JBC RESEARCH ARTICLE

# Prioritizing cardiovascular disease-associated variants altering NKX2-5 and TBX5 binding through an integrative computational approach

Received for publication, September 13, 2023, and in revised form, October 18, 2023 Published, Papers in Press, November 4, 2023,
https://doi.org/10.1016/j.jbc.2023.105423

Edwin G. Peña-Martínez[1,‡], Diego A. Pomales-Matos[1,‡], Alejandro Rivera-Madera[2], Jean L. Messon-Bird[1], Joshua G. Medina-Feliciano[1], Leandro Sanabria-Alberto[1], Adriana C. Barreiro-Rosario[1], Jeancarlos Rivera-Del Valle[1], Jessica M. Rodríguez-Ríos[1], and José A. Rodríguez-Martínez[1,*]

*From the [1]Department of Biology, University of Puerto Rico Río Piedras Campus, San Juan, Puerto Rico; [2]Department of Biology, University of Puerto Rico Cayey Campus, Cayey, Puerto Rico*

Reviewed by members of the JBC Editorial Board. Edited by Brian D. Strahl

**Cardiovascular diseases (CVDs) are the leading cause of death worldwide and are heavily influenced by genetic factors. Genome-wide association studies have mapped >90% of CVD-associated variants within the noncoding genome, which can alter the function of regulatory proteins, such as transcription factors (TFs). However, due to the overwhelming number of single-nucleotide polymorphisms (SNPs) (>500,000) in genome-wide association studies, prioritizing variants for *in vitro* analysis remains challenging. In this work, we implemented a computational approach that considers support vector machine (SVM)-based TF binding site classification and cardiac expression quantitative trait loci (eQTL) analysis to identify and prioritize potential CVD-causing SNPs. We identified 1535 CVD-associated SNPs within TF footprints and putative cardiac enhancers plus 14,218 variants in linkage disequilibrium with genotype-dependent gene expression in cardiac tissues. Using ChIP-seq data from two cardiac TFs (NKX2-5 and TBX5) in human-induced pluripotent stem cell-derived cardiomyocytes, we trained a large-scale gapped k-mer SVM model to identify CVD-associated SNPs that altered NKX2-5 and TBX5 binding. The model was tested by scoring human heart TF genomic footprints within putative enhancers and measuring *in vitro* binding through electrophoretic mobility shift assay. Five variants predicted to alter NKX2-5 (rs59310144, rs6715570, and rs61872084) and TBX5 (rs7612445 and rs7790964) binding were prioritized for *in vitro* validation based on the magnitude of the predicted change in binding and are in cardiac tissue eQTLs. All five variants altered NKX2-5 and TBX5 DNA binding. We present a bioinformatic approach that considers tissue-specific eQTL analysis and SVM-based TF binding site classification to prioritize CVD-associated variants for *in vitro* analysis.**

Cardiovascular diseases (CVDs) are the leading cause of death worldwide and encompass multiple disorders (coronary artery disease, congenital heart disease, stroke, etc.), many of which are heritable (1–5). Genome-wide associations studies (GWAS) have mapped over 90% of CVD-associated variants within noncoding regions of the genome (6, 7). Noncoding single nucleotide polymorphisms (SNPs) can impact phenotype by altering gene regulatory mechanisms, such as transcription factor (TF)-DNA binding and gene expression (8–11). NKX2-5 and TBX5 are cardiac TFs that regulate gene expression in the developing heart (12–17). Previous research has identified CVD-associated SNPs that alter cardiac TF-DNA binding, but further research is required to establish causality (18–22). However, with the overwhelming number of GWAS SNPs (>500,000), prioritizing candidate CVD-causing variants for experimental validation remains challenging.

One approach to address this challenge is the integration of functional genomic datasets with predictive models to identify variants that create or disrupt TF binding sites (TFBSs) (23–26). Large-scale gapped k-mer (LS-GKM) support vector machine (SVM) models can be trained to predict TFBSs by using *in vitro* genome-wide binding data from chromatin immunoprecipitation followed by DNA sequencing (ChIP-seq). LS-GKM-SVM models have several advantages over traditional position weight matrix (PWM)-based methods, by considering complex sequence features like dinucleotide interactions, longer/gapped k-mers, and intracellular patterns (27–30). LS-GKM-SVM models can be trained with ChIP-seq data from specific cell lines or tissue to integrate relevant epigenomic and regulatory context (23).

In this work, we prioritize noncoding variants associated with CVDs and CV traits using public data from the GWAS catalog (31), genome tissue expression (GTEx) portal (32), ENCODE (33), ChIP-Atlas (34), and Remap (35). We compiled a list of CVD-associated SNPs linked with a genotype-dependent gene expression in cardiac tissue and trained a LS-GKM-SVM model with ChIP-seq data from NKX2-5 and TBX5 in human-induced PSC-derived cardiomyocytes (HiPSC-CMs). Both models were used to score previously identified heart DNase I hypersensitivity genomic footprints (DGFs) (36) that colocalize within putative cardiac enhancers

[‡] These authors contributed equally to this work.
[*] For correspondence: José A. Rodríguez-Martínez, jose.rodriguez233@upr.edu.

(37) and tested them through *in vitro* binding by electrophoretic mobility shift assay (EMSA). Our predictive model was successful at identifying NKX2-5 and TBX5 binding sites and distinguishing between DNA sequences with different binding affinities.

Having validated DGF scored by the GKM-SVM model, we calculated the impact of CVD-associated SNPs in NKX2-5 and TBX5 binding. Five variants (rs59310144, rs6715570, rs61872084, rs7612445, and rs7790964) predicted to alter NKX2-5 and TBX5 binding were prioritized for *in vitro* validation based on the predicted change in binding and expression quantitative trait loci (eQTLs) in cardiac tissue. The five variants were validated through EMSA and resulted in changes in NKX2-5 and TBX5 DNA binding. In short, we present a bioinformatic approach that considers tissue-specific eQTL analysis and SVM-based TFBS classification to prioritize functional CVD-associated SNPs.

## Results

### Identification of CVD-associated SNPs in cardiac eQTL

To identify potential CVD-causing SNPs, we downloaded the GWAS catalog and filtered the data to keep cardiovascular disease or trait-associated SNPs (*e.g.*, congenital heart defects, cardiomyocyte differentiation, stroke, arrhythmia, etc.; full list of SNPs in File S1). We then intersected the CVD-associated SNPs with a catalog of putative fetal and adult heart enhancers and genomic footprints of fetal hearts, resulting in 1535 genomic variants. The CVD-associated SNP set was expanded to include SNPs in linkage disequilibrium (LD $r^2 > 0.8$) from diverse populations (EUR, AFR, SAS, EAS, and AMR) and resulted in 14,218 unique SNPs occurring in one or more populations. To evaluate the potential of these SNPs to be biologically relevant in cardiovascular biology, we analyzed gene expression patterns in cardiac tissue with the previously identified variants in the GTEx portal. We found 792 genes with genotype-dependent activity in the heart atrial appendage or left ventricle associated with the previously identified SNPs. The workflow is illustrated in Fig. 1*A*, and the list of SNPs associated with differentially expressed genes in cardiac tissue is in File S1. We identified chromosomes with a higher frequency of CVD-associated SNPs, with >1000 variants (chromosomes 1 and 6) and ∼500 (chromosomes 2, 3, 7, 10, 11, 12, 15, 16, 17, 19, and 22), including those in LD (Fig. 1*B*). Chromosomes with a high SNP frequency may have variants evenly distributed among them, like chromosomes 1 and 2, while others contain multiple variants in the same (or near) loci, like chromosomes 6, 10, 15, and 22 (Fig. 1*C*). We also analyzed data from the GTEx database and found genes with genotype-dependent expression in cardiac tissue (heart atrial appendage and left ventricle) containing the identified CVD-associated SNPs or the variants in LD. We identified 31,122 SNP-gene pairs (792 unique genes) with genotype-dependent expression in cardiac tissue (Fig. 1*D*).

### Training LS-GKM SVM model to score NKX2-5 and TBX5 binding sites

We trained a LS-GKM SVM model to prioritize CVD-associated SNPs that alter DNA binding by TFs known to play important roles in heart development and biology. The models were trained using NKX2-5 and TBX5 ChIP-seq data from HiPSC-CMs. The 1000 top-scoring ChIP-seq peaks were used as a positive training set, while unbound sequences of the same length, GC content, and chromosome were used as negative training (Fig. 2*A*) The best-performing LS-GKM SVM classifier model trained with NKX2-5 ChIP-seq data (SRX9284027) (38) obtained an AUROC value of 0.955 and an AUPRC value of 0.954. The best TBX5 (SRX2023721) (39) model obtained an AUROC value of 0.921 and an AUPRC value of 0.912 (Fig. S1, *A* and *B*). The models were used to score all possible 2,097,152 nonredundant 11 bp oligomers (11-mers). The 11-mer scores were sorted, and the 1000 top-scoring sequences were used to generate a PWM using multiple Em for motif elicitation and observed similar motifs to those previously reported for NKX2-5 (40, 41) and TBX5 (42) (Fig. S1, *C* and *D*). We proceeded to score ∼520,000 fetal heart DGF that occur heart enhancers to identify genomic loci potentially bound by NKX2-5 or TBX5 (Fig. 2*B*). We then chose the DNA sequences with the highest, middle, and lowest scores to test for *in vitro* binding through EMSA (Figs. 2*C* and S2). As predicted by the GKM SVM model, sequences with a higher score resulted in increased NKX2-5 and TBX5-DNA binding.

### CVD-associated SNPs alter NKX2-5 and TBX5-DNA binding

After successful training and validation of the LS-GKM SVM predictive model, we proceeded to score the 14,218 SNPs to prioritize variants that may directly contribute to CVDs. Both reference and alternate allele sequences were scored to predict fold change (deltaSVM score) of TF-DNA binding. We selected five SNPs (rs59310144, rs6715570, rs61872084, rs7612445, and rs7790964) that deltaSVM predicted significant change in NKX2-5 or TBX5 binding and are associated with genotype-dependent expression in cardiac tissue (Fig. 3*A* and Table S2). When evaluated through EMSA, we observed a difference in NKX2-5 and TBX5 DNA binding between reference and alternate for all five SNPs (Figs. 3, *B* and *C* and S3). Variants rs59310144 and rs61872084 resulted in a decrease in NKX2-5 DNA binding, while rs6715570 increased binding. Variant rs7612445 resulted in an increase in TBX5 binding, while rs7790964 decreased binding. We found that all five SNPs were in eQTLs described in cardiac tissue and identified five genes with genotype-dependent expression in the heart atrial appendage or left ventricle (Fig. 3*D*).

## Discussion

As we continue to research the genetic basis for human disease, the number of identified functional/causal noncoding SNPs continues to grow. Understanding and prioritizing SNPs that contribute to the disease phenotypes is essential. However, there is a lack of consensus or bioinformatic protocol to

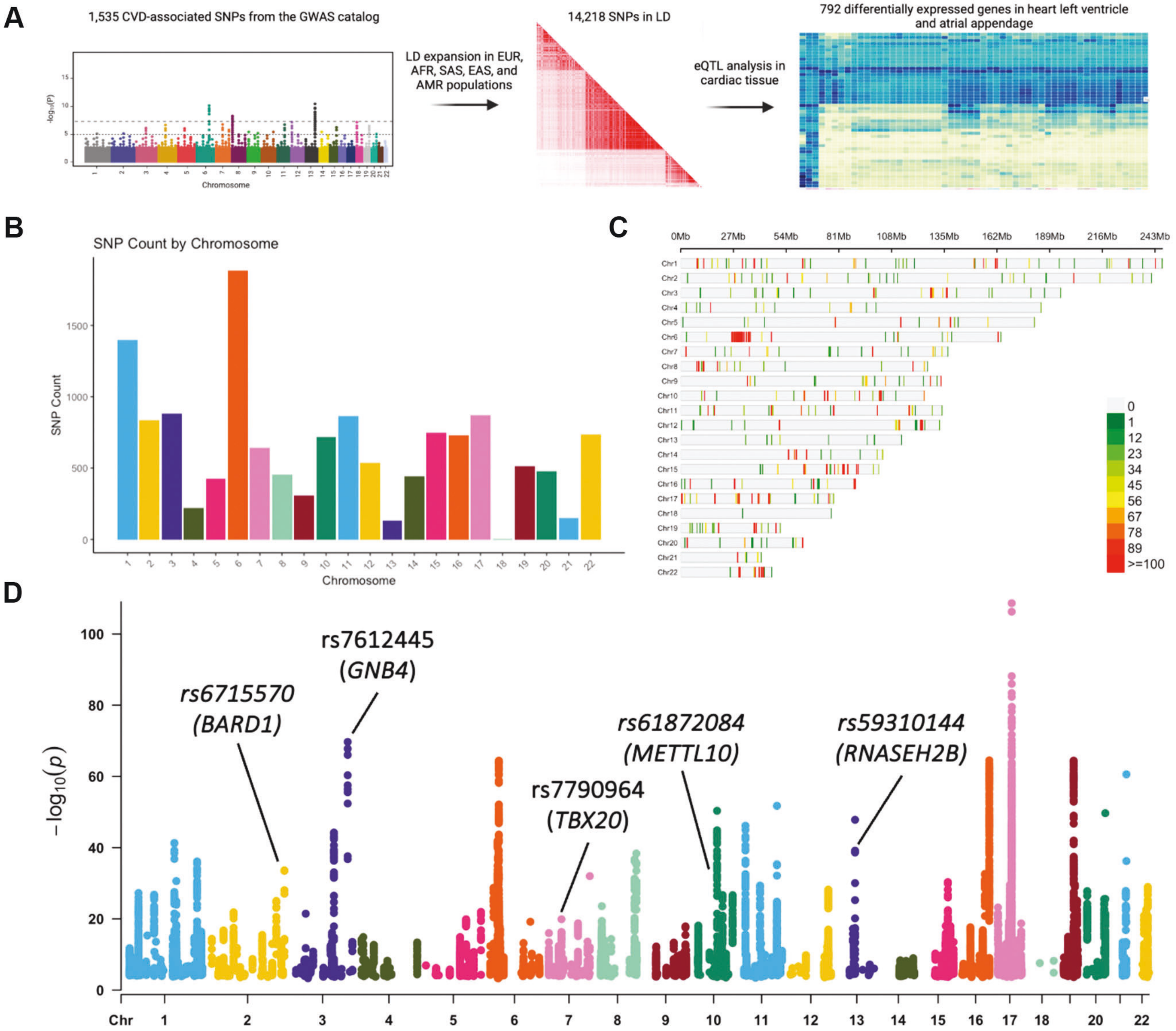

**Figure 1. Identification of CVD-associated SNPs.** *A*, pipeline to identify potential CVD-causing SNPs. *B*, number of CVD-associated SNPs per chromosome. *C*, distribution of SNP frequency within autosomal chromosome, binned by 1 Mb windows. *D*, SNP-gene pairs with genotype-dependent expression in cardiac tissue. Each dot represents a SNP-gene pair with genotype-dependent expression in the heart atrial appendage or left ventricle in one or more populations. rs6715570-BARD1, rs61872084-METTL10, rs59310144-RNASEH2B, rs7612445-GNB4, and rs7790964-TBX20 are SNP-gene pairs that were evaluated *in vitro* in Figure 3. CVD, cardiovascular disease; LD, linkage disequilibrium; SNP, single nucleotide polymorphism.

prioritize noncoding SNPs that are biologically relevant in the development of human diseases (25). To address this challenge, we applied a GKM-SVM–based model to identify and prioritize potential CVD-causing variants for experimental validation. We leveraged on public data from the GWAS catalog and extracted SNPs that were associated with cardiovascular disease or traits and included variants in LD from multiple populations (EUR, AFR, SAS, EAS, and AMR). The distribution of the identified CVD-associated SNPs was not uniform throughout the genome (Fig. 1, *B* and *C*). This suggests that certain chromosomes, or specific loci, are enriched with CVD-associated SNPs and may contribute to the cardiac phenotype. SNPs in eQTL were identified to prioritize variants with genotype-dependent expression in cardiac tissue (left ventricle and atrial appendage). Through this approach, we aimed to narrow the extensive list of noncoding variants and identify SNPs that may contribute to CVDs.

To further prioritize variants that may contribute to a cardiac phenotype, we trained a GKM-SVM classifier to identify SNPs that disrupt NKX2-5 and TBX5-DNA binding. Because TFs regulate gene expression in a tissue-specific manner (43, 44), we trained the predictive model with ChIP-seq data from NKX2-5 and TBX5 collected in HIPSC-CM. After testing the model by scoring all possible 11-mers, we generated PWMs for both models resulting in DNA binding motifs in agreement with previously described motifs for NKX2-5 and TBX5 (Fig. S1) (40–42). The GKM-SVM model was successful at identifying binding sites within heart DGF for both NKX2-5 DBD and full-length TBX5 (Fig. 2, *B* and *C*). Our results suggest that the LS-GKM SVM model will be able to

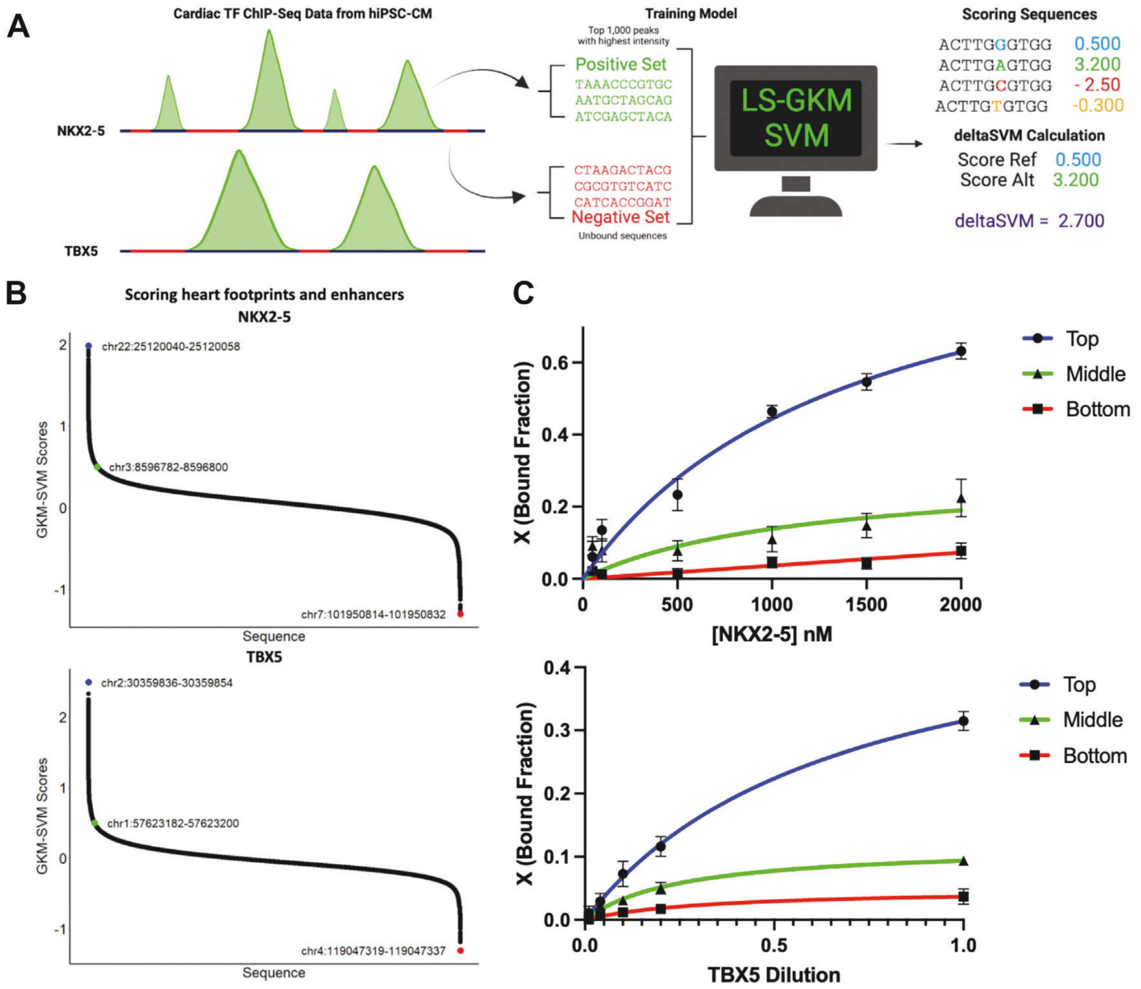

**Figure 2. Training and testing of LS-GKM SVM predictive model**. *A*, schematic of model training with NKX2-5 and TBX5 ChIP-seq data from HiPSC-CM. *B*, scoring of ∼520,000 DGF that occur in heart enhancers with the NKX2-5 (*top*) and TBX5 (*bottom*) predictive models. *C*, *in vitro* testing of predictive model for highest, middle, and lowest scored sequences for NKX2-5 (*top*) and TBX5 (*bottom*). For NKX2-5, we tested chr22:25,120,040 to 25,120,058 (*circle* with *blue line*), chr3:8,596,782 to 8,596,800 (*triangle* with *green line*), and chr7:101,950,814 to 101,950,832 (*square* with *red line*). For TBX5, we tested chr2:30,359,836 to 30,359,854 (*circle* with *blue lines*), chr1:57,623,182 to 57,623,200 (*triangle* with *green line*), and chr4:119,047,319 to 119,047,337 (*square* with *red line*). HiPSC-CM, human-induced PSC-derived cardiomyocytes; LS-GKM, large-scale gapped k-mer; SVM, support vector machine; TF, transcription factor.

successfully predict changes in binding affinity between reference and variant DNA sequences that alter cardiac TF-DNA binding. Functional genomics studies have identified millions of putative enhancers, but the identity of most bound TFs is unknown (33). Our study can contribute to identifying transcriptional regulators active in specific tissues or environments.

We tested five SNPs that altered NKX2-5- (rs59310144, rs6715570, and rs61872084) and TBX5-(rs7612445 and rs7790964) DNA binding activity and associated with genotype-dependent expression (*RNASEH2B*, *BARD1*, *METTL10*, *GNB4*, and *TBX20*, respectively) in cardiac tissue (Fig. 3, *C* and *D*). Four of the variants (rs6715570, rs61872084, rs76122445, and rs7790964) effect allele's impact on gene expression were proportional to the changes in TF-DNA binding activity. Inversely, the effect allele of variant rs59310144 increased gene expression, while TF-DNA binding decreased. *RNASEH2B*, *BARD1*, *GNB4*, and *TBX20* have been previously identified to be differentially expressed in the heart atrial appendage when variants rs59310144, rs6715570, rs7612445, and rs7790964 (respectively) occur. *RNASEH2B*, which has been previously found to be differentially expressed in CVD risk events, is upregulated when the alternate allele of variant rs59310144 is present (45). NKX2-5 can act as either an activator or repressor (46–48). Benaglio *et al.* (2019) used a combination of EMSA, luciferase reporter assays, and CRISPRi to describe how the stronger TF-binding allele (rs590041) was associated with reduced *SSBP3* expression. In contrast, the stronger TF binding allele (rs3807989) increased CAV1 and CAV2 expression (19). This suggests that disruption of the NKX2-5 binding site by rs59310144 could increase *RNASEH2B* expression. *BARD1* has also been identified as upregulated when the alternate allele of variant rs6715570 occurs in the heart atrial appendage. Copy number alterations in the *BARD1* locus have been associated with developmental delays, including coarctation of the aorta during early organogenesis and heart development (49). Variant rs61872084 has been identified in the heart's left ventricle when *METTL10* (methyltransferase-like protein 10) is downregulated when the alternate allele occurs. Accumulation of METTL10 methylated

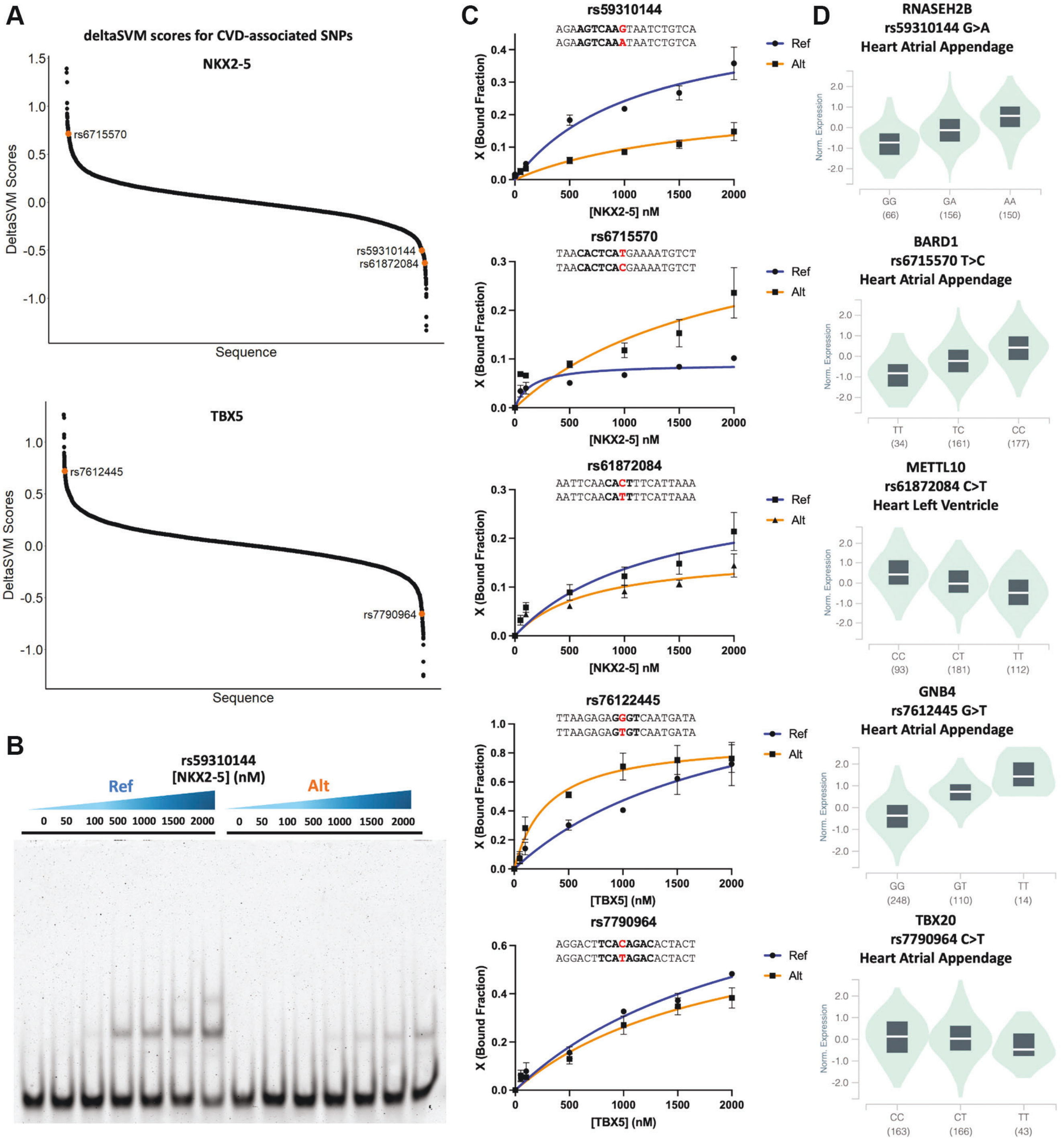

**Figure 3. CVD-associated SNPs alter NKX2-5 and TBX5 *in vitro* binding**. *A*, DelstaSVM score distribution of the 14,218 CVD-associated SNPs for NKX2-5 (*top*) and TBX5 (*bottom*). *B*, representative EMSA gel for rs59310144 reference/noneffect (Ref) and alternate/effect (Alt) alleles. Source image is available in Fig. S3. *C*, binding curves of rs59310144, rs6715570, rs61872084, rs76122445, and rs7790964. Experiments were performed in triplicates and binding curves show average bound fraction (X) and error bars are standard error. *D*, cardiac tissue eQTL analysis of *RNASEH2B*, *BARD1*, *METTL10*, *GNB4*, and *TBX20* expressed in heart atrial appendage or left ventricle when rs59310144, rs6715570, rs61872084, rs76122445, and rs7790964 occur, respectively. CVD, cardiovascular disease; SNP, single nucleotide polymorphism.

products, such as S-adenosyl-L-methionine, S-adenosyl-L-homocysteine, and homocysteine have been correlated with kidney dysfunction and CVDs in patients with type 2 diabetes (50). G protein subunit beta 4 (GNB4) has been previously associated with the regulation of heart rate stability and identified in patients with higher cardiovascular mortality risk (51). TBX20, a cardiac TF regulated by TBX5, is essential for proper heart development with mutations that have been associated with congenital heart diseases (52).

Our findings suggest that NKX2-5 regulation of the *RNASEH2B* (repression), *BARD1* (activation), and *METTL10* (activation), and TBX5 regulation of *GNB4* (activation) and *TBX20* (activation), genes are possible mechanisms that can be further explored to establish rs59310144, rs6715570, rs61872084, rs76122445, and rs7790964 as causal CVD risk-variants. Although the etiology of human diseases is complex and multifactorial, this approach can provide crucial information that can be implemented during *in vivo* experiments or clinical research to address genetic diseases caused by noncoding SNPs. In summary, we believe this bioinformatic approach, which considers tissue-specific eQTL analysis and SVM-based TFBS classification, is a scalable method that can be applied to multiple types of human diseases.

## Experimental procedures

### Data

ChIP-seq data sets for NKX2-5 and TBX5 from HiPSC-CMs were collected from the ChIP-Atlas (34) and Remap (35) databases. DNase I hypersensitivity footprints for fetal heart tissue (left atrium, right ventricle), heart fibroblast, and differentiated cardiomyocytes were obtained from ENCODE (ENCSR764UYH) (33). Putative heart enhancers were downloaded from the supplementary files from Dickel *et al.* (37) Disease or trait-associated SNPs were downloaded from the GWAS catalog (gwas_catalog_v1.0-associations_e0_r2022–11–29.tsv).

### Model training

LS-GKM was implemented to perform predictions on TF-DNA binding affinity for NKX2-5 and TBX5 (53, 54). LS-GKM was downloaded through the Comprehensive R Archive Network for Linux, Mac OS, and Windows platforms. For each TF ChIP-seq bed file, peaks were sorted by intensity, and the top 1000 peaks were used as a positive set for training the models. The *genNullSeqs* function from the gkmSVM package in R was used to generate negative training by selecting unbound sequences of the same length, chromosome, and GC content as the positive training file. The *gkmtrain* function was used to train the SVM classifiers. The following parameters were used to train the model using a 5-fold cross-validation: word length ($l$) = 11 and the number of informative positions ($k$) = 7 (gkmtrain -x 5 -L 11 -k 7 -d 3 -C 1 -t 2 -e 0.005). Model performance was assessed *via* receiver operator characteristic and precision-recall curves area under the curve using the gkmSVM package in R.

### Sequence scoring

The models for each TF were used to predict TF-DNA binding through weighted scoring. The gkmpredict function was used to score 18 bp sequences within 519,540 DGF from cardiac tissue that were found within previously identified human heart enhancers. These sequences were identified by intersecting genomic coordinates of ∼1.6 million DGFs from cardiac tissue with ∼80,000 putative enhancers active in fetal and adult human hearts (36, 37). Parameters were set to their default values, and *gkmpredict* was used to generate an output file listing all sequences and their respective assigned scores by the classifier model for NKX2-5 and TBX5 binding predictions. Positive scores predicted TF-DNA binding, while negative scores predicted no binding activity.

### Position weight matrix from LS-GKM models

We scored and sorted every possible 11-mers and selected the top 1000 sequences for the generation of a PWM using the multiple Em for motif elicitation (55) web-based tool default parameters to generate a logo.

### Cardiovascular disease-associated risk-variants identification

Variants from the GWAS catalog (gwas_catalog_v1.0-associations_e0_r2022–11–29.tsv) were downloaded and filtered to identify CVD or trait-associated SNPs. Insertions, deletions, and incomplete entries from the GWAS catalog were removed. Variants were filtered from the "DISEASE/TRAIT" column using the following function:

*grepl('heart|cardiac|aortic|atrial|ventric|cardio|vascular|artery|coronary|myocardial|valve|cardio|cardium|stroke', `DISEASE/TRAIT`)*

CVD SNPs were intersected with human putative enhancers active in the human heart and DGF from the fetal heart. CVD-associated SNPs that occur within human heart enhancers and footprints were expanded to include variants in LD using the LDLinkR package (56). CVD-associated SNPs and variants in LD with genotype-dependent expression in cardiac tissue (heart atrial appendage and left ventricle) were identified through the GTEx Portal database. The GTEx portal reports permutation-adjusted $p$-values for each gene for the most significant SNP per gene. The *r2d_threshold* was set to >0.8, and the p_*threshold* argument was left at default ($p$-value $< 0.1$).

### NKX2-5 and TBX5 expression and purification

The NKX2-5 homeodomain gene (Asp16 to Leu96) was cloned in pET-51(+) expression vector containing an N-terminal Strep•Tag II and a C-terminal 10 × His•Tag through Gibson Cloning and purified through Ni-NTA affinity chromatography, as previously described (18). The TBX5 T-box domain gene (Met51 to Ser248) was cloned into a pET expression vector with a 6× His•Tag (VectorBuilder Inc) and purified through Ni-NTA affinity chromatography, as previously described (18) (Fig. S4*A*). The human TBX5 gene (Clone ID HsCD00079979, DNASU Plasmid Repository) was cloned in pEU-E01-GST-TEV-MCS-N1 (Cambridge Isotope Laboratories, Inc CFS-PEU-V1.0) vectors using Gibson Assembly (New England Biolabs, Inc). Clones were verified by Sanger Sequencing from the University of Wisconsin Biotechnology Center DNA Sequencing Facility. TBX5 was expressed using the Wheat Germ Cell-Free Protein Expression from the Cell-Free Sciences Co following the manufacturer's protocol. Protein expression was confirmed through SDS-PAGE followed by Western Blot using Anti-GST HRP-conjugated (dilution: 1:10,000) NB100–63173) antibody (Novus Biological) (Fig. S4*B*).

### Electrophoretic mobility shift assay

NKX2-5 and TBX5 binding were evaluated using 20 bp genomic sequences centered on the candidate SNP plus 20 bp constant sequence for primer binding and extension. The primer contained IR-700 fluorophore. All oligonucleotides were purchased from Integrated DNA Technologies, Inc and sequences are in Table S1. The IR-700 fluorophore was added through a primer extension reaction and purified using EconoTaq Plus (Lucigen, 30,035–2), the QIAquick PCR Purification Kit (Qiagen 28,106). Twenty microliter binding reactions were performed in binding buffer (50 mM NaCl, 10 mM Tris-HCl (pH 8.0), and 10% glycerol) and 5 nM fluorescently labeled dsDNA. Five concentration points were used for purified NKX2-5 homeodomain and TBX5 T-box ranging from 50 nM to 2000 nM. Cell-free expressed TBX5-DNA binding was evaluated using four TBX5 dilutions (1, one-fifth, 1/10, and 1/25) of the cell-free extract. Binding reactions were incubated for 30 min at 30 °C followed by 30 min at room temperature before loading onto a 6% polyacrylamide gel in 0.5× TBE (89 mM Tris/89 mM boric acid/2 mM EDTA, pH 8.4). The gel was pre-ran at 85 V for 15 min, loaded at 30 V, and resolved at 75 V for 1.5 h at 4 °C. Gels were imaged with Azure Sapphire Bio-molecular Imager with 658 nm excitation and 710 nm emission.

Binding curves were generated by first quantifying the fluorescence signal in each DNA band using ImageJ (57). Background intensities obtained from blank regions of the gel were subtracted from the band intensities. The fraction of bound DNA was determined using Equation 1. The fraction of bound DNA was plotted *versus* the TF concentration. Binding curves were obtained by "one-site–specific binding" nonlinear regression using Prism software.

Equation 1. Binding affinity from the integrated density of bound and unbound bands.

$$\textbf{\textit{Fraction bound}} = \frac{\textbf{\textit{bound}}}{(\textbf{\textit{bound}}+\textbf{\textit{unbound}})}$$

## Data availability

All data and supplementary material generated for this study are publicly available at https://github.com/joshuagmedina/cardioDisease_riskVariants.

*Supporting information*—This article contains supporting information.

*Author contributions*—E. G. P.-M. and J. A. R.-M. conceptualization; E. G. P.-M., D. A. P.-M., J. G. M.-F. data curation; E. G. P.-M., D. A. P.-M., J. G. M.-F. formal analysis; E. G. P.-M., D. A. P.-M., A. R.-M., J. L. M.-B., L. S.-A., A. C. B.-R., J. R.-D. V., J. M. R.-R. investigation; E. G. P.-M., D. A. P.-M methodology; E. G. P.-M., D. A. P.-M., A. R.-M., J. L. M.-B., L. S.-A., A. C. B.-R., J. R.-D. V. validation; E. G. P.-M. and J. A. R.-M. project administration; E. G. P.-M., A. R.-M. supervision; E. G. P.-M., D. A. P.-M., A. R.-M., and J. A. R.-M. writing-original draft; D. A. P.-M., J. G. M.-F. software; J. G. M.-F. writing-review and editing; J. M. R.-R. resources; J. A. R.-M. funding acquisition.

*Funding and additional information*—This project was supported by NIH-SC1GM127231, NSF [1736026], University of Puerto Rico Rio Piedras Institutional Funds (FIPI), Puerto Rico Science, Technology, and Research Trust, and NIH Institutional Development Award (IDeA) INBRE [P20GM103475]. E. G. P. M., D. A. P. M., A. C. B. R., J. G. M. F., and J. M. R. R. were funded by the NIH RISE Fellowship (5R25GM061151–20). D. A. P. M. was funded by NSF [IQ BIOREU 1852259]. E. G. P. M. and J. M. R. R. were funded by the NSF BioXFEL Fellowship (STC-1231306). A. R. M. were funded by NSF HRD-2008186. A. R. M. was funded by NSF REU: PR-CLIMB Program (2,050,493) and NIH 1T34GM145404 to 01A1. L. S. A. was funded by NIH ID-GENE Fellowship (1R25HG012702–01). J. M. R. R. was funded NSF Graduate Research Fellowship (1744619). Figures 1*A*, and 2*A* were created in BioRender.

*Conflict of interest*—The authors declare no conflicts of interest with the contents of this article.

*Abbreviations*—The abbreviations used are: CVD, cardiovascular disease; DGFS, DNase I hypersensitivity genomic footprints; EMSA, electrophoretic mobility shift assay; GTEx, genome tissue expression; GWAS, genome-wide association studies; LD, linkage disequilibrium; LS-GKM, large-scale gapped k-mer; SVM, support vector machine; TF, transcription factor.

## References

1. Kathiresan, S., and Srivastava, D. (2012) Genetics of human cardiovascular disease. *Cell* **148**, 1242–1257
2. Ma, L. Y., Chen, W. W., Gao, R. L., Liu, L. S., Zhu, M. L., Wang, Y. J., *et al.* (2020) China cardiovascular diseases report 2018: an updated summary. *J. Geriatr. Cardiol.* **17**, 1–8
3. Mensah, G. A., Roth, G. A., and Fuster, V. (2019) The global burden of cardiovascular diseases and risk factors: 2020 and beyond. *J. Am. Coll. Cardiol.* **74**, 2529–2532
4. Yuyun, M. F., Sliwa, K., Kengne, A. P., Mocumbi, A. O., and Bukhman, G. (2020) Cardiovascular diseases in sub-saharan Africa compared to high-income countries: an epidemiological perspective. *Glob. Heart* **15**, 15
5. Townsend, N., Kazakiewicz, D., Lucy Wright, F., Timmis, A., Huculeci, R., Torbica, A., *et al.* (2022) Epidemiology of cardiovascular disease in Europe. *Nat. Rev. Cardiol.* **19**, 133–143
6. Heshmatzad, K., Naderi, N., Maleki, M., Abbasi, S., Ghasemi, S., Ashrafi, N., *et al.* (2023) Role of non-coding variants in cardiovascular disease. *J. Cell Mol. Med.* https://doi.org/10.1111/jcmm.17762
7. Villar, D., Frost, S., Deloukas, P., and Tinker, A. (2020) The contribution of non-coding regulatory elements to cardiovascular disease. *Open Biol.* **10**, 200088
8. Zhang, F., and Lupski, J. R. (2015) Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24**, R102–R110
9. Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M. A., and Gerstein, M. (2016) Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.* **17**, 93–108
10. Deplancke, B., Alpern, D., and Gardeux, V. (2016) The genetics of transcription factor DNA binding variation. *Cell* **166**, 538–554
11. Jiang, X., Li, T., Liu, S., Fu, Q., Li, F., Chen, S., *et al.* (2021) Variants in a *cis*-regulatory element of TBX1 in conotruncal heart defect patients impair GATA6-mediated transactivation. *Orphanet J. Rare Dis.* **16**, 334
12. Hiroi, Y., Kudoh, S., Monzen, K., Ikeda, Y., Yazaki, Y., Nagai, R., and Komuro, I. (2001) Tbx5 associates with Nkx2-5 and synergistically promotes cardiomyocyte differentiation. *Nat. Genet.* **28**, 276–280
13. Mori, A. D., Zhu, Y., Vahora, I., Nieman, B., Koshiba-Takeuchi, K., Davidson, L., *et al.* (2006) Tbx5-dependent rheostatic control of cardiac gene expression and morphogenesis. *Dev. Biol.* **297**, 566–586
14. Horb, M. E., and Thomsen, G. H. (1999) Tbx5 is essential for heart development. *Development* **126**, 1739–1751
15. Warren, S. A., Terada, R., Briggs, L. E., Cole-Jeffrey, C. T., Chien, W. M., Seki, T., *et al.* (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
16. Small, E. M., and Krieg, P. A. (2003) Transgenic analysis of the atrial-natriuretic factor (ANF) promoter: Nkx2-5 and GATA-4 binding sites are required for atrial specific expression of ANF. *Dev. Biol.* **261**, 116–131
17. Vincentz, J. W., Barnes, R. M., Firulli, B. A., Conway, S. J., and Firulli, A. B. (2008) Cooperative interaction of Nkx2.5 and Mef2c transcription factors during heart development. *Dev. Dyn.* **237**, 3809–3819
18. Peña-Martínez, E. G., Rivera-Madera, A., Pomales-Matos, D. A., Sanabria-Alberto, L., Rosario-Cañuelas, B. M., Rodríguez-Ríos, J. M., *et al.* (2023) Disease-associated non-coding variants alter NKX2-5 DNA-binding affinity. *Biochim. Biophys. Acta Gene Regul. Mech.* **1866**, 194906
19. Benaglio, P., D'Antonio-Chronowska, A., Ma, W., Yang, F., Young Greenwald, W. W., Donovan, M. K. R., *et al.* (2019) Allele-specific NKX2-5 binding underlies multiple genetic associations with human electrocardiographic traits. *Nat. Genet.* **51**, 1506–1517
20. van Weerd, J. H., Mohan, R. A., van Duijvenboden, K., Hooijkaas, I. B., Wakker, V., Boukens, B. J., *et al.* (2020) Trait-associated noncoding variant regions affect tbx3 regulation and cardiac conduction. *Elife* **9**, e56697
21. Smemo, S., Campos, L. C., Moskowitz, I. P., Krieger, J. E., Pereira, A. C., and Nobrega, M. A. (2012) Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. *Hum. Mol. Genet.* **21**, 3255–3263
22. Ghosh, T. K., Packham, E. A., Bonser, A. J., Robinson, T. E., Cross, S. J., and Brook, J. D. (2001) Characterization of the TBX5 binding site and analysis of mutations that cause Holt-Oram syndrome. *Hum. Mol. Genet.* **10**, 1983–1994
23. VandenBosch, L. S., Luu, K., Timms, A. E., Challam, S., Wu, Y., Lee, A. Y., and Cherry, T. J. (2022) Machine learning prediction of non-coding variant impact in human retinal *cis* -regulatory elements. *Transl. Vis. Sci. Technol.* **11**, 16
24. Zheng, A., Lamkin, M., Zhao, H., Wu, C., Su, H., and Gymrek, M. (2021) Deep neural networks identify sequence context features predictive of transcription factor binding. *Nat. Mach Intell.* **3**, 172–180
25. Felício, D., Alves-Ferreira, M., Santos, M., Quintas, M., Lopes, A. M., Lemos, C., *et al.* (2023) Integrating functional scoring and regulatory data to predict the effect of non-coding SNPs in a complex neurological disease. *Brief Funct. Genomics*. https://doi.org/10.1093/bfgp/elad020
26. Claussnitzer, M., Dankel, S. N., Kim, K. H., Quon, G., Meuleman, W., Haugen, C., *et al.* (2015) FTO obesity variant circuitry and adipocyte browning in humans. *New Engl. J. Med.* **373**, 895–907
27. Siddharthan, R. (2010) Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. *PLoS One* **5**, e9722
28. Tomovic, A., and Oakeley, E. J. (2007) Position dependencies in transcription factor binding sites. *Bioinformatics* **23**, 933–941
29. Bulyk, M. L., Johnson, P. L. F., and Church, G. M. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.* **30**, 1255–1261
30. Inukai, S., Kock, K. H., and Bulyk, M. L. (2017) Transcription factor–DNA binding: beyond binding site motifs. *Curr. Opin. Genet. Dev.* **43**, 110–119
31. Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., *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
32. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., *et al.* (2013) The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580–585
33. Dunham, I., *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74
34. Zou, Z., Ohta, T., Miura, F., and Oki, S. (2022) ChIP-Atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data. *Nucleic Acids Res.* **50**, W175–W182
35. Hammal, F., De Langen, P., Bergon, A., Lopez, F., and Ballester, B. (2022) ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Res.* **50**, D316–D325
36. Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., *et al.* (2020) Global reference mapping of human transcription factor footprints. *Nature* **583**, 729–736
37. Dickel, D. E., Barozzi, I., Zhu, Y., Fukuda-Yuzawa, Y., Osterwalder, M., Mannion, B. J., *et al.* (2016) Genome-wide compendium and functional assessment of *in vivo* heart enhancers. *Nat. Commun.* **7**, 12923
38. Gonzalez-Teran, B., Pittman, M., Felix, F., Thomas, R., Richmond-Buccola, D., Hüttenhain, R., *et al.* (2022) Transcription factor protein interactomes reveal genetic determinants in heart disease. *Cell* **185**, 794–814.e30
39. Ang, Y. S., Rivas, R. N., Ribeiro, A. J. S., Srivas, R., Rivera, J., Stone, N. R., *et al.* (2016) Disease model of GATA4 mutation reveals transcription factor cooperativity in human cardiogenesis. *Cell* **167**, 1734–1749.e22
40. Barrera, L. A., Vedenko, A., Kurland, J. V., Rogers, J. M., Gisselbrecht, S. S., Rossin, E. J., *et al.* (2016) Survey of variation in human transcription factors reveals prevalent DNA binding changes. *Science* **351**, 1450–1454
41. Pradhan, L., Genis, C., Scone, P., Weinberg, E. O., Kasahara, H., and Nam, H. J. (2012) Crystal structure of the human NKX2.5 Homeodomain in complex with DNA target. *Biochemistry* **51**, 6312–6319
42. Luna-Zurita, L., Stirnimann, C. U., Glatt, S., Kaynak, B. L., Thomas, S., Baudin, F., *et al.* (2016) Complex interdependence regulates heterotypic transcription factor distribution and coordinates cardiogenesis. *Cell* **164**, 999–1014
43. Ballester, B., Medina-Rivera, A., Schmidt, D., Gonzàlez-Porta, M., Carlucci, M., Chen, X., *et al.* (2014) Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways. *Elife* **3**, e02626
44. Dong, S., and Boyle, A. P. (2022) Prioritization of regulatory variants with tissue-specific function in the non-coding regions of human genome. *Nucleic Acids Res.* **50**, e6
45. Li, C., Qu, L., Matz, A. J., Murphy, P. A., Liu, Y., Manichaikul, A. W., *et al.* (2022) AtheroSpectrum reveals novel macrophage foam cell gene signatures associated with atherosclerotic cardiovascular disease risk. *Circulation* **145**, 206–218
46. Carter, D. R., Buckle, A. D., Tanaka, K., Perdomo, J., and Chong, B. H. (2014) Art27 interacts with GATA4, FOG2 and NKX2.5 and is a novel co-repressor of cardiac genes. *PLoS One* **9**, e95253
47. Robbe, Z. L., Shi, W., Wasson, L. K., Scialdone, A. P., Wilczewski, C. M., Sheng, X., *et al.* (2022) CHD4 is recruited by GATA4 and NKX2-5 to repress noncardiac gene programs in the developing heart. *Genes Dev.* **36**, 468–482
48. Habets, P. E. M. H., Moorman, A. F. M., Clout, D. E. W., van Roon, M. A., Lingbeek, M., van Lohuizen, M., *et al.* (2002) Cooperative action of Tbx2 and Nkx2.5 inhibits ANF expression in the atrioventricular canal: implications for cardiac chamber formation. *Genes Dev.* **16**, 1234–1246
49. Testori, A., Lasorsa, V. A., Cimmino, F., Cantalupo, S., Cardinale, A., Avitabile, M., *et al.* (2019) Exploring shared susceptibility between two neural crest cells originating conditions: neuroblastoma and congenital heart disease. *Genes (Basel)* **10**
50. Srihardyastutie, A., Soeatmadji, D. W., Fatchiyah, and Aulanni'Am. (2018) The proteomic analysis of pancreatic exocrine insufficiency protein marker in type 2 diabetes mellitus patients. In **, 299**. *IOP Conference Series: Materials Science and Engineering*. IOS Press, Amsterdam, NL-NH
51. Lodder, E. M. M., De Nittis, P., Koopman, C., Wiszniewski, W., Moura de Souza, C., Lahrouchi, N., *et al.* (2016) GNB5 mutations cause an autosomal-recessive multisystem syndrome with sinus bradycardia and cognitive disability. *Am. J. Hum. Genet.* **99**, 704–710

52. Bruneau, B. G. (2013) Signaling and transcriptional networks in heart development and regeneration. *Cold Spring Harb. Perspect. Biol.* **5**, a008292

53. Ghandi, M., Mohammad-Noori, M., Ghareghani, N., Lee, D., Garraway, L., and Beer, M. A. (2016) GkmSVM: an R package for gapped-kmer SVM. *Bioinformatics* **32**, 2205–2207

54. Lee, D. (2016) LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics* **32**, 2196–2198

55. Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (2015) The MEME suite. *Nucleic Acids Res.* **43**, W39–W49

56. Myers, T. A., Chanock, S. J., and Machiela, M. J. (2020) LDlinkR: an R package for rapidly calculating linkage disequilibrium statistics in diverse populations. *Front. Genet.* **11**

57. Schneider, C. A., Rasband, W. S., and Eliceiri, K. W. (2012) NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675