Prioritizing Cardiovascular Disease-Associated Variants Altering NKX2-5 Binding through an Integrative Computational Approach

Edwin G. Peña-Martínez*", Diego A. Pomales-Matos*", Alejandro Rivera-Madera⁻
Jean J. Massam Rive^l, Jackson G. Madina Eslisiana¹, Jacquelus Gambria Alberta¹ $\overline{}$ Jean L. Messon-Bird , Joshua G. Medina-Feliciano , Leandro Sanabria-Alberto
Advisos G. Permine Pessois¹, Jessice M. Pedriause Pisa¹, end Jesé A. Pedríause $\frac{1}{\sqrt{2}}$ Adriana C. Barreiro-Rosario , Jessica M. Rodríguez-Rios , and José A. Rodríguez-
Martíasz¹ Martínez 1 1

² University of Puerto Rico-Cayey Campus
buted equally to this work
Campus \overline{a}

University of Puerto Rico-Cayey Campus
ted equally to this work *These authors contributed equally to this work

Abstract
Cardiova:
influence
of CVD-a
regulator
number d
in vitro a
that cons
expressic
causing influenced by genetic factors. Genome-wide association studies (GWAS) have mapped and CVD-associated variants within the non-coding genome, which can alter the functive regulatory proteins, like transcription factors (TFs) of CVD-associated variants within the non-coding genome, which can alter the funct
regulatory proteins, like transcription factors (TFs). However, due to the overwhe
number of GWAS single nucleotide polymorphisms (SNPs) (> regulatory proteins, like transcription factors (TFs). However, due to the overwhe
number of GWAS single nucleotide polymorphisms (SNPs) (>500,000), prioritizing varian
in vitro analysis remains challenging. In this work, rumber of GWAS single nucleotide polymorphisms (SNPs) (>500,000), prioritizing varia
in vitro analysis remains challenging. In this work, we implemented a computational applemation and that considers support vector machine in vitro analysis remains challenging. In this work, we implemented a computational applemat considers support vector machine (SVM)-based TF binding site classification and expression quantitative trait loci (eQTL) analysi that considers support vector machine (SVM)-based TF binding site classification and conspression quantitative trait loci (eQTL) analysis to identify and prioritize potential causing SNPs. We identified 1,535 CVD-associate expression quantitative trait loci (eQTL) analysis to identify and prioritize potential
causing SNPs. We identified 1,535 CVD-associated SNPs that occur within human
footprints/enhancers and 9,309 variants in linkage diseq causing SNPs. We identified 1,535 CVD-associated SNPs that occur within human
footprints/enhancers and 9,309 variants in linkage disequilibrium (LD) with differentia
expression profiles in cardiac tissue. Using hiPSC-CM Ch > 90%
tion of
elming
ints for
proach
cardiac
al CVD-
ineart
al gene ion of

elming

nts for

proach

ardiac

CVD-

heart

l gene

TBX5, elming
Ints for
proach
Cardiac
Il CVD-
Inteart
Inteart
Inteart
Inteart
Intearty
Intearty elming

nts for

proach

cardiac

l CVD-

heart

l gene

TBX5,

k-mer proach
cardiac
| CVD-
| eart
| gene
| TBX5,
| k-mer
pciated r
| CVD-
| heart
| gene
| TBX5,
| k-mer
| pciated
| nts and al CVD-
al CVD-
al gene
al TBX5,
d k-mer
ociated
nts and
variants heart
I gene
TBX5,
I k-mer
pciated
ariants
ariants
sed on

footprints/enhancers and 9,309 variants in linkage disequilibrium (LD) with differentia
expression profiles in cardiac tissue. Using hiPSC-CM ChIP-seq data from NKX2-5 and
two cardiac TFs essential for proper heart develop expression profiles in cardiac tissue. Using hiPSC-CM ChlP-seq data from NKX2-5 and
two cardiac TFs essential for proper heart development, we trained a large-scale gapped
SVM
(LS-GKM-SVM) predictive model that can identif two cardiac TFs essential for proper heart development, we trained a large-scale gapped
SVM
(LS-GKM-SVM) predictive model that can identify binding sites altered by CVD-assc
SNPs. The computational predictive model was tes SVM
(LS-GKM-SVM) predictive model that can identify binding sites altered by CVD-asso
SNPs. The computational predictive model was tested by scoring human heart footprin
enhancers in vitro through electrophoretic mobility (LS-G
(LS-G
SNPs
enha
enheir
varial
consi
priori SNPs. The computational predictive model was tested by scoring human heart footprinenhancers in vitro through electrophoretic mobility shift assay (EMSA). Three variancers in vitro through electrophoretic mobility shift as enhancers in vitro through electrophoretic mobility shift assay (EMSA). Three varies (rs59310144, rs6715570, and rs61872084) were prioritized for in vitro validation bases
their eQTL in cardiac tissue and LS-GKM-SVM predic (rs59310144, rs6715570, and rs61872084) were prioritized for in vitro validation bases their eQTL in cardiac tissue and LS-GKM-SVM prediction to alter NKX2-5 DNA binding. Alvariants altered NKX2-5 DNA binding. In summary, (resp. Alternation to alter NKX2-5 DNA binding. All variants altered NKX2-5 DNA binding. In summary, we present a bioinformatic approac
considers tissue-specific eQTL analysis and SVM-based TF binding site classificati
pri Variants altered NKX2-5 DNA binding. In summary, we present a bioinformatic approace
considers tissue-specific eQTL analysis and SVM-based TF binding site classificati
prioritize CVD-associated variants for in vitro experi considers tissue-specific eQTL analysis and SVM-based TF binding site classificati
prioritize CVD-associated variants for in vitro experimental analysis.
Keywords: transcription factors, non-coding variants, gene regula l gene
TBX5,
k-mer
pciated
ts and
ariants
sed on
l three d TBX5,
d k-mer
ociated
nts and
variants
used on
ll three
ich that k-mer

k-mer

ociated

ts and

ariants

sed on

I three

th that

ion to sociated
ints and
variants
ased on
All three
ach that
tion to nts and
variants
sed on
II three
ch that
tion to
seases, variants
ised on
ill three
ich that
tion to
iseases, ariance
sed on
Il three
ch that
ion to
seases, ll three
ch that
iion to
seases, ch that
ion to
seases,
EMSA ch to
seases,
and the that

Keywords: transcription factors, non-coding variants, gene regulation, cardiovascular diseases,

eQTL analysis of variants in cardiac tissue

Graphical Abstract

່
2
}
} Introduction
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 non-c which are heritable. ^{1–5} Genome-wide associations studies (GWAS) have mapped over 90% of
CVD-associated variants within non-coding regions of the genome (promoters, enhancers,
introns, 5 $\mathbb{E}/3\mathbb{E}$ UTRs, etc.), ^{6,7} which are heritable. ² Genome-wide associations studies (GWAS) have mapped over 90% of
CVD-associated variants within non-coding regions of the genome (promoters, enhancers,
introns, 5 $\mathbb{Z}/3\mathbb{Z}$ UTRs, etc.). ^{6,7} introns, 5 $\overline{2}/3\overline{2}$ UTRs, etc.). ^{6,7} Non-coding single nucleotide polymorphisms (SNPs) can impact
phenotype by altering gene regulatory mechanisms, such as transcription factor (TF)-DNA
binding and gene expression introns, 5, 5, 13, UTRs, etc.). ^{5,7} Non-coding single nucleotide polymorphisms (SNPs) can impact
phenotype by altering gene regulatory mechanisms, such as transcription factor (TF)-DNA
binding and gene expression. ^{8–11} 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, bu 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, bu expression in the developing heart. 44 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 overwhelm

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 potential
CVD-causing variants for experime CVD-causing variants for experimental validation remains challenging.

One approach to address this challenge is implementing predictive models to identify

variants that create or disrupt TF binding sites (TFBS). ^{23–25} One approach to address this challenge is implementing predivariants that create or disrupt TF binding sites (TFBS). ^{23–25} Large-scale support vector machine (SVM) predictive models can be trained to in vitro or in vivo ts that create or disrupt TF binding sites (TFBS). $^{23-25}$ Large-scale gapped k-mer (LS-GKM)
rt vector machine (SVM) predictive models can be trained to identify TFBS by using in
pr in vivo DNA-binding data, such as chr variants that create or disrupt TF binding sites (TFBS). ²³ Large-scale gapped k-mer (LS-GKM) support vector machine (SVM) predictive models can be trained to identify TFBS by using in vitro or in vivo DNA-binding data, supper or in vivo DNA-binding data, such as chromatin immunoprecipitation followed by sequencing (ChIP-seq). LS-GKM-SVM models outperform traditional approaches, such as position weight matrix (PWM)-based methods, by cons sequencing (ChIP-seq). LS-GKM-SVM models outperform traditional approaches, such as
position weight matrix (PWM)-based methods, by considering complex sequence features like
dinucleotide interactions, longer/gapped k-mers position weight matrix (PWM)-based methods, by considering complex sequence features like
dinucleotide interactions, longer/gapped k-mers, and intracellular patterns. ^{26–29} LS-GKM-SVM
predictive models can be trained wi

dinucleotide interactions, longer/gapped k-mers, and intracellular patterns. ²⁶⁻²⁹ LS-GKM-SVM
predictive models can be trained with ChIP-seq data from specific cell lines or tissue to
integrate relevant epigenomic and r dinucleotide interactions, longer/gapped k-mers, and intracellular patterns. ²⁶ 29 LS-GKM-SVM
predictive models can be trained with ChIP-seq data from specific cell lines or tissue to
integrate relevant epigenomic and r Integrate relevant epigenomic and regulatory context.²³
In this work, we present an integrative approach to prioritize functional non-coding
variants that can contribute to the biology of CVDs. Using publicly accessible integrate relevant epigenomic and regulatory context. ²³
In this work, we present an integrative approac
variants that can contribute to the biology of CVDs. Us
GWAS catalog ³⁰, GTEx Portal ³¹, ENCODE ³², ChIP-Atla In this that can contribute to the biology of CVDs. Using publicly accessible data from the catalog 30 , GTEx Portal 31 , ENCODE 32 , ChIP-Atlas 33 , and Remap 34 , we compiled a list of ssociated SNPs link GWAS catalog ³⁰, GTEx Portal ³¹, ENCODE ³², ChIP-Atlas ³³, and Remap ³⁴, we compiled a list of
CVD-associated SNPs linked with a differentially expressed gene in cardiac tissue. We trained a
LS-GKM-SVM predictiv GWAS catalog ³⁰, GTEx Portal ³¹, ENCODE ³², ChIP-Atlas ³³, and Remap ³⁴, we compiled a list of
CVD-associated SNPs linked with a differentially expressed gene in cardiac tissue. We trained a
LS-GKM-SVM predictiv LS-GKM-SVM predictive model with ChIP-seq data from NKX2-5 and TBX5 in human-induced
PSC-derived cardiomyocytes (hiPSC-CM). Both models were used to score previously identified
heart DNase 1 hypersensitivity genomic footp PSC-derived cardiomyocytes (hiPSC-CM). Both models were used to score previously identified
heart DNase I hypersensitivity genomic footprints (DGF) 35 that colocalize within putative
cardiac enhancers 36 and teste heart DNase I hypersensitivity genomic footprints (DGF) ³⁵ that colocalize within putative
cardiac enhancers ³⁶ and tested them through in vitro binding by electrophoretic mobility shift
assay (EMSA). Our predictive mo

cardiac enhancers 36 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 cardiac enhancers ³⁶ 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 DN sites and distinguishing between DNA sequences with different binding affinities.

Having validated DGF scored by the predictive model, we scored all CVD-associated

SNPs to alter NKX2-5 DNA binding. We chose three variant SNPs to alter NKX2-5 DNA binding. We chose three variants (rs59310144, rs6715570, and rs61872084) to prioritize for in vitro validation based on their expression quantitative trait loci (eQTL) in cardiac tissue and LS-GKMto alter NKX2-5 DNA binding. We chose three variants (rs59310144, rs6715570, and
72084) to prioritize for in vitro validation based on their expression quantitative trait loci
1 in cardiac tissue and LS-GKM-SVM prediction rs61872084) to prioritize for in vitro validation based on their expression quantitative trait loci (eQTL) in cardiac tissue and LS-GKM-SVM prediction to alter NKX2-5 DNA binding. All three variants were validated through (eQTL) in cardiac tissue and LS-GKM-SVM prediction to alter NKX2-5 DNA binding. All three
variants were validated through EMSA and resulted in changes on NKX2-5 DNA binding. In
short, we present a bioinformatic approach th variants were validated through EMSA and resulted in changes on NKX2-5 DNA binding. In
short, we present a bioinformatic approach that considers tissue-specific eQTL analysis and
SVM-based TF binding site classification to short, we present a bioinformatic approach that considers tissue-specific eQTL analysis and sVM-based TF binding site classification to prioritize functional CVD-associated SNPs.

SVM-based TF binding site classification to prioritize functional CVD-associated SNPs. SVM-based TF binding site classification to prioritize functional CVD-associated SNPs.

Methods

|
|
|
|
| 【(
(
1
3 Data
ChIP-seg data sets for NKX2-5 and TBX5 from human induced pluripotent stem cell-derived cardiomyocytes (HiPSC-CM) were collected from the ChIP-Atlas ³³ and Remap ³⁴ databases.
Dnase I hypersensitivity footprints for fetal heart tissue (left atrium, right ventricle), heart
fibroblast, and differentiated c cardiomyocytes (HiPSC-CM) were collected from the ChIP-Atlas ³³ and Remap ³⁴ databases.
Dnase I hypersensitivity footprints for fetal heart tissue (left atrium, right ventricle), heart
fibroblast, and differentiated ca Fibroblast, and differentiated cardiomyocytes were obtained from ENCODE (ENCSR764UYH).

³²Heart enhancers were downloaded from the supplementary files from Dickel et al. ³⁶ Disease

or trait-associated SNPs were downl ³²Heart enhancers were downloaded from the supplementary files from Dickel et al. ³³ Disease
or trait-associated SNPs were downloaded from the GWAS catalog (gwas_catalog_v1.0-
associations_e0_r2022-11-29.tsv).
Model

associations_e0_r2022-11-29.tsv).
 Model training

Large-scale gapped k-mer (LS-GKM) was implemented to perform predictions on TF-DNA

binding affinity for NKX2-5 and TBX5. $37,38$ LS-GKM was downloaded through the

Com Model training
Large-scale gapped k-mer (LS-GK
binding affinity for NKX2-5 an
Comprehensive R Archive Networ
each TF ChIP-seq bed file, peaks w 【 】 】 【
((() Model training binding affinity for NKX2-5 and TBX5. $37,38$ LS-GKM was downloaded through the Comprehensive R Archive Network (CRAN), for Linux, Mac OS, and Windows platforms. For each TF ChIP-seq bed file, peaks were sorted by intensi binding affinity for NKX2-5 and TBX5. $37,38$ LS-GKM was downloaded through the Comprehensive R Archive Network (CRAN), for Linux, Mac OS, and Windows platforms. For each TF ChIP-seq bed file, peaks were sorted by intensi each TF ChIP-seq bed file, peaks were sorted by intensity and the top 1,000 peaks were used
as a positive set for training the predictive models. The *genNullSeqs()* function from the
gkmSVM package in R was used to gener as a positive set for training the predictive 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 as a positive set for training the predictive 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 general 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 fivefold function was used to train the SVM classifiers. The following parameters were used to train
the model using a fivefold cross-validation: word length (*I*) = 11 and the number of informative
positions (*k*) = 7 (gkmtrain the model using a fivefold 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 operato 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 (ROC) and precision-recall curves (PCR) area under the curve (AUC) using the gkmSVM

positions (k) = 7 (gkmtrain x 5 -L 11 - k 7 -d 5 -C 1 + 2 -C 0.005). Model performance was
assessed via receiver operator characteristic (ROC) and precision-recall curves (PCR) area
under the curve (AUC) using the gkmSV under the curve (AUC) 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 Sequence Scoring
The models for each TF were used to predict TF-DNA
gkmpredict() function was used to score 18 bp sequentissue that were found within previously identified hur
were identified by intersecting genomic coordi s. 「 ()
. ()
. () sequence scoring 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 co tissue that were found within previously identified human heart enhancers. These sequences
were identified by intersecting genomic coordinates of $^{\sim}1.6$ million DHFs from cardiac tissue
with >80,000 putative enhancers were identified by intersecting genomic coordinates of \sim 1.6 million DHFs from cardiac tissue with >80,000 putative enhancers active in fetal and adult human hearts that were identified
through ChIP-seq. All function parameters were set to their default values and *gkmpredict()*
was used to generate an output file through ChIP-seq. All function parameters were set to their default values and *gkmpredict()* was used to generate an output file listing all sequences and their respective assigned scores

Motif Extraction from LS-GKM Models

Whe classifier model for NKX2-5 and TBX5 binding predictions. Positive scores predicted TF-
DNA binding, while negative scores predicted no binding activity.
Motif Extraction from LS-GKM Models
We scored and sorted every p by the classifier model for NKX2-5 and TBX5 binding predictions. Fositive scores predicted TF-
DNA binding, while negative scores predicted no binding activity.
Motif Extraction from LS-GKM Models
We scored and sorted ever $(DEME)^{39}$ web-based tool default parameters to generate a logo. **『** 〜
〜 8(generation of a Position Weight Matrix (PWM) using the Multiple Em for Motif Elicitation generation of a Poster of a Poster of a Poster Matrix (MEME)³⁹ web-based tool default parameters to generate a logo. (MEME)³⁹ web-based tool default parameters to generate a logo.
MEME)⁹⁹ web-based tool default parameters to generate a logo.
.

(\
}
} Cardiovascular disease-associated risk-variants Identification
Variants from the GWAS catalog were downloaded and filtered to identify CVD or trait-

associated SNPs. Variants were filtered from the state of the following the following distribution:
grepl('heart|cardiac|aortic|atrial|ventric|cardio|vascular|artery|coronary|myocardial|valve|
cardio|cardium|stroke', `DISE

CVD SNPs were intersected with human putative enhancers active in the human heart and 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
DG }
(
【
什 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
DG cardio|cardium|stroke', `DISEASE/TRAIT`)
CVD SNPs were intersected with human
DGF from the fetal heart. CVD-associated
footprints were expanded to include varia
package. ⁴⁰ CVD-associated SNPs and va
appendage and left v DGF from the fetal heart. CVD-associated SNPs that occur within human heart enhancers and
footprints were expanded to include variants in linkage disequilibrium (LD) using the LDLinkR
package. ⁴⁰ CVD-associated SNPs and footprints were expanded to include variants in linkage disequilibrium (LD) using the LDLinkR
package. ⁴⁰ CVD-associated SNPs and variants in LD found in cardiac tissue (heart atrial
appendage and left ventricle) with di package. ⁴⁰ CVD-associated SNPs and variants in LD found in cardiac tissue (heart atrial
appendage and left ventricle) with differentially expressed genes were identified through the
Genome Tissue Expression (GTEx) Porta appendage and left ventricle) with differentially expressed genes were identified through the Genome Tissue Expression (GTEx) Portal database.
NKX2-5 and TBX5 expression and purification

Genome Tissue Expression (GTEx) Portal database.
 NKX2-5 and TBX5 expression and purification

The NKX2-5 homeodomain (HD) gene (Asp16 to Leu96) was cloned in pET-51(+) expression

vector containing an N-terminal Strep•T **NKX2-5 and TBX5 expression and purification**
The NKX2-5 homeodomain (HD) gene (Asp16 to I
vector containing an N-terminal Strep•Tag II® and
Cloning and purified through Ni-NTA affinity chron
human TBX5 gene (Clone ID HsCD 「ヽヽ」 The NKX2-5 homeodomain (HD) 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
human TBX5 gene (Clone ID HsCD000799 Cloning and purified through Ni-NTA affinity chromatography, as previously described. ¹⁸ The
human TBX5 gene (Clone ID HsCD00079979, DNASU Plasmid Repository, AZ) was cloned in
pEU-E01-GST-TEV-MCS-N1 (Cambridge Isotope L Cloning and purified through Ni-NTA affinity chromatography, as previously described. ²⁶ The
human TBX5 gene (Clone ID HsCD00079979, DNASU Plasmid Repository, AZ) was cloned in
pEU-E01-GST-TEV-MCS-N1 (Cambridge Isotope L 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 period Assembly (New England Biolabs, Inc). Clones were verified by Sanger Sequencing from
the University of Wisconsin Biotechnology Center DNA Sequencing Facility. Protein expression
was made using the Wheat Germ Cell-Fre The University of Wisconsin Biotechnology Center DNA Sequencing Facility. Protein expression
was made using the Wheat Germ Cell-Free Protein Expression from the CellFree Sciences Co
following the manufacturer's protocol. P was made using the Wheat Germ Cell-Free Protein Expression from the CellFree Sciences Co
following the manufacturer's protocol. Protein expression was confirmed through an SDS-
PAGE followed by Western Blot using Anti-GST PAGE followed by Western Blot using Anti-GST HRP-conjugated (NB100-63173) antibody
(Novus Biological).
Electrophoretic mobility shift assay

Electrophoretic mobility shift assay

enhancers were evaluated using 20 bp sequences that contained an additional 20 bp constant Patter op Source Side follower Side for IR-700 fluorescent Data follows Side terms and their testing to their respective scored sequences of human heart footprints and enhancers were evaluated using 20 bp sequences that co (November 1997)
Electrophoretic monder and TBX5 b
Enhancers were ever
Sequence for IR-70
Available in Supple 【【】 (c) (c) (nancers were evaluated using 20 bp sequences that contained an additional 20 bp constant
sequence for IR-700 fluorescent marking (IDT). All sequences were ordered in IDT and are
available in **Supplementary Table 1**. The IR sequence for IR-700 fluorescent marking (IDT). All sequences were ordered in IDT and are
available in **Supplementary Table 1**. The IR-700 fluorophore was added to all the sequences
through a primer extension reaction and p available in **Supplementary Table 1**. The IR-700 fluorophore was added to all the sequences
through a primer extension reaction and purified using the QIAquick® PCR Purification Kit
(Qiagen 28106). Binding reactions were available in **Supplementary Table 1.** The IR 700 haotophore was added to all the sequences through a primer extension reaction and purified using the QIAquick® PCR Purification Kit (Qiagen 28106). Binding reactions were p (Qiagen 28106). 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 employed for purified NKX2-5 H DNA binding was evaluated using four TBX5 dilutions (1, $1/5$, $1/10$, and $1/25$) of the cell-free DNA binding was evaluated using four TBX5 dilutions (1, 1/5, 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% po 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 0.5x TBE (89 mM Tris/89 mM
M temperature before loading onto a 6% polyacrylamide gel in 0.5x TBE (89 mM Tris/89 mM temperature before loading onto a 6% polyacrylamide gel in 0.5x TBE (89 mM Tris/89 mM

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 For 1.5 am excitation and 710 nm emission.
Binding curves were generated by first quantifying the fluorescence signal in each DNA band
using Imagel. Background intensities obtained from blank regions of the gel were subtra Binding curves were generated by first quantifying the fluorescence signal in each DNA band using ImageJ. Background intensities obtained from blank regions of the gel were subtracted from the band intensities. The fractio - 1
- 1
- 1
- 1 using ImageJ. 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 vers 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" non-li fraction of bound DNA was plotted versus the TF concentration. Binding curves were obtained
by "one-site specific binding" non-linear regression using Prism software.
Equation 1. Binding Affinity from the integrated den by "one-site specific binding" non-linear regression using Prism software.
 Equation 1. Binding Affinity from the integrated density of bound and unbound bands.
 Fraction bound = $\frac{bound}{(bound + unbound)}$

Equation 1. Binding Affinity from the integrated density of bound are
 $Fraction\,bound = \frac{bound}{(bound + unbound)}$. Equation 1. Binding Affilm, i.e. integrated density of bound and disposition bands.
Fraction bound $=$ $\frac{bound}{(bound + unbound)}$ Γ action bound Γ $\frac{1}{1}$ $(bound + unbound)$

\overline{a} Results and Discussion

Figure 1: Identification of Tunctional CVD associated SNPs. A) Figure to identify potential
CVD-causing SNPs. **B)** Number of CVD-associated SNPs per chromosome. **C)** Distribution of SNP
frequency within autosomal chromosom CVD causing SNPs. B) Number of CVD associated SNPs per emombsome. C) Distribution of SNP
frequency within autosomal chromosome, binned by 1Mb windows. D) SNP-Gene pairs with
differentially expressed in heart atrial appenda frequency within autosomal emoniosome, binned by 1Mb windows. D) SNP Gene pairs with
differentially expressed in heart atrial appendage or left ventricle in one or more populations.
rs6715570-BARD1, rs61872084-METTL10 and differentially expressed in heart atrial appendage or left ventricle in one or more populations.
rs6715570-BARD1, rs61872084-METTL10 and rs59310144-RNASEH2B are SNP-Gene pairs that
were evaluated in vitro in Figure 3. differentially expressed in heart atrial appendage or left ventricle in one or more populations. rsection of and a, research and research and research in the rate of the pair and
were evaluated in vitro in Figure 3. were evaluated in vitro in Figure 3.

|
|
|
|
|
| data to keep cardiovascular disease or trait-associated SNPs (e.g., congenital heart defects, cardiomyocyte differentiation, stroke, arrhythmia, etc.; full list of SNPs in **Supplementary File**
1). We then intersected the cardiomyocyte differentiation, stroke, arrhythmia, etc.; full list of SNPs in **Supplementary File**
1). We then intersected the CVD-associated SNPs with a catalog of putative fetal and adult
heart enhancers and genomic foo 1). 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 1,535 genomic variants. The CVD-associated SNP set was expand 1). We then intersected the CVD associated SMPs with a catalog of putative fetal and adult
heart enhancers and genomic footprints of fetal hearts, resulting in 1,535 genomic variants.
The CVD-associated SNP set was expand 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 9,309 unique SNPs
occurring in one or more populatio The CVD-associated SNP set was expanded to include SNPs in linkage disequilibrium (LD r⁻>0.8)
from diverse populations (EUR, AFR, SAS, EAS, and AMR) and resulted in 9,309 unique SNPs
occurring in one or more populations. 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 var biologically relevant in cardiovascular biology, we analyzed gene expression patterns in cardiac
tissue with the previously identified variants in the GTex portal. We found 636 differentially
expressed genes associated wit bissue with the previously identified variants in the GTex portal. We found 636 differentially expressed genes associated with the previously identified SNPs in the heart atrial appendage or left ventricle. The workflow i expressed genes associated with the previously identified SNPs in the heart atrial appendage
or left ventricle. The workflow is illustrated in **Figure 1A** and the list of SNPs associated with
differentially expressed genes or left ventricle. The workflow is illustrated in **Figure 1A** and the list of SNPs associated with differentially expressed genes in cardiac tissue is found in **Supplementary File 1**. The distribution of CVD-associated SN or left ventricle. The workflow is individual in Figure 1A and the list of SNPs associated with
differentially expressed genes in cardiac tissue is found in **Supplementary File 1**. The
distribution of CVD-associated SNPs i distribution of CVD-associated SNPs is not uniform throughout the genome. We identified
chromosomes with a higher frequency of CVD-associated SNPs which contain >1,000 variants
(chromosomes 1 and 6) and ~500 (chromosomes 2 chromosomes with a higher frequency of CVD-associated SNPs which contain >1,000 variants (chromosomes 1 and 6) and ~500 (chromosomes 2, 3, 7, 15, 17, and 22), including those in LD (Figure 1B). Chromosomes with a high SNP (chromosomes 1 and 6) and ~500 (chromosomes 2, 3, 7, 15, 17, and 22), including those in LD
(Figure 1B). Chromosomes with a high SNP frequency may have variants evenly distributed
among them, like chromosomes 1 and 2, whi (Figure 1B). 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, a near) loci, like chromosomes 6, 10, 15, and 22 (Figure 1C). This suggests that certain near) loci, like chromosomes 6, 10, 15, and 22 (**Figure 1C**). This suggests that certain
chromosomes, or specific loci, are enriched with CVD-associated SNPs and contribute to the
cardiac phenotype. We also analyzed data f near) loci, like chromosomes 6, 10, 15, and 22 (Figure 1c). This suggests that ecrtain
chromosomes, or specific loci, are enriched with CVD-associated SNPs and contribute to the
cardiac phenotype. We also analyzed data fro cardiac phenotype. We also analyzed data from the GTEx database to find genes that are differentially expressed in cardiac tissue (heart atrial appendage and left ventricle) containing the identified CVD-associated SNPs or differentially expressed in cardiac tissue (heart atrial appendage and left ventricle) containing
the identified CVD-associated SNPs or the variants in LD. We identified 25,479 SNP-Gene pairs
(636 unique genes) that were s the identified CVD-associated SNPs or the variants in LD. We identified 25,479 SNP-Gene pairs
(636 unique genes) that were significantly differentially expressed in cardiac tissue (**Figure 1D**).
Through this approach, we a (636 unique genes) that were significantly differentially expressed in cardiac tissue (**Figure 1D**).
Through this approach, we aimed to narrow the extensive list of non-coding variants and
identify functional SNPs that con (636 unique genes) that were significantly differentially expressed in cardiac tissue (Figure 1D).
Through this approach, we aimed to narrow the extensive list of non-coding variants and
identify functional SNPs that contr Through this approach, we annually the extensive linear to narrow the extensive linearing variants and identify functional SNPs that contribute to CVD. identify functional SNPs that contribute to CVD.

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 NKX2training with NKX2-5 and TBX5 Cim-seq data from Him-se-CM. B) scoling 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 hig occur in heart emfancers 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, w (top) and TBX5 (bottom). For NKX2-5, we tested chr22:25120040-25120058 (circle with blue
line), chr3:8596782-8596800 (triangle with green line), and chr7:101950814-101950832
(square with red line). For TBX5, we tested chr2 (top), chr3:8596782-8596800 (triangle with green line), and chr7:101950814-101950832
(square with red line). For TBX5, we tested chr2:30359836-30359854 (circle with blue lines),
chr1:57623182-57623200 (triangle with green line), the with red line). For TBX5, we tested chr2:30359836-30359854 (circle with blue lines), chr1:57623182-57623200 (triangle with green line), and chr4:119047319-119047337 (square with red line). $\frac{1}{2}$ chr1:57623182-57623200 (triangle with green line), and chr4:119047319-119047337 (square
with red line).
with red line). with red line).
With red line), $\frac{1}{2}$ with red line).

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 human induced pluripotent stem cel The NKX2-5 and TBX5. The 1,000 top-scoring ChiP-seq peaks were used as a positive training
for NKX2-5 and TBX5. The 1,000 top-scoring ChiP-seq peaks were used as a positive training
set, while unbound sequences of the same for NKX2-5 and TBX5. The 1,000 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 (**Figure 2A)** The bestset, while unbound sequences of the same length, GC content, and chromosome were used as negative training (**Figure 2A)** The best-performing LS-GKM SVM classifier model trained with NKX2-5 ChIP-seq data (SRX9284027) ⁴¹ negative training (**Figure 2A)** The best-performing LS-GKM SVM classifier model trained with NKX2-5 ChIP-seq data (SRX9284027) ⁴¹ obtained an AUROC value of 0.955 and an AUPRC value of 0.954. The best TBX5 (SRX2023721) NKX2-5 ChIP-seq data (SRX9284027) ⁴¹ obtained an AUROC value of 0.955 and an AUPRC value
of 0.954. The best TBX5 (SRX2023721) ⁴² model obtained an AUROC value of 0.921 and an
AUPRC value of 0.912 (Supplementary Figure NKX2-5 ChIP-seq data (SRX9284027)⁻¹ obtained an AUROC value of 0.955 and an AUPRC value
of 0.954. The best TBX5 (SRX2023721)⁴² model obtained an AUROC value of 0.921 and an
AUPRC value of 0.912 (Supplementary Figure 1 of 0.954. The best TBX5 (SRX2023721) ⁴² model obtained an AUROC value of 0.921 and an AUPRC value of 0.912 (Supplementary Figure 1A-B). The models were used to score all possible 2,097,152 non-redundant 11 bp oligomers possible 2,097,152 non-redundant 11 bp oligomers (11-mers). The 11-mer scores were sorted
and the 1,000 top-scoring sequences were used to generate Position Weight Matrix (PWM)
using MEME (Supplementary Figure 1C-D). The P and the 1,000 top-scoring sequences were used to generate Position Weight Matrix (PWM) using MEME (Supplementary Figure 1C-D). The PWMs for both models resulted in DNA binding motifs in agreement with previously described and MEME (Supplementary Figure 1C-D). The PWMs for both models resulted in DNA
binding motifs in agreement with previously described models for NKX2-5 and TBX5. $43-45$ We
proceeded to score ~520,000 fetal heart DGF that using WEWE (Supplementary Figure 1C-D). The PWMs for both models resulted in DNA
binding motifs in agreement with previously described models for NKX2-5 and TBX5. ⁴³⁻⁴⁵ We
proceeded to score ~520,000 fetal heart DGF that proceeded to score ~520,000 fetal heart DGF that occur heart enhancers to identify genomic
loci potentially bound by NKX2-5 or TBX5 (Figure 2B). We then chose the DNA sequences with
the highest, middle, and lowest scores t proci potentially bound by NKX2-5 or TBX5 (Figure 2B). We then chose the DNA sequences with
the highest, middle, and lowest scores to test for in vitro binding through EMSA (Figure 2C,
Supplementary Figure 2). There was ag loci potentially bound by NKX2-5 or TBX5 (Figure 2B). We then chose the DNA sequences with
the highest, middle, and lowest scores to test for in vitro binding through EMSA (Figure 2C,
Supplementary Figure 2). There was agr Supplementary Figure 2). There was agreement between LS-GKM SVM scores and extent of in vitro binding activity for both, NKX2-5 and TBX5. Our results suggest that our LS-GKM SVM with the will be able to successfully predict changes in binding affinity between reference and variant DNA sequences that alter cardiac TF-DNA binding. model will be about the control of the successfully predict changes in binding and the successfully between reference and the successfully between reference and the successfully between reference and the successfully betwe variant DNA sequences that alter cardiac TF-DNA binding.

Figure 3: CVD-associated SNPs alter NKX2-5 in vitro binding. A) DelstaSVM score distribution of
the 9,309 CVD-associated SNPs. B) Representative EMSA gel for rs59310144 reference (Ref)
and alternate (Alt) alleles. C) Bindi and alternate (Alt) alleles. C) Binding curves for reference (Ref) and variant (Alt) alleles of and alternate (Alt) alleles. C) Binding curves for reference (Ref) and variant (Alt) alleles or
rs59310144 (top), rs6715570 (middle), and rs61872084 (bottom). Experiments were
performed in triplicates and binding curves sh performed in triplicates and binding curves show average bound fractio (X) and error bars are
standard error. **D)** Cardiac tissue eQTL analysis of *RNASEH2B* (top), *BARD1* (middle), and
METTL10 (bottom) expressed in hea performed in triplem in triplem is transfer and be triplicated in text and METTL10 (bottom) expressed in heart atrial appendage or left ventricle when rs59310144, rs6715570, and rs61872084 occur, respectively. Standard error. **D)** Cardiac tissue eQTL analysis of *RNASEH2B* (top), BARD1 (imdate), and
METTL10 (bottom) expressed in heart atrial appendage or left ventricle when rs59310144,
rs6715570, and rs61872084 occur, respecti METTL10 (bottom) expressed in heart atrial appendage or left ventricle when 1399310144,
rs6715570, and rs61872084 occur, respectively. rs671570, and respectively.
671570, and respectively.
671570, and respectively.

score the 9,309 SNPs to prioritize functional variants. Both reference and alternate allele
sequences were scored to predict fold change (deltaSVM score) of TF-DNA binding. We
selected three SNPs (rs59310144, rs6715570, an sequences were scored to predict fold change (deltaSVM score) of TF-DNA binding. We
selected three SNPs (rs59310144, rs6715570, and rs61872084) that deltaSVM predicted
significant change in NKX2-5 binding and are associate selected three SNPs (rs59310144, rs6715570, and rs61872084) that deltaSVM predicted
significant change in NKX2-5 binding and are associated with a differentially expressed gene in
cardiac tissue (Figure 3A, Supplementary T significant change in NKX2-5 binding and are associated with a differentially expressed gene in
cardiac tissue (**Figure 3A, Supplementary Table 2**). When evaluated through EMSA, we
observed a differences in NKX2-5 DNA bin cardiac tissue (**Figure 3A, Supplementary Table 2**). When evaluated through EMSA, we observed a differences in NKX2-5 DNA binding between reference and alternate for all three SNPs (**Figure 3B-C, Supplementary Figure 3**). cardiac tissue (Figure 3A, Supplementary Table 2). When evaluated through EMSA, we
observed a differences in NKX2-5 DNA binding between reference and alternate for all three
SNPs (Figure 3B-C, Supplementary Figure 3). Vari

SNPs (Figure 3B-C, Supplementary Figure 3). Variants rs59310144 and rs61872084 resulted in
a decrease in NKX2-5 DNA binding, while rs6715570 increased binding.
We found that all three SNPs were in eQTLs described in cardia SNPS (Figure 3B-C, Supplementary Figure 3). Variants 13333101444 and 1301672064 resulted in
a decrease in NKX2-5 DNA binding, while rs6715570 increased binding.
We found that all three SNPs were in eQTLs described in cardi We found that all three SNPs were in eQTLs described in cardiac tiss
genes that are differentially expressed when these variants occur in the
or left ventricle (Figure 3D). RNASEH2B and BARD1 have been prev
differentially \
{
{
(genes that are differentially expressed when these variants occur in the heart atrial appendage
or left ventricle (**Figure 3D**). RNASEH2B and BARD1 have been previously identified to be
differentially expressed in the hear gene in that are differentially expressed in the heart atrial appendage when variants rs59310144 and rs6715570 (respectively) occur. RNASEH2B, which has been previously found to be differentially expressed in CVD risk eve differentially expressed in the heart atrial appendage when variants rs59310144 and rs6715570 (respectively) occur. *RNASEH2B*, which has been previously found to be differentially expressed in CVD risk events, is upregul rs6715570 (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. ⁴⁶ *BARD1* has al rs6715570 (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. ⁴⁶ BARD1 has also rs59310144 is present. ⁴⁶ *BARD1* has also been identified as upregulated when the alternate allele of variant rs6715570 occurs in the heart atrial appendage. Copy number alternate allele of variant rs6715570 occurs in rs59310144 is present. ⁴⁶ BARD1 has also been identified as upregulated when the alternate allele of variant rs6715570 occurs in the heart atrial appendage. Copy number alternations in the BARD1 locus have been associat the *BARD1* locus have been associated with developmental delays, including coarctation of the
aorta during early organogenesis and heart development. ⁴⁷ Variant rs61872084 has been
identified in the heart's left ventric and during early organogenesis and heart developmental delays, including coarctation of the
aorta during early organogenesis and heart development. ⁴⁷ Variant rs61872084 has been
identified in the heart's left ventricle aorta during early organogenesis and heart development. ⁴⁷ Variant rs61872084 has been
identified in the heart's left ventricle when *METTL10* (Methyltransferase like protein 10) is
downregulated when the alternate allel identified in the heart's left ventricle when *METTL10* (Methyltransferase like protein 10) is
downregulated when the alternate allele occurs. Accumulation of METTL10 methylated
products, such as S-adenosyl-L-methionine, S products, such as S-adenosyl-L-methionine, S-adenosyl-L-homocysteine, and homocystein
have been correlated with kidney dysfunction and CVD in patients with type 2 diabetes. ⁴⁸ This
suggests that NKX2-5 regulation of the provide been correlated with kidney dysfunction and CVD in patients with type 2 diabetes. ⁴⁸ This suggests that NKX2-5 regulation of the *RNASEH2B* (inhibition), *BARD1* (activation), and *METTL10* (activation) genes are suggests that NKX2-5 regulation of the *RNASEH2B* (inhibition), *BARD1* (activation), and *METTL10* (activation) genes are possible mechanisms that can be further explored to establish rs59310144, rs6715570, and rs61872084 $METTL10$ (activation) genes are possible mechanisms that can be further explored to establish METTLIO (activation) genes are possible mechanisms that can be further explored to establish
rs59310144, rs6715570, and rs61872084 as causal CVD risk-variants. rs59310144, rs6715570, and rs61872084 as causal CVD risk-variants.

Conclusion
As we continue to research the genetic basis for human disease, the number of identified Functional/causal non-coding SNPs continues to grow. Understanding and prioritizing SNPs
that contribute to the disease phenotypes is essential. However, we lack a consensus or
bioinformatic protocol to prioritize non-codi that contribute to the disease phenotypes is essential. However, we lack a consensus or
bioinformatic protocol to prioritize non-coding SNPs that are biologically relevant in the
development of human diseases. ²⁵ To add bioinformatic protocol to prioritize non-coding 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 potentia 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 GW 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 GW everaged 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). These SNPs we cardiovascular disease or traits and included variants in LD from multiple populations (EUR, AFR, SAS, EAS, and AMR). These SNPs were analyzed with data from the GTEx database to identify genes that are differentially expr AFR, SAS, EAS, and AMR). These SNPs were analyzed with data from the GTEx database to
identify genes that are differentially expressed when these variants are present in cardiac
tissue. We tested three SNPs (rs59310144, rs Matrify genes that are differentially expressed when these variants are present in cardiac
tissue. We tested three SNPs (rs59310144, rs6715570, and rs61872084) associated with a
differentially expressed gene (RNASEH2B, BAR itissue. We tested three SNPs (rs59310144, rs6715570, and rs61872084) associated with a differentially expressed gene (RNASEH2B, BARD1, and METTL10 respectively) in cardiac tissue that resulted in changes in NKX2-5 DNA bin differentially expressed gene (*RNASEH2B, BARD1,* and *METTL10* respectively) in cardiac tissue
that resulted in changes in NKX2-5 DNA binding activity. Our findings open the possibility that
NKX2-5 regulation of *RNASEH2B* that resulted in changes in NKX2-5 DNA binding activity. Our findings open the possibility that NKX2-5 regulation of *RNASEH2B*, *BARD1*, and *METTL10* is a possible mechanism that can be further researched to determine th NKX2-5 regulation of *RNASEH2B*, *BARD1*, and *METTL10* is a possible mechanism that can be further researched to determine the causality of CVD-risk variants. Although the etiology of human diseases is complex and multifa NKX2-5 regulation of *HWASEH2B*, *BARD1*, and *METTL10* is a possible mechanism that can be further researched to determine the causality of CVD-risk variants. Although the etiology of human diseases is complex and multifa further 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 non-coding SNPs. I that can be implemented during in vivo experiments or clinical research to address genetic
diseases caused by non-coding SNPs. In summary, we believe this bioinformatic approach,
which considers tissue-specific eQTL analys diseases caused by non-coding SNPs. In summary, we believe this bioinformatic approach,
which considers tissue-specific eQTL analysis and SVM-based TF binding site classification, is a
scalable method that can be applied t which considers tissue-specific eQTL analysis and SVM-based TF binding site classification, is a
scalable method that can be applied to multiple types of human diseases.
 $\frac{1}{2}$ which considers the expectation of the specific equation of the state of human diseases.
Scalable method that can be applied to multiple types of human diseases. scalable method that can be applied to multiple types of human diseases.

Acknowledgments/Funding
This project was supported by NIH-SC1GM127231, NSF [1736026], NSF LSAMP [HRD-2008186], This project was supported by the Piedras Institutional Funds (FIPI), Puerto Rico Science,
Technology, and Research Trust, and NIH Institutional Development Award (IDeA) INBRE
[P20GM103475]. EGPM, DAPM, ACBR, JGMF and JMRR Technology, and Research Trust, and NIH Institutional Development Award (IDeA) INBRE
[P20GM103475]. EGPM, DAPM, ACBR, JGMF and JMRR were funded by the NIH RISE
Fellowship (5R25GM061151-20). DAPM was funded by NSF [IQ BIORE [P20GM103475]. EGPM, DAPM, ACBR, JGMF and JMRR were funded by the NIH RISE
Fellowship (5R25GM061151-20). DAPM was funded by NSF [IQ BIOREU 1852259]. EGPM and
JMRR were funded by the NSF BioXFEL Fellowship (STC-1231306). AR Eellowship (5R25GM061151-20). DAPM was funded by NSF [IQ BIOREU 1852259]. EGPM and
JMRR were funded by the NSF BioXFEL Fellowship (STC-1231306). ARM and JLMB were funded
by NSF HRD-2008186. ARM was funded by NSF REU: PR-CL JMRR were funded by the NSF BioXFEL Fellowship (STC-1231306). ARM and JLMB were funded
by NSF HRD-2008186. ARM was funded by NSF REU: PR-CLIMB Program (2050493) and NIH
1T34GM145404-01A1. LSA was funded by NIH ID-GENE Fell by NSF HRD-2008186. ARM was funded by NSF REU: PR-CLIMB Program (2050493) and NIH
1T34GM145404-01A1. LSA was funded by NIH ID-GENE Fellowship (1R25HG012702-01). JMRR
was funded NSF Graduate Research Fellowship (1744619). G 1T34GM145404-01A1. LSA was funded by NIH ID-GENE Fellowship (1R25HG012702-01). JMRR
was funded NSF Graduate Research Fellowship (1744619). Graphical abstract, Figure 1A, and
Figure 2A were created in Biorender®. 1T34GM145404-01A1. LSA was funded by NIH ID-GENE Fellowship (1R25HG012702-01). JMRR Was funded NSF Graduate Research Fellowship (1744619). Graphical abstract, Figure 2A were created in Biorender®. Figure 2A were created in Biorender®.

Data Availability

All data generated for this study is publicly available at https://github.com/joshuagmedina/cardioDisease riskVariants (accessed on 29 August 2023). https://github.com/joshuagmedina/cardiodisease_riskUardioDisease_riskUardioDisease_risk
University
Disease and 29 August 2023

- References
1. Kathiresan, S. & Srivastava, D. Genetics of human cardiovascular disease. Cell vol. 148 1242–1257 1. Kathiresan, S. & Srivastava, D. Genetics of human cardiovascular disease. Cell vol. 148 1242–1257
(2012).
2. Ma, L. Y. *et al.* China cardiovascular diseases report 2018: An updated summary. *Journal of Geriatric*
- (2022).
Ma, L. Y
Cardiolo 2. Ma, L. Y. et al. China cardiovascular diseases report 2018. An updated summary. Journal by Geriatric
Cardiology vol. 17 1–8 (2020).
3. Mensah, G. A., Roth, G. A. & Fuster, V. The Global Burden of Cardiovascular Diseases
- Cardiology vol. 17 1–8 (2020).
Mensah, G. A., Roth, G. A. & F
2020 and Beyond. Journal of th 3. Mensah, G. A., Roth, G. A., R
3. 3. Yuyun, M. F., Sliwa, K., Kengne, A. P., Mocumbi, A. O. & Bukhman, G. Cardiovascular diseases in sub-
- 2020 and Beyond. Journal of the American College of Cardiology vol. 74 2525–2552 (2015).
Yuyun, M. F., Sliwa, K., Kengne, A. P., Mocumbi, A. O. & Bukhman, G. Cardiovascular dis
saharan Africa compared to high-income countr (2020) . saharan Africa compared to high-income countries. An epidemiological perspective. Glob Heart 15,
(2020).
Townsend, N. *et al.* Epidemiology of cardiovascular disease in Europe. Nature Reviews Cardiology vol.
- .
Townse
19 133-5. Townsend, N. *et al*. Epidemiology of cardiovascular disease in Europe. *Nature Reviews Cardiology* vol.
19 133–143 (2022).
6. Heshmatzad, K. *et al*. Role of non-coding variants in cardiovascular disease. *Journal*
- Molecular Medicine (2023). 6. Heshmatzad, K. et al. Role of non-coding variants in cardiovascular disease. Journal of Cellular and
Molecular Medicine (2023).
7. Villar, D., Frost, S., Deloukas, P. & Tinker, A. The contribution of non-coding regu
- *Molecular Medicine (2023).*
Villar, D., Frost, S., Delouka
cardiovascular disease. *Opel* 7. Villar, D., Frost, S., Deloukas, P. & Tinker, A. The contribution of non-coding regulatory elements to
- cardiovascular disease. Open Biol 10, 200088 (2020).
Zhang, F. & Lupski, J. R. Non-coding genetic variants
(2015). 8. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum Mol Genet* 24, R102–F
(2015). (2015). (2015). Khurana, E. *et al.* Role of non-coding sequence variants in cancer. *Nat Rev Genet* 17, 93–108
-
- (2020).
Khurana
Depland 9. Khurana, E. et al. Role of non-coding sequence variants in cancer. Nut Rev Genet 17, 99–108 (2016).
10. Deplancke, B., Alpern, D. & Gardeux, V. The Genetics of Transcription Factor DNA Binding Variation.
166, 538–554 10. Deplancke, B., Alpern, D. & Gardeux, V. The Genetics of Transcription Factor DNA Binding Variation. *Cell*
166, 538–554 (2016).
11. Jiang, X. *et al.* Variants in a cis-regulatory element of TBX1 in conotruncal heart d
- GATA6-mediated transactivation. Orphanet J Rare Dis 16, (2021). 11. Jiang, X. et *al.* Variants in a cis-regulatory element of TBX1 in conotration heart defect patients impair
GATA6-mediated transactivation. *Orphanet J Rare Dis* 16, (2021).
12. Hiroi, Y. *et al.* Tbx5 associates with
- GATA6 mediated transactivation. Orphanet 3 Kare Dis 16, (2021).
Hiroi, Y. *et al.* Tbx5 associates with Nkx2-5 and synergistically pre
Nat Genet 28, 276–280 (2001). 12. Hiroi, Y. et al. 19x5 associates with NKX2-5 and synergistically promotes cardiomyocyte differentiation.
Nat Genet 28, 276–280 (2001).
13. Mori, A. D. et al. Tbx5-dependent rheostatic control of cardiac gene expression
- Nat Genet **28**, 276–288 (2001).
Mori, A. D. *et al.* Tbx5-depende
Biol **297**, 566–586 (2006). 13. Mori, A. D. et al. Tbx5-dependent rheostatic control of cardiac gene expression and morphogenesis. Dev
Biol 297, 566–586 (2006).
14. Horb, M. E. & Thomsen, G. H. Tbx5 is essential for heart development. Development 126
- Biol **297**, 566–586 (2006).
Horb, M. E. & Thomsen, (
(1999). 14. Horb, M. E. & Hiomsen, G. H. Tbx5 is essential for heart development. Development 126, 1755–1751
(1999).
15. Warren, S. A. *et al.* Differential Role of Nkx2-5 in Activation of the Atrial Natriuretic Factor Gene in the
- 、
Warren
Develop 15. Warren, S. A. et al. 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 (2011).
16. Small, E. M. & Krieg, P. A. Transge
- Developing versus Failing Heart. Mor Cell Biol 31, 4633–4645 (2011).
Small, E. M. & Krieg, P. A. Transgenic analysis of the atrialnatriuretic
GATA-4 binding sites are required for atrial specific expression of AN 16. Small, E. M. & Krieg, P. A. Transgenic analysis of the atrialnatriuretic factor (ANF) promoter: Nkx2-5 and
6ATA-4 binding sites are required for atrial specific expression of ANF. *Dev Biol* 261, 116–131 (2003).
17. Vi
- GATA-4 binding sites are required for atrial specific expression of ANF. Dev Biol 201, 116–131 (2003).
Vincentz, J. W., Barnes, R. M., Firulli, B. A., Conway, S. J. & Firulli, A. B. Cooperative interaction of Nkx
and Mef2c 17. Vincentz, M.M., Barnes, R. M., Firm, Firm, Firm, Premisit, Premisin, B. Cooperative interaction of Names
and Mef2c transcription factors during heart development. Developmental Dynamics 237, 3809–3819
(2008). and Merze transcription factors during heart development. *Developmental Dynamics* 237, 3809–3819
(2008).
Peña-Martínez, E. G. *et al.* Disease-associated non-coding variants alter NKX2-5 DNA-binding affinity.
- 、
Peña-M
Biochim 18. Peña-Martínez, E. G. et al. Disease-associated non-coding variants alter NKX2-5 DNA-binding affinity.
Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms 1866, 194906 (2023).
19. Benaglio, P. et al. Allele
- Biochimica et Biophysica Acta (BBA) Gene Regulatory Mechanisms 1866, 194906 (2029).
Benaglio, P. *et al.* Allele-specific NKX2-5 binding underlies multiple genetic association
electrocardiographic traits. Nat Genet 51, 1 19. Benaglio, P. et *di. Allele-specific NKX2-5 binding* underlies multiple genetic associations with human
electrocardiographic traits. Nat Genet 51, 1506–1517 (2019).
20. van Weerd, J. H. et al. Trait-associated noncodin
- electrocardiographic traits. Nat Genet 51, 1506–1517 (2015).
van Weerd, J. H. *et al.* Trait-associated noncoding variant r
conduction. *Elife* 9, 1–26 (2020). 20. van Weerd, J. H. et al. Trait-associated noncoding variant regions affect tbx3 regulation and cardiac
conduction. *Elife* 9, 1–26 (2020).
21. Smemo, S. *et al.* Regulatory variation in a TBX5 enhancer leads to isolated
- conduction. *Elife 9, 1–2*0 (2020).
Smemo, S. *et al.* Regulatory var
H*um Mol Genet* **21**, 3255–3263 (21. Smemo, S. et al. Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease.
Hum Mol Genet 21, 3255–3263 (2012).
22. Ghosh, T. K. Characterization of the TBX5 binding site and analysis of mutati
- Hum Mor Genet 21, 3233–3263 (2012).
Ghosh, T. K. Characterization of the TB
syndrome. Hum Mol Genet **10**, 1983–19 syndrome. Hum Mol Genet 10, 1983–1994 (2001).
23. VandenBosch, L. S. et al. Machine Learning Prediction of Non-Coding Variant Impact in Human Retinal
- syndrome. *Hum Mor* Genet **10**, 1983–1994 (2001).
VandenBosch, L. S. *et al. Machine Learning Predic*
cis -Regulatory Elements. *Transl Vis Sci Technol* 11, 23. VandenBosch, L. S. et al. Machine Learning Prediction of Non-
cis - Regulatory Elements. Transl Vis Sci Technol 11, 16 (2022). cis Regulatory Elements. Transl Vis Sci Technol 11, 16 (2022).

- 24. Zheng, A. et al. Deep neural networks identify sequence context reatures predictive or transcription
factor binding. Nat Mach Intell 3, 172–180 (2021).
25. Felício, D. et al. Integrating functional scoring and regulato
- ractor binding. *Nat Mach Intell* 3, 172–180 (2021).
Felício, D. *et al.* Integrating functional scoring and
SNPs in a complex neurological disease. *Brief Funct* 25. Felicio, D. et al. Integrating functional scoring and regulatory data to predict the effect of non-coding
26. Siddharthan, R. Dinucleotide weight matrices for predicting transcription factor binding sites:
- SNPS in a complex neurological disease. *Brief Funct Genomics* (2023)
Siddharthan, R. Dinucleotide weight matrices for predicting ti
Generalizing the position weight matrix. *PLoS One* 5, (2010). 22. Siddhartham, R. Dinucleotide Weight matrices for predicting transcription factor and my Generalizing the position weight matrix. PLoS One 5, (2010).
27. Tomovic, A. & Oakeley, E. J. Position dependencies in transcripti
- Generalizing the position weight matrix. *PLOS One 5*, (2010).
Tomovic, A. & Oakeley, E. J. Position dependencies in trans
23, 933–941 (2007). 27. Tomovic, A. & Oakeley, E. J. Position dependencies in transcription factor binding sites. Bioinformatics
23, 933–941 (2007).
28. Bulyk, M. L., Johnson, P. L. F. & Church, G. M. Nucleotides of transcription factor bindi
- 23, 933–941 (2007).
Bulyk, M. L., Johnso
interdependent effec 28. Bulyk, M. L., Johnson, P. L. F. & Church, G. M. Nucleotides of transcription factor binding sites exert
interdependent effects on the binding affinities of transcription factors. Nucleic Acids Research vol. 30
(2002). interdependent effects on the binding affinities of transcription factors. Nucleic Acids Research vol. 30
(2002).
Inukai, S., Kock, K. H. & Bulyk, M. L. Transcription factor–DNA binding: beyond binding site motifs. *Curr*
- (2002).
Inukai, :
Opin Ge 29. Inukai, S., Kock, K. H. & Bulyk, M. L. Transcription factor–DNA binding: beyond binding site motifs. Curr
Opin Genet Dev 43, 110–119 (2017).
30. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide a
- arrays and summary statistics 2019. Nucleic Acids Res 47, D1005-D1012 (2019). 30. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted
arrays and summary statistics 2019. Nucleic Acids Res 47, D1005–D1012 (2019).
31. Lonsdale, J. *et al.* The Genotype
- arrays and summary statistics 2019. Nucleic Acids Res 47, D1005–D1012 (2019).
Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. Nature Ge
Preprint at https://doi.org/10.1038/ng.2653 (2013). 31. Lonsdale, J. et al. The Genotype-Tissue Expression (GTEx) project. Nature Genetics vol. 45–580–585
Preprint at https://doi.org/10.1038/ng.2653 (2013).
32. Dunham, I. *et al.* An integrated encyclopedia of DNA elements
- prepries at https://doi.org/10.1038/ng.
74 (2012). 32. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–
74 (2012).
33. Zou, Z., Ohta, T., Miura, F. & Oki, S. ChIP-Atlas 2021 update: a data-mining suite for exploring
- zou, Z., O
epigenomi epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data. Nucleic Acids Res
50, W175–W182 (2022). epigenomic landscapes by fully integrating ChiP-seq, ATAC-seq and Bisulfite-seq data. Nucleic Acids Res
50, W175–W182 (2022).
Hammal, F., De Langen, P., Bergon, A., Lopez, F. & Ballester, B. ReMap 2022: A database of Human
- 50, Willie Wilder, Capach,
Hammal, F., De Langen,
Mouse, Drosophila and 34. Hammal, F., De Langen, P., Bergon, A., Lopez, F. & Ballester, B. ReMap 2022: A database of Human,
Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding
sequencing experiments. Mouse, Drosophila and Arabidopsis regionsity regions from an integrative analysis of DNA-binding
Sequencing experiments. Nu*cleic Acids Res* 50, D316–D325 (2022).
Vierstra, J. *et al.* Global reference mapping of human tra
- sequencing experiments. *Nucleic Acids Res 50*, D316–D325 (2022).
Vierstra, J. *et al.* Global reference mapping of human transcription
736 (2020). 35. Vierstra, J. et al. Global reference mapping of human transcription factor footprints. Nature 383, 723–
736 (2020).
36. Dickel, D. E. *et al.* Genome-wide compendium and functional assessment of in vivo heart enhancers
- 7377.
Dickel, D. E.
Nat Commu 36. Dickel, D. E. et al. Genome-wide compendium and functional assessment of in vivo heart enhancers.
Nat Commun 7, 1–13 (2016).
37. Ghandi, M. et al. GkmSVM: An R package for gapped-kmer SVM. Bioinformatics 32, 2205–2207
- Nat Comman 7, 1–13 (2010).
Ghandi, M. *et al.* GkmSVM: A
Lee, D. LS-GKM: A new gkm-S
-
- 37. Chandi, M. et di. GkmSVM: An R package for gapped-kmer SVM. *Biomformatics 32, 2203–2207 (2010).*
38. Lee, D. LS-GKM: A new gkm-SVM for large-scale datasets. *Bioinformatics* **32**, 2196–2198 (2016).
39. Bailey, T. L., 38. Lee, D. LS-GKM: A new gkm-SVM for large-scale datasets. *Biomyormatics 32*, 2196–2198 (2016).
39. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. *Nucleic Acids Res* 43, W
(2015). 39. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. Nucleic Acids Res 43, W39–W49–
(2015).
40. Myers, T. A., Chanock, S. J. & Machiela, M. J. LDlinkR: An R Package for Rapidly Calculating Linkage
- 、
Myers,
Disequil 140. Myers, T. A., Chansen, S. A., Chanock, M. J. Lemma for An Range for Papus, Saramang Linkage
Disequilibrium Statistics in Diverse Populations. Front Genet 11, (2020).
41. Gonzalez-Teran, B. et al. Transcription factor
- Disequilibrium Statistics in Diverse Populations. Front Genet 11, (2020).
Gonzalez-Teran, B. *et al.* Transcription factor protein interactomes reve
disease. *Cell* 185, 794-814.e30 (2022). 41. Gonzalez-Teran, B. et al. Transcription factor protein interactomes reveal genetic determinants in heart
disease. *Cell* 185, 794-814.e30 (2022).
42. Ang, Y. S. *et al.* Disease Model of GATA4 Mutation Reveals Transcri
- alsease. Cell **185**, 794-814.e50 (2022).
Ang, Y. S. *et al.* Disease Model of GAT/
Cardiogenesis. *Cell* **167**, 1734-1749.e2. 42. Ang, Y. S. et al. Disease Model of GATA4 Mutation Reveals Transcription Factor Cooperativity in Human
Cardiogenesis. Cell 167, 1734-1749.e22 (2016).
43. Luna-Zurita, L. et al. Complex Interdependence Regulates Heteroty
- Cardiogenesis. Cell 167, 1734-1749.e22 (2016).
Luna-Zurita, L. *et al.* Complex Interdependence
and Coordinates Cardiogenesis. *Cell* 164, 999–1 43. Luna-Zurita, L. et *ul.* Complex Interdependence Regulates Heterotypic Transcription Factor Distribution
and Coordinates Cardiogenesis. *Cell* 164, 999–1014 (2016).
44. Barrera, L. A. *et al.* Survey of variation in hu
- and Coordinates Cardiogenesis. Cell 164, 999–1014 (2016).
Barrera, L. A. *et al.* Survey of variation in human transcri_l
changes. Science (1979) **351**, 1450–1454 (2016). 44. Barrera, L. A. et al. Survey of variation in human transcription factors reveals prevalent DNA binding
changes. Science (1979) 351, 1450–1454 (2016).
45. Pradhan, L. et al. Crystal structure of the human NKX2.5 Homeodo
- changes. Science (1979) 331, 1450–1454 (2010).
Pradhan, L. *et al.* Crystal structure of the huma
Biochemistry 51, 6312–6319 (2012). 45. Pradhan, L. et al. Crystal structure of the human NKX2.5 Homeodomain in complex with DNA target.
Biochemistry 51, 6312–6319 (2012).
46. Li, C. et al. AtheroSpectrum Reveals Novel Macrophage Foam Cell Gene Signatures As
- Biochemistry 51, 6312–6313 (2012).
Li, C. *et al.* AtheroSpectrum Reveal
Atherosclerotic Cardiovascular Disea 46. Li, C. et al. Atherospectrum Reveals Novel Macrophage Foam Cell Gene Signatures Associated With
Atherosclerotic Cardiovascular Disease Risk. *Circulation* 145, 206–218 (2022).
47. Testori, A. *et al.* Exploring shared
- Atherosclerotic cardiovascular Disease Risk. C*irculation* 145, 200–218 (2022).
Testori, A. *et al.* Exploring shared susceptibility between two neural crest content of Neuroblastoma and congenital heart disease. *Genes (B* 47. Testori, A. *et al.* Exploring shared susceptibility between two neural creations and congenital heart disease. *Genes (Basel)* **10**, (2019). Neuroblastoma and congenital heart disease. Genes (Basel) 10, (2019).

48. Srihardyastutie, A., Soeatmadji, D. W., Fatchiyah & Aulanni'Am. The Proteomic Analysis of Pancreatic
Exocrine Insufficiency Protein Marker in Type 2 Diabetes Mellitus Patients. in *IOP Conference Series:* Exocrine Insufficiency Protein Marker in Type 2 Diabetes Mellits Patients. in IOP Conference Series:
Materials Science and Engineering vol. 299 (Institute of Physics Publishing, 2018).

Supplementary Material

 $\frac{1}{2}$ Table 1: Oligonucleotides used in this work.

Supplementary Figure 1: Performance parameters and motif analysis. Performance parameters of **A**) NKX2-5 and **B**) TBX5 as determined by their receiver operating characteristics parameters of A) NKX2-5 and B) TBX5 as determined by their receiver operating characters.
(BOC) and precision recell curves. Binding matif for C) NKV3-F and D) TBVF after seen (ROC) and precision-recall curves. Binding motif for C) NKX2-5 and D) TBX5 after scor possible 11-mers and generating a PWM logo. Performance eristics ring all

Supplementary Figure 2: EMSA analysis of heart footprint and enhancers for NKX2-5 (le TBX5 (right). All EMSA were performed and triplicates and regions within dashed were u generate binding curves. ft) and sed to

֦ Supplementary Figure 3: EMSA analysis of three CVD-associated SNPs.

