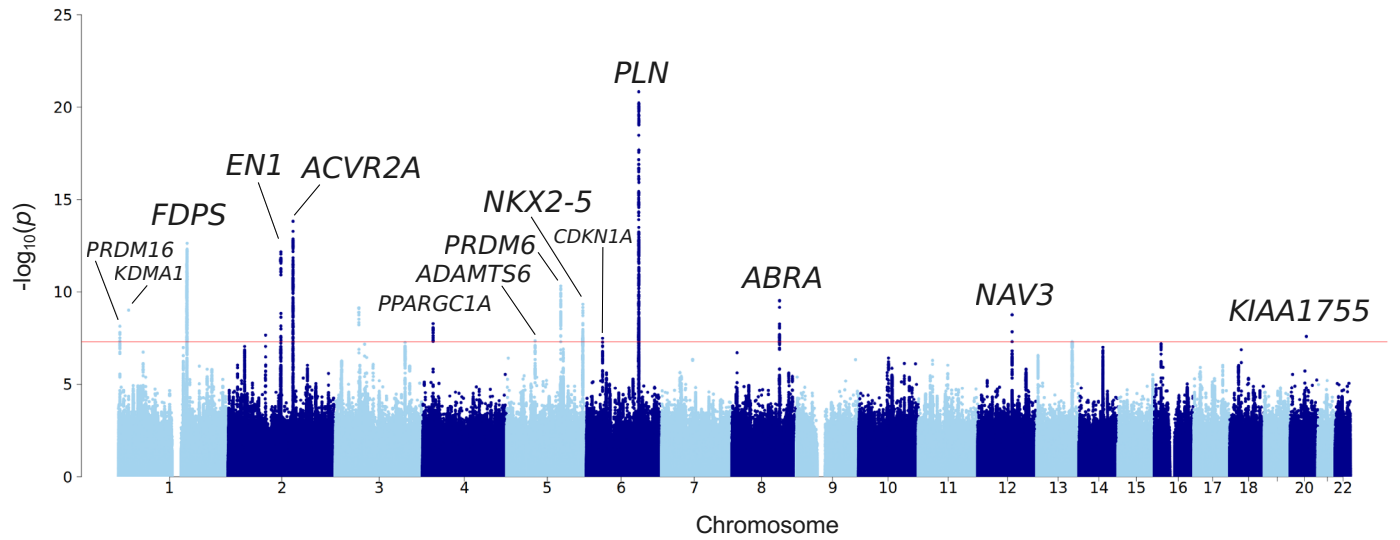
Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this

article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

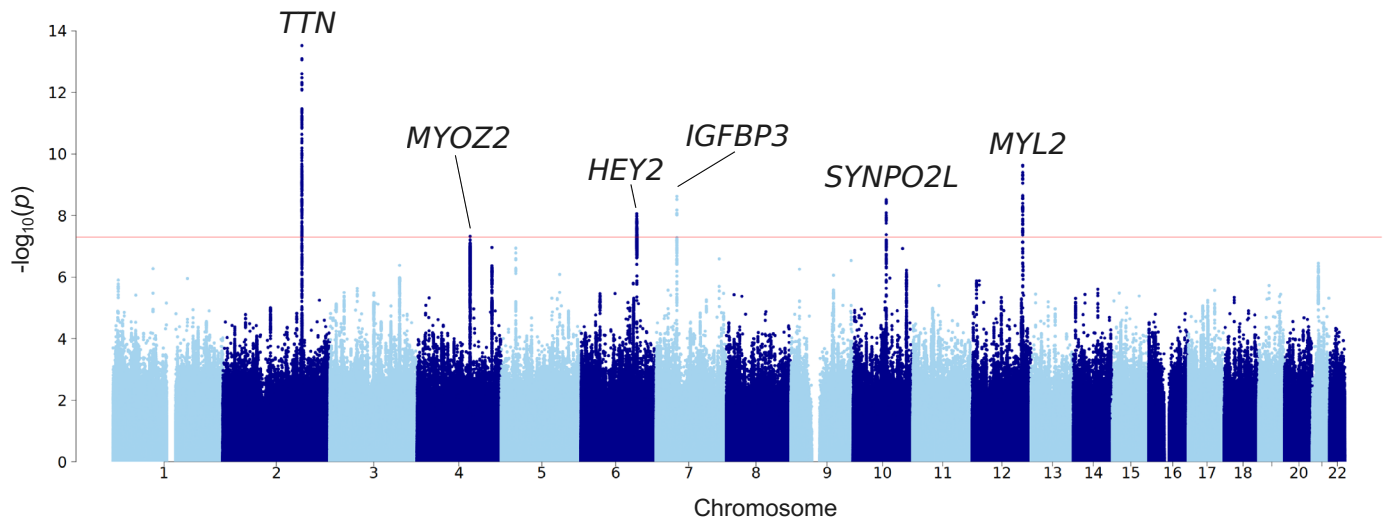
© The Author(s) 2024



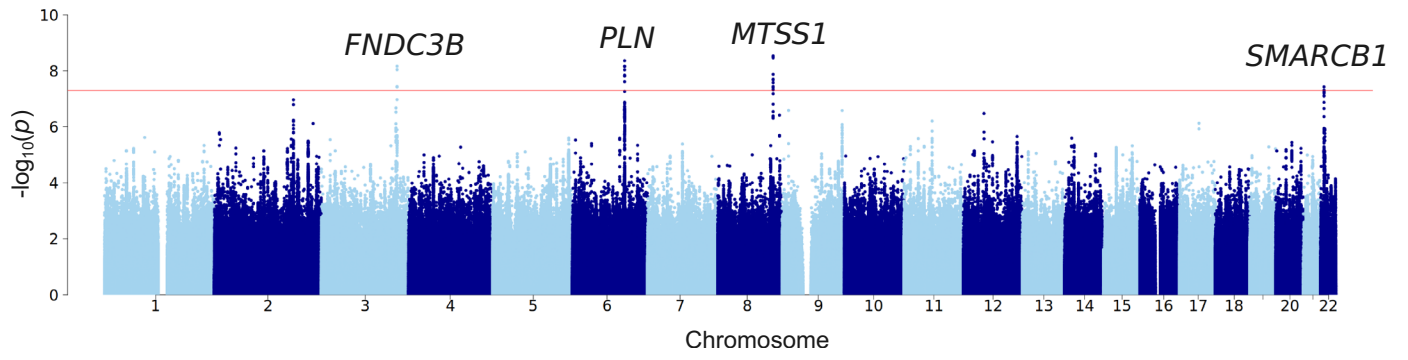
Extended Data Fig. 1 | Manhattan plot for GWAS of LVEDV. Manhattan plot of GWAS of left-ventricular end-diastolic volume (LVEDV).



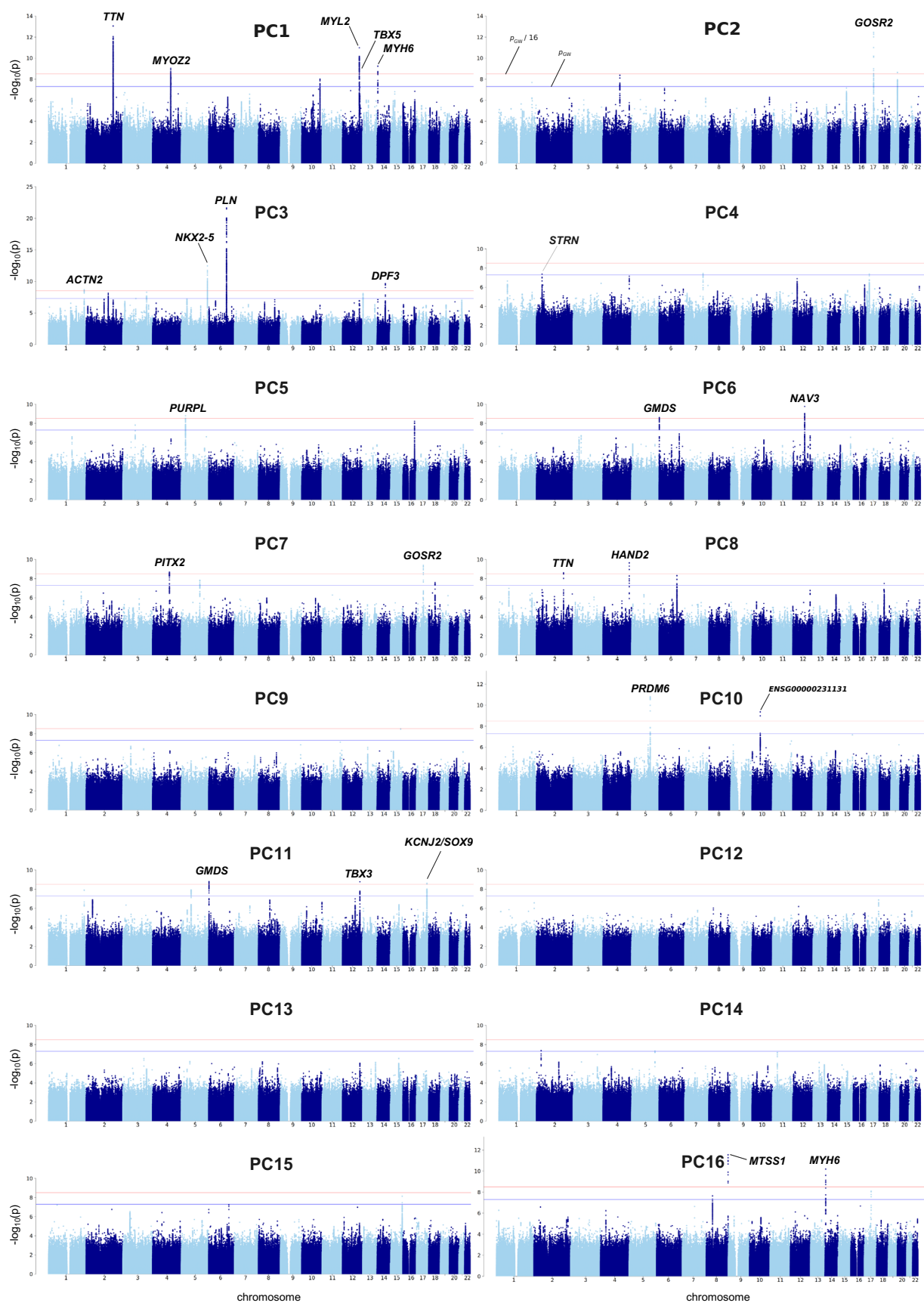
Extended Data Fig. 2 | Manhattan plot for GWAS of LVEDSph. Manhattan plot of GWAS of left-ventricular end-diastolic sphericity index (LVEDSph).



Extended Data Fig. 3 | Manhattan plot for GWAS of LVM. Manhattan plot of GWAS of left-ventricular myocardial mass (LVM).



Extended Data Fig. 4 | Manhattan plot for GWAS of LVMVR. Manhattan plot of GWAS of left-ventricular mass-to-volume ratio (LVMVR).



Extended Data Fig. 5 | Manhattan plots for GWAS on 16 first shape PCs. Manhattan plot of GWAS for the first 16 shape PCs of the left-ventricular shapes.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Data was downloaded from UK Biobank using tools provided by UK Biobank. 3D cardiac mesh data was produced from CMR imaging data using custom code (deep learning code written in Python/PyTorch).

Data analysis

Deep learning model training and inference was performed using custom code written in Python/PyTorch. The analysis was performed with custom Python and R code (tested on Python 3.9 and R 4.1). Genomic PCs were computed using flashpca. The Manhattan plots were generated using the R package qqman (v0.1.9). GWAS was performed using the BGENIE tool, and manipulations on genetic data were performed using qctool. TWAS was performed using the S-PrediXcan tool. All of these are open-source tools.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Data for performing the GWAS in this work comes in its integrity from the UK Biobank. UK Biobank Accession code for this application was 11350. Individual-level data are protected and therefore need to be downloaded from the UK Biobank. 3D mesh data have been produced by ourselves via segmentation of the UKBB CMR imaging data. Interested researchers authorized by UKBB can be advised on how to reproduce these mesh data upon request.

Publicly available datasets used for GWAS downstream analyses have been queried for this work: the Ensembl Biomart database (www.ensembl.org), the IEU Open GWAS Project (gwas.mrcieu.ac.uk) for GWAS summary statistics, g:Profiler (biit.cs.ut.ee/gprofiler) for gene ontology terms, and predictdb.org for GTE-based prediction models and SNP covariance matrices needed to run S-PrediXcan. In all cases, the date of last access was August 12, 2023.

For comparison, GWAS summary statistics were downloaded from <http://ftp.ebi.ac.uk> using the following study accession codes: GCST009393 through GCST009397 for Nay Aung et al. (2019), GCST010125 through GCST010131 for Pirruccello et al. (2020), GCST90000287 through GCST90000295 for Meyer et al. (2020) and GCST90162626 for Levin et al. (2022).

Relevant output data from this study has been uploaded to Zenodo: network weights for the ensemble of 36 autoencoders (<https://zenodo.org/doi/10.5281/zenodo.10536836>) and the GWAS summary statistics for the traditional indices (LVEDV, LVEDSph, LVM and LVMVR) and for the first 16 shape PCs (<https://zenodo.org/doi/10.5281/zenodo.10537202>).

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender	Genetic association analysis was performed on both sexes. Performing a separate study for each sex would limit greatly the statistical power of our study.
Population characteristics	Participants were 40-69 at the time of recruitment, which took place between 2006 and 2010. The CMR images were taken starting in 2014, until the date of last download (June 2023).
Recruitment	The UK Biobank sample aims to be representative of the UK population within the age range mentioned above. However, it is not exempt of 'healthy volunteer bias', however this is not expected to impact the results in this work.
Ethics oversight	UK Biobank has approval from the North West Multi-centre Research Ethics Committee (MREC) as a Research Tissue Bank (RTB) approval. This approval means that researchers do not require separate ethical clearance and can operate under the RTB approval (there are certain exceptions to this which are set out in the Access Procedures, such as re-contact applications).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Full sample size: 54,121. For the discovery set: 48,651. For the replication set (see below): 5,470. Since the purpose of this study is gene discovery, the full of set of CMR-scanned subjects available by 2023/06/09, was employed. This sample size was expected to be sufficient given the expected effect sizes for genetic variants discovered in recent genetic studies of LV image-derived phenotypes, which were conducted on smaller sample sizes.
Data exclusions	Subjects who did not self-report their ancestry as British. Individuals were excluded based on high missingness rate in SNP data. Finally, related subjects up to a third degree (one of each pair) were removed before GWAS.
Replication	Replication of the findings from the discovery phase was conducted using a left-out set of 5,470 subjects of British ancestry, also from the UK

Replication	Biobank (for which the whole pipeline was executed identically). This replication study aims to provide evidence of lack of data dredging during the discovery phase.
Randomization	Not relevant for this study, as there are no different groups defined.
Blinding	Not relevant for this study, as this is an observational study.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).
Research sample	State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.
Sampling strategy	Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.
Data collection	Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.
Timing	Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.
Data exclusions	If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.
Non-participation	State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.
Randomization	If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.
Research sample	Describe the research sample (e.g. a group of tagged <i>Passer domesticus</i> , all <i>Stenocereus thurberi</i> within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.
Sampling strategy	Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.
Data collection	Describe the data collection procedure, including who recorded the data and how.
Timing and spatial scale	Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken
Data exclusions	If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.
Reproducibility	Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.
Randomization	Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.

Blinding

Did the study involve field work? Yes No

Field work, collection and transport

Field conditions

Location

Access & import/export

Disturbance

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Included in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Included in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used

Validation

Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s)

Authentication

Mycoplasma contamination

Commonly misidentified lines (See [ICLAC](#) register)

Palaeontology and Archaeology

Specimen provenance

Specimen deposition

Indicate where the specimens have been deposited to permit free access by other researchers.

Dating methods

If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided. Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight

Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research, and [Sex and Gender in Research](#)

Laboratory animals

For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.

Wild animals

Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.

Reporting on sex

Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.

Field-collected samples

For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.

Ethics oversight

Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about [clinical studies](#)All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration

Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.

Study protocol

Note where the full trial protocol can be accessed OR if not available, explain why.

Data collection

Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.

Outcomes

Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.

Dual use research of concern

Policy information about [dual use research of concern](#)

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

- | No | Yes | |
|--------------------------|--------------------------|----------------------------|
| <input type="checkbox"/> | <input type="checkbox"/> | Public health |
| <input type="checkbox"/> | <input type="checkbox"/> | National security |
| <input type="checkbox"/> | <input type="checkbox"/> | Crops and/or livestock |
| <input type="checkbox"/> | <input type="checkbox"/> | Ecosystems |
| <input type="checkbox"/> | <input type="checkbox"/> | Any other significant area |

Experiments of concern

Does the work involve any of these experiments of concern:

- | No | Yes | |
|--------------------------|--------------------------|---|
| <input type="checkbox"/> | <input type="checkbox"/> | Demonstrate how to render a vaccine ineffective |
| <input type="checkbox"/> | <input type="checkbox"/> | Confer resistance to therapeutically useful antibiotics or antiviral agents |
| <input type="checkbox"/> | <input type="checkbox"/> | Enhance the virulence of a pathogen or render a nonpathogen virulent |
| <input type="checkbox"/> | <input type="checkbox"/> | Increase transmissibility of a pathogen |
| <input type="checkbox"/> | <input type="checkbox"/> | Alter the host range of a pathogen |
| <input type="checkbox"/> | <input type="checkbox"/> | Enable evasion of diagnostic/detection modalities |
| <input type="checkbox"/> | <input type="checkbox"/> | Enable the weaponization of a biological agent or toxin |
| <input type="checkbox"/> | <input type="checkbox"/> | Any other potentially harmful combination of experiments and agents |

ChIP-seq

Data deposition

- Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).
- Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links

May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

Files in database submission

Provide a list of all files available in the database submission.

Genome browser session

(e.g. [UCSC](#))

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates

Describe the experimental replicates, specifying number, type and replicate agreement.

Sequencing depth

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

Antibodies

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

Peak calling parameters

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

Software

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots

Confirm that:

- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.

Instrument

Identify the instrument used for data collection, specifying make and model number.

Software *Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*

Cell population abundance *Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*

Gating strategy *Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type *Indicate task or resting state; event-related or block design.*

Design specifications *Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*

Behavioral performance measures *State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s) *Specify: functional, structural, diffusion, perfusion.*

Field strength *Specify in Tesla*

Sequence & imaging parameters *Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*

Area of acquisition *State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI Used Not used

Preprocessing

Preprocessing software *Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*

Normalization *If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*

Normalization template *Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*

Noise and artifact removal *Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring *Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

Statistical modeling & inference

Model type and settings *Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).*

Effect(s) tested *Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*

Specify type of analysis: Whole brain ROI-based Both

Statistic type for inference (See [Eklund et al. 2016](#)) *Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*

Correction *Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).*

Models & analysis

- n/a | Involved in the study
- Functional and/or effective connectivity
- Graph analysis
- Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.