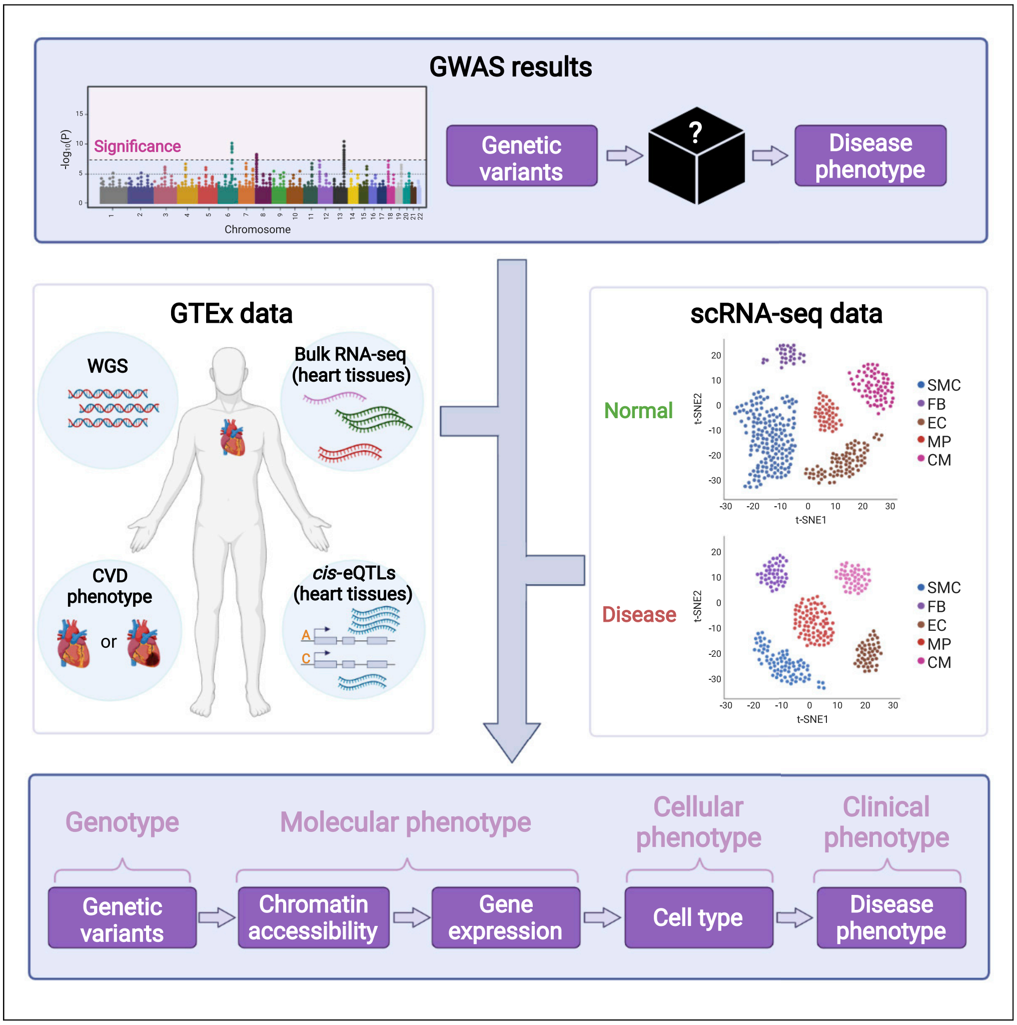
iScience

CellPress
OPEN ACCESS

Article

Unfolding the genotype-to-phenotype black box of cardiovascular diseases through cross-scale modeling



Xi Xi, Haochen Li, Shengquan Chen, ..., Ping Zhang, Wing Hung Wong, Xuegong Zhang

zhangxg@tsinghua.edu.cn

Highlights

We defined one type of cross-scale genotype-to-phenotype regulation path

We designed a framework GRPath to uncover putative regulation paths for diseases

GRPath helped uncover molecular mechanisms for two major types of heart failure

Xi et al., iScience 25, 104790
August 19, 2022 © 2022 The Author(s).
https://doi.org/10.1016/j.isci.2022.104790

iScience

CellPress
OPEN ACCESS

Article

Unfolding the genotype-to-phenotype black box of cardiovascular diseases through cross-scale modeling

Xi Xi,[1] Haochen Li,[2] Shengquan Chen,[1] Tingting Lv,[3] Tianxing Ma,[1] Rui Jiang,[1] Ping Zhang,[3] Wing Hung Wong,[4] and Xuegong Zhang[1,2,5,*]

SUMMARY

Complex traits such as cardiovascular diseases (CVD) are the results of complicated processes jointly affected by genetic and environmental factors. Genome-wide association studies (GWAS) identified genetic variants associated with diseases but usually did not reveal the underlying mechanisms. There could be many intermediate steps at epigenetic, transcriptomic, and cellular scales inside the black box of genotype-phenotype associations. In this article, we present a machine-learning-based cross-scale framework GRPath to decipher putative causal paths (pcPaths) from genetic variants to disease phenotypes by integrating multiple omics data. Applying GRPath on CVD, we identified 646 and 549 pcPaths linking putative causal regions, variants, and gene expressions in specific cell types for two types of heart failure, respectively. The findings suggest new understandings of coronary heart disease. Our work promoted the modeling of tissue- and cell type-specific cross-scale regulation to uncover mechanisms behind disease-associated variants, and provided new findings on the molecular mechanisms of CVD.

INTRODUCTION

Genome-wide association studies (GWAS) helped reveal the genotype-phenotype relations by finding associations between single-nucleotide polymorphisms (SNPs) and disease or trait (Edwards, 2005; The Wellcome Trust Case Control Consortium, 2007), but suffered from two limitations: 1) owing to the linkage disequilibrium (LD) structure in the human genome, many significantly associated SNPs identified by GWAS are tag SNPs rather than causal SNPs (Ding and Kullo, 2007; MacArthur et al., 2014; Stram, 2004); 2) GWAS studies only detect "black box" associations between genotype and disease phenotype and cannot explain how these SNPs influence disease risk (Neumeyer et al., 2020).

Inside the "black box" associations, the genotype-to-phenotype regulations are usually cross-scale multistep processes. From micro to macro, there are multiple levels of phenotypes: 1) molecular phenotype, the direct effect of a molecular-level variant (Wierbowski et al., 2018), such as transcriptome factor (TF) binding efficiency and change in gene expression; 2) cellular phenotype, the conglomerate of multiple cellular processes involving gene and protein expression that result in the elaboration of the particular morphology and function of a cell (Sul et al., 2009), which can appear as different cell types and specialized pathways; and 3) clinical phenotype (simply referred as phenotype), observable characteristic or trait of a disease for a given individual. It can be morphology, physiological properties, or behavior. The genotype-to-phenotype regulation process can involve many intermediate steps in different phenotype levels at epigenetic, transcriptomic, and cellular scales (MacRae and Vasan, 2016; Wang et al., 2018a).

Several works tried to open the "black box" behind genotype-phenotype associations. Integrating GWAS summary statistics with multi-omics data, researchers tried to find functional variants (Amlie-Wolf et al., 2018; Kircher et al., 2014; Li et al., 2020; Lu et al., 2017; Ritchie et al., 2014; Ward and Kellis, 2016) and disease-related cell types or driver cell types (Calderon et al., 2017; Gasperi et al., 2021; Shang et al., 2020; The Brainstorm Consortium et al., 2018; Watanabe et al., 2019). These works moved steps further—from GWAS associations to functional interpretations, which deepened our understanding of heredity and molecular mechanisms in diseases. But to the best of our knowledge, there is still few work that can link multiple steps together to achieve the whole regulation path from genotype to phenotype.

[1]MOE Key Laboratory of Bioinformatics and Bioinformatics Division, BNRIST / Department of Automation, Tsinghua University, Beijing 100084, China

[2]School of Medicine, Tsinghua University, Beijing 100084, China

[3]Department of Cardiology, Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua University, Beijing 102218, China

[4]Departments of Statistics and Biomedical Data Science, Stanford University, Stanford, CA 94305, USA

[5]Lead contact

*Correspondence: zhangxg@tsinghua.edu.cn

https://doi.org/10.1016/j.isci.2022.104790

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Take cardiovascular diseases (CVD) as an example. CVD is a broad class of diseases that involve the heart or blood vessels (Geneva: World Health Organization, 2011). It is the leading cause of death globally (Geneva: World Health Organization, 2011), and genetic factors contribute greatly to it (Kathiresan and Srivastava, 2012; Khera and Kathiresan, 2017). In coronary heart disease, for example, the heritability was estimated to be between 40 and 60% (McPherson and Tybjaerg-Hansen, 2016; Vinkhuyzen et al., 2013); in patients with dilated cardiomyopathy, 25-30% were estimated to have familial influence (Burkett and Hershberger, 2005; Grünig et al., 1998; Hershberger and Siegfried, 2011; Michels et al., 1992; Rosenbaum et al., 2020). Although GWAS have found a large number of genomic variations associated with different types of CVD, the underlying mechanisms are still unclear, hindering scientists' effort in finding medical solutions at the clinical level (Mattson and Liang, 2017). Closing the gap between genotype and clinical phenotype should be the future of cardiovascular genetics and genomics research and would be essential for precision medicine (MacRae and Vasan, 2016).

In this work, we explicitly defined one type of gene regulation path that depicts the impact of non-coding variants on chromatin accessibility, downstream gene expressions in specific cell types and on the disease phenotype, and designed an interpretable computational framework Gene Regulation Path (GRPath) to decipher such paths. In the framework, we incorporated GWAS summary statistics, tissue expression quantitative trait loci (eQTLs), whole-genome sequencing (WGS), RNA sequencing (RNA-seq), single-cell RNA sequencing (scRNA-seq) data and individual-level disease phenotype information, and utilized statistical modeling and machine learning techniques to find the paths. Applying the framework to two CVD subtypes, heart failure caused by dilated cardiomyopathy (dHF) and by coronary heart disease (cHF), we identified a set of multi-step regulation paths underlying dHF and cHF. Some findings were well supported by evidence in the literature, and some suggested new discoveries on previously unknown molecular mechanisms of the disease. The work provides new understandings of putative regulation paths of CVD, and demonstrated the potential of unfolding multi-step tissue- and cell type-specific regulation paths inside the genotype-phenotype association black boxes by mining data of multiple types in the public domain.

RESULTS

Modeling the gene regulation process

We proposed an interpretable computational framework GRPath to model the multi-layer gene regulation process that links genetic variants, chromatin accessibility, gene expression, cell type, and disease phenotype (Figure 1A). This framework first identified putative causal regions (pcRegions) and putative causal variants (pcVariants) of disease from personal genomes, transcriptomes, and clinical phenotypes, then added putative causal genes (pcGenes) and cell types into the link by incorporating eQTLs and scRNA-seq data (STAR Methods, Figure S4). Through this framework, we obtained putative gene regulation paths in the form of "pcRegion-pcVariant-pcGene-noteworthy cell type-disease phenotype" (Figure 1B).

To study the mechanisms of CVD, we collected relevant omics data including GWAS, GTEx, and scRNA-seq data.

Genome-wide association studies summary statistics

There are overall 340 statistically significant SNPs under the term "cardiovascular disease" or "cardiovascular disease risk factors" in GWAS summary statistics downloaded from GWAS Catalog v1.0 (Buniello et al., 2019). These SNPs are significantly associated with CVD, but may not be the causal variants.

GTEx

In GTEx v7, there are 357 donors with RNA-seq data obtained from heart tissue (left heart ventricle or atrial appendage). Each donor was also provided with WGS data and disease phenotype information. In addition, GTEx calculated single-tissue *cis*-eQTLs from corresponding WGS and RNA-seq data as well.

Single-cell RNA sequencing

We collected scRNA-seq data of dHF and cHF from the work of (Wang et al., 2020), which include smooth muscle cells (SMCs), endothelial cells (ECs), fibroblasts (FBs), macrophages (MPs), and cardiomyocytes (CMs) in the left ventricle and left atrial appendage. After quality control, there are 7,418 cells from normal heart samples, 2,728 cells from dHF samples, and 1,386 cells from cHF samples remaining.

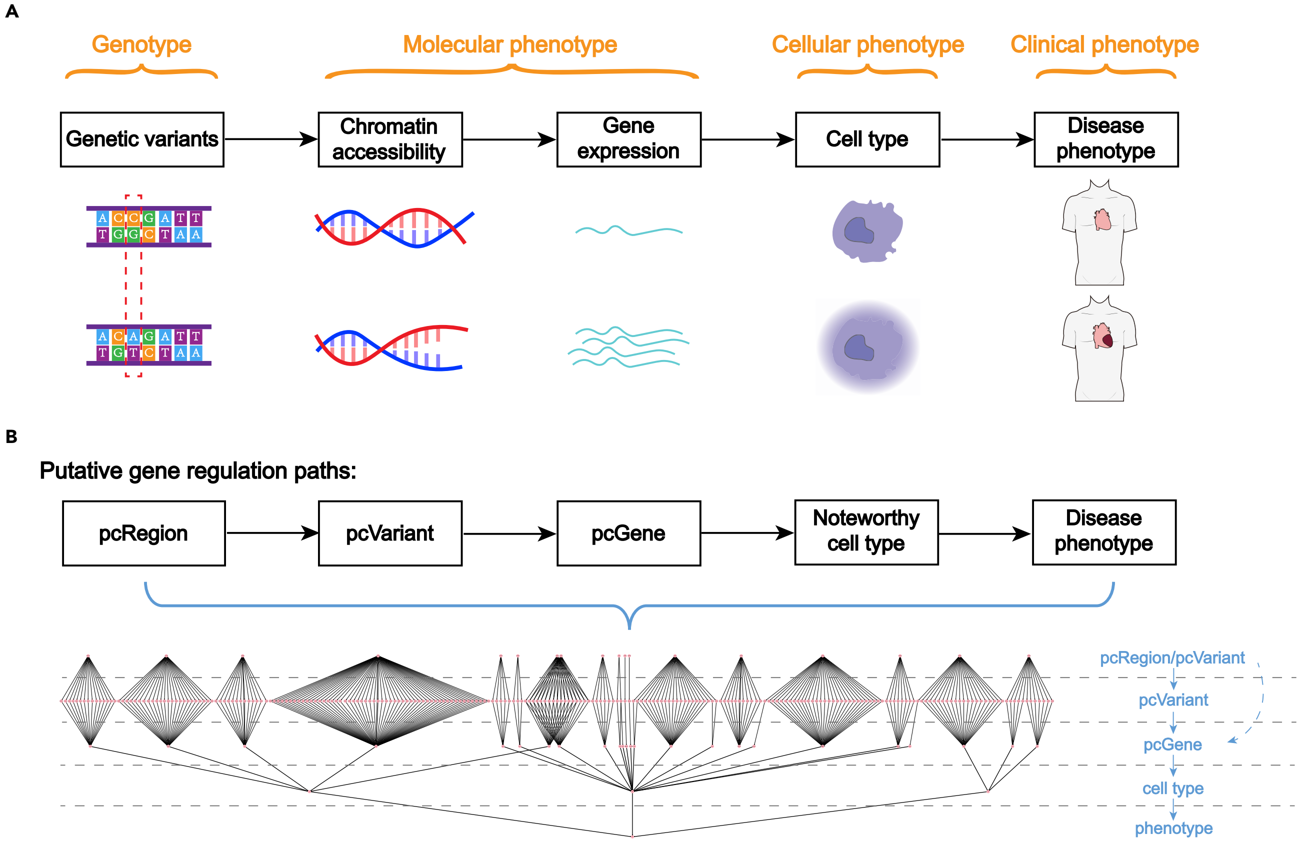


Figure 1. The multi-layer genotype-to-phenotype gene regulation process and the cross-scale gene regulation path

(A) Gene regulation process along genetic variants, chromatin accessibility, gene expression, cell type, and disease state.

(B) Putative gene regulation path in the form of "pcRegion-pcVariant-pcGene-noteworthy cell type-disease phenotype" illustrated by the hierarchical multi-layer structure.

We performed the GRPath analysis of CVD based on the data described above. The modeling procedure can be divided into three major steps. First, we predicted heart-tissue-specific openness scores for quantifying the chromatin accessibility state of pre-defined regulatory elements (REs) in the genome (STAR Methods). Second, we defined 338 candidate genomic regions (each 200 kb in size) for CVD according to GWAS summary statistics, and further identified pcRegions and top-ranked pcVariants within (STAR Methods). Third, we found target genes of the pcVariants, which we named as pcGenes, and incorporated scRNA-seq data of a CVD subtype (dHF or cHF) to obtain the most noteworthy cell type for each pcGene in the disease (STAR Methods). Through this model, we identified pcRegions and pcVariants for CVD in step two, pcGenes and noteworthy cell types in step three, and corresponding regulation paths for the two CVD subtypes.

pcRegions, pcVariants, and pcGenes of cardiovascular diseases

First, we introduce the pcRegions, pcVariants, and pcGenes found for CVD. We defined regions that show statistically significant influence on CVD risk as "pcRegions," and variants in these regions with relatively high causal effects on CVD risk as "pcVariants" (STAR Methods). In GTEx population, 192 of the 338 candidate genomic regions were identified as pcRegions, with FDR<1 × 10^{-9}. Specifically, 100 of them are alternate-allele-pathogenic regions (Table S1), where alternate alleles of top-ranked variants in this region increase the overall disease risk; 92 of them are reference-allele-pathogenic regions (Table S2), where reference alleles of top-ranked variants in this region increase the overall disease risk (Figure 2A).

We identified pcGenes of the pcVariants by referring to *cis*-eQTLs in heart tissues. We define the genes whose expressions are significantly associated with variation at a pcVariant as "pcGenes." In the 192 pcRegions, we found 295 and 249 pcGenes in alternate- and reference-allele-pathogenic regions,

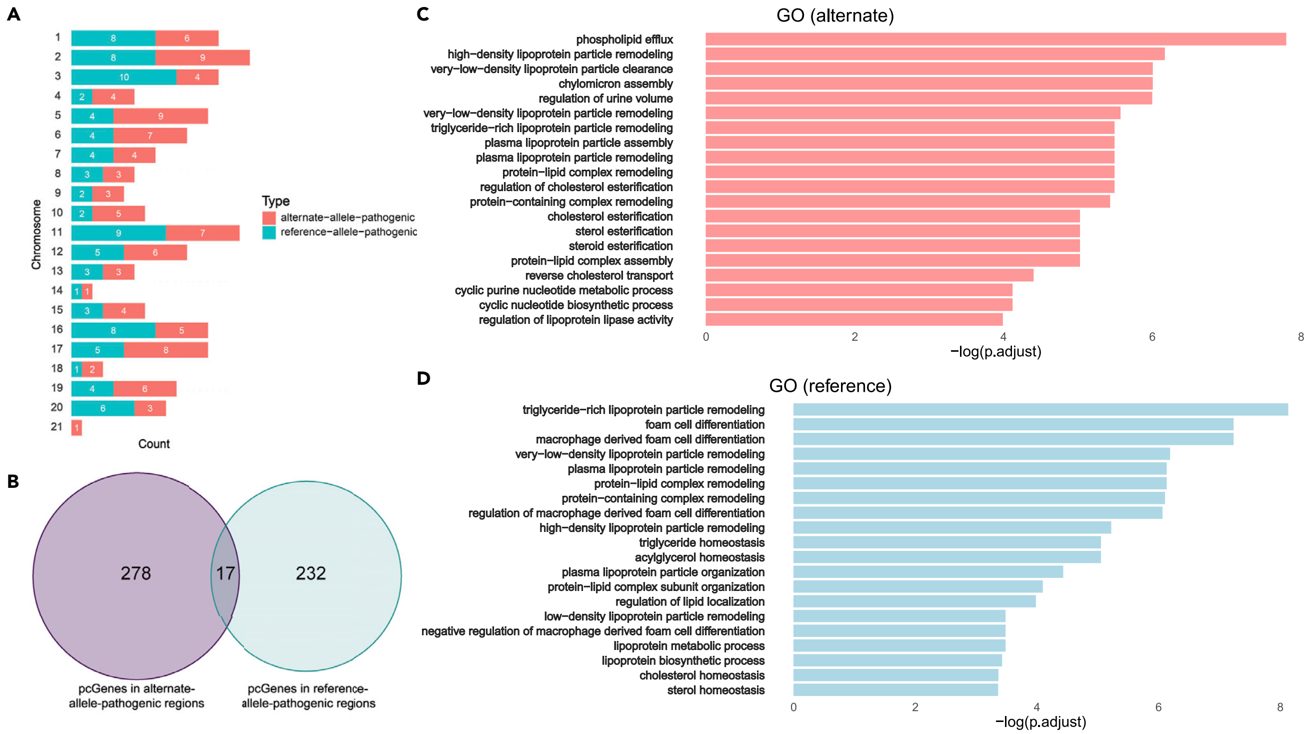


Figure 2. pcRegions and pcGenes of CVD

(A) the distribution of alternate- and reference-allele pathogenic regions on different chromosomes.

(B) Venn plot for the number of pcGenes in the two types of pcRegions, which share 17 pcGenes. Top-20 significant GO enrichment terms (FDR<0.05) for the CVD pcGenes in (C), alternate-allele-pathogenic regions and (D), reference-allele-pathogenic regions.

respectively, with 17 genes overlapping (Figure 2B). After that, we performed enrichment analysis on these pcGenes. Analysis of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways showed that the pcGenes in alternate- and reference-allele-pathogenic regions are both significantly enriched in cholesterol metabolism (hsa04979) (FDR<0.02). Analysis of Gene Ontology (GO) showed that there are 40 and 26 significant enrichment terms for pcGenes in the two scenarios, respectively (FDR<0.05) (Figures 2C and 2D). They have 9 shared enrichment terms, all of which are related to lipids or lipoproteins metabolic processes such as triglyceride-rich lipoprotein particle remodeling (GO:0034370), plasma lipoprotein particle organization (GO:0071827), and triglyceride catabolic process (GO:0019433), whose close associations with atherosclerosis have been well studied (Ishibashi, 2001; Libby et al., 2019; Musunuru and Kathiresan, 2016). In addition, pcGenes in the two types of regions also have their unique functions. It is worth noticing that in alternate-allele-pathogenic regions, the pcGenes are specially enriched in cyclic-nucleotide-related processes such as cyclic purine nucleotide metabolic process (GO:0052652) and cGMP biosynthetic process (GO:0006182), where genes *ADCY6*, *NPR2*, *NPPA* and *NPPB* that are not pcGenes in reference-allele-pathogenic regions participate. Likewise, pcGenes in reference-allele-pathogenic regions are uniquely and significantly enriched in terms relative to foam cell differentiation, for example, macrophage-derived foam cell differentiation (GO:0010742), where pcGenes *NR1H3*, *APOB*, *LPL*, *ITGB3*, *CETP*, and *SELENOK* participate. These results suggested that although there are large overlaps in the biological processes and functions that the pcVariants in alternate- and reference-allele-pathogenic regions are involved in, the two types of regions have different power over certain functions.

Noteworthy cell types and gene regulation paths for dilated cardiomyopathy

The above analysis was based on heart tissues, whereas the roles of different heart cell types in CVD were still unclear. Thus, after obtaining pcRegions, pcVariants, and corresponding pcGenes, we tried to identify the most noteworthy cell types for these pcGenes, and reveal the complete gene regulation paths regarding pcRegion, pcVariant, pcGene, the most noteworthy cell type and disease phenotype in a more specific CVD phenotype.

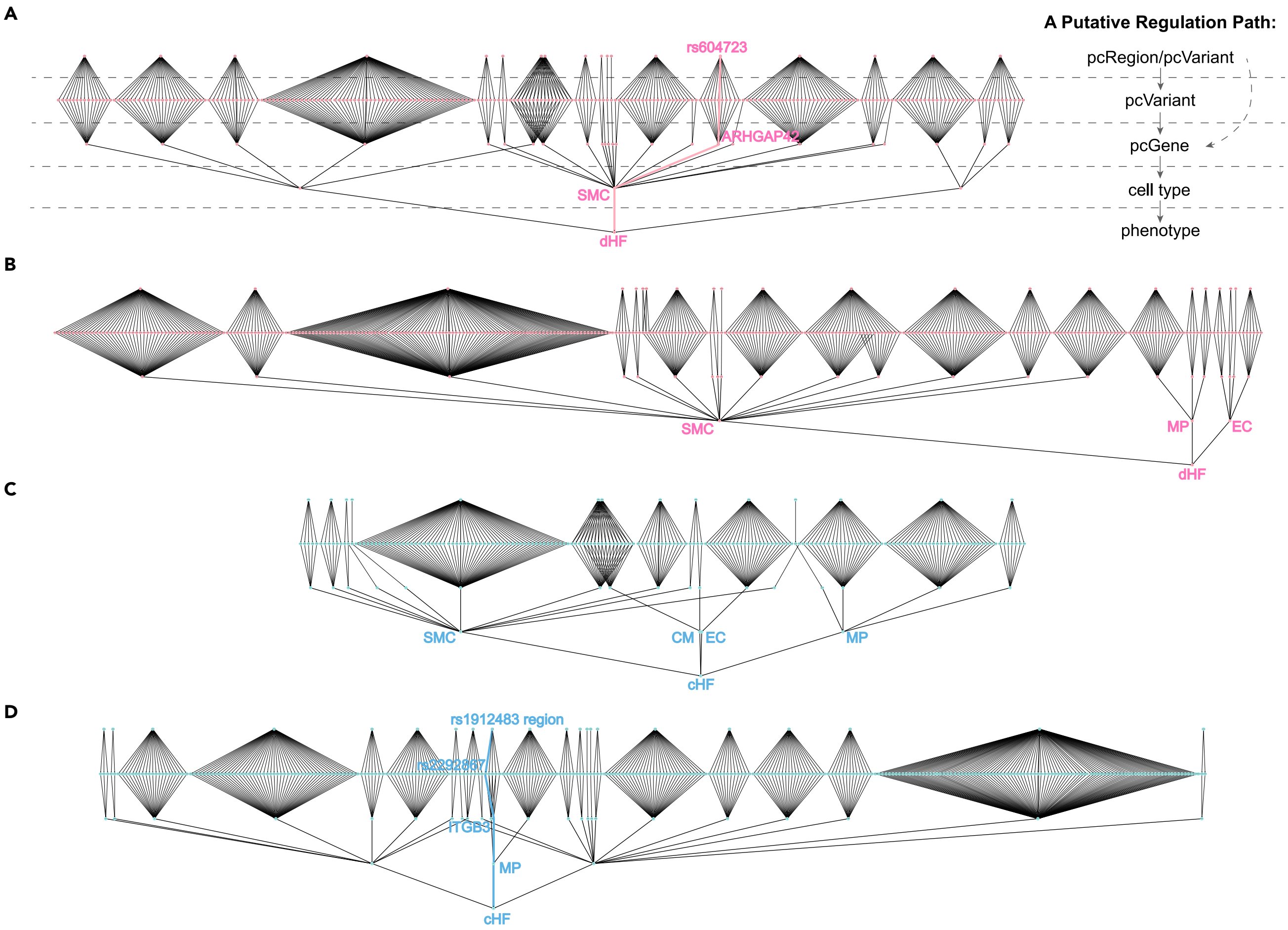


Figure 3. pcPaths for dHF and cHF

pcPaths from (A) alternate-allele-pathogenic regions and (B) reference-allele-pathogenic regions in dHF, or from (C) alternate-allele-pathogenic regions and (D) reference-allele-pathogenic regions in cHF. There are 5 layers in the paths. From top to bottom, each layer contains nodes representing pcRegions, pcVariants, pcGenes, cell types, and disease phenotype, respectively.

(A and D) showcase examples of the pcPath "rs604723 surrounding region-rs604723-*ARHGAP42*-SMCs-dHF" in dHF and "rs1912483 surrounding region-rs2292867-*ITGB3*-MPs-cHF" in cHF.

Here, we first focused on dHF. We performed quality control, normalization, highly variable genes selection, batch correction, and integration for scRNA-seq data of SMCs, ECs, FBs, MPs, and CMs from 7,418 normal and 2,728 dHF cells. Of the top-2000 highly variable genes, 44 genes were among the found pcGenes, which we named as highly variable pcGenes. We then identified the most noteworthy cell type for each highly variable pcGene with the highest AUROC in a classification-based method (STAR Methods). Linking the pcVariant, the corresponding pcGenes, and the most noteworthy cell type, we obtained 646 putative causal paths (pcPaths) in the form of "pcRegion-pcVariant-pcGene-cell type-disease" for dHF. Among them, 293 paths were derived from variants in alternate-allele pathogenic regions and 353 were from reference-allele-pathogenic regions (Figures 3A and 3B, Table S3).

To better focus on and interpret some key regulatory genes and regulation paths, we performed GO enrichment analysis on the 44 highly variable pcGenes, and narrowed down to 8 that are involved in the 23 significant enrichment terms (FDR<0.05) (Figure 4C). We then analyzed the corresponding pcRegions and cell types regarding the 8 genes (Figures 4A and 4B), and focused on the gene *ARHGAP42*. *ARHGAP42* is involved in top-ranked biological process "negative regulation of systematic arterial blood pressure (GO:0003085)" in the enrichment analysis. Its pcRegion that was centered around SNP rs604723 also has relatively high significance, ranking 25 among the 100 alternate-allele-pathogenic regions, and the central

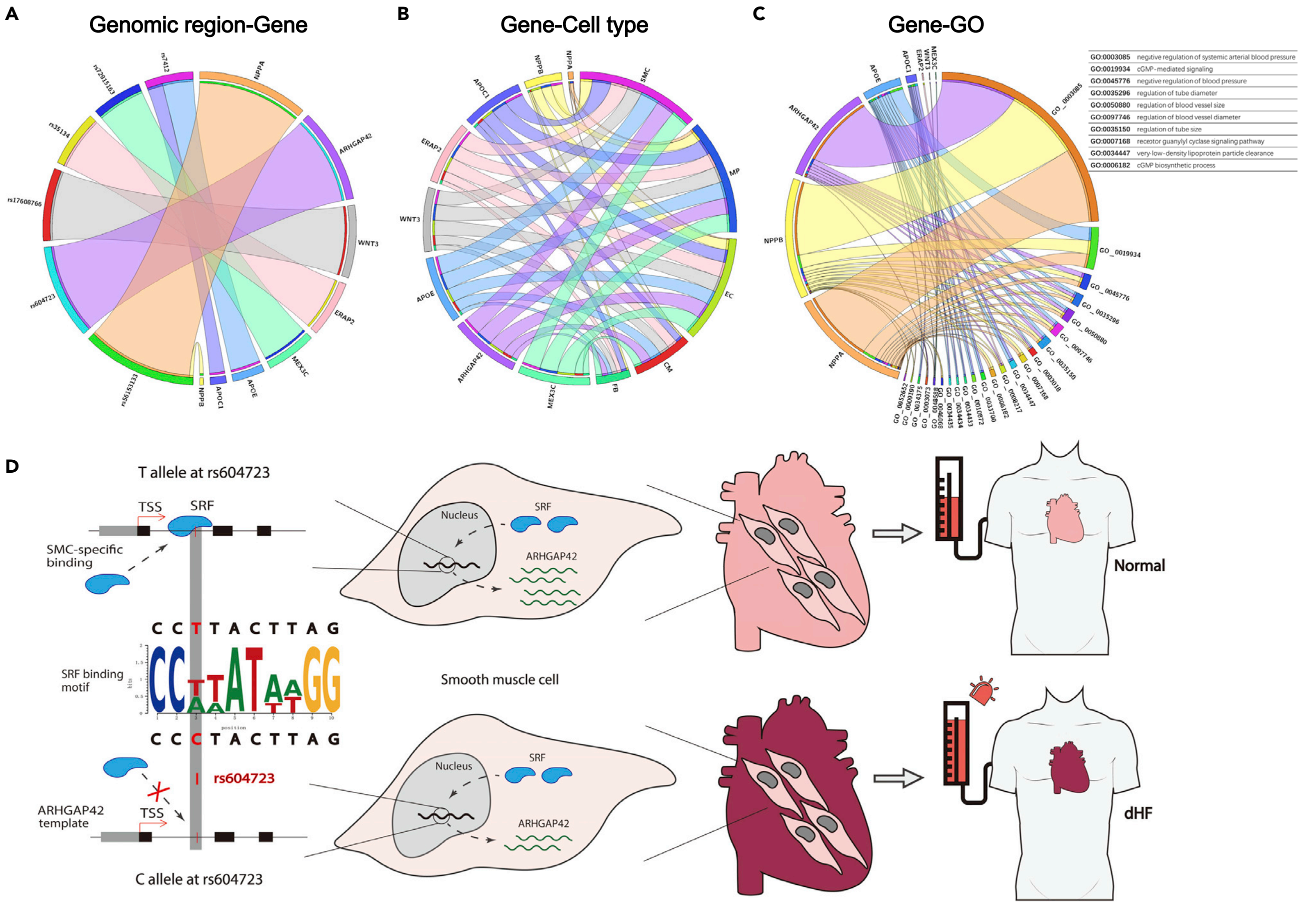


Figure 4. Overview of the key regulation paths in dHF and detailed explanation of an example

(A) Interaction between the 8 highly variable pcGenes and their corresponding pcRegions. The ribbon width represents combined influence of the significance of the pcRegion and the number of eQTLs that are pcVariants of the respective gene in that region. Thicker ribbon means higher significance of the corresponding region and more pcVariants of the gene.

(B) Interaction between the 8 highly variable pcGenes and the 5 heart cell types. Thicker ribbon represents higher AUROC.

(C) Interaction between the 8 highly variable pcGenes and the 23 significantly enriched GO terms. Thicker ribbon represents higher significance of the corresponding GO term. The same gene in (A-C) is in the same color.

(D) An example of the complete genotype-to-phenotype gene regulation process regarding rs604723, SRF binding, *ARHGAP42*, SMCs, blood pressure, and dHF: T allele at rs604723 promotes SMC-specific SRF binding, which increases *ARHGAP42* expression. Upregulation of *ARHGAP42* helps lower blood pressure, and thus reduces the risk of dHF.

SNP rs604723 itself was found to be a CVD pcVariant. Furthermore, *ARHGAP42* has the highest AUROC in SMCs (0.986 ± 0.002).

Linking the key components together, we obtained a pcPath "rs604723 surrounding region-rs604723-*ARHGAP42*-SMCs-dHF" (Figure 3A), which was very well supported by previous findings (Bai et al., 2017; Kasper et al., 1994; Messerli et al., 2017). The SNP rs604723 (chr11:100,610,546, GRCh37/hg19) is located on the first intron of *ARHGAP42* (chr11:100,558,019-100,864,672). Its reference (minor) allele is T, and alternate (major) allele is C. Bai et al. showed that T allele at rs604723 increases *ARHGAP42* expression by promoting a TF—serum response factor (SRF) binding to its located 600 bp DNase hypersensitivity site, and this process is specific to SMCs (Bai et al., 2017). *ARHGAP42* is involved in the Rho GTPase RhoA signaling pathway, where the upregulation of this gene helps lowering blood pressure, which in turn decreases the risk of dilated cardiomyopathy (Kasper et al., 1994, p. 67; Messerli et al., 2017) (Figure 4D). Our more detailed computational results further confirmed the knowledge suggested by the literature (Bai et al., 2017). From scRNA-seq data of normal and dHF samples, we observed that compared with SMCs in the dHF status, the *ARHGAP42* expression level was higher in normal SMCs. However, the result would be opposite if combining all cell types together (Figure S1A), which can be confirmed by GTEx bulk RNA-seq data (Figure S1B). This provided strong evidence that the increase of *ARHGAP42* expression level in dHF hearts is specific for SMCs. Combining the above data and knowledge, we summarize the gene regulation process regarding rs604723, *ARHGAP42*, SRF binding, SMCs, blood pressure, and dHF as follows: minor T allele at rs604723 promotes SRF binding to its located region, which is specific to SMCs. This in turn increases the *ARHGAP42* expression level in SMCs, which helps lowering blood pressure, thus reducing the risk of dHF.

The other 7 highly variable pcGenes involved in the significant enrichment terms also proved to be highly relevant for the cardiovascular system. For example, the genes *NPPA* and *NPPB* are widely used diagnostic and prognostic biomarkers for a spectrum of CVD (Chow et al., 2017; Goetze et al., 2020; Knowlton et al., 1995; Lee et al., 2019; Man et al., 2021; Ponikowski et al., 2016; Sergeeva et al., 2016; Warren et al., 2011). They are involved in a series of cardiac biological processes such as cardiac development and cardiorenal homeostasis and are implicated in response to cardiac injury and stress, especially in heart failure and cardiac hypertrophy (Chow et al., 2017; Knowlton et al., 1995; Lee et al., 2019; Man et al., 2021; Matsuoka et al., 2014; Ponikowski et al., 2016; Warren et al., 2011). Another gene, *WNT3*, participates in the Wnt signaling pathway, which is quiescent under normal conditions, but activated on pathological stress of the heart, such as chronic afterload increase (Malekar et al., 2010). The activation of Wnt signaling is critical for maladaptive cardiac hypertrophy and cardiomyopathy (Malekar et al., 2010).

Noteworthy cell types and gene regulation paths for coronary heart disease

Next, we focused on the other CVD subtype—cHF. Similarly, we analyzed scRNA-seq data of 7,418 normal cells and 1,386 cHF cells, and identified 39 pcGenes among the top-2000 highly variable genes. From these 39 highly variable pcGenes, we obtained 549 pcPaths for cHF, with 229 paths derived from variants in alternate-allele pathogenic regions and 320 from reference-allele-pathogenic regions (Figures 3C and 3D, Table S4). We further performed GO enrichment analysis on the 39 genes, and narrowed down to 11 highly variable pcGenes involved in the 29 significant enrichment terms (FDR<0.05) (Figure 5C).

One of the highly variable pcGenes, *ITGB3*, particularly attracted us. *ITGB3* participates in the macrophage-derived foam cell differentiation (GO:0010742) pathway, in which the CVD pcGenes in reference-allele-pathogenic regions were previously found to be enriched. From the AUROC results, we found that *ITGB3,* indeed, has the highest score in MPs (0.971 ± 0.005), closely followed by SMCs (0.970± 0.003) and other cell types (<0.940) (Figure 5B). Furthermore, the related pcRegion of *ITGB3* centered around SNP rs1912483, and it ranked 38 among the 92 reference-allele-pathogenic regions (Figure 5A). Together, these findings suggested pcPaths for cHF regarding the pcVariants in rs1912483 surrounding region, *ITGB3,* and MPs. This path has not been as systematically studied as the one found for dHF. It may reveal new biological insights. We, therefore, proposed hypothetical explanations for it by combining the computational results with current knowledge.

Five pcVariants (rs2292867, rs2292866, rs3785872, rs12940207, rs11868894) were found in the rs1912483 surrounding region, one of which is rs2292867. The SNP rs2292867 (chr17:45357489 C>T) is an intron variant located on the 2nd intron of *ITGB3* (+26.3 kb of TSS). According to HaploReg v4.1 (Ward and Kellis, 2016), this locus is predicted as an enhancer regulatory element in heart tissues. We can see that its variation from reference allele C to alternate allele T would increase the binding affinity of the transcriptional activator POU2F2 that binds to this position on the negative strand, according to the position weight matrix (PWM). Hi-C datasets in the heart left ventricle and macrophage further confirmed the active interaction between rs2292867 and *ITGB3* promoter region (Leung et al., 2015; Phanstiel et al., 2017; Wang et al., 2018b) (Figures S2A-S2B). These evidences suggested that the SNP rs2292867 located region should function as an enhancer regulatory element that actively interacts with the *ITGB3* promoter region. The T allele at rs2292867 increases the *ITGB3* expression level by increasing the transcriptional activator POU2F2 binding affinity (Figure 5D).

We further inferred the relationships between *ITGB3*, MPs, and cHF. It has been demonstrated that macrophage β3 integrin (ITGB3) suppresses the expression of TNFα, which impairs IL-6 cytokine and inflammation caused by hyperlipidemia (Schneider et al., 2007). Transplantation with β3-deficient marrow could increase mice atherosclerosis (Schneider et al., 2007). Another study reported that *ITGB3* is critical for regulating SMC proliferation and clonality, which is closely related to atherosclerosis development

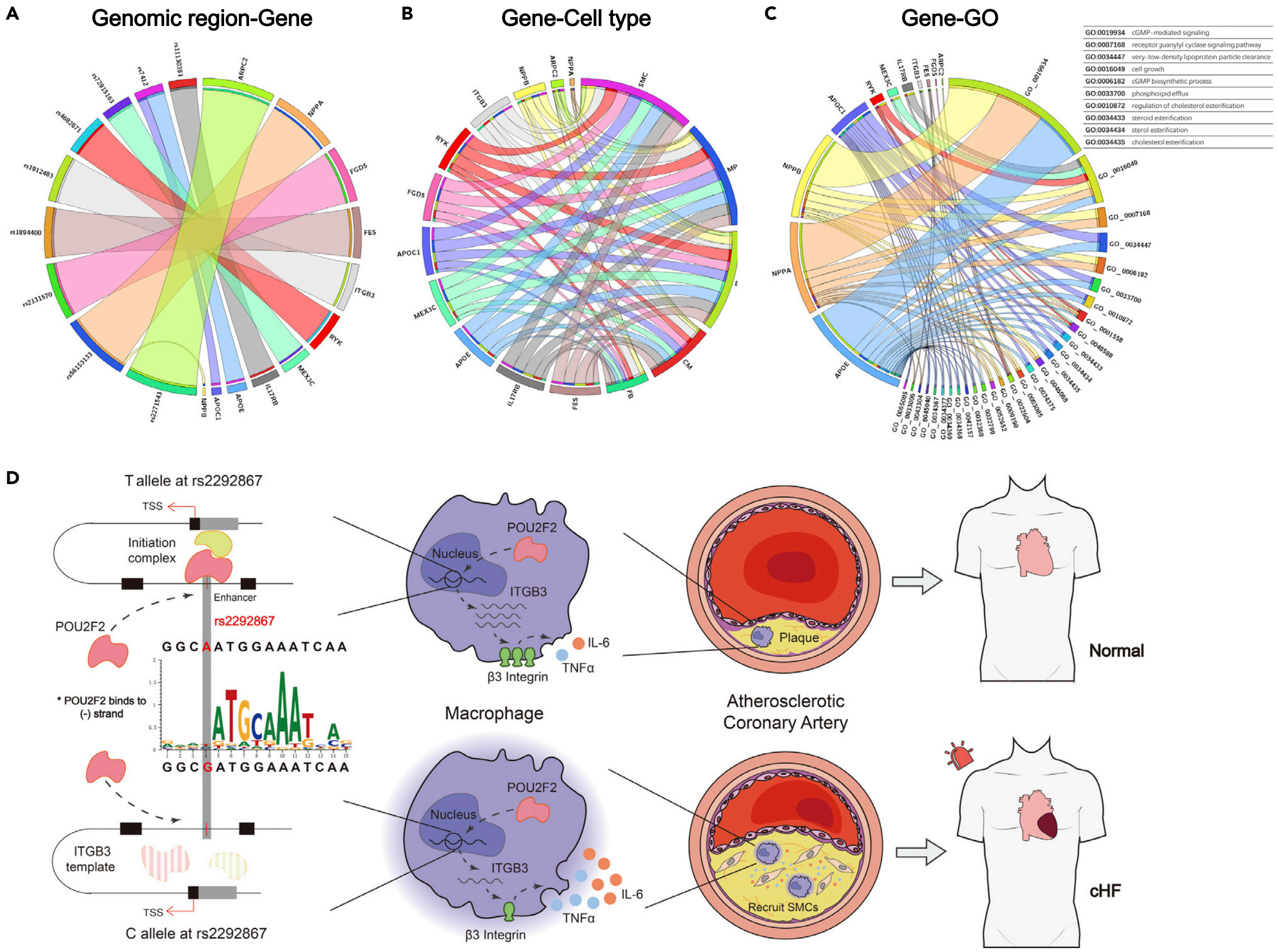


Figure 5. Overview of the key regulation paths in cHF and the detailed explanation of an example

(A) Interaction between the 11 highly variable pcGenes and their corresponding pcRegions. The ribbon width represents combined influence of the significance of the pcRegion and the number of eQTLs that are pcVariants of the respective gene in that region. A thicker ribbon means higher significance of the corresponding region and more pcVariants of the gene.

(B) Interaction between the 11 highly variable pcGenes and the 5 heart cell types. A thicker ribbon represents higher AUROC.

(C) Interaction between the 11 highly variable pcGenes and the 29 significantly enriched GO terms. A thicker ribbon represents higher significance of the corresponding GO term. The same gene in (A-C) is in the same color.

(D) Example of a proposed pcPath regarding rs2292867, *ITGB3*, MPs, and cHF: alternate allele T at rs2292867 promotes transcriptional activator POU2F2 binding, which increases *ITGB3* expression. Upregulation of *ITGB3* in MPs protects against atherosclerosis progression by suppressing TNFα expression, impairing IL-6, inhibiting SMCs migrating into the plaque, and reduces the risk of cHF.

(Misra et al., 2018). Their experiments showed that *ITGB3*-deficient MPs induce multiple SMCs to migrate into a plaque; SMCs, then, clonally expand within the plaque with limited migration, which accelerates atherosclerosis progression (Misra et al., 2018). Associating these two studies with the reports that IL-6 stimulates smooth muscle cell migration (Chava et al., 2009; Ikeda et al., 1991; Lee et al., 2012; Wang and Newman, 2003), we inferred that the upregulation of *ITGB3* gene in MPs decreases TNFα expression and inhibits IL-6. This impairs the migration and proliferation of SMCs, delays atherosclerosis progression, and thus reduces the risk of cHF (Figure 5D). The scRNA-seq data of normal and cHF samples support this explanation: MPs in normal samples have higher *ITGB3* gene expression levels compared with MPs in cHF samples (Figure S2C).

In summary, we proposed the pcPath "rs1912483 surrounding region-rs2292867-*ITGB3*-MPs-cHF" for cHF (Figures 3D and 5D): the alternate allele T at rs2292867 enhances the binding affinity of POU2F2 transcriptional activator, which increases *ITGB3* gene expression level. Then, the upregulation of *ITGB3* in MPs inhibits IL-6 cytokine and thus restricts the proliferation and migration of SMCs. This process delays atherosclerosis progression and reduces the risk of cHF.

Except for *ITGB3*, 10 other highly variable pcGenes are also involved in the significant enrichment terms. It is worth noticing that *NPPA*, *NPPB*, *APOE*, and *APOC1* are actively involved in both dHF and cHF. Especially, *APOE* has higher AUROC in ECs in cHF (0.981 ± 0.003) than that in dHF (0.960± 0.004), which is probably because *APOE* modulates EC functions that are more important to coronary heart disease than to dilated cardiomyopathy. Studies found that *APOE* is central to the transport and metabolism of lipids, which is closely related to atherosclerogenesis (Huang and Mahley, 2014; Marais, 2019; Satizabal et al., 2018). It plays a novel role in modulating the cav-1-eNOS interaction in ECs, which is associated with a number of CVDs such as atherosclerosis and hypertension (Yue et al., 2012). The *APOC1* gene is located very close (~5 kb downstream) to the *APOE* gene. It is also involved in lipoprotein metabolism and might inhibit the *APOE*-mediated uptake of very-low-density lipoprotein particles (Zhou et al., 2019; Shachter, 2001).

DISCUSSION

GWAS revealed the genotype-phenotype associations but was lack of functional interpretations. In this work, we proposed a cross-scale computational framework GRPath to open the ''black box'' associations between genotype and multi-layer phenotypes. Starting from personal genomes, GRPath links genetic variants, chromatin accessibility, gene expression, cell type, and individual-level disease state together, and can uncover putative gene regulation paths of a specific disease in the form of ''pcRegion-pcVariant-pcGene-noteworthy cell type-disease phenotype'' paths. It bridges the gap between statistical associations and biological mechanisms and can be used to study heredity and micro-to-macro mechanisms of complex diseases with strong genetic effects.

We showcased the use of GRPath on CVD, a disease where genetic factors contribute greatly. We revealed a list of pcRegions, pcVariants, and pcGenes of CVD, and identified 646 and 549 pcPaths for two CVD subtypes dHF and cHF, respectively. We studied two example paths in detail. One example is the ''rs604723 surrounding region-rs604723-*ARHGAP42*-SMCs-dHF'' path which can be well validated by existing works. The other is a new discovery on the putative path for cHF involving rs1912483 surrounding region, rs2292867, *ITGB3,* and MPs. The findings illustrated the power of the proposed method and brought new understandings of the possible mechanism underlying the studied disease.

The example paths we analyzed focused only on highly variable pcGenes in significantly enriched GO terms. There can be multiple regulatory paths in which the highly variable pcGenes may not be enriched for a specific GO annotation. For example, *USP36* is a highly variable pcGene for cHF, and it has the highest AUROC in ECs. It has been reported that *USP36* might be related to the formation of a circular RNA hsa_circ_0003204. The circRNA influences the development of atherosclerosis as it functions through the miR-330-5p/Nod2 axis that promotes oxidative stress and apoptosisand worsens endothelial cell injury induced by low-density lipoprotein (Zhang et al., 2021). Our analysis showed that the SNP rs1057040, a pcVariant in rs1044486 surrounding region, may affect the expression of *USP36* and hsa_circ_0003204 through TF binding (Figure S3). Further investigation on the ''rs1044486 surrounding region-rs1057040-*USP36*-ECs-cHF'' path should be able to bring new biological insights.

The regulatory paths we found can explain possible biological mechanisms that cause the associations between genotypes and the disease phenotype. However, stringent causality cannot be established based only on the static data collected from multiple studies. Mendelian randomization (MR) methods have been used to identify possible causal genes for GWAS results, but they were not designed to unfold the black-box, and they have either the tissue non-specificity problem or interpretability problem (Burgess and Thompson, 2015; Neumeyer et al., 2020). Our framework, on the other hand, can be specified to disease-relevant tissues. Each step in the inference process was transparent and interpretable. Moreover, our framework may capture some cohort-specific signals derived from personal genomes, which may be missed by summary-data MR methods.

The cross-scale computational framework we proposed can be applied to any diseases that have genetic effects. To use GRPath, users should prepare GWAS SNPs of the disease, personalized chromatin openness scores from disease-relevant tissues and corresponding donor disease phenotypes, tissue-eQTLs, and scRNA-seq data from disease and control samples. Taking these data as input, GRPath can then predict putative gene regulation paths for the disease. The source code and usage description of GRPath are freely available online.

Limitations of the study

There are several aspects that our work can be further improved. In the current work, we only considered one type of gene regulation path through gene expression. But phenotypes can be affected by many other biological processes such as DNA methylation, histone modification, alternative splicing, and so forth. These factors should be considered in future versions of GRPath models. Another limitation of the current work is that the data we used were unpaired. Being able to utilize scattered data from multiple sources is an advantage of the proposed method, but if we can use data collected from better-coordinated studies, the findings can be more accurate. Another aspect that can further improve the method is to decipher the link between gene expression properties in cells and the disease phenotype. Ideally, we should be able to quantify the deviations of certain types of cells from their normal states and model the quantitative influences of such deviations on the disease phenotypes. This would require the complete characterization of molecular and functional features of all major types of cells, which is the goal of building comprehensive cell atlases of healthy references (Chen et al., 2021; Regev et al., 2017).

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Predict heart-specific openness scores
  - Decide pcRegions and pcVariants for CVD
  - Define candidate genomic regions
  - Evaluate causal effects of variants in each candidate genomic region
  - Evaluate causal effects of candidate genomic regions
  - Call pcRegions and pcVariants
  - Find noteworthy cell types for two types of heart failure
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.104790.

ACKNOWLEDGMENTS

We thank Dr. Jinmeng Jia and colleagues Fanhong Li, Yixin Chen, Qiuchen Meng, Jinyang Liu for helpful discussions.

The work of X.Z and R.J. was supported by the National Key Research and Development Program of China (2021YFF1200901 and 2021YFF1200902) and NSFC grants 61721003, 62050178. The work of W.H.W was supported by NIH grants HG007735 and HG010359.

AUTHOR CONTRIBUTIONS

Conceptualization, X.Z.; Methodology, X.X.; Investigation, X.X. and H.L.; Writing – Original Draft, X.X. and H.L.; Writing – Review & Editing, X.X., H.L., S.C., T.L., T.M. and X.Z.; Supervision, X.Z., W.H.W., R.J. and P.Z.; Funding Acquisition, X.Z., W.H.W., and R.J.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 5, 2021
Revised: January 26, 2022
Accepted: July 14, 2022
Published: August 19, 2022

REFERENCES

Amlie-Wolf, A., Tang, M., Mlynarski, E.E., Kuksa, P.P., Valladares, O., Katanic, Z., Tsuang, D., Brown, C.D., Schellenberg, G.D., and Wang, L.-S. (2018). INFERNO: inferring the molecular mechanisms of noncoding genetic variants. Nucleic Acids Res. *46*, 8740–8753. https://doi.org/10.1093/nar/gky686.

Bai, X., Mangum, K.D., Dee, R.A., Stouffer, G.A., Lee, C.R., Oni-Orisan, A., Patterson, C., Schisler, J.C., Viera, A.J., Taylor, J.M., and Mack, C.P. (2017). Blood pressure–associated polymorphism controls ARHGAP42 expression via serum response factor DNA binding. J. Clin. Invest. *127*, 670–680. https://doi.org/10.1172/JCI88899.

Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47*, D1005–D1012. https://doi.org/10.1093/nar/gky1120.

Burgess, S., and Thompson, S.G. (2015). Multivariable mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. Am. J. Epidemiol. *181*, 251–260. https://doi.org/10.1093/aje/kwu283.

Burkett, E.L., and Hershberger, R.E. (2005). Clinical and genetic issues in familial dilated cardiomyopathy. J. Am. Coll. Cardiol. *45*, 969–981. https://doi.org/10.1016/j.jacc.2004.11.066.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. *36*, 411–420. https://doi.org/10.1038/nbt.4096.

Calderon, D., Bhaskar, A., Knowles, D.A., Golan, D., Raj, T., Fu, A.Q., and Pritchard, J.K. (2017). Inferring relevant cell types for complex traits by using single-cell gene expression. Am. J. Hum. Genet. *101*, 686–699. https://doi.org/10.1016/j.ajhg.2017.09.009.

Chava, K.R., Karpurapu, M., Wang, D., Bhanoori, M., Kundumani-Sridharan, V., Zhang, Q., Ichiki, T., Glasgow, W.C., and Rao, G.N. (2009). CREB-mediated IL-6 expression is required for 15(S)-Hydroxyeicosatetraenoic acid–induced vascular smooth muscle cell migration. Arterioscler. Thromb. Vasc. Biol. *29*, 809–815. https://doi.org/10.1161/ATVBAHA.109.185777.

Chen, S., Luo, Y., Gao, H., Li, F., Li, J., Chen, Y., You, R., Lv, H., Hua, K., Jiang, R., and Zhang, X. (2021). Toward a unified information framework for cell atlas assembly. Natl. Sci. Rev. *9*, nwab179. https://doi.org/10.1093/nsr/nwab179.

Chow, S.L., Maisel, A.S., Anand, I., Bozkurt, B., de Boer, R.A., Felker, G.M., Fonarow, G.C., Greenberg, B., Januzzi, J.L., Kiernan, M.S., et al. (2017). Role of biomarkers for the prevention, assessment, and management of heart failure: a scientific statement from the American heart association. Circulation *135*, e1054–e1091. https://doi.org/10.1161/CIR.0000000000000490.

Ding, K., and Kullo, I.J. (2007). Methods for the selection of tagging SNPs: a comparison of tagging efficiency and performance. Eur. J. Hum. Genet. *15*, 228–236. https://doi.org/10.1038/sj.ejhg.5201755.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. *4*, 1184–1191. https://doi.org/10.1038/nprot.2009.97.

Edwards, A.O., Ritter, R., 3rd, Abel, K.J., Manning, A., Panhuysen, C., and Farrer, L.A. (2005). Complement factor H polymorphism and age-related macular degeneration. Science *308*, 421–424. https://doi.org/10.1126/science.1110189.

Gasperi, C., Chun, S., Sunyaev, S.R., and Cotsapas, C. (2021). Shared associations identify causal relationships between gene expression and immune cell phenotypes. Commun. Biol. *4*, 279. https://doi.org/10.1038/s42003-021-01823-w.

Geneva: World Health Organization (2011). Global Atlas on Cardiovascular Disease Prevention and Control.

Goetze, J.P., Bruneau, B.G., Ramos, H.R., Ogawa, T., de Bold, M.K., and de Bold, A.J. (2020). Cardiac natriuretic peptides. Nat. Rev. Cardiol. *17*, 698–717. https://doi.org/10.1038/s41569-020-0381-0.

Grünig, E., Tasman, J.A., Kücherer, H., Franz, W., Kübler, W., and Katus, H.A. (1998). Frequency and phenotypes of familial dilated cardiomyopathy. J. Am. Coll. Cardiol. *31*, 186–194. https://doi.org/10.1016/S0735-1097(97)00434-8.

Hagberg, A.A., Schult, D.A., and Swart, P.J. (2008). Exploring network structure, dynamics, and function using NetworkX. In Proceedings of the 7th Python in Science Conference, G. Varoquaux, T. Vaught, and J. Millman, eds., pp. 11–15.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. Nature *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2.

Hershberger, R.E., and Siegfried, J.D. (2011). Update 2011: clinical and genetic issues in familial dilated cardiomyopathy. J. Am. Coll. Cardiol. *57*, 1641–1649. https://doi.org/10.1016/j.jacc.2011.01.015.

Huang, Y., and Mahley, R.W. (2014). Apolipoprotein E: structure and function in lipid metabolism, neurobiology, and Alzheimer's diseases. Neurobiol. Dis. *72 Pt A*, 3–12. https://doi.org/10.1016/j.nbd.2014.08.025.

Ikeda, U., Ikeda, M., Oohara, T., Oguchi, A., Kamitani, T., Tsuruya, Y., and Kano, S. (1991). Interleukin 6 stimulates growth of vascular smooth muscle cells in a PDGF-dependent manner. Am. J. Physiol. *260*, H1713–H1717. https://doi.org/10.1152/ajpheart.1991.260.5.H1713.

Ishibashi, S. (2001). Lipoprotein(a) and atherosclerosis. Arterioscler. Thromb. Vasc. Biol. *21*, 1–2. https://doi.org/10.1161/01.ATV.21.1.1.

Kasper, E.K., Agema, W.R., Hutchins, G.M., Deckers, J.W., Hare, J.M., and Baughman, K.L. (1994). The causes of dilated cardiomyopathy: a clinicopathologic review of 673 consecutive patients. J. Am. Coll. Cardiol. *23*, 586–590. https://doi.org/10.1016/0735-1097(94)90740-4.

Kathiresan, S., and Srivastava, D. (2012). Genetics of human cardiovascular disease. Cell *148*, 1242–1257. https://doi.org/10.1016/j.cell.2012.03.001.

Khera, A.V., and Kathiresan, S. (2017). Genetics of coronary artery disease: discovery, biology and clinical translation. Nat. Rev. Genet. *18*, 331–344. https://doi.org/10.1038/nrg.2016.160.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315. https://doi.org/10.1038/ng.2892.

Knowlton, K.U., Rockman, H.A., Itani, M., Vovan, A., Seidman, C.E., and Chien, K.R. (1995). Divergent pathways mediate the induction of ANF transgenes in neonatal and hypertrophic ventricular myocardium. J. Clin. Invest. *96*, 1311–1318. https://doi.org/10.1172/JCI118166.

Krzywinski, M., Schein, J., Birol, İ., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res. *19*, 1639–1645. https://doi.org/10.1101/gr.092759.109.

Lee, D.P., Tan, W.L.W., Anene-Nzelu, C.G., Lee, C.J.M., Li, P.Y., Luu, T.D.A., Chan, C.X., Tiang, Z., Ng, S.L., Huang, X., et al. (2019). Robust CTCF-based chromatin architecture underpins epigenetic changes in the heart failure stress–gene response. Circulation *139*, 1937–1956. https://doi.org/10.1161/CIRCULATIONAHA.118.036726.

Lee, G.-L., Chang, Y.-W., Wu, J.-Y., Wu, M.-L., Wu, K.K., Yet, S.-F., and Kuo, C.-C. (2012). TLR 2 induces vascular smooth muscle cell migration through cAMP response Element−Binding Protein−Mediated interleukin-6 production. Arterioscler. Thromb. Vasc. Biol. *32*, 2751–2760. https://doi.org/10.1161/ATVBAHA.112.300302.

Leung, D., Jung, I., Rajagopal, N., Schmitt, A., Selvaraj, S., Lee, A.Y., Yen, C.-A., Lin, S., Lin, Y., Qiu, Y., et al. (2015). Integrative analysis of haplotype-resolved epigenomes across human tissues. Nature *518*, 350–354. https://doi.org/10.1038/nature14217.

Li, W., Duren, Z., Jiang, R., and Wong, W.H. (2020). A method for scoring the cell type-specific impacts of noncoding variants in personal genomes. Proc. Natl. Acad. Sci. USA *117*, 21364–21372. https://doi.org/10.1073/pnas.1922703117.

Libby, P., Buring, J.E., Badimon, L., Hansson, G.K., Deanfield, J., Bittencourt, M.S., Tokgözoğlu, L., and Lewis, E.F. (2019). Atherosclerosis. Nat. Rev. Dis. Primers *5*, 56. https://doi.org/10.1038/s41572-019-0106-z.

Lu, Q., Powles, R.L., Abdallah, S., Ou, D., Wang, Q., Hu, Y., Lu, Y., Liu, W., Li, B., Mukherjee, S., et al. (2017). Systematic tissue-specific functional annotation of the human genome highlights immune-related DNA elements for late-onset Alzheimer's disease. PLoS Genet. *13*, e1006933. https://doi.org/10.1371/journal.pgen.1006933.

MacArthur, D.G., Manolio, T.A., Dimmock, D.P., Rehm, H.L., Shendure, J., Abecasis, G.R., Adams, D.R., Altman, R.B., Antonarakis, S.E., Ashley, E.A., et al. (2014). Guidelines for investigating causality of sequence variants in human disease. Nature *508*, 469–476. https://doi.org/10.1038/nature13127.

MacRae, C.A., and Vasan, R.S. (2016). The future of genetics and genomics: closing the phenotype gap in precision medicine. Circulation *133*, 2634–2639. https://doi.org/10.1161/CIRCULATIONAHA.116.022547.

Malekar, P., Hagenmueller, M., Anyanwu, A., Buss, S., Streit, M.R., Weiss, C.S., Wolf, D., Riffel, J., Bauer, A., Katus, H.A., and Hardt, S.E. (2010). Wnt signaling is critical for maladaptive cardiac hypertrophy and accelerates myocardial remodeling. Hypertension *55*, 939–945. https://doi.org/10.1161/HYPERTENSIONAHA.109.141127.

Man, J.C.K., van Duijvenboden, K., Krijger, P.H.L., Hooijkaas, I.B., van der Made, I., de Gier-de Vries, C., Wakker, V., Creemers, E.E., de Laat, W., Boukens, B.J., and Christoffels, V.M. (2021). Genetic dissection of a super enhancer controlling the *nppa-nppb* cluster in the heart. Circ. Res. *128*, 115–129. https://doi.org/10.1161/CIRCRESAHA.120.317045.

Marais, A.D. (2019). Apolipoprotein E in lipoprotein metabolism, health and cardiovascular disease. Pathology *51*, 165–176. https://doi.org/10.1016/j.pathol.2018.11.002.

Matsuoka, K., Asano, Y., Higo, S., Tsukamoto, O., Yan, Y., Yamazaki, S., Matsuzaki, T., Kioka, H., Kato, H., Uno, Y., et al. (2014). Noninvasive and quantitative live imaging reveals a potential stress-responsive enhancer in the failing heart. Faseb. J. *28*, 1870–1879. https://doi.org/10.1096/fj.13-245522.

Mattson, D.L., and Liang, M. (2017). From GWAS to functional genomics-based precision medicine. Nat. Rev. Nephrol. *13*, 195–196. https://doi.org/10.1038/nrneph.2017.21.

McKerns, M.M., Strand, L., Sullivan, T., Fang, A., and Aivazis, M.A.G. (2012). Building a framework for predictive science. Preprint at arXiv. https://doi.org/10.48550/arXiv.1202.1056.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. Presented at the Python in Science Conference, pp. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a.

McPherson, R., and Tybjaerg-Hansen, A. (2016). Genetics of coronary artery disease. Circ. Res. *118*, 564–578. https://doi.org/10.1161/CIRCRESAHA.115.306566.

Messerli, F.H., Rimoldi, S.F., and Bangalore, S. (2017). The transition from hypertension to heart failure. JACC. Heart Fail. *5*, 543–551. https://doi.org/10.1016/j.jchf.2017.04.012.

Michels, V.V., Moll, P.P., Miller, F.A., Tajik, A.J., Chu, J.S., Driscoll, D.J., Burnett, J.C., Rodeheffer, R.J., Chesebro, J.H., and Tazelaar, H.D. (1992). The frequency of familial dilated cardiomyopathy in a series of patients with idiopathic dilated cardiomyopathy. N. Engl. J. Med. *326*, 77–82. https://doi.org/10.1056/NEJM199201093260201.

Misra, A., Feng, Z., Chandran, R.R., Kabir, I., Rotllan, N., Aryal, B., Sheikh, A.Q., Ding, L., Qin, L., Fernández-Hernando, C., et al. (2018). Integrin beta3 regulates clonality and fate of smooth muscle-derived atherosclerotic plaque cells. Nat. Commun. *9*, 2073. https://doi.org/10.1038/s41467-018-04447-7.

Musunuru, K., and Kathiresan, S. (2016). Surprises from genetic analyses of lipid risk factors for atherosclerosis. Circ. Res. *118*, 579–585. https://doi.org/10.1161/CIRCRESAHA.115.306398.

Neumeyer, S., Hemani, G., and Zeggini, E. (2020). Strengthening causal inference for complex disease using molecular quantitative trait loci. Trends Mol. Med. *26*, 232–241. https://doi.org/10.1016/j.molmed.2019.10.004.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python (Presented at the J. Mach. Learn. Res.), pp. 2825–2830.

Phanstiel, D.H., Van Bortle, K., Spacek, D., Hess, G.T., Shamim, M.S., Machol, I., Love, M.I., Aiden, E.L., Bassik, M.C., and Snyder, M.P. (2017). Static and dynamic DNA loops form AP-1-bound activation hubs during macrophage development. Mol. Cell *67*, 1037–1048.e6. https://doi.org/10.1016/j.molcel.2017.08.006.

Ponikowski, P., Voors, A.A., Anker, S.D., Bueno, H., Cleland, J.G.F., Coats, A.J.S., Falk, V., González-Juanatey, J.R., Harjola, V.-P., Jankowska, E.A., et al.; ESC Scientific Document Group (2016). 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC)Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. Eur. Heart J. *37*, 2129–2200. https://doi.org/10.1093/eurheartj/ehw128.

R Development Core Team. (2010). A Language and Environment for Statistical Computing: Reference Index (R Foundation for Statistical Computing).

Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., et al.; Human Cell Atlas Meeting Participants (2017). The human cell atlas. Elife *6*, e27041. https://doi.org/10.7554/eLife.27041.

Ritchie, G.R.S., Dunham, I., Zeggini, E., and Flicek, P. (2014). Functional annotation of noncoding sequence variants. Nat. Methods *11*, 294–296. https://doi.org/10.1038/nmeth.2832.

Rosenbaum, A.N., Agre, K.E., and Pereira, N.L. (2020). Genetics of dilated cardiomyopathy: practical implications for heart failure management. Nat. Rev. Cardiol. *17*, 286–297. https://doi.org/10.1038/s41569-019-0284-0.

Satizabal, C.L., Samieri, C., Davis-Plourde, K.L., Voetsch, B., Aparicio, H.J., Pase, M.P., Romero, J.R., Helmer, C., Vasan, R.S., Kase, C.S., et al. (2018). *APOE* and the association of fatty acids with the risk of stroke, coronary heart disease, and mortality. Stroke *49*, 2822–2829. https://doi.org/10.1161/STROKEAHA.118.022132.

Schneider, J.G., Zhu, Y., Coleman, T., and Semenkovich, C.F. (2007). Macrophage β3 integrin suppresses hyperlipidemia-induced inflammation by modulating TNFα expression. Arterioscler. Thromb. Vasc. Biol. *27*, 2699–2706. https://doi.org/10.1161/ATVBAHA.107.153650.

Sergeeva, I.A., Hooijkaas, I.B., Ruijter, J.M., van der Made, I., de Groot, N.E., van de Werken, H.J.G., Creemers, E.E., and Christoffels, V.M. (2016). Identification of a regulatory domain controlling the Nppa-Nppb gene cluster during heart development and stress. Development (Camb.) *143*, 2135–2146. https://doi.org/10.1242/dev.132019.

Shachter, N.S. (2001). Apolipoproteins C-I and C-III as important modulators of lipoprotein metabolism. Curr. Opin. Lipidol. *12*, 297–304. https://doi.org/10.1097/00041433-200106000-00009.

Shang, L., Smith, J.A., and Zhou, X. (2020). Leveraging gene co-expression patterns to infer trait-relevant tissues in genome-wide association studies. PLoS Genet. *16*, e1008734. https://doi.org/10.1371/journal.pgen.1008734.

Skinnider, M.A., Squair, J.W., Kathe, C., Anderson, M.A., Gautier, M., Matson, K.J.E., Milano, M., Hutson, T.H., Barraud, Q., Phillips, A.A., et al. (2021). Cell type prioritization in single-cell data. Nat. Biotechnol. *39*, 30–34. https://doi.org/10.1038/s41587-020-0605-1.

Stram, D.O. (2004). Tag SNP selection for association studies. Genet. Epidemiol. *27*, 365–374. https://doi.org/10.1002/gepi.20028.

Sul, J.-Y., Wu, C. -w. K., Zeng, F., Jochems, J., Lee, M.T., Kim, T.K., Peritz, T., Buckley, P., Cappelleri, D.J., Maronski, M., et al. (2009). Transcriptome transfer produces a predictable cellular phenotype. Proc. Natl. Acad. Sci. USA *106*, 7624–7629. https://doi.org/10.1073/pnas.0902161106.

Finucane, H.K., Reshef, Y.A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P.R., Lareau, C., Shoresh, N., Genovese, G., Saunders, A., Macosko, E., Pollack, S., Brainstorm

Consortium, Perry, J.R.B., Buenrostro, J.D., Bernstein, B.E., Raychaudhuri, S., McCarroll, S., Neale, B.M., and Price, A.L. (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat. Genet. *50*, 621–629. https://doi.org/10.1038/s41588-018-0081-4.

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14, 000 cases of seven common diseases and 3, 000 shared controls. Nature *447*, 661–678. https://doi.org/10.1038/nature05911.

Vinkhuyzen, A.A.E., Wray, N.R., Yang, J., Goddard, M.E., and Visscher, P.M. (2013). Estimation and partition of heritability in human populations using whole-genome analysis methods. Annu. Rev. Genet. *47*, 75–95. https://doi.org/10.1146/annurev-genet-111212-133258.

Virtanen and Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F.C.P., Clarke, D., Gu, M., Emani, P., Yang, Y.T., et al. (2018a). Comprehensive functional genomic resource and integrative model for the human brain. Science *362*, eaat8464. https://doi.org/10.1126/science.aat8464.

Wang, L., Yu, P., Zhou, B., Song, J., Li, Z., Zhang, M., Guo, G., Wang, Y., Chen, X., Han, L., and Hu, S. (2020). Single-cell reconstruction of the adult human heart during heart failure and recovery reveals the cellular landscape underlying cardiac function. Nat. Cell Biol. *22*, 108–119. https://doi.org/10.1038/s41556-019-0446-7.

Wang, Y., Song, F., Zhang, B., Zhang, L., Xu, J., Kuang, D., Li, D., Choudhary, M.N.K., Li, Y., Hu, M., et al. (2018b). The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. Genome Biol. *19*, 151. https://doi.org/10.1186/s13059-018-1519-9.

Wang, Z., and Newman, W.H. (2003). Smooth muscle cell migration stimulated by interleukin 6 is associated with cytoskeletal reorganization. J. Surg. Res. *111*, 261–266. https://doi.org/10.1016/S0022-4804(03)00087-8.

Ward, L.D., and Kellis, M. (2016). HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. Nucleic Acids Res. *44*, D877–D881. https://doi.org/10.1093/nar/gkv1340.

Warren, S.A., Terada, R., Briggs, L.E., Cole-Jeffrey, C.T., Chien, W.-M., Seki, T., Weinberg, E.O., Yang, T.P., Chin, M.T., Bungert, J., and Kasahara, H. (2011). Differential role of nkx2-5 in activation of the atrial natriuretic factor gene in the developing versus failing heart. Mol. Cell Biol. *31*, 4633–4645. https://doi.org/10.1128/MCB.05940-11.

Watanabe, K., Umićević Mirkov, M., de Leeuw, C.A., van den Heuvel, M.P., and Posthuma, D. (2019). Genetic mapping of cell type specificity for complex traits. Nat. Commun. *10*, 3222. https://doi.org/10.1038/s41467-019-11181-1.

Wierbowski, S.D., Fragoza, R., Liang, S., and Yu, H. (2018). Extracting complementary insights from molecular phenotypes for prioritization of disease-associated mutations. Curr. Opin. Syst. Biol. *11*, 107–116. https://doi.org/10.1016/j.coisb.2018.09.006.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for comparing biological themes among gene clusters. OMICS A J. Integr. Biol. *16*, 284–287. https://doi.org/10.1089/omi.2011.0118.

Yue, L., Bian, J.-T., Grizelj, I., Cavka, A., Phillips, S.A., Makino, A., and Mazzone, T. (2012). Apolipoprotein E enhances endothelial-NO production by modulating caveolin 1 interaction with endothelial NO synthase. Hypertension *60*, 1040–1046. https://doi.org/10.1161/HYPERTENSIONAHA.112.196667.

Zhang, B., Zhang, Y., Li, R., Li, Y., and Yan, W. (2021). Knockdown of circular RNA hsa_circ_0003204 inhibits oxidative stress and apoptosis through the miR-330-5p/Nod2 axis to ameliorate endothelial cell injury induced by low-density lipoprotein. Cent. Eur. J. Immunol. *46*, 140–151. https://doi.org/10.5114/ceji.2021.108174.

Zhou, X., Chen, Y., Mok, K.Y., Kwok, T.C.Y., Mok, V.C.T., Guo, Q., Ip, F.C., Chen, Y., Mullapudi, N.; Alzheimer's Disease Neuroimaging Initiative, and Sullivan, P.F., et al. (2019). Non-coding variability at the APOE locus contributes to the Alzheimer's risk. Nat. Commun. *10*, 3310. https://doi.org/10.1038/s41467-019-10945-z.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
GTEx data	dbGaP	phs000424.v7.p2
GWAS summary statistics of CVD	GWAS Catalog	https://www.ebi.ac.uk/gwas/api/search/downloads/full
scRNA-seq data from normal heart tissues	Gene Expression Omnibus	GEO: GSE109816
scRNA-seq data from disease heart tissues	Gene Expression Omnibus	GEO: GSE121893
Software and algorithms		
GRPath	This paper	https://github.com/xixi-cathy/GRPath
OpenCausal	Li et al. (2020)	https://github.com/liwenran/OpenCausal
biomaRt (Ensembl GRCh37 release 105)	Durinck et al. (2009)	https://grch37.ensembl.org/biomart/martview/5504fcd92011bb08f7d23da941f8bd54
Augur	Skinnider et al. (2021)	https://github.com/neurorestore/Augur
numpy 1.19.4	Harris et al. (2020)	https://numpy.org/
pandas 1.1.5	McKinney (2010)	https://pandas.pydata.org/
multiprocess 0.70.12.2	McKerns et al., 2012	https://uqfoundation.github.io/project/pathos
scikit-learn 0.24.1	Pedregosa et al. (2011)	https://scikit-learn.org/
scipy 1.6.0	Virtanen and Gommers et al. (2020)	https://scipy.org/
networkx 2.1	Hagberg et al. (2008)	https://networkx.org/
stats 3.6.1	R Development Core Team (2010)	http://www.R-project.org/
Seurat 3.2.3	Butler et al. (2018)	https://satijalab.org/seurat/
clusterProfiler 3.14.3	Yu et al. (2012)	https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html
HaploReg v4.1	Ward and Kellis (2016)	https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php
3D Genome Browser	Wang et al. (2018a), 2018b	http://3dgenome.fsm.northwestern.edu/view.php
Circos	Krzywinski et al. (2009)	http://circos.ca/
BioRender	BioRender	https://biorender.com/

RESOURCE AVAILABILITY

Lead contact

Further information and requests should be directed to and will be fulfilled by the lead contact, Xuegong Zhang (zhangxg@tsinghua.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- All original code has been deposited on GitHub and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information is available from the lead contact upon reasonable request.

METHOD DETAILS

Predict heart-specific openness scores

Since GTEx did not provide chromatin accessibility information, we need to first predict chromatin accessibility state from current WGS and RNA-seq data. Li et al. predicted openness scores of 2,965,129 REs by first training a regression model based on paired RNA-seq and ATAC-seq data from the ENCODE project, then applying the trained model on GTEx samples (Li et al., 2020). Each RE was a 500 bp genomic region centering around a peak from ENCODE project, and its corresponding background was a 1 Mb genomic region centering around the same peak. The openness score was a continuous value, quantifying the relative openness of a 500 bp peak region compared with its corresponding 1 Mb background region. We applied this model on 635 GTEx heart samples from 357 donors, and obtained the predicted openness scores. More specifically, we constructed an average TF expression profile for each of the two tissue types (left heart ventricle or atrial appendage), and then combined with the WGS to predict accessibility for each sample in each donor.

Decide pcRegions and pcVariants for CVD

The second step was to decide pcRegions that contain pcVariants of CVD, instead of regions that are just associated with CVD. This step was further separated into some sub-steps. We first defined some candidate genomic regions according to prior knowledge. Then, we randomly separated the labeled donors into two groups to evaluate the causal effect of variants in each candidate genomic region as well as the causal effect of these candidate genomic regions on CVD. Finally, we repeated the random sampling process, and defined pcRegions and pcVariants.

Define candidate genomic regions

There are 340 SNPs significantly associated with CVD according to GWAS summary statistics. We first expanded each SNP to a 200 kb region centering around it, and then selected regions containing variants that are *cis*-eQTLs in heart left ventricle or atrial appendage. There were 338 candidate genomic regions that satisfied the above criteria, and we regarded them as candidate genomic regions for further analysis. We used web-based software *biomaRt* (Durinck et al., 2009) to convert human reference genome version in GWAS from GRCh38 to GRCh37 to be compatible with GTEx samples.

Evaluate causal effects of variants in each candidate genomic region

After we defined candidate genomic regions, we quantified and prioritized the causal effect of the variants on CVD. We used Python packages *numpy* (Harris et al., 2020), *pandas* (McKinney, 2010) and *scikit-learn* (Pedregosa et al., 2011) in this part.

First, we selected a subset of donors for this task. We referred to "DTHCOD" (the direct cause of death) and "DTHDUCOD" (the first underlying cause of death) information of GTEx donors, and labeled 90 donors whose deaths were related to CVD (e.g., cardiovascular disease, myocardial infarction, etc.) as "CVD positive", and another 60 donors whose deaths were caused by accidents (e.g., motor vehicle accident, drug intoxication, etc.) as "CVD negative". Since other donors have other diseases, which might introduce some confounding factors into the model, we abandoned these samples to minimize potential influence.

Then, in each candidate genomic region, we filtered out variants with less than 10 reference or alternate allele donors, and calculated *VCS* score for each remaining variant. The *k*th variant's *VCS* score is defined as

$$VCS_k = |\omega_k \cdot \lambda_k|.$$

To avoid information leakage, we randomly separated all labeled donors into five folds. We used two folds of the labeled donors to obtain the weights $\boldsymbol{\omega}$, and the other three folds to calculate $\boldsymbol{\lambda}$ and ***VCS*** score.

λ_k is defined as

$$\lambda_k = \left| \frac{\sum_{d=1}^{D_1} O_d}{D_1} - \frac{\sum_{d=1}^{D_2} O_d}{D_2} \right|,$$

O_d is the openness score of RE where variant k locates, D_1 is the number of donors with alternate alleles, and D_2 is the number of donors with reference allele. Intuitively, λ_k quantifies the influence of variant k on chromatin accessibility of corresponding RE.

To define ω_k, let $\boldsymbol{\omega}$ be the weight vector of all features in a logistic regression model, where we used the openness scores of the REs in this candidate genomic region to classify donors with or without CVD. The objective function of this logistic regression model is

$$\min_{\omega,\ c} \frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega}+C\sum_{d=1}^{D}\log\left(\exp\left(-\ y_d\left(\boldsymbol{O}_d{}^T\boldsymbol{\omega}+c\right)\right)+1\right),$$

where D is the total number of the 2/5 labeled donors, $\boldsymbol{O}_d$ is the vector of RE openness scores in this candidate genomic region, and y_d is the binary label (CVD positive/CVD negative) of donor d. ω_k is the component in $\boldsymbol{\omega}$ which corresponds to the RE where the variant k locates. Intuitively, ω_k quantifies the influence of chromatin accessibility of corresponding RE on CVD.

In summary, the *VCS* score integrates the effect of the variant on RE openness and the effect of RE openness on CVD, to evaluate the relative causal effect of the variant within the candidate genomic region. Variants with higher *VCS* scores should have a stronger causal effect on CVD.

It is worth noticing that calculating λ and *VCS* score does not require disease phenotype information, so we could still save the 3/5 labeled donors for downstream analysis.

Evaluate causal effects of candidate genomic regions

Based on variants prioritization results, we further evaluated the causal effect of the candidate genomic regions on CVD.

For each variant, we could obtain the following statistics:

	CVD positive	CVD negative
Number of alternate-allele donors	a	b
Number of reference-allele donors	c	d

Then, we could calculate the odds ratio (OR) for each variant, which is defined as

$$\text{OR} = \frac{a/b}{c/d}.$$

OR quantifies the causal effect of each candidate genomic region on CVD, and we used the 3/5 labeled donors whose disease phenotype information had not been used in the previous step to calculate OR.

We reason that if a candidate region includes one or more REs mediating the causal effects in a tissue-specific manner, then ORs of the variants in these REs should be different from those outside. Thus, if the ORs of top-ranked variants are significantly higher or lower than bottom-ranked variants in a candidate genomic region, causal variants should exist in this region, which shows alternate-allele-pathogenic or reference-allele-pathogenic effect on CVD, respectively. Based on this hypothesis, we performed one-tail Wilcoxon rank-sum tests to compare ORs of top-ranked variants and the same number of bottom-ranked variants in each candidate genomic region. We tested for "ORs of top-ranked variants are higher than the same number of bottom-ranked variants" and "ORs of top-ranked variants are lower than the same number of bottom-ranked variants", which correspond to alternate-allele-pathogenic and reference-allele-pathogenic scenario respectively, and calculated corresponding p-values. Since the number of causal variants in each region should be different, we performed hypothesis testing for top- and bottom- 2%~50% variants separately, and kept the lowest p-value in each scenario as the quantification of alternate-allele-pathogenic and reference-allele-pathogenic effect on CVD of this region. We also kept the top-ranked variants corresponding to the lowest p-value.

Call pcRegions and pcVariants

To call pcRegions and pcVariants of CVD, we repeated the above two steps ("Evaluate causal effects of variants in each candidate genomic region" and "Evaluate causal effects of candidate genomic regions") for

30 times. In each time, we resampled the labeled donors, and obtained p-values with corresponding top-ranked variants. We used Fisher's method to combine the 30 p-values in alternate-allele-pathogenic and reference-allele-pathogenic cases, respectively. Then, we performed Bonferroni test on Fisher's p-values across all candidate genomic regions in multiple test correction, regarded the regions with false discovery rate (FDR) lower than 1×10^{-9} as pcRegions, and further distinguished them as alternate-allele-pathogenic or reference-allele-pathogenic regions. In pcRegions, we defined variants which appeared more than one time in the top-ranked variants list in the repetitive tests as pcVariants. We used Python package *multiprocess* (McKerns et al., 2012) to accelerate computational process, Python package *scipy* (Virtanen and Gommers et al., 2020) and R package *stats* (R Development Core Team, 2010) to perform statistical analysis.

Find noteworthy cell types for two types of heart failure

The third step was to identify disease-noteworthy cell types, and to find out the roles that pcVariants play in these cell types. We bridged the gap between pcVariants and cell types through gene expression. By leveraging eQTL studies and scRNA-seq data, we were able to first relate pcVariants to gene expressions of eGenes, then relate these pcGenes to certain heart cell types in a more specific type of CVD.

We collected SMCs, ECs, FBs, MPs, and CMs from normal, dHF, and cHF donors. By comparing gene expression changes between normal and disease cells, we observed which cell types were most affected, and the importance of the pcGenes in different cell types in this disease. Following the idea proposed by Skinnider et al. (Skinnider et al. (2021), we turned this biological problem into a classification problem. We assumed that if cells in normal and disease status from the same cell type were better classified, this cell type should be more affected by this disease, and vice versa. We applied a random forest model for classification, and used AUROC to prioritize cell types. The comparison of AUROCs among different cell types revealed which cell type was more affected, and the corresponding feature coefficients in each cell type suggested the importance of each gene in these cell types. If we only use one feature (gene) to do classification, then the AUROC quantifies the importance of this gene to different cell types in this disease. This part was implemented using Python package *scikit-learn* (Pedregosa et al., 2011).

Take dHF as an example, after quality control and normalization, we integrated 7,418 normal cells and 2,728 dHF cells from five cell types. Based on R package Seurat (Butler et al., 2018), 2000 highly variable genes were selected for batch correction and data integration. Then, taking each highly variable pcGene as the feature, we applied random forest classifiers on the processed gene expression matrix for each cell type, and obtained the classification (dHF versus normal) AUROCs, which compared the importance of each pcGene in different cell types. We repeated the classification process for 10 times per highly variable pcGene, and took their average as the final AUROC.

QUANTIFICATION AND STATISTICAL ANALYSIS

In deciding pcRegions, we used one-tail Wilcoxon rank-sum tests to compare ORs of top-ranked variants and the same number of bottom-ranked variants in each candidate genomic region. We then used Fisher's method to combine the 30 p-values for each region in repetitive tests, and performed Bonferroni test on Fisher's p-values across all candidate genomic regions in multiple test correction. We defined significance as adjusted p-value less than 1×10^{-9}. Python package *scipy* (Virtanen and Gommers et al., 2020) was used in one-tail Wilcoxon rank-sum tests and in combining p-values. R package *stats* (R Development Core Team, 2010) was used in multiple test correction.

In downstream enrichment analysis, we used R package *clusterProfiler* (Yu et al., 2012), and defined significance as FDR less than 0.05. The web-based visualization tool *Circos* (Krzywinski et al., 2009) and Python package *networkx* (Hagberg et al., 2008) were used for visualization.