Global Evaluation of Congenital Heart Disease-Associated Non-Coding Variants

Edwin G. Peña-Martínez[1,*,#], Shreya Sharma[2], Joshua G. Medina-Feliciano[1], Elise Root[3], Lois G. Parks[4,5,6], Marissa Granitto[4,5], Diego A. Pomales-Matos[1], Jean L. Messon-Bird[1], Adriana C. Barreiro-Rosario[1], Leandro Sanabria-Alberto[1], Alejandro Rivera-Madera[7], Jessica M. Rodríguez-Ríos[1], Rosalba Velázquez-Roig[1], Juan A. Figueroa-Rosado[8], Mackenzie Noon[3], Omer A. Donmez[4,5], Carmy Forney[4,5], Hayley K. Hesse[4,5], Katelyn A. Dunn[4,5], Xiaoting Chen[4,5], Matthew R. Hass[4,5], Lucinda P. Lawson[4,5], Matthew T. Weirauch[4,5,9], Leah C. Kottyan[4,5,9], Steven K. Reilly[3], Devesh Bhimsaria[2,†], and José A. Rodríguez-Martínez[1,†,*]

[1]Department of Biology, University of Puerto Rico-Río Piedras, San Juan, PR 00931, USA
[2]Department of Biosciences and Bioengineering, Indian Institute of Technology Roorkee, Roorkee, 247667, India
[3]Department of Genetics, Yale University, New Haven, CT, 06510, USA
[4]Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA
[5]Division of Allergy and Immunology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, 45229, USA
[6]Immunology Graduate Program, University of Cincinnati College of Medicine, Cincinnati, Ohio, 45229, USA
[7]Department of Biology, University of Puerto Rico-Cayey, Cayey, PR 00736, USA
[8]Department of Computer Engineering, University of Puerto Rico-Mayagüez, Mayagüez, PR 00681, USA
[9]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH 45229, USA
[*]Corresponding authors
[†]Joint supervised work
[#]Current address: epea@wustl.edu

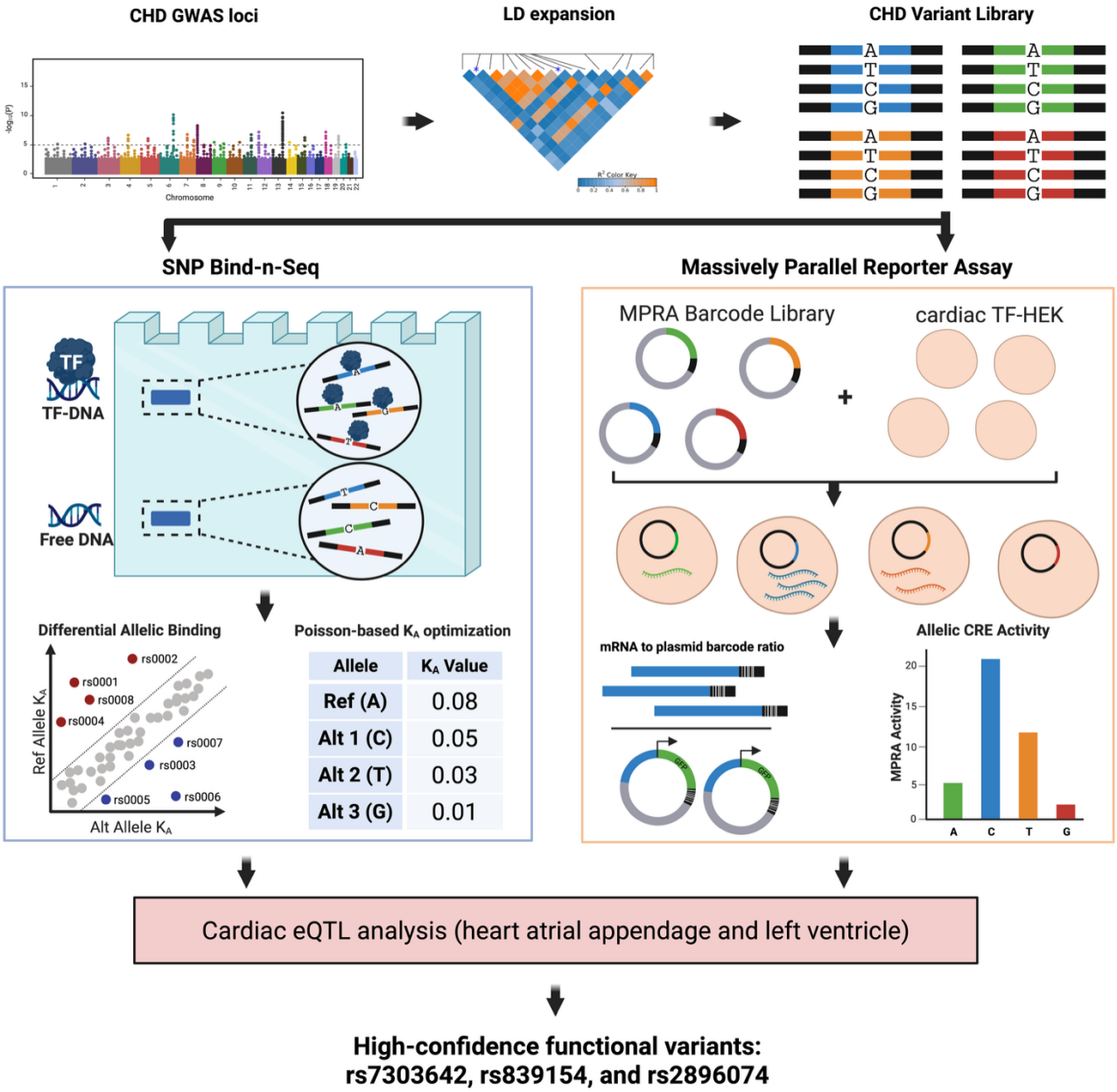
Abstract (Summary)
Genome-wide association studies (GWAS) have mapped hundreds of congenital heart disease (CHD)-associated variants within non-coding regions of the genome. Non-coding variants can alter regulatory mechanisms, such as transcription factor (TF) binding control of gene expression, potentially contributing human diseases. However, with the increasing number of disease-associated variants, comprehensive functional validation remains a significant challenge. In this work, we developed a novel method called SNP Bind-n-Seq to evaluate >3,000 CHD-risk variants for allelic binding for the cardiac TFs NKX2-5, GATA4, and TBX5 in a high-throughput manner. These binding affinity data sets were coupled with a massively parallel reporter assay (MPRA) to screen CHD-risk variant genotype-dependent regulatory activity. We identified 170 variants that exhibit allelic TF binding and 187 that modulate gene expression. Combining both approaches revealed three high-confidence variants with genotype-dependent TF binding, genotype-dependent transcriptional activity, and eQTL behavior in cardiac cells. Collectively, this study provides the first combined high-throughput biochemical and functional genomic evaluation of thousands of CHD-risk variants.

Highlights:

- Allelic binding affinity measurements of ~9,600 variants for NKX2-5, GATA4, and TBX5
- Evaluation of >3,000 CHD-risk variants for genotype-dependent regulatory activity
- Interaction networks identify functional variants and genes involving cardiac eQTLs

Keywords: transcription factors, non-coding variants, MPRA, DNA-binding, gene regulation, GWAS, congenital heart disease, genotype-dependent biology

Graphical Abstract



Cardiac eQTL analysis (heart atrial appendage and left ventricle)

High-confidence functional variants:
rs7303642, rs839154, and rs2896074

Introduction

Congenital heart diseases (CHDs) are the most common birth defect, affecting 1 in every 100 liveborn infants. CHDs are characterized by structural abnormalities in the heart and vessels that cause thousands of deaths each year, mainly in infants under one year of age. [1–3] Among the multiple causes of CHDs are genetic variants that alter regulatory mechanisms during heart development. [4–6] Genome-wide association studies (GWAS) have identified hundreds of CHD-associated variants, with the overwhelming majority (>90%) mapping to non-coding regions of the genome. [7–11] Non-coding genetic variants, such as single nucleotide polymorphisms (SNPs), have been suggested to contribute to CHD etiology by altering the functions of cardiac transcription factors (TFs), which bind to DNA to regulate cardiac differentiation. [2,12–16]

NKX2-5, GATA4, and TBX5 are evolutionarily conserved cardiac TFs necessary for heart development in vertebrates, and mutations impairing their function have been implicated in multiple types of CHDs. [12,17–23] For example, previous work has identified coding variants within their DNA-binding domains (DBDs) to be causal for CHDs. [21,24–30] Each of these cardiac TFs belongs to its own TF family: Homeobox, GATA, and T-box family, respectively, with each TF displaying unique DNA binding preferences. [18,31] NKX2-5, GATA4, and TBX5 work synergistically and cooperatively to regulate cardiac genes needed to initiate cardiogenesis and differentiation of multiple tissues within the heart. [24,28,32–34] Non-coding variants alter the DNA-binding affinity of cardiac TFs, potentially leading to multiple types of cardiovascular diseases. [8,15,20,35–44] However, with a continuously increasing number of CHD-risk variants being discovered, testing each of them to identify functional variants remains challenging.

In this work, we systematically evaluated over 3,000 CHD-risk variants for allele-biased regulatory function involving cardiac TF-DNA binding and transcriptional activity. Building upon previous methods [45,46], we developed SNP Bind-n-Seq, a high-throughput gel shift-based assay to quantify differential allelic binding events. Using this assay, we constructed allelic enrichment curves and, through a Poisson-based optimization framework, derived binding affinities for the cardiac TFs NKX2-5, GATA4, and TBX5. To assess gene regulatory effects, we also performed a massively parallel reporter assay (MPRA) [47–50] to identify CHD-risk variants that exhibited genotype-dependent transcriptional activity. Using a combined experimental and computational approach, we identified 352 CHD-risk variants with allelic regulatory functions for either cardiac TF binding (170 variants) or *cis*-regulatory element (CRE) activity (187 variants). Additionally, we identified three high-confidence variants with an allelic skew in both TF binding and transcriptional activity, and are in cardiac expression quantitative trait loci (eQTL), which can be further explored for roles in CHD disease mechanisms. In summary, this study provides the first combined biochemical and functional genomic evaluation of thousands of CHD-risk variants, offering a scalable approach that can be applied broadly to complex human diseases.

Results

Quantifying allelic cardiac TF binding through SNP Bind-n-Seq

To evaluate the impact of non-coding variants on cardiac TF binding, we first collected all CHD-associated variants from the GWAS catalog [7] available in November 2022. We identified 121 CHD-associated SNPs from six GWASs [51–56] which were expanded to include variants in linkage disequilibrium (LD) in four ancestries (EUR, AFR, EAS, SAS; $R^2 > 0.80$), totaling 3,232 unique variants. An oligonucleotide library of the 3,232 variants containing every possible variant allele was synthesized (12,928 unique sequences, **Supplementary Data 1**). Variants were centered on a 40 bp DNA sequence of genomic context with constant regions for downstream barcoding and sequencing (102 bp total, **Supplementary Figure 1A**).

To evaluate the impact of CHD-risk variants on TF-DNA binding, we developed SNP Bind-n-Seq, a high-throughput gel shift-based assay coupled to DNA sequencing (**Figure 1A**). In SNP Bind-n-Seq, a TF is equilibrated with the oligonucleotide library containing thousands of sequence variants. TF-bound and -unbound sequences are separated in a native polyacrylamide gel, and both fractions are sequenced. SNP Bind-n-Seq was performed with the purified DBD of NKX2-5, GATA4, and TBX5 at seven concentration points ranging from 0 nM to 3,000 nM. Experiments were performed in duplicates for each TF. The oligonucleotide library was modified to have a fluorescent probe through primer extension reaction as previously described [57], which allowed gel excision of bound and unbound fractions at each concentration. Bound and unbound fractions were individually barcoded with a unique identifier for pooling and downstream computational analysis.

To quantify TF binding through SNP Bind-n-Seq, we employed a statistical framework for all CHD-risk variants centered within the 40-bp sequence. Assuming non-cooperative, site-independent binding, we compute occupancy probabilities based on relative affinities at varying protein concentrations. To estimate these affinities, a joint model fits bound and unbound read counts across concentrations, inferring a concentration-independent binding parameter K_A for each sequence (see Method for details). K_A is conceptually analogous to a binding constant, which is defined as the molecular interaction strength at equilibrium. Read counts were modelled as a Poisson distribution, and K_A is optimized to maximize the likelihood of the observed data. Enrichment was then calculated by comparing bound counts to a pseudo-unbound estimate, derived by scaling total unbound reads in each replicate by a factor of 2.5 and normalizing to a million. The resulting K_A values were used to reconstruct expected counts and enrichment across concentrations.

Using this approach, we quantified TF binding enrichment and affinity for all 12,928 sequences (3,232 base-permutated variants; **Figure 1B**). Duplicates across all concentration points, excluding 0 nM, showed strong correlation for all three TFs ($R^2 > 0.75$), indicating strong experimental reproducibility (**Supplementary Figure 1B and C**). TF motif derivation for the top 200 bound sequences produced position weight matrix (PWM) logos similar to those previously described for NKX2-5, GATA4, and TBX5 (**Figure 1B**, **Supplementary Figure 2**). [15,16] Additionally, the K_a values showed strong correlation with the TF binding enrichment measurements for NKX2-5 ($r = 0.92$), GATA4 ($r = 0.89$),

and TBX5 (r = 0.94; **Figure 1C**). Collectively, these results indicate that SNP Bind-N-Seq experimental data are highly reproducible and recapitulate known TF DNA binding specificities.

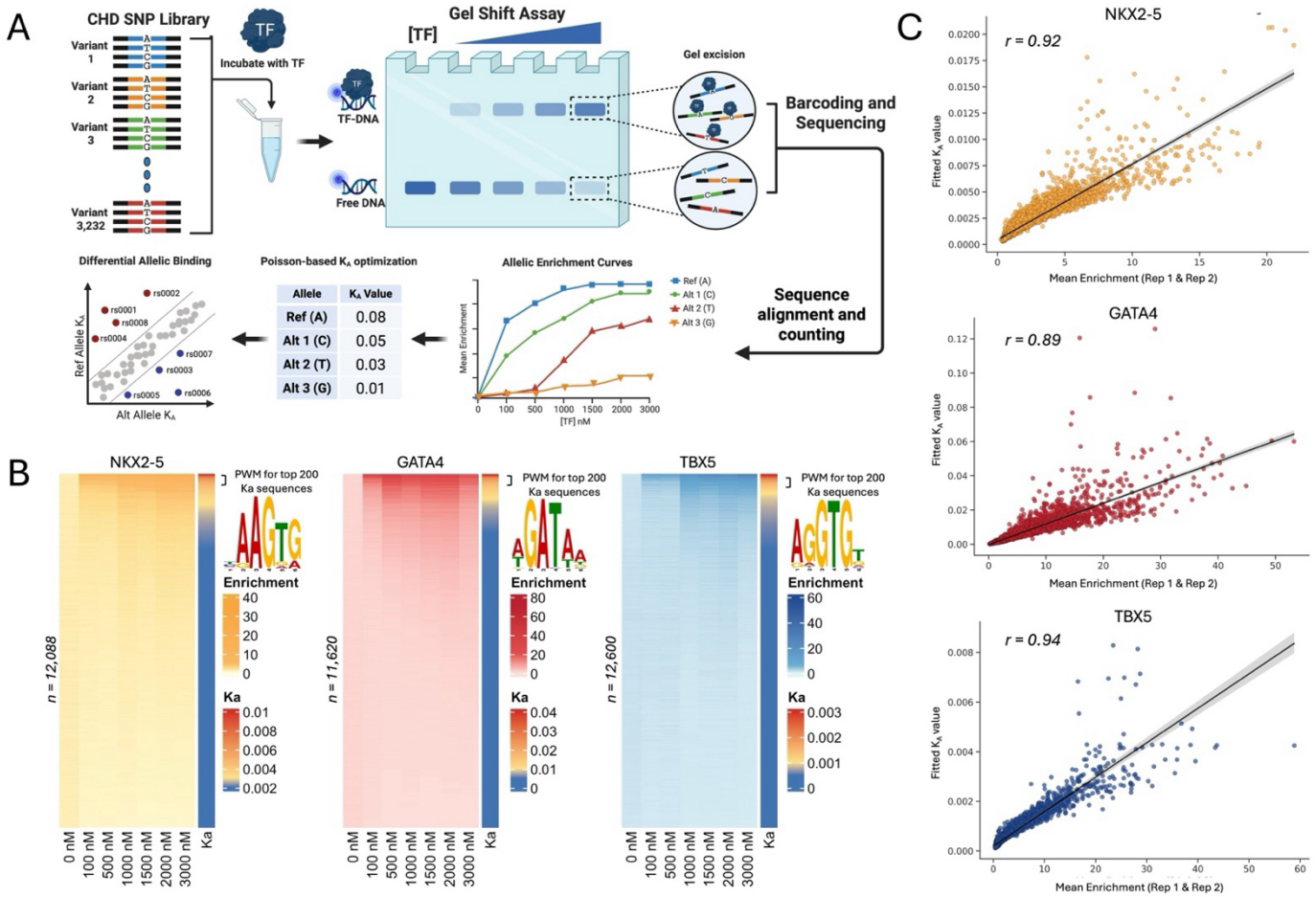


Figure 1: High-throughput evaluation of TF binding through SNP Bind-n-Seq. **A)** Overview of SNP Bind-n-Seq experimental approach and computational analysis. **B)** Sequence enrichment and binding affinity measurements for CHD-associated variants. Enrichment was calculated for all sequences at seven concentration points ranging from 0 nM to 3,000 nM for NKX2-5 (left), GATA4 (middle), and TBX5 (right). PWM logos generated from the top 200 sequences with the highest binding affinity (K_A). **C)** Correlation between mean enrichment scores and fitted K_A values. Enrichment at 3,000 nM is displayed for NKX2-5 (top), GATA4 (middle), and TBX5 (bottom).

SNP Bind-n-Seq identifies CHD-risk variants with allele-biased TF binding

Using SNP Bind-n-Seq, we constructed allelic enrichment curves and quantified binding affinity for 3,232 CHD-risk variants for all three cardiac TFs. In doing so, we identified 170 CHD-risk variants (~5% of all tested variants) with >2-fold differential allelic binding compared to the reference genome allele for NKX2-5 (54 SNPs; 31 increase, 23 decrease), GATA4 (58 SNPs; 30 increase, 28 decrease), or TBX5 (62 SNPs; 35 increase, 27 decrease) (**Figure 2A** and **Supplementary Figure 3A**). As controls, we included variants with differential binding for NKX2-5 that we previously described through electrophoretic mobility shift assay (EMSA), and observed a significant correlation (R^2 = 0.95, *p*-value = 0.012, **Supplementary Figure 3B**). [35]

Next, we proceeded to evaluate biochemical mechanisms that could be driving allelic binding events of cardiac TFs. We identified four variants (rs863392, rs2465147, rs28394479, and rs77931854) that altered the binding of two TFs (**Figure 2B**). For example, variant rs2465147(T>C) had a significant decrease in NKX2-5 binding affinity, but an increase for TBX5. When observing the risk (C) and non-risk (T) alleles, variant rs2465147 disrupts the NKX2-5 binding motif (5'- CACTT -3') while simultaneously creating a consensus TBX5 binding motif (5'- ACACCT -3') (**Figure 2C, Supplementary Figure 4)**. Likewise, variant rs863392(G>A) disrupted two binding motifs, decreasing the binding affinity of TBX5 (5'- AGGTGT -3') and GATA4 (5'- AGATAA -3') (**Figure 2D**). Our experiment included all four nucleotides in the central base, providing additional information on alleles with no disease or trait association to date. We observed multiple variants where an alternate non-risk allele had a larger effect on TF-DNA binding (e.g., rs57527611 for NKX2-5, rs13353548 for GATA4, and rs3911240 for TBX5, **Figure 2E; Supplementary Figure 5**).

Finally, we explored biochemical mechanisms behind SNPs with allelic binding. Specifically, we asked whether variants with allelic binding directly create or disrupt TF binding motifs. From the 170 SNPs that exhibited allelic differences in binding in the SNP Bind-N-Seq experiments, 126 (74%) directly created (71/126, 56%) or disrupted (55/126, 44%) binding motifs for one of three tested TFs (**Figure 2F**, **Supplementary Figure 6**). The remaining 44 variants (26%) did not create or disrupt a core TF binding motif. Some variants with differential binding occurred in the flanking regions of the TF binding site, which have been previously shown to contribute to TF regulation. [58,59] Another subset of variants created low-affinity binding sites adjacent to core binding motifs of NKX2-5, GATA4, and TBX5. This suggests that for approximately one-quarter of variants that experimentally exhibit allelic binding, changes in TF binding are likely not predicted using motif-dependent models. However, for all three TFs, binding motifs were more likely to be created rather than disrupted (**Figure 2G**).

The experimental scope of this work was limited to only three TFs, which may not fully describe the complete regulatory mechanisms behind CHD etiology. To identify other TFs potentially impacted by the identified CHD-associated variants, we performed a HOMER motif discovery analysis to identify allelic TF binding sites that are being created or disrupted. From the 170 variants that exhibited allelic binding, we identified 931 unique TF binding motifs that occur in either the reference or risk allele (**Supplementary Figure 7A-B**). Of those, 602 (65%) created new TF binding motifs that were absent in the reference allele; the remaining 329 (35%) disrupted motifs in the

reference allele and are absent in the alternate sequence. Motifs were created for 21 TF families (213 TF motifs) and disrupted for 18 TF families (133 TF motifs; **Supplementary Figure 7C**). These findings suggest that for the 170 SNPs identified in our work, motifs are more likely to be created than disrupted, which is consistent with our experimental results for NKX2-5, GATA4, and TBX5. As expected, when SNPs altered binding for NKX2-5, GATA4, and TBX5, we observed created and disrupted motifs from TFs of the same family (e.g., NKX2-2, NKX3-1, GATA3, GATA6, TBX6, TBX21, etc.), which are also involved in the GRNs of heart development (**Supplementary Figure 7D**). [5,14–16,60]

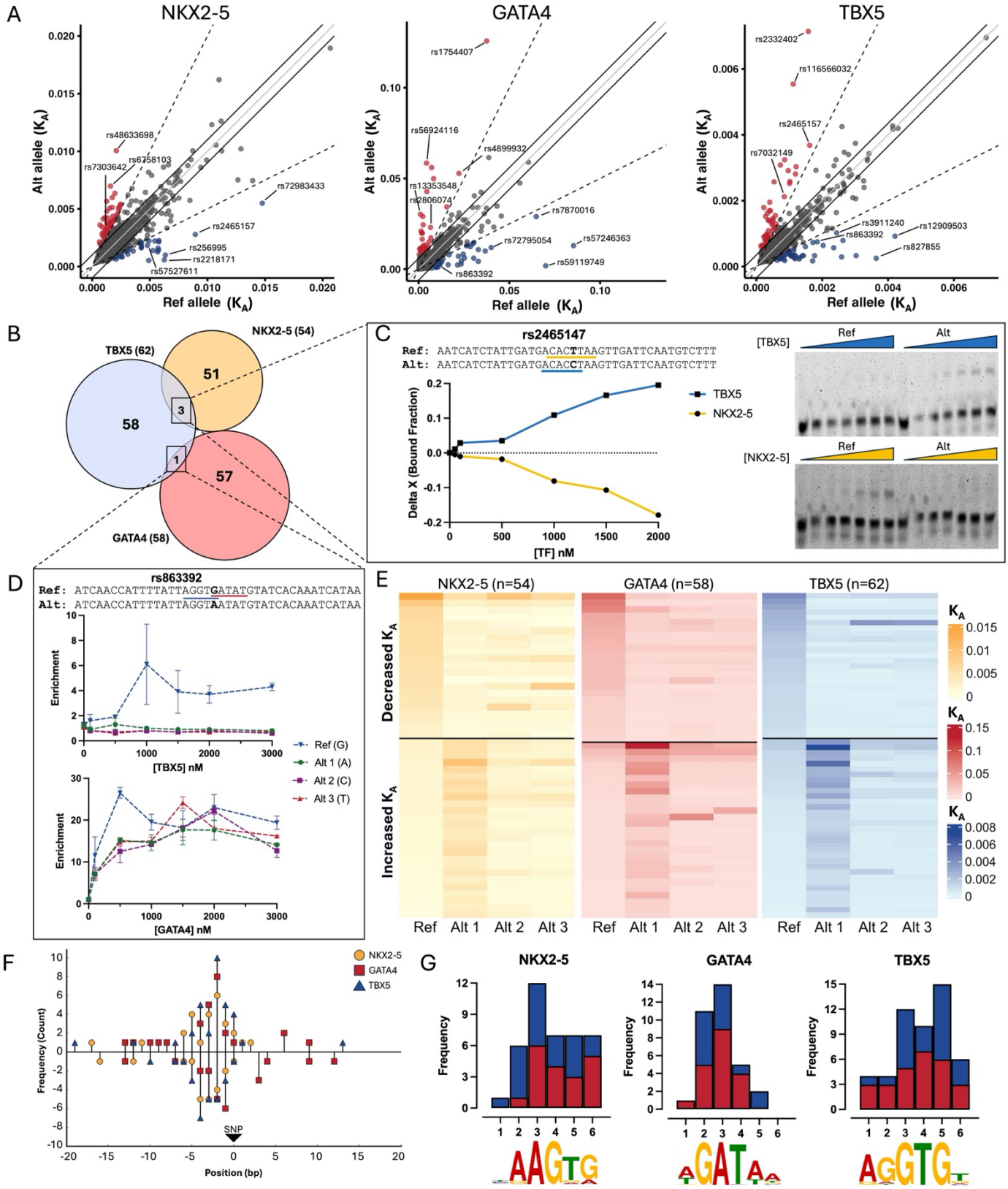


Figure 2: CHD-associated variants exhibit allele-biased binding for cardiac TFs. **A)** Differential binding affinity analysis between reference and CHD-risk alleles. Variants with increased binding affinity (higher K_A value for the alternate allele) are represented in red, while a decrease in binding affinity (higher K_A value for the reference allele) is represented in blue. The solid gray line represents the Y = X intercept with a slope of 1. The dashed line has a 15º angle from the solid line and represents a 2-fold change in binding affinity between the reference and alternate allele. **B)** Venn diagram of CHD-risk variants with differential allelic binding. Overlaps between diagrams represent variants that altered DNA-binding for multiple TFs. **C)** *In vitro* validation of rs2465147 through

EMSA for TBX5 (top) and NKX2-5 (bottom). Binding sites in the reference sequence are underlined in yellow for NKX2-5 and in blue for TBX5 in the alternate sequence. **D)** Allelic enrichment curve of rs863392 for TBX5 (top) and GATA4 (bottom). Reference alleles (Ref) are represented in blue, and tag-SNP alleles from the GWAS catalog (Alt 1) are represented in green. Permutated alleles (alternate non-risk; Alt 2 and Alt 3) are represented in red and purple, respectively. **E)** Heatmaps illustrating allele-specific PrOBEX fitted K-values for SNPs predicted to alter transcription factor binding of NKX2-5 (left), GATA4 (center), and TBX5 (right). Each row corresponds to an individual SNP (rsID), with columns representing the reference allele and all possible alternative nucleotides. “Alt1” denotes the observed alternative allele reported in the GWAS catalog, while “Alt2” and “Alt3” correspond to the remaining permutated alleles. Cell color intensity reflects the magnitude of the fitted K-value, with warmer colors indicating stronger predicted binding affinity. The upper and lower panels display SNPs associated with decreased and increased binding affinity, respectively. **F)** Distribution of TF binding motifs relative to the position of the SNP with allelic binding. Dots represent the number of motifs created or disrupted for NXK2-5 (yellow circles), GATA4 (red square), and TBX5 (blue triangle). The X-axis represents genomic coordinates, a 40 bp window in the SNP-Bind-n-Seq assay. The arrow represents the SNP location at X = 0. **G)** Nucleotide contribution of variants that directly create or disrupt TF binding motifs for NKX2-5 (left), GATA4 (middle), and TBX5 (right). The contribution of created motif counts is presented in red, and disrupted motifs are shown in blue. The motif used to scan variant contribution is displayed below the X-axis, where the value represents position within the motif. The bars in the plot are overlapping, not stacked.

Computational prediction of GWAS SNPs on cardiac TF binding

To further evaluate the potential impact of disease-associated SNPs, we trained three computational models (MinSeqS [61], LS-GKM-SVM[62,63], and MEME[64]) with the NKX2-5, GATA4, and TBX5 *in vitro* data generated from SNP Bind-n-Seq, and scored every variant from the GWAS catalog. [7] For this, we used the top 500 sequences with the highest K_A values of each TF and trained the three models to predict changes in NKX2-5, GATA4, and TBX5 DNA binding (**Figure 3A**). [61,62,64] As a negative set we used random genomic sequences within the same chromosomes that match length and GC content. Performance parameters were determined by a 60-40 split of the dataset, training the model with 60% of the data and scoring the remaining 40% in multiple iteration to determine the area under the receiver operating curve (AUROC; **Figure 3B** and **Supplementary Figure 8A**).

After scoring all disease-associated SNPs from the GWAS Catalog (n = 235,773), we identified 709 variants predicted to alter NKX2-5, GATA4, or TBX5 DNA binding (**Figure 3C**). From those predicted SNPs, 243 (~34 %, Odds ratio = 4.69, p-value = 3.3×10^{-68}) were associated with cardiovascular diseases (CVDs) and traits, mainly blood pressure, CHDs, heart function/structure, and cardiac cell traits (**Figure 3D** and **Supplementary Figure 8B**). Variants with cardiovascular phenotypes had the highest counts for all three TFs, which is unsurprising considering their extensive research and contributions to cardiovascular trait mechanisms. [17,21,22,30,36,65,66] Additionally, we identified non-CVD associated variants, such as neurological and immune diseases and traits (**Supplementary Figure 8B**). Although these TFs are mainly studied in the context of cardiovascular genetics, there is evidence of their role in the regulatory mechanisms of other tissues. For example, GATA4 has been identified to play a key role in liver development and immune responses, and TBX5 has been shown to be involved in limb development. [67–71] These findings could open potential research areas towards establishing functional mechanisms of NXK2-5, GATA4, and TBX5 outside of CVDs and other previously identified developmental pathways.

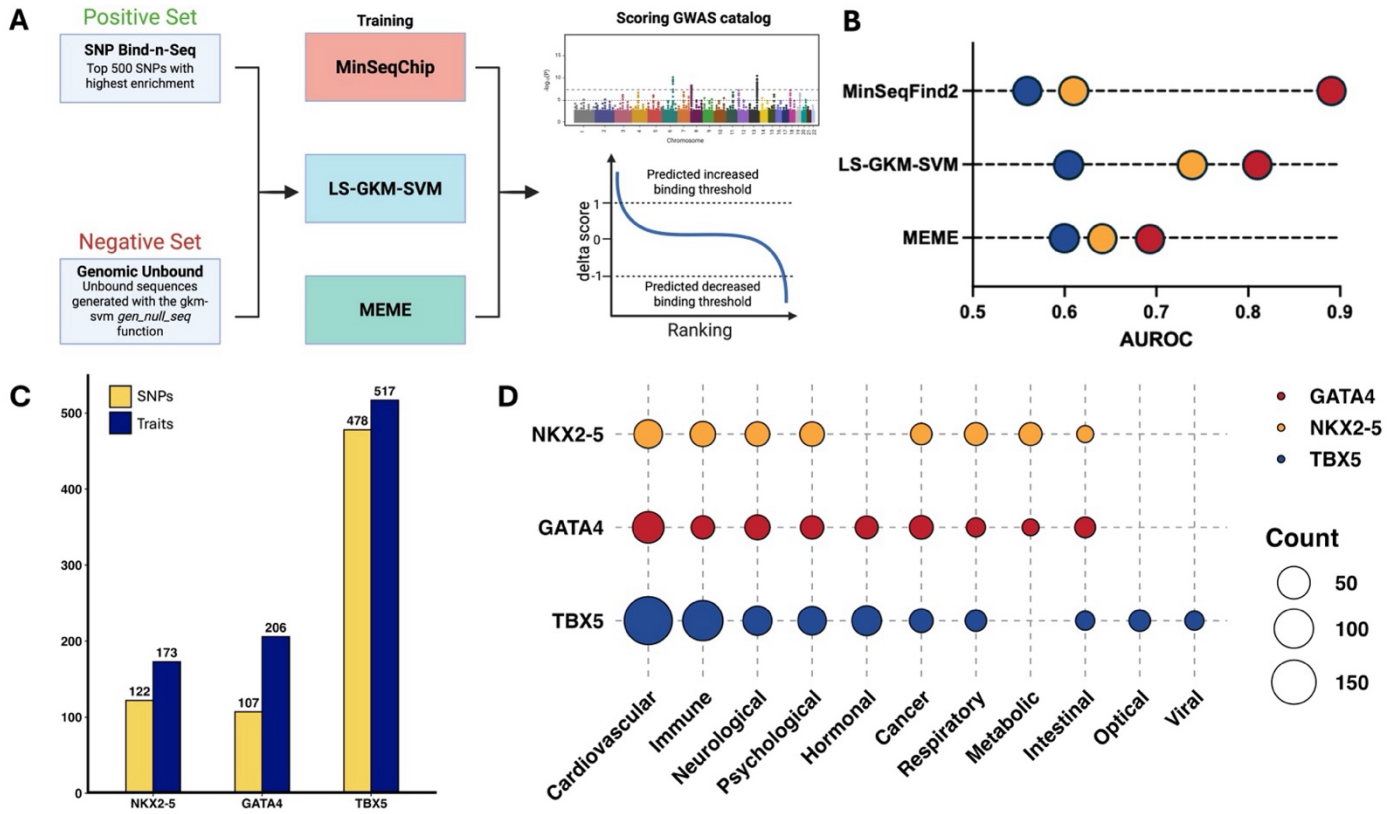


Figure 3: Computational prediction of GWAS catalog variants on cardiac TF-DNA binding. **A)** Schematic of model training using SNP Bind-n-Seq binding data. **B)** Performance parameters of three predictive models trained on SNP Bind-n-Seq binding data. **C)** Number of SNPs (yellow) with traits (blue) from the GWAS catalog predicted to alter NKX2-5, GATA4, and TBX5 binding. **D)** Number of disease-associated SNPs divided by trait parent term predicted to alter NKX2-5, GATA4, and TBX5 binding. Dots in the figure represent NKX2-5 (yellow), GATA4 (red), and TBX5 (blue).

CHD-risk variants in cardiac regulatory elements have transcriptional activity

To identify CHD-risk variants regulating gene expression, we performed a massively parallel reporter assay (MPRA; **Supplementary Figure 9**) in Flp-In cell lines that were modified to stably express NKX2-5, GATA4, and TBX5 (**Supplementary Figure 10**). For the MPRA library, we compiled all CHD-risk associations (n = 157) listed in the GWAS catalog as of February 2024. For each of the tag SNPs, we performed an LD expansion ($r^2 \geq 0.8$) in each ancestry of the initial genetic association(s). In total, we identified 5,431 variants, totaling 14,524 unique sequences with every reported variant allele. Oligos of 170 bp centered on the variant were cloned, synthesized, and labeled with degenerate 20-mer barcode sequences using PCR followed by cloning upstream of an *eGFP* gene under the control of a minimal promoter.

For downstream analysis after performing MPRA, we considered variants with at least 10 unique barcodes, resulting in 14,114 oligos (97.2% of assayed variants; **Supplementary Figure 11A**). The normalized *eGFP* mRNA to plasmid control barcode ratio was used to quantify CRE activity driven by each oligonucleotide.

Using the finalized CHD MPRA library, we first identified variants capable of driving CRE activity. A variant was considered to have CRE activity if any of its alleles had a significant increase in transcriptional activity (*eGFP* mRNA: plasmid DNA ratio, **Supplementary Figure 11B**). Experiments were performed in triplicate with strong correlation of MPRA activity (log2 fold change RNA/DNA) between replicates (R^2 >0.99, **Supplementary Figure 11C**). We therefore defined that a sequence in the MPRA library had CRE activity if the transcriptional activity was significant (p_{adj} <0.05) and increased by at least 50 % (≥1.5 fold-change) when compared to their corresponding barcode count in the plasmid controls. Based on these criteria, 10.6% of CHD-risk variants (574 variants) exhibited CRE activity, which we refer to as "expression modulating variants" (emVars) and "expression modulating alleles" (emAlleles), respectively (**Figure 4A**).

We next examined the potential of the identified emVars to be meaningful in cardiovascular biology by evaluating overlap with putative enhancers and DNase I genomic footprints (DGF) active during heart development and in the adult heart. [72,73] Significance of the overlap between emVars and cardiac CREs was determined by a Fisher's Exact Test using non-emAlleles (remaining oligos from the MPRA library) as a comparison (**Supplementary Figure 12**). We identified 57 emVars (~10%; p-value = 2.10×10^{-29} and Odds Ratio = 4.98) that occur in cardiac enhancers, 48 emVars (~10%; p-value = 1.57×10^{-21} and Odds Ratio = 3.73) in DGF in heart tissue, and 11 emVars (~2%; p-value = 1.61×10^{-23} and Odds Ratio = infinite [i.e., none occurred in the non-emAllele null set]) that occur in both types of cardiac cis-regulatory elements (CREs, **Figure 4B**). From the entire MPRA library, the fact that only emVars overlapped with both types of cardiac CREs suggests a likely important role for these variants in cardiac gene regulation.

Next, we identified functional genomic features (e.g., TF and histone mark ChIP-seq peaks) enriched within emVars in cardiac CREs relative to non-emVars using the RELI algorithm. [74] We observed significant enrichment for ChIP-seq peaks of regulatory proteins involved in cardiovascular processes and diseases, such as RAD21 (p_{adj} = 1.59×10^{-215})[75–77], KLF10 (p_{adj} = 8.39×10^{-107})[78–80], SMAD2 (p_{adj} = 1.60×10^{-70})[81], SRF (p_{adj} = 1.11×10^{-48})[82,83], and members of the GATA family (p_{adj} < 1.11×10^{-26})[84–86]. In particular, RAD21, SRF, GATA3, and GATA1 were among the top 15 peak overlaps and have been

previously implicated in the development of CHDs (**Figure 4C**). [87–89] Additionally, we evaluated TF motif enrichment for emVars in cardiac CREs using HOMER. [90] Through this analysis, we observed enrichment of multiple TF families known to play a role in heart development and cardiac diseases, such as KLF ($p_{adj} < 1.0x10^{-11}$)[91,92], ETS ($p_{adj} < 1.0x10^{-11}$)[44], NK2 ($p_{adj} < 1.0x10^{-10}$)[22,93], CREB ($p_{adj} < 1.0x10^{-9}$)[94,95], and T-box ($p_{adj} < 1.0x10^{-8}$)[17,41,42,96] (**Figure 4D**). Many of the TF families with enriched motifs in emVars also share enriched ChIP-seq peaks from our RELI analysis (**Supplementary Data 3**). Collectively, these analyses identified CHD-risk variants that can drive gene expression from cardiac CREs, which are enriched for ChIP-seq peaks and motifs for TFs with known roles in cardiovascular development and disease progression.

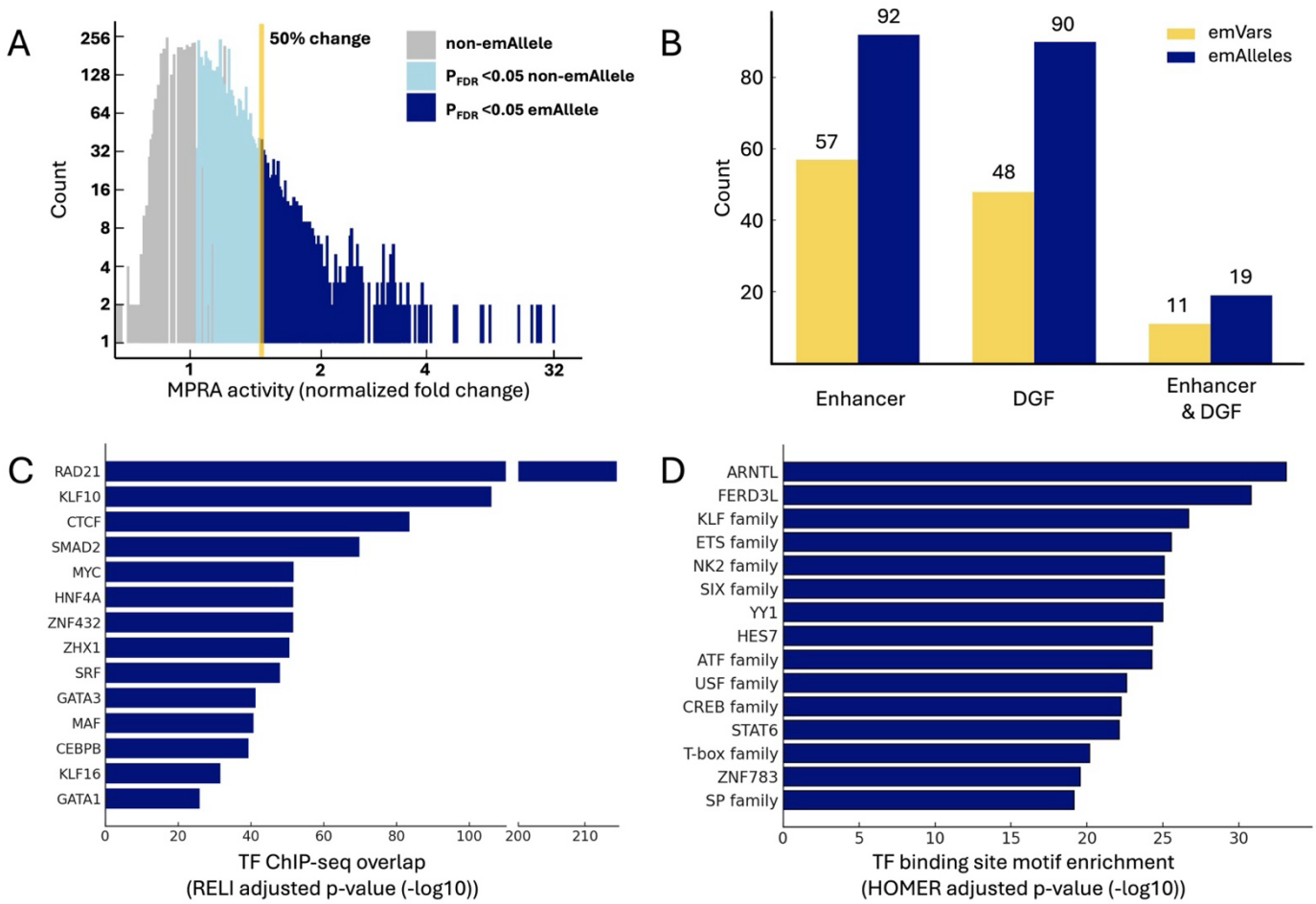


Figure 4: Regulatory activity of CHD-risk emVars. **A)** Distribution of MPRA regulatory activity. The normalized fold change of MPRA activity relative to plasmid control (X-axis) was calculated using DESeq2 (n = 3 biological replicates). Expression modulating alleles (emAlleles; dark blue) were identified as those alleles with significant activity relative to control ($p_{FDR} < 0.05$) and at least a 50% increase in activity. Full results are provided in **Supplementary Data 3**. **B)** Overlap between emVars and cardiac regulatory elements active during heart development and the adult heart. **C)** Enrichment of regulatory protein and TFs binding at emVars. Enrichments were calculated compared to non-emVars. p-values were estimated by a one-sided z-test with Bonferroni multiple testing correction using RELI. The top 15 regulatory proteins and TFs (based on RELI p-values) that overlap at least 10% of emVars are shown. Full results are provided in **Supplementary Data 3**. **D)** TF binding site motif enrichment for emVars compared to non-emVars. p-values were estimated by one-sided hypergeometric test with Benjamini–Hochberg multiple testing correction by HOMER using the full oligo sequences of emVars and non-emVars. The top 15 enriched TF motif families are shown. Full results are provided in **Supplementary Data 3**.

MPRA identifies 187 CHD-risk variants with allelic CRE activity

We next identified CHD-risk variants that exhibited allelic CRE activity, henceforth referred to as "allelic emVars". Variants were considered to have allelic CRE activity if at least one allele was an emAllele, there was a significant change in CRE activity (p_{adj} <0.05), and if there was at least a 25% change in expression compared to the reference allele. Following these criteria, we identified 187 CHD-risk variants (33% of emVars, 3.4% of all CHD-risk variants) as allelic emVars (**Figure 5A**). Of the 187 allelic emVars, 169 (90.4%) had an increased CRE activity compared to their corresponding reference allele, while only 18 (9.6%) resulted in decreased activity. We then intersected the allelic emVars with the putative cardiac CREs to identify potential functional variants in cardiovascular genomics. From the 187 allelic emVars, 19 (10.2%) overlapped with cardiac enhancers, and 7 (3.7%) overlapped with heart DGFs (**Figure 5B**). From the allelic emVars, rs559405101 C>T was the only one to occur in a genomic sequence that is both a cardiac enhancer and DGF, with a 1.94-fold increase (rank 29th in allelic emVars fold-change) in transcriptional activity (**Figure 5B-C**).

As a possible mechanism of action, rs559405101 is flanked by three TBX5 and two GATA4 predicted binding sites, which could be key regulators of *MYOM1* in cardiac tissue (**Figure 5D**). *MYOM1* is overexpressed in muscle-skeletal and cardiac tissue (**Supplementary Figure 13**) and is known to play a role in cardiac muscle function. [97] Additionally, *MYOM1* has been identified as a key gene in multiple types of cardiomyopathies through cardiac gene dysregulation and alternative splicing events. [98–100] To further understand how rs559405101 can regulate gene expression, we identified allelic binding events of TFs (computed by MARIO using ChIP-seq data) that can potentially drive these regulatory mechanisms (**Figure 5E**). Among the top allelic binding events, we identified TFs that play established roles in cardiac development, such as IRX1/2 (heart development and function) [101–103] and POU3F1 (cardiac mesoderm differentiation) [104], which are predicted to preferentially bind to the rs559405101-T risk allele. Conversely, binding motifs of other cardiac TFs involved in heart development, such MECP2 [105–108] and FOXH1 [109,110], are disrupted by the rs559405101-T risk allele. Together, our findings suggest a possible mechanism by which the allelic emVar rs559405101 can alter *MYOM1* expression by altering biochemical interactions with multiple cardiac TFs.

Finally, we identified TFs that are predicted to have allelic binding to the remaining allelic emVars. We identified 376 unique TF motifs that only occur in either the risk or non-risk allele (**Supplementary Data 4**). However, TF-DNA interactions are not always directly disrupted by the variants, but could instead disrupt a TF binding partner or flanking region. [15,58,111,112] To identify which TFs tend to be variant overlapping (within 10 bps of the variant, odds ratio >1.5) versus variant adjacent (**Figure 5F**), we calculated the frequency of binding sites for each TF relative to the emVars and compared these frequencies to random expectation using a proportions test. We identified 26 high-frequency variant overlapping TFs associated with cardiac phenotypes, such as cardiac-specific immune responses (e.g., ASCL2 and IRF3) [113–117], cardiovascular diseases (MAFK, BACH2, and IRF8) [118–124], and heart function and development (OCT6 and TCF21) [125–130] (**Figure 5G**). Among the variant adjacent, many cardiac developmental transcription factors (TFs) were among the most frequent, such as members of the Pitx,

LHX, MEF2, NK2, GATA, and T-box families (**Figure 5H**). [96,131–136] Altogether, integrated computational analyses identified many CHD-risk variants with genotype-dependant expression, along with regulatory genomic features and TFs that could be driving these transcriptional changes.

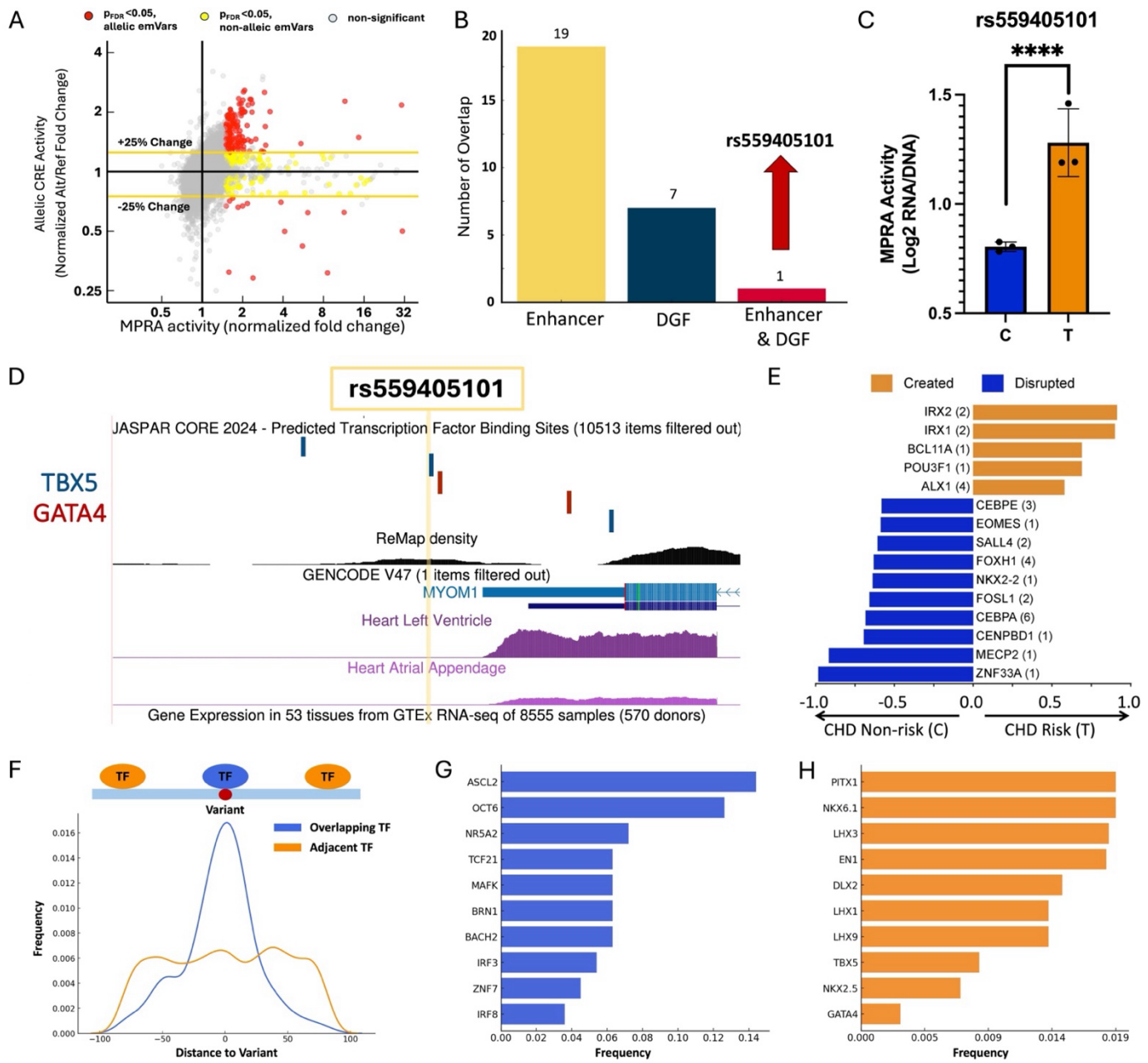


Figure 5: Regulatory activity and mechanisms of allelic emVars. **A)** Identification of variants with allelic CRE activity. Allelic CRE activity (Y-axis) is defined as the normalized fold change of MPRA activity between the non-reference and reference alleles (n = 3 biological replicates). MPRA activity (X-axis) is the normalized fold change of MPRA activity for any allele of the variant. Allelic emVars (red) were defined as variants with a significant difference in MPRA activity ($p_{FDR} < 0.05$) between any pair of alleles and at least a 25% change in activity difference compared to the reference allele. Full results are provided in **Supplementary Data 4**. **B)** Overlap between allelic emVars and cardiac regulatory elements active during heart development and the adult heart. **C)** Normalized MPRA CRE activity of each experimental replicate for rs559405101. **D)** Genome browser map of a 2 kb window centered on rs559405101. Binding sites for GATA4 and TBX5 are displayed as blue and red rectangles, respectively. rs559405101 is upstream of *MYOM1* and is upregulated in the heart left ventricle and atrial appendage. **E)** Genotype-dependent TF binding events predicted for rs559405101. The X-axis indicates the preferred allele, along with a value indicating the strength of the allelic behavior (MARIO ARS value > 0.4), calculated as one minus the ratio of the weak to strong read counts

(e.g., 0.5 indicates the strong allele has twice the reads of the weak allele). Significance (p-value < 0.05) was determined relative to binding events found in non-emVars sequences. Values in parentheses next to the TF name are the number of binding events created or disrupted by that specific TF. **F)** TF binding site location distribution for variant overlapping (blue) and variant adjacent (orange) TFs, relative to all allelic emVars. **G-H)** Motif enriched for TFs categorized as **G)** variant overlapping (Odds Ratio > 1.5, blue) and **H)** adjacent (Odds ratio < 1.5, orange) to the allelic emVars. Full results for figures 5F-H are provided in **Supplementary Data 4**.

CHD-risk variants exhibit allelic binding and expression

We next intersected our list of 187 allelic emVars with the 170 SNPs that exhibited allelic binding with cardiac TFs to identify potential CHD-causal variants. After overlapping the datasets, we identified five allelic emVars with allelic binding: NKX2-5 (rs2137643 and rs7303642), GATA4 (rs2896074 and rs839154), and TBX5 (rs7032149) (**Figure 6A**). The overlap between both datasets was significant by Fishers Exact Test (p-value = 0.032), suggesting that the association is likely not random. Thus, these variants were selected to be further explored as important functional CHD allelic mechanisms. The correlation between allelic binding and allelic MPRA activity throughout the CHD library was low for all three TFs (**Figure 6B**), which is not surprising considering the experimental differences between SNP Bind-n-Seq (*in vitro* with recombinant TFs) and the MPRA (cell-based assay). Additionally, less than 5% of the library actually has motifs for NKX2-5, GATA4, or TBX5. Even with motif matches, gene regulation is a highly complex process where TF binding does not always lead to transcriptional changes, as is the case with transcriptional repressor proteins and silencer elements. [137–140] Particularly relevant to this work, NKX2-5, GATA4, and TBX5 have all been previously reported to have dual activating and repressive functions. [65,86,141,142] However, variants rs2137643, rs7303642, rs2896074, and rs7032149 all resulted in a significant increase in both TF binding and expression (**Figure 6C-D**). Variant rs839154 was the only SNP to have opposite effects in regulatory mechanisms, with an increase in expression but a decrease in GATA4 binding, suggesting a repressing role for GATA4 at this locus.

Next, we proceeded to identify variants with differential allelic binding (170 SNPs) and transcriptional activity (187 allelic emVars) that are eQTL in cardiac tissue (heart left ventricle and atrial appendage). From the CHD-risk variants with allelic binding, 44 (26%) were in cardiac eQTL affecting 19 genes, with over 60% sharing direction of the effect size (e.g. increase for both TF binding and eQTL effect size; **Supplementary Figure 14A**). From the allelic emVars, 34 variants were also in cardiac eQTL with 33 genes (**Supplementary Figure 14B**). We identified seven eQTL genes that overlapped between the SNP Bind-n-Seq and MPRA variants and constructed interaction networks with both sets **(Figure 6E and Supplementary Figure 14C-D**). An interaction network revealed that rs7303642 and rs2896074 were eQTL for two overlapping genes, *MGAT4C* and *GALC*, respectively. MGAT4C (Mannosyl Alpha-1,3-Glycoprotein Beta-1,4-N Acetylglucosaminyltransferase) is an enzyme that is highly expressed in the heart and has been linked to CHDs, such as conotruncal heart defects. [143,144] Additionally, 14 eQTL variants for *MGAT4C* altered the binding affinity of NKX2-5 (4 SNPs), GATA4 (5 SNPs), and TBX5 (5 SNPs) when biochemically evaluated through SNP Bind-n-Seq (**Figure 6F-G**, **Supplementary Figure 15**). GALC (Galactosylceramidase) is an enzyme involved in galactose metabolism that has been implicated in several diseases, particularly respiratory and neurological diseases. [145,146] Although having modest expression in cardiac tissue, to date, there is no *GALC* implication with specific cardiovascular diseases or phenotypes.

Finally, the interaction network revealed additional cardiac eQTL genes with eQTLs that displayed altered binding of all three TFs tested, such as *CNOT6L*, *VARS2*, and *DDR1*. For example, 20 SNPs that are *CNOT6L* eQTL also that exhibited allelic binding for the three TFs (**Supplementary Figure 16**). CNOT6L is a protein subunit of the

CCR4-NOT complex, which regulates mRNA deadenylation and has been implicated in proper heart functioning. [147–151] In an interesting case, *VARS2* and *DDR1* are eQTL genes of the same eight variants with allelic binding, where *VARS2* consistently had an increased effect size and *DDR1* a decreased effect size (**Supplementary Figure 17**). *DDR1* is a discoid domain receptor that is involved in cardiovascular calcification, while *VARS2* is a tRNA synthetase that has been linked to heart failures when mutated. [152–155] In summary, our combined findings provide a high-confidence list of possible mechanisms of CHD-risk variants that exhibit allelic binding, transcriptional activity, and cardiac eQTL.

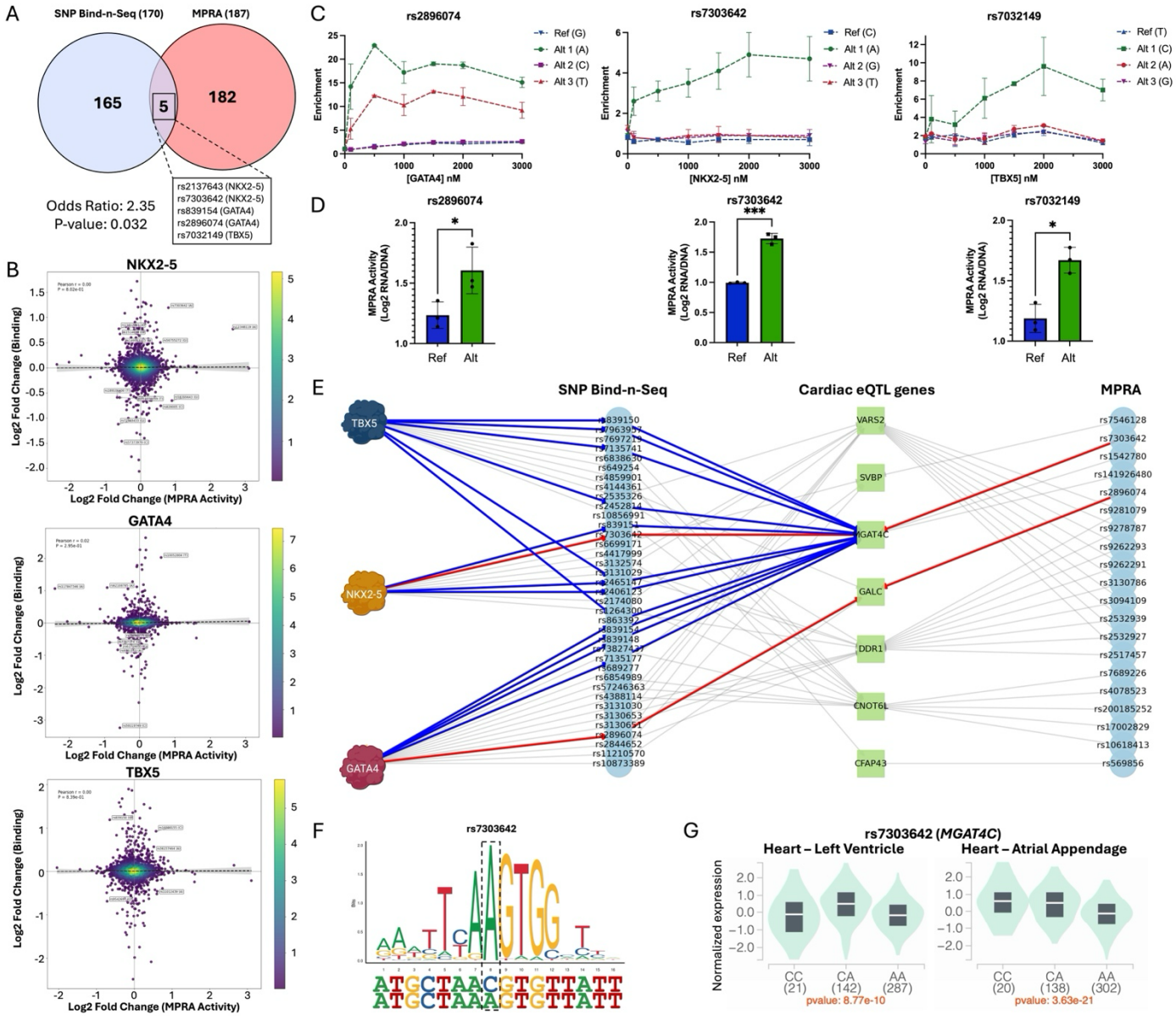


Figure 6: CHD-risk variants with allelic binding and regulatory activity. **A)** Venn diagram of common variants with allelic TF binding (SNP Bind-n-Seq, blue) and transcriptional activity (MPRA, red). Five common variants are displayed with the TF that showed differential binding. Significance in the association between allelic binding and gene expression was determined by Fisher's Exact Test (p-value < 0.005). **B)** Density plot correlating expression fold change of the allelic emVars with binding fold change of NKX2-5 (left), GATA4 (middle), and TBX5 (right). Variants with a 25 % change in both binding and expression are labeled on the plot. **C)** Representative binding curve for each TF for a variant that altered the binding of NKX2-5 (top), GATA4 (middle), and TBX5 (bottom). Reference alleles (Ref) are represented in blue, and tag-SNP alleles from the GWAS catalog (Alt 1) are represented in green. Permutated alleles (alternate non-risk; Alt 2 and Alt 3) are represented in red and purple, respectively. **D)** MPRA activity for the variants used for binding curves in (**C**). **E)** Interaction networks of cardiac eQTL genes with variants exhibit allelic binding and/or gene expression. Interactions highlighted in red indicate cardiac eQTL variants that exhibited allelic behavior in binding (SNP Bind-n-Seq) and expression (MPRA). Interactions highlighted in blue indicate eQTL variants for *MGAT4C* that altered binding for all three TFs. **F)** DNA-binding motif logos are shown for NKX2-5 in the context of the DNA sequence surrounding rs7303642. **G)** eQTL indicating

MGAT4C expression dependent on rs7303642 genotype in the heart left ventricle (data from the GTEx portal).

Discussion

Genome-wide association studies (GWAS) have linked thousands of non-coding genomic loci to congenital heart disease (CHDs).[51–55,156–159] However, translating these associations into meaningful biological mechanisms remains a challenge. This work provides the first combined genome-wide evaluation of cardiac TF binding and transcriptional activity, encompassing over 3,000 CHD-risk variants. Our approach provides crucial information needed to understand biochemical and genetic mechanisms underlying CHD etiology, and ultimately for identifying causal variants.

To date, over 90% of CHD-associated variants have been mapped to the non-coding genome and thus could potentially alter the function of cardiac TFs during heart development.[2,4,6,10] With SNP Bind-n-Seq, we build upon previous experimental approaches[45,46] to assay thousands of variants in a high-throughput manner. Comparted to other techniques to profile TF binding, SNP Bind-n-Seq can quantify enrichment across all variant alleles simultaneously and determine allelic binding constants, which has previously not been done. As proof of concept, we chose NKX2-5, GATA4, and TBX5, three cardiac TFs crucial for heart development that have been implicated in CHD etiology, to evaluate CHD-risk variants for allelic binding events.[17,18,20–22,26,66,160] Using SNP Bind-n-Seq, we constructed allelic binding curves for 3,232 CHD-risk variants (12,928 alleles) for all three cardiac TFs. Our findings provide ~38,400 unique DNA binding affinity measurements, with 5.3% of variants exhibiting allelic binding for at least one of the three TFs tested in this study. This approach also provided mechanistic insight into the observed allelic binding events, with only 74% of the examined SNPs directly creating or disrupting the binding motifs of NKX2-5, GATA4, and TBX5. The remaining 26% altered cardiac TF binding by creating low-affinity binding sites adjacent to core binding motifs or occurred within TF binding flanking regions.

The data generated through SNP Bind-n-Seq were subsequently used to train computational models to predict changes in TF-DNA binding. In doing so, we scored all of the disease-associated variants contained in the GWAS catalog to predict allelic binding events for NKX2-5, GATA4, and TBX5. We identified over 1,000 variants predicted to alter the binding of these three TFs that are associated with multiple types of diseases, not just cardiovascular.. Thus, this study provides insight into potential roles for NKX2-5, GATA4, and TBX5 in disease etiology outside of CHDs. SNP Bind-n-Seq is a scalable approach that can be implemented with other TFs or DNA-binding proteins to study biochemical mechanisms of multiple genetic diseases associated with non-coding variants.

We complemented our *in vitro* binding affinity data with transcriptional activity in a cellular context using an MPRA to quantify changes in transcriptional activity. Since our previous approach was based on three cardiac TFs involved in heart development, we performed the MPRA in a Flp-In 293 system that was genetically modified to stably express NKX2-5, GATA4, and TBX5. Our results showed that 574 emVars (10.6% of tested variants) drove CRE activity, and of those, 16.4% occur within known cardiac CREs that are active during heart development.[72,73] We identified 187 CHD-risk variants (3.5% of tested variants) with allelic CRE activity, of which 13% occurred in cardiac CREs. Our downstream computational analyses also provide mechanistic insight into the likely regulatory proteins driving these transcriptional changes, along with their

affected target genes. Our MPRA findings have some limitations, including the cell line of choice, which lacks cardiac-specific chromatin features and other cardiac TFs beyond those within the scope of this work. However, previous studies on the genotype dependence of cis-eQTLS have shown that variants exhibit a bimodal distribution, where expression patterns are either single tissue-specific or shared across most tissues and cell types. [161–163] Additionally, previous work has shown that MPRA activity is highly correlated across multiple cell types. [164]

Finally, our study reveals possible mechanisms of CHD-risk variant that may make important functional contributions to CHD. After combining our findings from the SNP Bind-n-Seq and MPRA assays, we identified five variants (rs2137643, rs7303642, rs2896074, rs839154, and rs7032149) that can be further studied for important roles in CHD. Notably, variants rs7303642, rs2896074, and rs839154 are eQTLs for genes expressed in cardiac tissue, which may further support their role in cardiovascular disease genomics. Additionally, the combined findings identified variants with allelic regulatory activity that were cardiac eQTL for seven genes. Among them, *MGAT4C*, *CNOT6L*, *VARS2*, and *DDR1* were the eQTL genes of multiple variants exhibiting allelic binding for NKX2-5, GATA4, and TBX5. Altogether, our combined findings through high-throughput TF-DNA binding and reporter assays provide the largest biochemical and genetic evaluation of thousands of CHD-risk variants.

With the advancement in sequencing technologies and rapid identification of variants through GWASs, identifying and validating disease-causing variants remains challenging. In this study, we present a functional framework for identifying and assessing thousands of variants that alter regulatory mechanisms, including TF binding and regulatory activity. We believe that SNP Bind-n-Seq, coupled with other high-throughput methods like MPRA, is a scalable approach to dissect the regulatory mechanisms of non-coding variants and understand the genetic etiology of many complex human diseases.

Limitations of Study

This work, although providing valuable insight into biochemical and genetic mechanism behind CHD-risk variants, has limitations that can be addressed in future experiments. First, *in vitro* binding assays were performed on purified TF DBDs from a bacterial expression system. This has limitations, such as proteins lacking post-translational modifications (PTMs) that occur in eukaryotic cells. Since binding is performed *in vitro*, biochemical interactions are performed out of cellular context without binding partners and co-factors. Additionally, binding is performed using short (40 bp) synthetic DNA sequences lacking sufficient genomic context. Second, the reporter assay was performed on genetically modified Flp-In 293 cell lines to express the three cardiac TF from the binding assays. This system lacks the cardiac-specific cellular context and gene expression profiles of biologically relevant cell lines, such as cardiomyocytes and cardiac progenitors. In the future, our framework can be adapted to perform binding assays using nuclear extracts of cardiac cell lines, as well as to conduct the MPRA at multiple developmental stages of cardiomyocyte differentiation.

Resource Availability

Lead contact: Jose.rodriguez233@upr.edu

Materials availability

The CHD libraries used for the SNP Bind-n-Seq and MPRA, and the genetically modified cell lines used in this work, are available upon request.

Data and code availability

The original code designed in this work for SNP Bind-n-Seq is available at: https://github.com/Shreya-droid/SNPoiss_bind_n_seq. Code for MPRA barcode count and analysis is available at: https://github.com/tewhey-lab/MPRAmodel. Code for MPRA plotting is available at: https://github.com/WeirauchLab/mpraprofiler. Code for RELI analysis is available at: https://github.com/WeirauchLab/RELI.

Declaration of interest

The authors declare no competing interests.

Author Contributions

EGPM and JARM designed and conceptualized the project. EGPM, DAPM, JLMB, LSA, ARM, and RVR performed protein purifications and TF-DNA binding assays (SNP Bind-n-Seq and validations through EMSA). SS, JGMF, JAFR, and DB performed computational analysis for the SNP Bind-n-Seq sequencing data. EGPM, ACBR, JMRR, HH, KD, and MTH performed Flp-In 293 genetic modification and cell culture. EGPM, ER, LGP, and MG performed MPRA experiments. EGPM, MN, OAD, ER, and XC performed computational analysis of MPRA data. EGPM and SS performed plotting and data visualization. LPL, CF, MTW, LCK, SKR, DB, and JARM supervised the work, provided mentoring, and secured funding for the project. EGPM, SS, DB, and JARM wrote and reviewed the original manuscript. All authors read, assisted in editing, and approved the final version of the manuscript.

Acknowledgments

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 an NIH Institutional Development Award (IDeA) INBRE (P20GM103475W). EGPM, JMRR, RVR, DAPM, ACBR, LSA, and NEMP were funded by the NIH RISE Fellowship (5R25GM061151–20). JLMB and ARM were funded by the NSF PR-LSAMP fellowship (HRD-2008186). DAPM was funded by NSF [IQ BIOREU 1852259]. EGPM and JMRR were funded by the NSF BioXFEL Fellowship (STC-1231306). ARM was funded by the NSF REU funded ARM: PR-CLIMB Program (2050493) and NIH 1T34GM145404. LSA was funded by the NIH ID-GENE Fellowship (1R25HG012702–01). JMRR was funded by an NSF graduate research fellowship (1744619). MN was supported by a Gruber fellowship. MTW and LCK were funded by NIH grants R01NS099068, R01AI024717, R01AI148276, U24 HG013078, U01 HG011172, and P30AR070549. SKR was funded by R01HG012872. The graphical abstract and Figure 1A were made using BioRender.

Materials and Methods

Variant selection and DNA sequence generation

For the SNP Bind-n-seq experiment, we downloaded 121 CHD-associated SNPs from six studies in several populations from the GWAS from the catalog. [51–56] Variants in linkage disequilibrium (LD) from four populations (EUR, AFR, EAS, SAS, R^2=0.80) were included using the TOP LD web tool. [165,166] Insertions, deletions, and incomplete entries from the GWAS catalog were removed. All variants were permutated to include every possible nucleotide for each CHD-associated variant. We retrieved 40 bps of hg38-flanking DNA sequences for every allele, with the variant located in the center (19 bps upstream and 20 bps downstream of the variant). Adapters and unique molecular identifiers (UMIs) were added to each sequence at either end (5'- TCCCTACACGACGCTCTTCCGATCT - NN - [40 bp oligo] - NN -GATCGGAAGAGCACACGTCTGAACTCCAGTCAC -3') to make a 102 bp DNA sequence. A total of 13,039 oligos (3,232 variants, 12,928 alleles, 91 cardiac DGF, and five control variants) were obtained from Custom Array.

For the MPRA, we downloaded 157 CHD-associated SNPs from the nine GWAS in several populations that were expended for variants in linkage disequilibrium (LD, R^2> 0.8) based on 1000 Genomes Data in the ancestry(ies) of the initial genetic association using PLINK(v1.90b) as previously described. [47,167] All expanded variants were updated to the dbSNP 155 table from the UCSC table browser based on either variant name or genomic location. Unmappable variants were discarded. For single nucleotide polymorphisms, we pulled 170 base pairs (bps) of hg38-flanking DNA sequences for every allele, with the variant located in the center (84 bps upstream and 85 bps downstream of the variant). For the other types of variants (indels), we designed the flanking sequences to ensure that the longest allele has 170 bps. Adapters (15 bps) were added to each sequence at either end (5'- ACTGGCCGCTTGACG - [170 bp oligo] - CACTGCGGCTCCTGC-3') to make a 200 bp DNA sequence. For all resulting sequences, we created a forward and reverse complement sequence to compensate for possible DNA synthesis errors. A total of 29,048 oligos (5,431 variants, 14,524 alleles) were obtained from Twist Bioscience.

NKX2-5, TBX5, and GATA4 expression and purification

The human NKX2-5 homeodomain (HD) coding sequence (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. [35] The coding sequence for the human TBX5 T-box domain (Met51 to Ser248) was cloned into a pET expression vector with a 6x His•Tag® (VectorBuilder Inc.) and purified through Ni-NTA affinity chromatography, as previously described. [35] Human GATA4 DNA-binding domain (DBD), also known as a zinc finger domain (ZF), gene (Uniprot: P43694, Met207 to Ala333) was cloned into the pET-28a(+) vector with an N-terminal 6x-Histag (Twist Bioscience).

SNP Bind-n-Seq

An oligopool of CHD-associated variants (Custom Array, Supplementary File) was amplified for 12 cycles by touchdown PCR (denaturing at 95 ºC for 20 sec, annealing at

70 - 0.5 ºC/cycle for 15 sec, and extension at 72 ºC for 15 sec) using KAPA HiFi HotStart ReadyMix (Roche Sequencing Store, #KK2602 07958935001). Sequences were amplified with IR700 fluorescent and biotinylated primers for subsequent gel excision and extraction, respectively.

Binding reactions for TBX5 and NKX2-5 were 20 µL (50 mM NaCl, 10 mM Tris-HCl (pH 8.0), 10% glycerol, 1000 ng pdI-dC, 0.5 µg BSA, 0.02% Tween-20, and 1 mM DTT) containing 50 ng of oligopool library DNA. For GATA4, binding reactions were 20 µL (10 mM HEPES (pH 8.0), 0.5 mM ZnA, 100 mM NaCl, 5% Glycerol, 1000 ng pdI-dC, 0.5 µg BSA, 0.02% Tween-20, and 1 mM DTT) containing 50 ng of oligopool library DNA. For each TF, six concentration points were used for binding assay: 0, 100, 500, 1000, 1500, 2000, and 3000 nM. Reactions were incubated for 30 min at 30 ºC, followed by a 30 min incubation at room temperature, and samples were loaded onto a 6% polyacrylamide gel in 0.5x TBE (89mM Tris/89mM boric acid/ 2mM EDTA, pH 8.4). The gel was pre-ran at 70 V for 15 min, loaded at 30 V, and resolved at 120 V for 2 hours at 4 ºC. The gels were revealed using Azure Sapphire Bio-molecular Imager with 658nm/ 710nm excitation and emission.

After EMSA, bound and unbound bands for each concentration point were cut, and DNA was extracted and left overnight in 500 µL of EB buffer (Qiagen, #19086) at 30 ºC shaking at 1200 rpm (Thermoshaker, BIOGRANT). After gel extraction, DNA was purified using Dynabeads M-280 Streptavidin (Invitrogen, #11205D) according to manufacturer procedures. Barcodes and sequencing adapters were added to bound and unbound sequences through 13 cycles of PCR (denaturing at 95 ºC for 30 sec, annealing at 64 ºC for 30 sec, and extension at 72 ºC for 30 sec).

SNP Bind-n-Seq datasets and preprocessing

The sequencing was performed using the Illumina platform, generating paired-end reads of 150 base pairs (PE150). This produces two reads per DNA fragment one from each end. For this analysis, only Read 1 (from the 1.fastq file) was used. Next, to extract our genomic fragment which is less than 150bp, we used the constant right flanking region from the library structure "GATCGGAAGAGCACACGTCTGAACTCCAGTCAC" to locate 44 bp DNA sequence that includes a SNP position. This is done with all the FASTQ files for each TF: NKX2-5, GATA4, and TBX5. Our library contained a 40 bp genomic DNA sequence along with a 4 bp Unique Molecular Identifier (UMI), two at each end. For each sequencing read, we generated a pattern where the SNP at the 20th position could be occupied by any of the four nucleotides: A, T, G, or C. We then calculated the frequency of each 4 possible nucleotide at the SNP position across all sequencing reads, giving us the counts for each SNP position for a sequence.

Enrichment Analysis

For each nucleotide at a given SNP position, enrichment was calculated relative to the unbound dataset. To avoid division by zero and reduce the impact of low-count noise, pseudo-counts were added to both bound and unbound read counts. The pseudo-counts were defined as a small background signal equivalent to 2.5 reads per million, scaled by the total number of reads in the respective bound or unbound dataset. Quantitatively, the pseudo-counts were computed as (2.5 X total reads)/10^6. Enrichment

for each nucleotide was then calculated by dividing the pseudo-count-adjusted count in the bound sample by the corresponding adjusted count in the unbound sample. This approach allows a robust comparison between bound and unbound conditions, generating enrichment profiles that reflect nucleotide-level binding preferences at SNP positions.

Before calculating the enrichment profile, we ensure data quality and consistency, and thus remove low-quality sequences based on specific thresholds designed to minimize deviations that could compromise data uniformity. To control for differences in sequencing depth between samples, the raw counts of each unique sequence, regardless of the nucleotide identity at the SNP position, were normalized by the total number of sequences in the corresponding sample. These normalized values were then scaled to parts per million (PPM). Then, the thresholds applied were as follows:

1. **Filtering Sequences from 0nM Unbound Condition**: To ensure a better analysis, a ratio threshold of **≤ 0.2 or ≥ 2** was applied to the 0nM Unbound-to-Library ratio.

$$\text{Ratio}_{\left[\frac{\text{unbound}}{\text{library}}\right]} = \frac{\text{fraction of sequence in 0 nM Unbound}}{\text{fraction of sequence in Library}} \quad (1)$$

   Samples with a 0 nM Unbound-to-Library ratio outside the ≤ 0.2 or ≥ 2 range were excluded from further analysis to minimize bias from gel-based irregularities.

2. **Low Counts in 0nM Unbound**: Sequences with read counts **less than 5 per million** (in both replicates, R1 and R2) in the Unbound fraction at 0nM were excluded, as this low representation indicates a lack of binding specificity at baseline.

3. **Low Counts in Library**: Sequences with read counts **less than 5 per million** in the original library sample were also removed to ensure that subsequent analyses focused only on sequences present in sufficient abundance.

4. **Low Ratio Threshold for 0nM Bound-to-Unbound Analysis**: A ratio threshold of **≤ 0.2 or ≥ 5** was applied to identify sequences highly enriched in the Unbound fraction (indicating low binding affinity) or overly represented in the Bound fraction (suggesting nonspecific binding at 0 nM).

5. **Sufficient Representation of Nucleotides**: Sequences with counts **below 5 per million** across all nucleotide bases (A/T/G/C) were excluded to ensure analyses included only sequences present in adequate quantities across all samples.

These criteria were essential for refining the dataset, focusing on sequences with meaningful binding behavior while eliminating potential artifacts or non-specific interactions. For those SNP sequences that have passed these thresholds, then the enrichment calculation can be expressed as the ratio of the relative frequencies of a

base in the sample data to its unbound sample. The Enrichment is defined as the ratio of relative fractions as following:

$$\mathrm{Enrichment}(S_{ij}) = \frac{f_{S_{ij}}^{bound}}{f_{S_{ij}}^{unbound}} = \frac{\left[\frac{n_{S_{ij}}^{bound}}{N_j^{bound}}\right]}{\left[\frac{n_{S_{ij}}^{unbound}}{N_j^{unbound}}\right]} \tag{2}$$

where,
$f_{S_{ij}}^{bound}$ fraction of sequence S_i at j^{th} concentration in the bound sample,
$n_{S_{ij}}^{bound}$ is the count of sequence S_i at j^{th}concentration in the bound sample,
N_j^{bound} is the total sequence count at the j^{th}concentration in the bound sample,
$f_{S_{ij}}^{unbound}$ fraction of sequence S_i at j^{th} concentration in the unbound sample,
$n_{S_{ij}}^{unbound}$ is the count of sequence S_i at j^{th}concentration in the unbound sample,
$N_j^{unbound}$ is the total sequence count at the j^{th}concentration in the unbound sample.

Enrichment indicates how much more (or less) frequent a sequence is occurring in the sample compared to its corresponding unbound sample. The significance of these calculations is that the enrichment value gives you an idea of how over- or under-represented a sequence is in your data compared to an unbound fraction. Enrichment was calculated for every sequence for each factor at all concentrations and for both the replicates, this was then used for plotting. After applying all the above thresholds, we generated a dataset that contains binary values (0s and 1s) for each SNP sequence (rsID's), indicating whether the respective sequence passed or failed each of the specified thresholds (see section). Only those rsID's that met all threshold criteria were retained for subsequent analyses, including model fitting.

Estimating Binding Affinity and Probability of Molecular Interaction

The probability of TF binding to DNA depends on the sequence-specific binding affinity, driven by direct molecular contacts like hydrogen bonds and hydrophobic interactions between TF amino acids and DNA bases. This affinity is quantified by the dissociation constant K_d, which reflects the free energy of binding. TF concentration influences binding via mass action, while chromatin accessibility modulated by nucleosome positioning and epigenetic modifications regulates physical DNA availability. Nonspecific binding results from weaker electrostatic interactions with the DNA phosphate backbone and bases outside consensus motifs, contributing to TF search efficiency and buffering TF concentration.

Here, we use a statistical mechanics approach to derive binding affinity and other equilibrium properties of the system. In our experiment the protein is exposed to a pool of 12940 SNP sequences simultaneously, each sequence consists of 44 nucleotides

containing up to n number of cognate binding sites. Assuming independent binding at each site the occupancy probability of sequence S_i at concentration j is calculated as

$$\mathrm{P}(\mathrm{S_{ij}}) = \frac{1}{1+\left(\frac{1}{\mathrm{K_i C_j}}\right)} = \frac{\mathrm{K_i C_j}}{1+\mathrm{K_i C_j}} \quad (3)$$

where, $\mathrm{K_i}$ is the association constant for sequence i at the $\mathrm{j^{th}}$ concentration, $\mathrm{C_j}$.

To account for DNA outside bound and bound fraction (e.g., stuck in the well of the gel, uneven gel migration, etc.), we introduced an additional parameter w_i, representing the probability that a given DNA molecule is *"stuck in the well"*. This decision further supported by insights from the enrichment line plots, which revealed that previously fitted K-values did not fully capture the behaviour of certain sequences. These deviations prompted us to introduce an additional parameter, w_i, to better model such effects and more accurately represent the binding behaviour under these conditions.

To substantiate the use of an additional parameter, w_i, we examine read count distributions across three experimental fractions: the original library (L), the unbound fraction (U), and the bound fraction (B). The library contains the initial representation of all DNA sequences before exposure to the TF, while the unbound fraction contains sequences that did not bind to the TF, and the bound fraction contains those that did. At 0nM TF, where no specific TF-DNA interactions should occur, we expect most sequences to remain in the unbound fraction. That is, for any sequence *i*, its unbound read count (U_i^0) fraction should be approximately equal to its initial library count (L_i) fraction:

$$U_i{}^0/U_{total} \cong L_i/L_{total} \quad (4)$$

However, when $U_i{}^0/U_{total} \cong L_i/L_{total}$, indicates that sequence *i* is unexpectedly depleted from the unbound fraction, even though no TF is present to cause this. This suggests the sequence might be subject to non-specific loss, perhaps binding to the experimental surface, interacting with background proteins, or being inefficiently recovered. More generally, if we observe that the total recovered reads $B_i/B_{total} + U_i/U_{total}$, are much smaller than the initial library pool (L_i), i.e.,

$$B_i/B_{total} + U_i/U_{total} << L_i/L_{total} \quad (5)$$

This supports the same conclusion that the sequence is systematically underrepresented and should be excluded to avoid introducing artifacts into the binding analysis.

In the updated equation we defined the range of $w_i \in [0,1]$, where w_i = 0 means the molecule is entirely unbound or completely stuck in the well and w_i = 1 completely bound with the TF, suggesting strong binding. Thus, we updated the probability of sequence to be found in the bound region of gel is given by equation (4)

$$P(S_{ij}) = \left(\frac{1}{1+\left(\frac{1}{K_i C_j}\right)} \right)(1 - w_i) \tag{6}$$

Our modelling approach integrates bound and unbound read counts along with library data to jointly infer the key kinetic parameters. Due to the competitive binding environment, these constants must be inferred jointly, as the observed binding reflects the relative affinities of all sequences in the pool. The model specifically optimizes the concentration-independent parameter K_i, which controls the observed SNP read counts across varying protein concentrations, thereby capturing the system's dynamic behaviour. In this framework:

Using the Poisson probability mass function:

$$P\left(n_{S_{ij}} | \lambda_{S_{ij}}\right) = \frac{\lambda_{S_{ij}}^{n_{S_{ij}}} e^{-\lambda_{S_{ij}}}}{n_{(S_{ij})}!} \tag{7}$$

given the observed counts $n_{S_{ij}}$, assumed to follow a Poisson distribution

$$n_{S_{ij}} \sim Poisson\left(\lambda_{S_{ij}}\right) \tag{8}$$

In our model, each sequence i is associated with a kinetic parameter K_i, which represents its effective association constant and determines its binding affinity in the competitive environment. The expected read count $\lambda_{S_{ij}}$ for sequence *i* at *j* is a function of K_i and the protein concentration, capturing the binding probability under equilibrium assumptions. Then the likelihood function is

$$L(K) = \prod_{S_{i,j}} P\left(n_{S_{ij}} | \lambda_{S_{ij}}\right) = \prod_{S_{i,j}} \frac{\lambda_{S_{ij}}^{n_{S_{ij}}} e^{-\lambda_{S_{ij}}}}{n_{(S_{ij})}!} \tag{9}$$

Taking the logarithm, log-likelihood function is:

$$log\ L(K) = \sum_{S_{ij}} \left[n_{S_{ij}}\, log\lambda_{ij} - \lambda_{ij} - log\,(n_{S_{ij}}!)\right] \tag{10}$$

The goal is to maximise the concentration independent parameter K, which influences the likelihood of observing read counts for SNPs at different concentrations.

Modelling the Expected Count $\lambda_{S_{ij}}$

The expected fraction of sequences S_i in the bound at concentration C_j can be estimated as:

$$f_{S_{ij}}^{B}(estimated) = \frac{f_{S_i}^{L} * P_B(S_{ij})}{D_j} \tag{11}$$

where,

$f_{S_i}^{L} = n_{S_{ij}}^{L}/N^{L}$ is the fraction of sequence S_i at j^{th} concentration in the library,

$P_B(S_{ij})$ = probability of bound sequence S_i at concentration j,

$n_{S_{ij}}^{L}$ is the count of sequence S_i at j^{th}concentration in the library,

N_L is the total sequence count in the library,

D_j is the denominator at j^{th} concentration in our enrichment equation. This is the sum of all weighted binding values over both replicates (see eq. 6). The denominator, D_j accounts for the total expected binding signal from all variants in the library at j^{th} concentration, this is critical for normalization. Initially these denominators were initialised with the (0.9/maximum enrichment), then a Poisson model without w_i parameter ran up to 1000 iterations. After 360th, we observed these denominators converge. So, we use 361st denominator values for the model equation. We can also write as:

$$f_{S_{ij}}^{B}(\text{estimated}) = \frac{\lambda_B(S_{ij})}{N_j^B} \tag{12}$$

where $\lambda_B(S_{ij})$ is the expected counts of sequence S_i in the bound fraction at the jth concentration, C_j. Therefore,

$$\frac{f_{S_i}^{L} * P_B(S_{ij})}{D_j} = \frac{\lambda_B(S_{ij})}{N_j^B} \tag{13}$$

$$\lambda_B(S_{i,j}) = \frac{N_j^B f_{S_i}^{L}}{D_j} P_B(S_{ij}) \tag{14}$$

$$\lambda_B(S_{i,j}) = \frac{N_j^B n_{L_i}}{D_j N_L} P_B(S_{i,j}) \tag{15}$$

Using equation (4) we get:

$$\lambda_B(S_{i,j}) = \frac{N_j^B n_{L_i}}{D_j N_L}\left(\frac{1}{1+\frac{1}{K_i.C_j}}\right)(1 - w_i) \tag{16}$$

Expected count estimate of sequences S_i at concentration C_j in case of unbound:

$$\lambda_U(S_{i,j}) = \frac{N_j^U n_{L_i}}{D_{u_j} N_L}\left(1 - \frac{1}{1+\frac{1}{K_i.C_j}}\right) \tag{17}$$

where, N_j^B and N_j^U are total sequence count at the j^{th}concentration in the bound and unbound sample respectively. D_j and D_{u_j} are the sum of all weighted binding values over both replicates in the bound and unbound sample respectively.

The total log-likelihood, as given in Equation (10), is expressed as:

$$\begin{aligned}\mathcal{L} = \sum_i \sum_j \left[n_{b_{S_{ij}}}^{R1} log\lambda_B^{R1}(S_{ij}) - \lambda_B^{R1}(S_{ij}) - \log\left(n_{b_{S_{ij}}}^{R1}!\right)\right] \\ + \sum_i \sum_j \left[n_{b_{S_{ij}}}^{R2} log\lambda_B^{R1}(S_{ij}) - \lambda_B^{R1}(S_{ij}) - \log\left(n_{b_{S_{ij}}}^{R1}!\right)\right] \\ + \sum_i \sum_j \left[n_{u_{S_{ij}}}^{R1} log\lambda_U^{R1}(S_{ij}) - \lambda_U^{R1}(S_{ij}) - \log\left(n_{u_{S_{ij}}}^{R1}!\right)\right] \\ + \sum_i \sum_j \left[n_{u_{S_{ij}}}^{R2} log\lambda_U^{R2}(S_{ij}) - \lambda_U^{R2}(S_{ij}) - \log\left(n_{u_{S_{ij}}}^{R2}!\right)\right] \\ + \sum_i \left[n_{L1_{Si}}^{R1} log\lambda_{L1}^{R1}(S_i) - \lambda_{L1}^{R1}(S_i) - \log\left(n_{L1_{Si}}^{R1}!\right)\right] \\ + \sum_i \left[n_{L2_{Si}}^{R2} log\lambda_{L2}^{R2}(S_i) - \lambda_{L2}^{R2}(S_i) - \log\left(n_{L2_{Si}}^{R2}!\right)\right]\end{aligned}$$

(18)

where, $n_{b_{S_{ij}}}^{R1}, n_{b_{S_{ij}}}^{R2}, n_{u_{S_{ij}}}^{R1}, n_{u_{S_{ij}}}^{R2}, n_{L1_{Si}}^{R1}$ *and* $n_{L2_{Si}}^{R2}$ are the count of sequence S_i at j^{th} concentration in the bound (b), unbound (u) and library (L) for replicate 1 (R1) and replicate 2 (R2). Note: $n_{L1_{Si}}^{R1}$ *and* $n_{L2_{Si}}^{R2}$ are not indexed by concentration, as the input library is shared across all concentrations. $\lambda_B^{R1}(S_{ij})$, $\lambda_B^{R2}(S_{ij})$, $\lambda_U^{R1}(S_{ij})$, and $\lambda_U^{R2}(S_{ij})$ represent the expected counts of sequence S_i in the bound (B) and unbound (U) respectively, for replicate 1 (R1) and replicate 2 (R2) at j^{th} concentration, as defined in equations (16) and (17). In contrast, $\lambda_{L1}^{R1}(S_i)$ and $\lambda_{L2}^{R2}(S_i)$ denote the expected counts of S_i in the input library (L) for R1 and R2, which are independent of concentration.

The term $\sum_{S_i}\sum_{S_j}\log\left(n_{b(S_{ij})}!\right)$, $\sum_{S_i}\sum_{S_j}\log\left(n_{u(S_{ij})}!\right)$, , and $\sum_{S_i}\sum_{S_j}\log\left(n_{L(S_{ij})}!\right)$, depends only on the observed data and not on the parameterλ . We therefore collect these data-dependent terms and denote them as a constant C :

$$C = -\sum_{S_i}[\sum_{S_j}[\log\left(n^{R1}_{b(S_{ij})}!\right)! + \log\left(n^{R2}_{b(S_{ij})}!\right) + \log\left(n^{R1}_{u(S_{ij})}!\right) + \log\left(n^{R2}_{u(S_{ij})}!\right)] + [\log(n^{R1}_{L1(S_i)}!) + \log(n^{R2}_{L2(S_i)}!)]] \quad (19)$$

Since the constant term C does not depend on the model parameters, we separate it from the full log-likelihood expression. The optimization is then performed by minimizing the negative log-likelihood excluding C, as it has no effect on the parameter estimates. Upon eliminating the constant term from the right-hand side, the equation reduces to:

$$\begin{aligned}\mathcal{L} - C = \sum_{S_{ij}} &\left[n^{R1}_{b_{S_{ij}}} \log \lambda^{R1}_B(S_{ij}) - \lambda^{R1}_B(S_{ij})\right] \\ &+ \sum_{S_{ij}} \left[n^{R2}_{b_{S_{ij}}} \log \lambda^{R2}_B(S_{ij}) - \lambda^{R2}_B(S_{ij})\right] + \sum_{S_{ij}} \left[n^{R1}_{u_{S_{ij}}} \log \lambda^{R1}_U(S_{ij}) - \lambda^{R1}_U(S_{ij})\right] \\ &+ \sum_{S_{ij}} \left[n^{R2}_{u_{S_{ij}}} \log \lambda^{R2}_U(S_{ij}) - \lambda^{R2}_U(S_{ij})\right] + \sum_{S_i} \left[n^{R1}_{L1_{Si}} \, log\lambda^{R1}_{L1}(S_i) - \lambda^{R1}_{L1}(S_{ij})\right] \\ &+ \sum_{S_i} \left[n^{R2}_{L2_{Si}} \, log\lambda^{R2}_{L2}(S_{ij}) - \lambda^{R2}_{L2}(S_{ij})\right]\end{aligned} \quad (20)$$

We minimize the negative log-likelihood to assess how well the model fits the observed data; this optimization simultaneously addresses a mathematical problem and holds biological significance in our context. Since optimizing all parameters at once is highly time-consuming, we instead adopt an iterative refinement procedure. To estimate parameters that minimize the negative log-likelihood, we used the *"minimize"* function from Python's *scipy.optimize* module. The model fitting proceeds iteratively to update denominators and parameters. First, in the iteration step: for each SNP sequence, we optimized parameters by minimizing the negative log-likelihood. Then using the fitted parameters, we updated the denominators to better reflect the current binding landscape. Parameter estimation was performed through 10 iterations of optimization. Starting with initial parameter guesses, likelihoods were computed from observed count data. Using the L-BFGS-B method, a quasi-Newton method that is efficient for large scale problems and capable of handling bound constraints, the model parameters were iteratively updated to maximize likelihood. At each iteration step, the updated values of $K, \lambda,$ and w_i were logged. Normalization terms were also updated accordingly. This iterative procedure progressively refined the estimates and enabled effective tracking of convergence across iterations.

We started this modelling approach to better understand the relationship between TF binding and concentration. By analyzing and maximizing the fit to the observed data, we aim to accurately capture the biological patterns underlying binding behavior. This is especially important because plotting enrichment values across different concentrations and between replicates reveals variability that is non-trivial, motivating a quantitative and rigorous model.

Note**:** The 500 nM concentration for TBX5 was excluded from binding curve fitting and subsequent plotting due to insufficient counts in the bound fraction.

MEME motif discovery

Upon initial receipt of the raw FASTQ files, a preliminary analysis was conducted to verify the presence of TF binding motifs within the sequencing data. To this end, we ran MEME on the complete dataset encompassing all three TFs: NKX2-5, GATA4, and TBX5. This analysis successfully identified the canonical motifs corresponding to each TF, confirming the validity and quality of the sequence data.

To further characterize motifs in individual samples relative to concentration, a more selective approach was used. For each TF, we selected the top 500 sequences with the highest enrichment scores across all four SNP variants (A, T, G, C). These subsets were then re-analysed using MEME to identify the most representative motifs under conditions of strong signal. MEME processes a group of sequences and outputs the requested number of motifs, using statistical modelling to automatically select the best motif width, number of occurrences, and description. To constrain the search and improve motif detection, specific flags were applied during MEME runs: the ***-meme-mod*** *anr* flag assumes each sequence can contain any number of non-overlapping occurrences of each motif, which is useful for detecting multiple repeats within a sequence and provides greater accuracy but requires about ten times more computational time and is less sensitive to weak non-repetitive motifs; the ***-minw*** *5* and ***-maxw*** *9* flags limit motif lengths to between 5 and 9 bases, focusing on short motifs; and the ***-meme-nmotifs*** *5* flag restricts the search to up to five distinct motifs. Together, these settings control the motif discovery process to identify a specific number of short, potentially repetitive patterns within the selected sequences.

Motif analysis by SNP position

To investigate SNP-driven disruptions in TF binding motifs, sequence extraction was initiated using the top 500 K-fitted values. These sequences served as input for the MEME suite, executed in “anr” mode with motif width parameters set between 5 and 6 nucleotides for NKX2-5. The analysis permitted the identification of up to 10 motif logos, considering both the forward and reverse complement strands. Among the identified motif logos, the highest scoring Position Weight Matrix (PWM), regardless of strand orientation, was selected for downstream analysis.

The SNP motif analysis aimed to quantify the impact of variants on TF binding, specifically assessing changes induced by alternate alleles. For each SNP sequence, unique molecular identifiers (UMIs) were removed from both sequence termini, and 40-nucleotide windows were constructed to incorporate either the reference or alternate allele. Reverse complement sequences were also generated. These sequences were scored using the selected PWM to assess motif affinity, and the highest-scoring subregions were identified for each case. SNPs were evaluated for their potential to disrupt or enhance transcription factor binding by examining whether the SNP position overlapped with the most affected k-mers. Affected and unaffected motifs were quantified separately for SNPs associated with increased or decreased binding affinity, enabling comparative analyses across conditions.

To visualize the positional influence of SNPs within motifs, bar plots were generated to depict the frequency of SNP overlap at each motif position, stratified by directionality of enrichment. Additionally, the relative contributions of SNPs to individual motif positions were assessed by aggregating the maximum scoring bins across increased and decreased binding categories. These contributions were further visualized using customized bar plots, with specific bin ranges (e.g., positions 16–20) renamed for interpretability (e.g., positions 1–5), facilitating focused inspection of key regions. This motif disruption analysis was conducted for the transcription factors GATA4, TBX5, and NKX2-5, providing detailed insights into the allelic effects of SNPs on transcription factor binding across these critical cardiac regulators.

Model training and testing

To evaluate the predictive performance of various models, we selected the top 500 SNPs based on their K-values. These SNPs were randomly shuffled and subsequently divided into 300 SNPs for training and 200 SNPs for testing. In parallel, we utilized ChIP-seq datasets for the cardiac transcription factors NKX2.5 (SRX9284027), GATA4 (SRX9284038), and TBX5 (SRX2023721). From each ChIP-seq dataset, the top 1000 genomic sequences were selected based on peak signal intensity scores and were similarly partitioned into training and testing sets.
We evaluated four cases:

1. **SNP-based model**: A classification model was trained using 300 positive and 300 negative SNP sequences, and evaluated on the remaining 200 positives and 200 negatives from a total of 500.

2. **ChIP-based model**: A model was trained using 600 positive and 600 negative ChIP sequences, and evaluated on the 400 positives and 400 negatives.

3. **ChIP-seq-trained model, SNP-seq tested**: A separate model was trained using 1000 ChIP-seq peak sequences (1000 positive and 1000 negative sequences) and tested on the 500 SNP dataset to assess cross-dataset generalizability.

4. **SNP-trained model, ChIP-seq-tested**: A model was trained on the SNP dataset: 500 positives and 500 negatives) and tested on the 1000 ChIP-seq-derived sequences.

For negative sequence generation in all scenarios, we used the “getNullSeqs” function from the LS-GKM toolkit to obtain matched genomic background sequences, ensuring comparable GC content and sequence length. Each of the three experimental conditions was evaluated using three distinct methodologies: MinSeqChIP, MEME, and LS-GKM, allowing comparative analysis across different feature learning and classification frameworks. We ran MEME using the same parameters described in the section “MEME”. For LS-GKM, we used the default mode. For MinSeqChIP, the following settings were used: minimum monomer size set to 2, maximum monomer size set to 6, and a maximum allowed gap of 1 between monomers (due to the short and ungapped

nature of the motifs). A Markov model of order 5 was applied in a non-gapped configuration. For PWM thresholding, we used the cutoff of 40 for the minimum number of sequences required to consider the library interpolated Markov model.

Generating Flp-In 293 with stable NKLX2-5, GATA4, and TBX5 expression

Flp-In 293 cells (Thermo Fisher, #R75007) were modified to express TBX5 by inserting gene according to manufacturer instructions. In short, cells were seeded at 200,000 cell/mL in 2.5 mL of media (DMEM+10%FBS+0.1% Anti-anti) 2 days prior to transfection. 30 µL of Lipofectamine 3000® Reagent (ThermoFisher, #L300015) was diluted in 1mL Opti-MEM® (ThermoFisher, #31985062) and incubated for 5 minutes (37°C/5% CO_2). A total of 2.5 µg of DNA (2.25 µg pOG44 + 0.25 µg of TBX5 gene insert) was diluted in 250 µL Opti-Mem. Both dilutions were combined to make a 1:1 ratio (250 µL Opti-MEM/Lipofectamine to 250 µL Opti-MEM/DNA) and incubated for 20 minutes (37°C/5% CO_2). The old medium was removed from the cells and 500 µL of Opti-MEM/DNA/Lipofectamine mixture was added and incubated for 5 hours (37°C/5% CO_2). Afterward, 1.5 mL of media was added to each well and incubated overnight. After reaching 80 % confluency, cells were switched to selection media containing hygromycin (50 µg/mL). Media was changed weekly to increase hygromycin until reaching 200 µg/mL. Once confluent, cells were transferred to a T75 plate until 80-90 % confluency for downstream experiments. Stocks were prepared at 3 million cells/mL in freezing media (HI-FBS (ThermoFisher, #A5670501) + 10% DMSO (Sigma Aldrich, #276855). Flp-In 293 cells already expressed NKX2-5 and GATA4 and were confirmed by RNA-seq and Western Blot.

Massively Parallel Reporter Assay

For the assembly of the MPRA library, we followed the procedure described by Lu et al. with minor modifications (See MPRA method overview in Figure 4.2). [47] In brief, the MPRA library was cloned into a previously generated pGL4.23ΔxbaΔluc empty vector from pGL4.23 [luc2/minP] (Promega, # E8411). First, 20 bps barcodes were added to the synthesized oligos through 24X PCR with 50 µL system, each containing 1.86 ng oligo, 25 µL NEBNext® Ultra™ II Q5® Master Mix (NEB, #M0544S), 1 µM MPRA_v3_F, and 1 µM MPRA_v3_201_R. PCR was performed under the following conditions: 98 °C for 2 min, 12 cycles of (98 °C for 10 s, 60 °C for 15 s, 72 °C for 45 s), 72 °C for 5 min. The amplified product was purified with 1.8X Ampure XP SPRI beads (Beckman Coulter, A63881) and cloned into SfiI-digested pGL4.23ΔxbaΔluc by Gibson assembly (1:2 molar ratio of vector:insert) at 50 °C for 1 h. The assembled backbone library was purified and then transformed into 10-β Escherichia coli (*E. coli*; NEB, #C3019H) through electroporation (2 kV, 200 ohm, 25 µF) according to manufacturer instructions. Electroporated *E. coli* was expanded in 200 mL of LB Broth supplemented with 100 µg/mL of carbenicillin at 37 °C for 12 to 16 h. Plasmid was then extracted using the QIAGEN Plasmid Maxi Kit (Qiagen, #12162).

To associate barcodes with oligo sequences, 200ng of the plasmid was amplified using NEBNext Q5® High-Fidelity 2X Master Mix (NEB, #M0492L), with primers TruSeq_Universal_Adapter and MPRA_v3_TruSeq_Amp2Sa_F (Table S7) using the

following conditions: 95°C for 20s, 6 cycles (95°C for 20s, 62°C for 15s, and 72°C for 30s), and 72°C for 2 minutes. The PCR product was purified using SPRI beads at 0.5X and then 1X. Next we performed additional 5 cycles of PCR to attached custom Illumina P7 indices purified using SPRI beads at 0.5X and then 1X. Samples were sequenced on a NovaSeq X Plus (2 × 150 bp) at Cincinnati Children's Hospital sequencing core to achieve a 30X read coverage. Identification of which barcodes were associated with which oligos was then conducted with the MPRAmatch pipeline (https://github.com/tewhey-lab/MPRAmatch).

A miniP + eGFP fragment was amplified from pGL4.23[eGFP/miniP] through 20X PCR with 50 µL system, each containing 1 ng plasmid, 25 µL NEBNext® Ultra™II Q5® Master Mix, 0.5 µM 200-MPRA_v3_GFP_Fusion_v2_F, and 201-MPRA_v3_GFP_Fusion_v2_R. PCR was performed under the following conditions: 98 °C for 2 min, 20 cycles of (98 °C for 10 s, 60 °C for 15 s, 72 °C for 45s), 72 °C for 5 min. The amplified product was purified with 1.8X SPRI and then inserted into AsiSI digested pGL4:23:ΔxbaΔluc backbone library through Gibson assembly at 50 °C for 1.5 h and purified at 1.5X SPRI. The resulting library was re-digested by RecBCD (NEB, #M0345S) and AsiSI (NEB ,#R063L) , purified at 1.5X SPRI, and then transformed into *E. coli* through electroporation (2 kV, 200 ohm, 25 µF). Transformed *E. coli* were cultured in 5 L of LB Broth supplemented with 100 µg/mL of carbenicillin at 37 °C for 12–16 h. The plasmid was then extracted using the QIAGEN Endo-free Plasmid Giga Kit (Qiagen, #12191).

Transfection of MPRA Library

Flp-In 293 cells (ThermoFisher, #R75007) were cultured in 30 mL Dulbecco's modified Eagle's medium (DMEM) (ThermoFisher, #10564) containing 10% fetal bovine serum (FBS) (ThermoFisher, #A3160401) and 1% penicillin-streptomycin (Pen-Strep) (ThermoFisher, #15140122). Five total replicates were transfected. For each replicate, cells were plated in two 15-cm plates and grown to a density of ~80 to 90% (~20 to 40 million cells per plate). Cells were then incubated with 80 µL of Lipofectamine 3000 (ThermoFisher, L3000015) and 20 µg of DNA for 24 hours. Then, transfected cells were split 1:3 into new 15-cm plates, keeping all transfected cells. After an additional 24 hours (48 hours after transfection), cells were pelleted by centrifugation, washed once with phosphate-buffered saline (PBS), flash-frozen using liquid nitrogen, and then stored at –80°C.

MPRA RNA Sample Processing

RNA processing was performed as previously described. [168] In short, total RNA was extracted from the cell pellets with the Qiagen Maxi RNeasy kit (Qiagen, #75162) with on-column DNase digest according to manufacturer's instructions. A DNase reaction was further performed to remove remaining MPRA library vectors using Turbo DNAse kit (ThermoFisher, #AM2238). The reaction was stopped with 0.1% SDS (Sigma Aldrich, #71736) and 0.05M EDTA (Sigma Aldrich, #03690). The GFP-transcripts in total RNA were then captured through a hybridization reaction with streptavidin beads (ThermoFisher, #65001) and three GFP-targeted biotinylated oligos (Table S7). RNA was then cleaned up with RNA SPRI (Beckman Coulter, #A63987) and converted to cDNA using a Superscript III (ThermoFisher, 18080044) reaction with primer MPRA_v3_Amp2Sc_R (Table 4.1). The

relative cDNA abundance was estimated through quantitative PCR along with serial dilutions of plasmid library serving as a standard curve (see Table S7 for primer sequences). The PCR conditions were: 98°C for 30s, 40x of (95°C for 20s, 65°C for 20s, and 72°C for 30s), and 72°C for 2 minutes. To minimize amplification bias, the Ct number reflecting the point at which the amplification just began to take off, subtracted by 1, was used to set up the first PCR for sequencing preparation. cDNA and plasmids were normalized to approximately the same concentration and cycled for 10 cycles using NEBNext Q5® High-Fidelity 2X Master Mix (NEB, #M0492L) and primers MPRA_v3_Illumina_GFP_F and TruSeq_Universal_Adapter (Table 4.1). The product was cleaned up with RNA SPRI at 1X, eluted in 30µL, 20µL of which was then subjected to another round of 6 cycles to attach custom p7 and p5 Illumina adapters with unique sample indices. Samples were sequenced on a NextSeq 2000 platform using the P3 100 cycle kit, with an average of around 100M reads per sample.

MPRA Data Analysis

Oligo counts were obtained via the MPRAcount pipeline (https://github.com/tewhey-lab/MPRA_oligo_barcode_pipeline). Oligos with at least 10 barcodes were retained for analysis and oligo counts were normalized for sequencing depth with the DESeq2 median of ratios method. DESeq2 was then used to estimate the fold change between plasmid DNA and cDNA with Wald's test and p-values were corrected for multiple hypothesis testing by Bonferroni's method. Significance threshold was determined at adjusted p-value less than 0.01 in either the reference or alternate allele in order to call a sequence as having a regulatory effect on expression. For identification of expression-modulating variants, only variants originating from sequences determined to have a regulatory effect were considered. Allelic skew was calculated by comparing the log ratios of the reference and alternative alleles using Wald's test. All skew p-values were adjusted with the Benjamini-Hochberg procedure and determined to be significant at 5% false discovery rate. The scripts for estimating variant activity and allelic skew are available on https://github.com/tewhey-lab/MPRAmodel. Scripts for plotting and visualization of MPRA analysis are available on https://github.com/WeirauchLab/mpraprofiler.

HOMER motif enrichment analysis

HOMER (v5.1) [90] was performed on 50 bp reference and alternate fasta file centered on the CVD-associated SNP using the findMotif.pl function. The background file was generated using the homer2 background function. Shared motifs between a reference allele and its corresponding alternate allele(s) were removed from the analysis. This ensured that only motifs that occur in only the reference or alternate allele were used for downstream analysis. Finally, TF and family counts were generated by counting unique instances of SNP-TF/family pairs for either the reference or alternate allele. Additionally, we used HOMER to calculate the enrichment of each motif in the sequence of emAlleles compared to the sequences of non-emAlleles.

Functional genomics dataset enrichment analysis with RELI

We used the Regulatory Element Locus Intersection (RELI) [74] algorithm to identify genomic features (TF binding events, histone marks, etc.) that coincide with emAlleles as previously described. [47] In short, RELI systematically intersects the coordinates of emAlleles with a large collection of ChIP-seq datasets and counts the number of times the input regions overlap with peaks in the dataset by at least one base. A p-value describing the significance of this overlap is estimated by comparing the input set with a negative set (variants with no CRE activity; 50% expression fold change). We performed the RELI analysis for all variants categorized as emAlleles, and for the individual subset that overlapped with cardiac CREs.

Key Resources Table:

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
α-NKX2-5	Cell Signaling	8792
α-GATA4	Cell Signaling	36966
α-FLAG	Sigma Aldrich	F1804-50UG
Bacterial and virus strains		
10-β E. Coli	NEB	C3019H
Chemicals, peptides, and recombinant proteins		
KAPA HiFi HotStart ReadyMix	Roche Sequencing Store	KK2602 07958935001
Sodium Chloride	Fisher Scientific	S671-3
UltraPure® 1M Tris-HCl, pH 8.0	Invitrogen	15568025
Glycerol	Sigma Aldrich	G5516
pdI-dC	Sigma Aldrich	P4929
BSA	Fisher Scientific	23210
Tween-20	Sigma Aldrich	P9416
DTT	Sigma Aldrich	D0632
HEPES	VWR	7365-45-9
Zinc Acetate	Sigma Aldrich	383058
10X TBE(Tris/Boric Acid/ EDTA)	Biorad	1610770
Buffer EB	Qiagen	19086
Dynabeads M-280 Streptavidin	ThermoFisher	11205D
DMEM	ThermoFisher	10564
FBS	ThermoFisher	A3160401
Opti-MEM	ThermoFisher	31985062
Lipofectamine 3000 Transfection Reagent	ThermoFisher	L300015
Hygromycin B	Sigma Aldrich	H3274
HI-FBS	ThermoFisher	A5670501
DMSO	Sigma Aldrich	276855
NEBNext® Ultra™ II Q5® Master Mix	NEB	M0544S
Gibson Assembly Master Mix	NEB	E2611L
XbaI	NEB	R0145S
NcoI	NEB	R0193S
RecBCD	NEB	M0345S
AsiSI	NEB	R063L
NEBNext Q5® High-Fidelity 2X Master Mix	NEB	M0492L
Penicilin-Streptomycin	ThermoFisher	15140122
SDS	Sigma Aldrich	71736
EDTA	Sigma Aldrich	03690
Streptavidin Beads	ThermoFisher	65001
RNA SPRI	Beckman Coutler	A63987
Superscript III	ThermoFisher	18080044
Critical commercial assays		
Q5 Site-directed Mutagenesis Kit	NEB	E0554S
Qiagen Plasmid Maxi Kit	Qiagen	12162
Qiagen Endo-free Plasmid Giga Kit	Qiagen	12191
Qiagen Maxi RNeasy	Qiagen	75162
Turbo DNAse Kit	ThermoFisher	AM2238
Deposited data		
CHD variant dataset from the GWAS catalog	Ref 49-54	N/A
ChIP-seq Dataset NKX2-5	Gonzalez-Terran et al Cell 2022	SRX9284027
ChIP-seq Dataset GATA4	Gonzalez-Terran et al Cell 2022	SRX9284038
ChIP-seq Dataset TBX5	Ang, Y.S. et al Cell 2016	SRX2023721
CHD variants MPRA dataset	GEO LINK	

CHD variants SNP Bind-n-Seq dataset	GEO LINK	
Experimental models: Cell lines		
Flp-In 293	ThermoFisher	R75007
Oligonucleotides		
MPRA_v3_F	IDT	N/A
MPRA_v3_201_R	IDT	N/A
GFP_seq_MS2_P65-HSF1_GFP_FWD	IDT	N/A
GFP_seq_MS2_P65-HSF1_GFP_REV	IDT	N/A
200-MPRA_v3_GFP_Fusion_v2_F	IDT	N/A
200-MPRA_v3_GFP_Fusion_v2_R	IDT	N/A
TruSeq_Universal_Adapter	IDT	N/A
MPRAv3_a2sa	IDT	N/A
MPRA_v3_Amp2Sc_R	IDT	N/A
MPRA_v3_Illumina_GFP_F	IDT	N/A
Recombinant DNA		
Oligopool of CHD-associated variants	Custom Array	N/A
MPRA Oligopool	Twist Bioscience	N/A
pET-NKX2.5-6XHisTag	VectorBuilder, Inc.	N/A
pET-TBX5-6XHisTag	VectorBuilder, Inc.	N/A
pET-28a(+)-6XHisTag-GATA4 DBD	Twist Bioscience	N/A
pOG44	ThermoFisher	V600520
pGL4.23 [Luc2/minP - firefly luciferase]	Promega	E8411
MS2-P65-HSF1_GFP	Addgene	61423
Software and algorithms		
TopLD Web Tool	Yun Li Lab http://topld.genetics.unc.edu/	N/A
Plink (v1.90b)		N/A
UCSC Genome Browser		N/A
Scipy.optimize (Python)		N/A
MEME Motif Discovery	https://meme-suite.org/meme/	N/A
MinSeqChIP		N/A
LS-GKM		N/A
MPRAmatch pipeline	Tewhey Lab	N/A
DESeq2		N/A
MPRAcount pipeline		N/A
Homer(v5.1)	Ref82	N/A
Regulatory Element Locus Intersection (RELI)	Ref66	N/A
Other		
Azure Zapphire Biomolecular Imager	Azure Biosystems	N/A
Thermoshaker	BIOGRANT	N/A

References

1. Pierpont, M.E., Brueckner, M., Chung, W.K., Garg, V., Lacro, R. v., McGuire, A.L., Mital, S., Priest, J.R., Pu, W.T., Roberts, A., et al. (2018). Genetic Basis for Congenital Heart Disease: Revisited: A Scientific Statement from the American Heart Association https://doi.org/10.1161/CIR.0000000000000606.
2. Zaidi, S., and Brueckner, M. (2017). Genetics and Genomics of Congenital Heart Disease. Circ Res *120*, 923–940. https://doi.org/10.1161/CIRCRESAHA.116.309140.
3. Zimmerman, M.S., Smith, A.G.C., Sable, C.A., Echko, M.M., Wilner, L.B., Olsen, H.E., Atalay, H.T., Awasthi, A., Bhutta, Z.A., Boucher, J.L.A., et al. (2020). Global, regional, and national burden of congenital heart disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet Child Adolesc Health *4*, 185–200. https://doi.org/10.1016/S2352-4642(19)30402-X.
4. Dallapiccola, B., Mingarelli, R., Digilio, M.C., Marino, B., and Novelli, G. (1994). Genetics of congenital heart diseases. G Ital Cardiol *24*, 155–166. https://doi.org/10.5005/jp/books/12075_6.
5. Bruneau, B.G. (2008). The developmental genetics of congenital heart disease. Nature *451*, 943–948. https://doi.org/10.1038/nature06801.
6. Morton, S.U., Quiat, D., Seidman, J.G., and Seidman, C.E. (2021). Genomic frontiers in congenital heart disease. Nat Rev Cardiol *0123456789*. https://doi.org/10.1038/s41569-021-00587-4.
7. Buniello, A., Macarthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res *47*, D1005–D1012. https://doi.org/10.1093/nar/gky1120.
8. Kathiresan, S., and Srivastava, D. (2012). Genetics of human cardiovascular disease. Cell *148*, 1242–1257. https://doi.org/10.1016/j.cell.2012.03.001.
9. Heshmatzad, K., Naderi, N., Maleki, M., Abbasi, S., Ghasemi, S., Ashrafi, N., Fazelifar, A.F., Mahdavi, M., and Kalayinia, S. (2023). Role of non-coding variants in cardiovascular disease. J Cell Mol Med. https://doi.org/10.1111/jcmm.17762.
10. Villar, D., Frost, S., Deloukas, P., and Tinker, A. (2020). The contribution of non-coding regulatory elements to cardiovascular disease. Open Biol *10*, 200088. https://doi.org/10.1098/rsob.200088.
11. Richter, F., Morton, S.U., Kim, S.W., Kitaygorodsky, A., Wasson, L.K., Chen, K.M., Zhou, J., Qi, H., Patel, N., DePalma, S.R., et al. (2020). Genomic analyses implicate noncoding de novo variants in congenital heart disease. Nat Genet *52*, 769–777. https://doi.org/10.1038/s41588-020-0652-z.
12. McCulley, D.J., and Black, B.L. (2012). Transcription Factor Pathways and Congenital Heart Disease (Elsevier Inc.) https://doi.org/10.1016/B978-0-12-387786-4.00008-7.
13. Xiao, F., Zhang, X., Morton, S.U., Kim, S.W., Fan, Y., Gorham, J.M., Zhang, H., Berkson, P.J., Mazumdar, N., Cao, Y., et al. (2024). Functional dissection of human cardiac enhancers and noncoding de novo variants in congenital heart disease. Nat Genet *56*, 420–430. https://doi.org/10.1038/s41588-024-01669-y.
14. Bruneau, B.G. (2013). Signaling and transcriptional networks in heart development and regeneration. Cold Spring Harb Perspect Biol *5*. https://doi.org/10.1101/cshperspect.a008292.
15. Gonzalez-Teran, B., Pittman, M., Felix, F., Thomas, R., Richmond-Buccola, D., Hüttenhain, R., Choudhary, K., Moroni, E., Costa, M.W., Huang, Y., et al. (2022). Transcription factor protein interactomes reveal genetic determinants in heart disease. Cell *185*, 794-814.e30. https://doi.org/10.1016/j.cell.2022.01.021.
16. Luna-Zurita, L., Stirnimann, C.U., Glatt, S., Kaynak, B.L., Thomas, S., Baudin, F., Samee, M.A.H., He, D., Small, E.M., Mileikovsky, M., et al. (2016). Complex Interdependence Regulates Heterotypic Transcription Factor Distribution and Coordinates Cardiogenesis. Cell *164*, 999–1014. https://doi.org/10.1016/j.cell.2016.01.004.

17. Steimle, J.D., and Moskowitz, I.P. (2017). TBX5: A Key Regulator of Heart Development. Curr Top Dev Biol *122*. https://doi.org/10.1016/bs.ctdb.2016.08.008.
18. Horb, M.E., and Thomsen, G.H. (1999). Tbx5 is essential for heart development. Development *126*, 1739–1751.
19. 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. https://doi.org/10.1038/90123.
20. 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. https://doi.org/10.1093/hmg/dds165.
21. Maitra, M., Schluterman, M.K., Nichols, H.A., Richardson, J.A., Lo, C.W., Srivastava, D., and Garg, V. (2009). Interaction of Gata4 and Gata6 with Tbx5 is critical for normal cardiac development. Dev Biol *326*, 368–377. https://doi.org/10.1016/j.ydbio.2008.11.004.
22. Chung, I.M., and Rajakumar, G. (2016). Genetics of congenital heart defects: The NKX2-5 gene, a key player. Genes (Basel) *7*. https://doi.org/10.3390/genes7020006.
23. Small, E.M., and Krieg, P.A. (2003). Transgenic analysis of the atrialnatriuretic factor (ANF) promoter: Nkx2-5 and GATA-4 binding sites are required for atrial specific expression of ANF. Dev Biol *261*, 116–131. https://doi.org/10.1016/S0012-1606(03)00306-3.
24. Molkentin, J.D., Lin, Q., Duncan, S.A., and Olson, E.N. (1997). Requirement of the transcription factor GATA4 for heart tube formation and ventral morphogenesis. Genes Dev *11*. https://doi.org/10.1101/gad.11.8.1061.
25. Lyons, I., Parsons, L.M., Hartley, L., Li, R., Andrews, J.E., Robb, L., Harvey, R.P., Walter, T., and Hall, E. (1995). Myogenic and morpho.genetic defects in the heart tubes of murlne embryos lacking the homeo box gene Nkx2-5. Genes Dev *9*, 1654–1666.
26. Bouveret, R., Waardenberg, A.J., Schonrock, N., Ramialison, M., Doan, T., de jong, D., Bondue, A., Kaur, G., Mohamed, S., Fonoudi, H., et al. (2015). NKX2-5 mutations causative for congenital heart disease retain functionality and are directed to hundreds of targets. Elife *4*, 1–30. https://doi.org/10.7554/eLife.06942.
27. Kirk, E.P., Sunde, M., Costa, M.W., Rankin, S.A., Wolstein, O., Castro, M.L., Butler, T.L., Hyun, C., Guo, G., Otway, R., et al. (2007). Mutations in crdiac T-box factor gene TBX20 are associated with diverse cardiac pathologies, including defects of septation and valvulogenesis and cardiomyopathy. Am J Hum Genet *81*, 280–291. https://doi.org/10.1086/519530.
28. Bruneau, B.G., Logan, M., Davis, N., Levi, T., Tabin, C.J., Seidman, J.G., and Seidman, C.E. (1999). Chamber-specific cardiac expression of Tbx5 and heart defects in Holt- Oram syndrome. Dev Biol *211*. https://doi.org/10.1006/dbio.1999.9298.
29. Bruneau, B.G., Nemer, G., Schmitt, J.P., Charron, F., Robitaille, L., Caron, S., Conner, D.A., Gessler, M., Nemer, M., Seidman, C.E., et al. (2001). A murine model of Holt-Oram syndrome defines roles of the T-Box transcription factor Tbx5 in cardiogenesis and disease. Cell *106*. https://doi.org/10.1016/S0092-8674(01)00493-7.
30. Ang, Y.S., Rivas, R.N., Ribeiro, A.J.S., Srivas, R., Rivera, J., Stone, N.R., Pratt, K., Mohamed, T.M.A., Fu, J.D., Spencer, C.I., et al. (2016). Disease Model of GATA4 Mutation Reveals Transcription Factor Cooperativity in Human Cardiogenesis. Cell *167*, 1734-1749.e22. https://doi.org/10.1016/j.cell.2016.11.033.
31. Fornes, O., Castro-Mondragon, J.A., Khan, A., Van Der Lee, R., Zhang, X., Richmond, P.A., Modi, B.P., Correard, S., Gheorghe, M., Baranašić, D., et al. (2020). JASPAR 2020: Update of the open-Access database of transcription factor binding profiles. Nucleic Acids Res *48*, D87–D92. https://doi.org/10.1093/nar/gkz1001.
32. Sam, J., Mercer, E.J., Torregroza, I., Banks, K.M., and Evans, T. (2020). Specificity, redundancy and dosage thresholds among gata4/5/6 genes during zebrafish cardiogenesis. Biol Open *9*. https://doi.org/10.1242/bio.053611.

33. Kuo, C.T., Morrisey, E.E., Anandappa, R., Sigrist, K., Lu, M.M., Parmacek, M.S., Soudais, C., and Leiden, J.M. (1997). GATA4 transcription factor is required for ventral morphogenesis and heart tube formation. Genes Dev *11*. https://doi.org/10.1101/gad.11.8.1048.
34. Tessadori, F., Tsingos, E., Colizzi, E.S., Kruse, F., Van Den Brink, S.C., Van Den Boogaard, M., Christoffels, V.M., Merks, R.M.H., and Bakkers, J. (2021). Twisting of the zebrafish heart tube during cardiac looping is a tbx5-dependent and tissue-intrinsic process. Elife *10*. https://doi.org/10.7554/eLife.61733.
35. Peña-Martínez, E.G., Rivera-Madera, A., Pomales-Matos, D.A., Sanabria-Alberto, L., Rosario-Cañuelas, B.M., Rodríguez-Ríos, J.M., Carrasquillo-Dones, E.A., and Rodríguez-Martínez, J.A. (2023). Disease-associated non-coding variants alter NKX2-5 DNA-binding affinity. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms *1866*, 194906. https://doi.org/10.1016/j.bbagrm.2023.194906.
36. Benaglio, P., D'Antonio-Chronowska, A., Ma, W., Yang, F., Young Greenwald, W.W., Donovan, M.K.R., DeBoever, C., Li, H., Drees, F., Singhal, S., et al. (2019). Allele-specific NKX2-5 binding underlies multiple genetic associations with human electrocardiographic traits. Nat Genet *51*, 1506–1517. https://doi.org/10.1038/s41588-019-0499-3.
37. Peña-Martínez, E.G., Pomales-Matos, D.A., Rivera-Madera, A., Messon-Bird, J.L., Medina-Feliciano, J.G., Sanabria-Alberto, L., Barreiro-Rosario, A.C., Rivera-Del Valle, J., Rodríguez-Ríos, J.M., and Rodríguez-Martínez, J.A. (2023). Prioritizing Cardiovascular Disease-Associated Variants Altering NKX2-5 and TBX5 Binding through an Integrative Computational Approach. Journal of Biological Chemistry *299*, 105423. https://doi.org/10.1016/j.jbc.2023.105423.
38. Peña-Martínez, E.G., Messon-Bird, J.L., Rodríguez-Ríos, J.M., Velázquez-Roig, R., Pomales-Matos, D.A., Rivera-Madera, A., Sanabria-Alberto, L., Barreiro-Rosario, A.C., Figueroa-Rosado, J.A., Rivera-Del Valle, J., et al. (2025). Cardiovascular disease-associated non-coding variants disrupt GATA4-DNA binding and regulatory functions. Human Genetics and Genomics Advances *6*, 100415. https://doi.org/10.1016/j.xhgg.2025.100415.
39. Yang, B., Zhou, W., Jiao, J., Nielsen, J.B., Mathis, M.R., Heydarpour, M., Lettre, G., Folkersen, L., Prakash, S., Schurmann, C., et al. (2017). Protein-altering and regulatory genetic variants near GATA4 implicated in bicuspid aortic valve. Nat Commun *8*. https://doi.org/10.1038/ncomms15481.
40. Ghosh, T.K. (2001). Characterization of the TBX5 binding site and analysis of mutations that cause Holt-Oram syndrome. Hum Mol Genet *10*, 1983–1994. https://doi.org/10.1093/hmg/10.18.1983.
41. Jiang, X., Li, T., Liu, S., Fu, Q., Li, F., Chen, S., Sun, K., Xu, R., and Xu, Y. (2021). Variants in a cis-regulatory element of TBX1 in conotruncal heart defect patients impair GATA6-mediated transactivation. Orphanet J Rare Dis *16*, 334. https://doi.org/10.1186/s13023-021-01981-4.
42. van Weerd, J.H., Mohan, R.A., Duijvenboden, K. van, Hooijkaas, I.B., Wakker, V., Boukens, B.J., Barnett, P., and Christoffels, V.M. (2020). Trait-associated noncoding variant regions affect tbx3 regulation and cardiac conduction. Elife *9*, 1–26. https://doi.org/10.7554/eLife.56697.
43. Beaudoin, M., Gupta, R.M., Won, H.H., Lo, K.S., Do, R., Henderson, C.A., Lavoie-St-Amour, C., Langlois, S., Rivas, D., Lehoux, S., et al. (2015). Myocardial Infarction-Associated SNP at 6p24 Interferes with MEF2 Binding and Associates with PHACTR1 Expression Levels in Human Coronary Arteries. Arterioscler Thromb Vasc Biol *35*, 1472–1479. https://doi.org/10.1161/ATVBAHA.115.305534.
44. Jindal, G.A., Bantle, A.T., Solvason, J.J., Grudzien, J.L., D'Antonio-Chronowska, A., Lim, F., Le, S.H., Song, B.P., Ragsac, M.F., Klie, A., et al. (2023). Single-nucleotide variants within heart enhancers increase binding affinity and disrupt heart development. Dev Cell *58*, 2206-2216.e5. https://doi.org/10.1016/j.devcel.2023.09.005.
45. Yan, J., Qiu, Y., Ribeiro dos Santos, A.M., Yin, Y., Li, Y.E., Vinckier, N., Nariai, N., Benaglio, P., Raman, A., Li, X., et al. (2021). Systematic analysis of binding of transcription factors to noncoding variants. Nature. https://doi.org/10.1038/s41586-021-03211-0.

46. Zykovich, A., Korf, I., and Segal, D.J. (2009). Bind-n-Seq: High-throughput analysis of in vitro protein DNA interactions using massively parallel sequencing. Nucleic Acids Res *37*. https://doi.org/10.1093/nar/gkp802.
47. Lu, X., Chen, X., Forney, C., Donmez, O., Miller, D., Parameswaran, S., Hong, T., Huang, Y., Pujato, M., Cazares, T., et al. (2021). Global discovery of lupus genetic risk variant allelic enhancer activity. Nat Commun *12*, 1611. https://doi.org/10.1038/s41467-021-21854-5.
48. Tewhey, R., Kotliar, D., Park, D.S., Liu, B., Winnicki, S., Reilly, S.K., Andersen, K.G., Mikkelsen, T.S., Lander, E.S., Schaffner, S.F., et al. (2016). Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. Cell *165*, 1519–1529. https://doi.org/10.1016/j.cell.2016.04.027.
49. Inoue, F., and Ahituv, N. (2015). Decoding enhancers using massively parallel reporter assays. Genomics *106*, 159–164. https://doi.org/10.1016/j.ygeno.2015.06.005.
50. Rong, S., Root, E., and Reilly, S.K. (2024). Massively parallel approaches for characterizing noncoding functional variation in human evolution. Curr Opin Genet Dev *88*. https://doi.org/10.1016/j.gde.2024.102256.
51. Agopian, A.J., Goldmuntz, E., Hakonarson, H., Sewda, A., Taylor, D., and Mitchell, L.E. (2017). Genome-Wide Association Studies and Meta-Analyses for Congenital Heart Defects. Circ Cardiovasc Genet *10*. https://doi.org/10.1161/CIRCGENETICS.116.001449.
52. Agopian, A.J., Mitchell, L.E., Glessner, J., Bhalla, A.D., Sewda, A., Hakonarson, H., and Goldmuntz, E. (2014). Genome-wide association study of maternal and inherited loci for conotruncal heart defects. PLoS One *9*. https://doi.org/10.1371/journal.pone.0096057.
53. Lahm, H., Jia, M., Dreßen, M., Wirth, F., Puluca, N., Gilsbach, R., Keavney, B.D., Cleuziou, J., Beck, N., Bondareva, O., et al. (2021). Congenital heart disease risk loci identified by genome-wide association study in European patients. Journal of Clinical Investigation *131*. https://doi.org/10.1172/JCI141837.
54. Oluwafemi, O.O., Musfee, F.I., Mitchell, L.E., Goldmuntz, E., Xie, H.M., Hakonarson, H., Morrow, B.E., Guo, T., Taylor, D.M., McDonald-Mcginn, D.M., et al. (2021). Genome-wide association studies of conotruncal heart defects with normally related great vessels in the United States. Genes (Basel) *12*. https://doi.org/10.3390/genes12071030.
55. Hu, Z., Shi, Y., Mo, X., Xu, J., Zhao, B., Lin, Y., Yang, S., Xu, Z., Dai, J., Pan, S., et al. (2013). A genome-wide association study identifies two risk loci for congenital heart malformations in Han Chinese populations. Nat Genet *45*, 818–821. https://doi.org/10.1038/ng.2636.
56. Rashkin, S.R., Cleves, M., Shaw, G.M., Nembhard, W.N., Nestoridi, E., Jenkins, M.M., Romitti, P.A., Lou, X.Y., Browne, M.L., Mitchell, L.E., et al. (2022). A genome-wide association study of obstructive heart defects among participants in the National Birth Defects Prevention Study. Am J Med Genet A *188*, 2303–2314. https://doi.org/10.1002/ajmg.a.62759.
57. Peña-Martínez, E.G., Rodríguez-Ríos, J.M., Messon-Bird, J.L., Barreiro-Rosario, A.C., Velázquez-Roig, R., Rivera-Madera, A., Peterson-Peguero, E.A., and Rodríguez-Martínez, J.A. (2025). Protocol for evaluating the impact of non-coding variants on transcription factor binding and gene expression. STAR Protoc. https://doi.org/10.1016/j.
58. Gordân, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R., and Bulyk, M.L. (2013). Genomic Regions Flanking E-Box Binding Sites Influence DNA Binding Specificity of bHLH Transcription Factors through DNA Shape. Cell Rep *3*, 1093–1104. https://doi.org/10.1016/j.celrep.2013.03.014.
59. Le, D.D., Shimko, T.C., Aditham, A.K., Keys, A.M., Longwell, S.A., Orenstein, Y., and Fordyce, P.M. (2018). Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding. Proc Natl Acad Sci U S A *115*, E3702–E3711. https://doi.org/10.1073/pnas.1715888115.
60. Olson, E.N. (2006). Gene regulatory networks in the evolution and development of the heart. Science (1979) *313*, 1922–1927. https://doi.org/10.1126/science.1132292.
61. Bhimsaria, D., Rodríguez-Martínez, J.A., Mendez-Johnson, J.L., Ghoshdastidar, D., Varadarajan, A., Bansal, M., Daniels, D.L., Ramanathan, P., and Ansari, A.Z. (2023). Hidden modes of DNA

binding by human nuclear receptors. Nat Commun *14*. https://doi.org/10.1038/s41467-023-39577-0.
62. Lee, D. (2016). LS-GKM: A new gkm-SVM for large-scale datasets. Bioinformatics *32*, 2196–2198. https://doi.org/10.1093/bioinformatics/btw142.
63. 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. https://doi.org/10.1093/bioinformatics/btw203.
64. Bailey, T.L., Johnson, J., Grant, C.E., and Noble, W.S. (2015). The MEME Suite. Nucleic Acids Res *43*, W39–W49. https://doi.org/10.1093/nar/gkv416.
65. Mori, A.D., Zhu, Y., Vahora, I., Nieman, B., Koshiba-Takeuchi, K., Davidson, L., Pizard, A., Seidman, J.G., Seidman, C.E., Chen, X.J., et al. (2006). Tbx5-dependent rheostatic control of cardiac gene expression and morphogenesis. Dev Biol *297*, 566–586. https://doi.org/10.1016/j.ydbio.2006.05.023.
66. Borok, M.J., Papaioannou, V.E., and Sussel, L. (2016). Unique functions of Gata4 in mouse liver induction and heart development. Dev Biol *410*. https://doi.org/10.1016/j.ydbio.2015.12.007.
67. Rodriguez-Esteban, C., Tsukul, T., Yonel, S., Magallon, J., Tamura, K., and Izpisua Belmonte, J.C. (1999). The T-box genes Tbx4 and Tbx5 regulate limb outgrowth and identity. Nature *398*. https://doi.org/10.1038/19769.
68. Patel, R.S., Romero, R., Watson, E. V., Liang, A.C., Burger, M., Westcott, P.M.K., Mercer, K.L., Bronson, R.T., Wooten, E.C., Bhutkar, A., et al. (2022). A GATA4-regulated secretory program suppresses tumors through recruitment of cytotoxic CD8 T cells. Nat Commun *13*. https://doi.org/10.1038/s41467-021-27731-5.
69. Earley, Z.M., Lisicka, W., Sifakis, J.J., Aguirre-Gamboa, R., Kowalczyk, A., Barlow, J.T., Shaw, D.G., Discepolo, V., Tan, I.L., Gona, S., et al. (2023). GATA4 controls regionalization of tissue immunity and commensal-driven immunopathology. Immunity *56*. https://doi.org/10.1016/j.immuni.2022.12.009.
70. Zheng, R., Rebolledo-Jaramillo, B., Zong, Y., Wang, L., Russo, P., Hancock, W., Stanger, B.Z., Hardison, R.C., and Blobel, G.A. (2013). Function of GATA factors in the adult mouse liver. PLoS One *8*. https://doi.org/10.1371/journal.pone.0083723.
71. Watt, A.J., Zhao, R., Li, J., and Duncan, S.A. (2007). Development of the mammalian liver and ventral pancreas is dependent on GATA4. BMC Dev Biol *7*. https://doi.org/10.1186/1471-213X-7-37.
72. Dickel, D.E., Barozzi, I., Zhu, Y., Fukuda-Yuzawa, Y., Osterwalder, M., Mannion, B.J., May, D., Spurrell, C.H., Plajzer-Frick, I., Pickle, C.S., et al. (2016). Genome-wide compendium and functional assessment of in vivo heart enhancers. Nat Commun *7*, 1–13. https://doi.org/10.1038/ncomms12923.
73. Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Haugen, E., et al. (2020). Global reference mapping of human transcription factor footprints. Nature *583*, 729–736. https://doi.org/10.1038/s41586-020-2528-x.
74. Harley, J.B., Chen, X., Pujato, M., Miller, D., Maddox, A., Forney, C., Magnusen, A.F., Lynch, A., Chetal, K., Yukawa, M., et al. (2018). Transcription factors operate across disease loci, with EBNA2 implicated in autoimmunity. Nat Genet *50*. https://doi.org/10.1038/s41588-018-0102-3.
75. Banerji, R., Skibbens, R. V., and Iovine, M.K. (2017). How many roads lead to cohesinopathies? Developmental Dynamics *246*. https://doi.org/10.1002/dvdy.24510.
76. Boyle, M.I., Jespersgaard, C., Nazaryan, L., Bisgaard, A.M., and Tümer, Z. (2017). A novel RAD21 variant associated with intrafamilial phenotypic variation in Cornelia de Lange syndrome – review of the literature. Clin Genet *91*. https://doi.org/10.1111/cge.12863.
77. Krab, L.C., Marcos-Alcalde, I., Assaf, M., Balasubramanian, M., Andersen, J.B., Bisgaard, A.M., Fitzpatrick, D.R., Gudmundsson, S., Huisman, S.A., Kalayci, T., et al. (2020). Delineation of phenotypes and genotypes related to cohesin structural protein RAD21. Hum Genet *139*. https://doi.org/10.1007/s00439-020-02138-2.

78. Subramaniam, M., Hawse, J.R., Rajamannan, N.M., Ingle, J.N., and Spelsberg, T.C. (2010). Functional role of KLF10 in multiple disease processes. BioFactors *36*. https://doi.org/10.1002/biof.67.
79. Memon, A., and Lee, W.K. (2018). KLF10 as a tumor suppressor gene and its TGF-β signaling. Cancers (Basel) *10*. https://doi.org/10.3390/cancers10060161.
80. Yang, J., Zhang, H., Wang, X., Guo, J., Wei, L., Song, Y., Luo, Y., Zhao, Y.X., Subramaniam, M., Spelsberg, T.C., et al. (2021). Kruppel-like factor 10 protects against acute viral myocarditis by negatively regulating cardiac MCP-1 expression. Cell Mol Immunol *18*. https://doi.org/10.1038/s41423-020-00539-x.
81. Bjørnstad, J.L., Skrbic, B., Marstein, H.S., Hasic, A., Sjaastad, I., Louch, W.E., Florholmen, G., Christensen, G., and Tønnessen, T. (2012). Inhibition of SMAD2 phosphorylation preserves cardiac function during pressure overload. Cardiovasc Res *93*. https://doi.org/10.1093/cvr/cvr294.
82. Anderson, D.M., Anderson, K.M., Nelson, B.R., McAnally, J.R., Bezprozvannaya, S., Shelton, J.M., Bassel-Duby, R., and Olson, E.N. (2021). A myocardin-adjacent lncRNA balances SRF-dependent gene transcription in the heart. Genes Dev *38*. https://doi.org/10.1101/gad.348304.121.
83. Guo, Y., Cao, Y., Jardin, B.D., Sethi, I., Ma, Q., Moghadaszadeh, B., Troiano, E.C., Mazumdar, N., Trembley, M.A., Small, E.M., et al. (2021). Sarcomeres regulate murine cardiomyocyte maturation through MRTF-SRF signaling. Proc Natl Acad Sci U S A *118*. https://doi.org/10.1073/pnas.2008861118.
84. Zodanu, G.K.E., Hwang, J.H., Mudery, J., Sisniega, C., Kang, X., Wang, L.K., Barsegian, A., Biniwale, R.M., Si, M.S., Halnon, N.J., et al. (2025). Whole-Exome Sequencing Identifies Novel GATA5/6 Variants in Right-Sided Congenital Heart Defects. Int J Mol Sci *26*. https://doi.org/10.3390/ijms26052115.
85. Bornhorst, D., Hejjaji, A. V., Steuter, L., Woodhead, N.M., Maier, P., Gentile, A., Alhajkadour, A., Santis Larrain, O., Weber, M., Kikhi, K., et al. (2024). The heart is a resident tissue for hematopoietic stem and progenitor cells in zebrafish. Nature Communications *15*. https://doi.org/10.1038/s41467-024-51920-7.
86. 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*. https://doi.org/10.1371/journal.pone.0095253.
87. Wei-Yan Chow, R., Fukui, H., Chan, W.X., Tan, K.S.J., Roth, S., Duchemin, A.L., Messaddeq, N., Nakajima, H., Liu, F., Faggianelli-Conrozier, N., et al. (2022). Cardiac forces regulate zebrafish heart valve delamination by modulating Nfat signaling. PLoS Biol *20*. https://doi.org/10.1371/journal.pbio.3001505.
88. Deshpande, A., Shetty, P.M.V., Frey, N., and Rangrez, A.Y. (2022). SRF: a seriously responsible factor in cardiac development and disease. J Biomed Sci *29*. https://doi.org/10.1186/s12929-022-00820-3.
89. Schuster, K., Leeke, B., Meier, M., Wang, Y., Newman, T., Burgess, S., and Horsfield, J.A. (2015). A neural crest origin for cohesinopathy heart defects. Hum Mol Genet *24*. https://doi.org/10.1093/hmg/ddv402.
90. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell *38*, 576–589. https://doi.org/10.1016/j.molcel.2010.05.004.
91. Hao, Y., Zhang, X., Ran, S., Li, Y., Ye, W., Wang, S., Li, X., Luo, Z., Zhao, J., Zong, J., et al. (2025). KLF1 Promotes Cardiomyocyte Proliferation and Heart Regeneration Through Regulation of Wnt/β-Catenin Signaling Pathway. Advanced Science. https://doi.org/10.1002/advs.202413964.
92. Santoyo-Suarez, M.G., Mares-Montemayor, J.D., Padilla-Rivas, G.R., Delgado-Gallegos, J.L., Quiroz-Reyes, A.G., Roacho-Perez, J.A., Benitez-Chao, D.F., Garza-Ocañas, L., Arevalo-

Martinez, G., Garza-Treviño, E.N., et al. (2023). The Involvement of Krüppel-like Factors in Cardiovascular Diseases. Life *13*. https://doi.org/10.3390/life13020420.

93. Warren, S.A., Terada, R., Briggs, L.E., Cole-Jeffrey, C.T., Chien, W.-M., Seki, T., Weinberg, E.O., Yang, T.P., Chin, M.T., Bungert, J., 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. https://doi.org/10.1128/mcb.05940-11.
94. Hall, E.J., Pal, S., Glennon, M.S., Shridhar, P., Satterfield, S.L., Weber, B., Zhang, Q., Salama, G., Lal, H., and Becker, J.R. (2022). Cardiac natriuretic peptide deficiency sensitizes the heart to stress-induced ventricular arrhythmias via impaired CREB signalling. Cardiovasc Res *118*. https://doi.org/10.1093/cvr/cvab257.
95. Dawuti, A., Sun, S., Wang, R., Gong, D., Liu, R., Kong, D., Yuan, T., Zhou, J., Lu, Y., Wang, S., et al. (2023). Salvianolic acid A alleviates heart failure with preserved ejection fraction via regulating TLR/Myd88/TRAF/NF-κB and p38MAPK/CREB signaling pathways. Biomedicine and Pharmacotherapy *168*. https://doi.org/10.1016/j.biopha.2023.115837.
96. Chapman, D.L., Garvey, N., Hancock, S., Alexiou, M., Agulnik, S.I., Gibson-Brown, J.J., Cebra-Thomas, J., Bollag, R.J., Silver, L.M., and Papaioannou, V.E. (1996). Expression of the T-box family genes, Tbx1-Tbx5, during early mouse development. Developmental Dynamics *206*. https://doi.org/10.1002/(SICI)1097-0177(199608)206:4<379::AID-AJA4>3.0.CO;2-F.
97. Wang, T.Y., Lee, D., Fox-Talbot, K., Arking, D.E., Chakravarti, A., and Halushka, M.K. (2018). Cardiomyocytes have mosaic patterns of protein expression. Cardiovascular Pathology *34*. https://doi.org/10.1016/j.carpath.2018.03.002.
98. Agarwal, R., Wakimoto, H., Paulo, J.A., Zhang, Q., Reichart, D., Toepfer, C., Sharma, A., Tai, A.C., Lun, M., Gorham, J., et al. (2022). Pathogenesis of Cardiomyopathy Caused by Variants in ALPK3, an Essential Pseudokinase in the Cardiomyocyte Nucleus and Sarcomere. Circulation *146*. https://doi.org/10.1161/CIRCULATIONAHA.122.059688.
99. Zhang, X., Wang, Z., Xu, Q., Chen, Y., Liu, W., Zhong, T., Li, H., Quan, C., Zhang, L., and Cui, C.P. (2021). Splicing factor Srsf5 deletion disrupts alternative splicing and causes noncompaction of ventricular myocardium. iScience *24*. https://doi.org/10.1016/j.isci.2021.103097.
100. Hang, C., Song, Y., Li, Y., Zhang, S., Chang, Y., Bai, R., Saleem, A., Jiang, M., Lu, W., Lan, F., et al. (2021). Knockout of MYOM1 in human cardiomyocytes leads to myocardial atrophy via impairing calcium homeostasis. J Cell Mol Med *25*. https://doi.org/10.1111/jcmm.16268.
101. Ma, Z.G., Yuan, Y.P., Fan, D., Zhang, X., Teng, T., Song, P., Kong, C.Y., Hu, C., Wei, W.Y., and Tang, Q.Z. (2023). IRX2 regulates angiotensin II-induced cardiac fibrosis by transcriptionally activating EGR1 in male mice. Nat Commun *14*. https://doi.org/10.1038/s41467-023-40639-6.
102. Kim, K.H., Rosen, A., Bruneau, B.G., Hui, C.C., and Backx, P.H. (2012). Iroquois homeodomain transcription factors in heart development and function. Circ Res *110*. https://doi.org/10.1161/CIRCRESAHA.112.265041.
103. Wang, K., Zhou, M., Zhang, Y., Du, Y., Li, P., Guan, C., and Huang, Z. (2023). IRX2 activated by jumonji domain-containing protein 2A is crucial for cardiac hypertrophy and dysfunction in response to the hypertrophic stimuli. Int J Cardiol *371*. https://doi.org/10.1016/j.ijcard.2022.09.070.
104. Hota, S.K., Rao, K.S., Blair, A.P., Khalilimeybodi, A., Hu, K.M., Thomas, R., So, K., Kameswaran, V., Xu, J., Polacco, B.J., et al. (2022). Brahma safeguards canalization of cardiac mesoderm differentiation. Nature *602*. https://doi.org/10.1038/s41586-021-04336-y.
105. Mayer, S.C., Gilsbach, R., Preissl, S., Monroy Ordonez, E.B., Schnick, T., Beetz, N., Lother, A., Rommel, C., Ihle, H., Bugger, H., et al. (2015). Adrenergic Repression of the Epigenetic Reader MeCP2 Facilitates Cardiac Adaptation in Chronic Heart Failure. Circ Res *117*. https://doi.org/10.1161/CIRCRESAHA.115.306721.
106. Wang, C., Wang, F., Cao, Q., Li, Z., Huang, L., and Chen, S. (2018). The effect of Mecp2 on heart failure. Cellular Physiology and Biochemistry *47*. https://doi.org/10.1159/000491610.

107. Alvarez-Saavedra, M., Carrasco, L., Sura-Trueba, S., Aiello, V.D., Walz, K., Neto, J.X., and Young, J.I. (2010). Elevated expression of MeCP2 in cardiac and skeletal tissues is detrimental for normal development. Hum Mol Genet *19*. https://doi.org/10.1093/hmg/ddq096.
108. Hara, M., Takahashi, T., Mitsumasu, C., Igata, S., Takano, M., Minami, T., Yasukawa, H., Okayama, S., Nakamura, K., Okabe, Y., et al. (2015). Disturbance of cardiac gene expression and cardiomyocyte structure predisposes Mecp2-null mice to arrhythmias. Sci Rep *5*. https://doi.org/10.1038/srep11204.
109. von Both, I., Silvestri, C., Erdemir, T., Lickert, H., Walls, J.R., Henkelman, R.M., Rossant, J., Harvey, R.P., Attisano, L., and Wrana, J.L. (2004). Foxh1 is essential for development of the anterior heart field. Dev Cell *7*. https://doi.org/10.1016/j.devcel.2004.07.023.
110. Lenhart, K.F., Holtzman, N.G., Williams, J.R., and Burdine, R.D. (2013). Integration of Nodal and BMP Signals in the Heart Requires FoxH1 to Create Left-Right Differences in Cell Migration Rates That Direct Cardiac Asymmetry. PLoS Genet *9*. https://doi.org/10.1371/journal.pgen.1003109.
111. Siersbæk, R., Rabiee, A., Nielsen, R., Sidoli, S., Traynor, S., Loft, A., Poulsen, L.L.C., Rogowska-Wrzesinska, A., Jensen, O.N., and Mandrup, S. (2014). Transcription factor cooperativity in early adipogenic hotspots and super-enhancers. Cell Rep *7*, 1443–1455. https://doi.org/10.1016/j.celrep.2014.04.042.
112. 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. https://doi.org/10.1016/j.gde.2017.02.007.
113. Visan, I. (2014). Ascl2 for TFH cells. Nat Immunol *15*. https://doi.org/10.1038/ni.2838.
114. Visan, I. (2014). Heart macrophages. Nat Immunol *15*. https://doi.org/10.1038/ni.2837.
115. Tsushima, K., Osawa, T., Yanai, H., Nakajima, A., Takaoka, A., Manabe, I., Ohba, Y., Imai, Y., Taniguchi, T., and Nagai, R. (2011). IRF3 regulates cardiac fibrosis but not hypertrophy in mice during angiotensin II-induced hypertension. The FASEB Journal *25*. https://doi.org/10.1096/fj.10-174615.
116. King, K.R., Aguirre, A.D., Ye, Y.X., Sun, Y., Roh, J.D., Ng, R.P., Kohler, R.H., Arlauckas, S.P., Yoshiko, V., Savo, A., et al. (2017). IRF3 and type i interferons fuel a fatal response to myocardial infarction. Nat Med *23*. https://doi.org/10.1038/nm.4428.
117. Sun, H., and Wang, Y. (2014). Interferon regulatory factors in heart: Stress response beyond inflammation. Hypertension *63*. https://doi.org/10.1161/HYPERTENSIONAHA.113.02795.
118. Leonard, D., Svenungsson, E., Sandling, J.K., Berggren, O., Jönsen, A., Bengtsson, C., Wang, C., Jensen-Urstad, K., Granstam, S.O., Bengtsson, A.A., et al. (2013). Coronary heart disease in systemic lupus erythematosus is associated with interferon regulatory factor-8 gene variants. Circ Cardiovasc Genet *6*. https://doi.org/10.1161/CIRCGENETICS.113.000044.
119. Jiang, D.S., Wei, X., Zhang, X.F., Liu, Y., Zhang, Y., Chen, K., Gao, L., Zhou, H., Zhu, X.H., Liu, P.P., et al. (2014). IRF8 suppresses pathological cardiac remodelling by inhibiting calcineurin signalling. Nat Commun *5*. https://doi.org/10.1038/ncomms4303.
120. Sun, H., and Wang, Y. (2014). Interferon Regulatory Factors in Heart. Hypertension *63*. https://doi.org/10.1161/hypertensionaha.113.02795.
121. Huang, R., Chen, X., Long, Y., and Chen, R. (2019). MIR-31 promotes Th22 differentiation through targeting Bach2 in coronary heart disease. Biosci Rep *39*. https://doi.org/10.1042/BSR20190986.
122. Jiang, X., Cao, M., Wu, J., Wang, X., Zhang, G., Yang, C., Gao, P., and Zou, Y. (2022). Protections of transcription factor BACH2 and natural product myricetin against pathological cardiac hypertrophy and dysfunction. Front Physiol *13*. https://doi.org/10.3389/fphys.2022.971424.
123. Tsai, F.C., Chang, G.J., Hsu, Y.J., Lin, Y.M., Lee, Y.S., Chen, W.J., Kuo, C.T., and Yeh, Y.H. (2016). Proinflammatory gene expression in patients undergoing mitral valve surgery and maze ablation for atrial fibrillation. Journal of Thoracic and Cardiovascular Surgery *151*. https://doi.org/10.1016/j.jtcvs.2015.12.003.
124. Katsuoka, F., Motohashi, H., Onodera, K., Suwabe, N., Engel, J.D., and Yamamoto, M. (2000). One enhancer mediates mafK transcriptional activation in both hematopoietic and cardiac muscle cells. EMBO Journal *19*. https://doi.org/10.1093/emboj/19.12.2980.

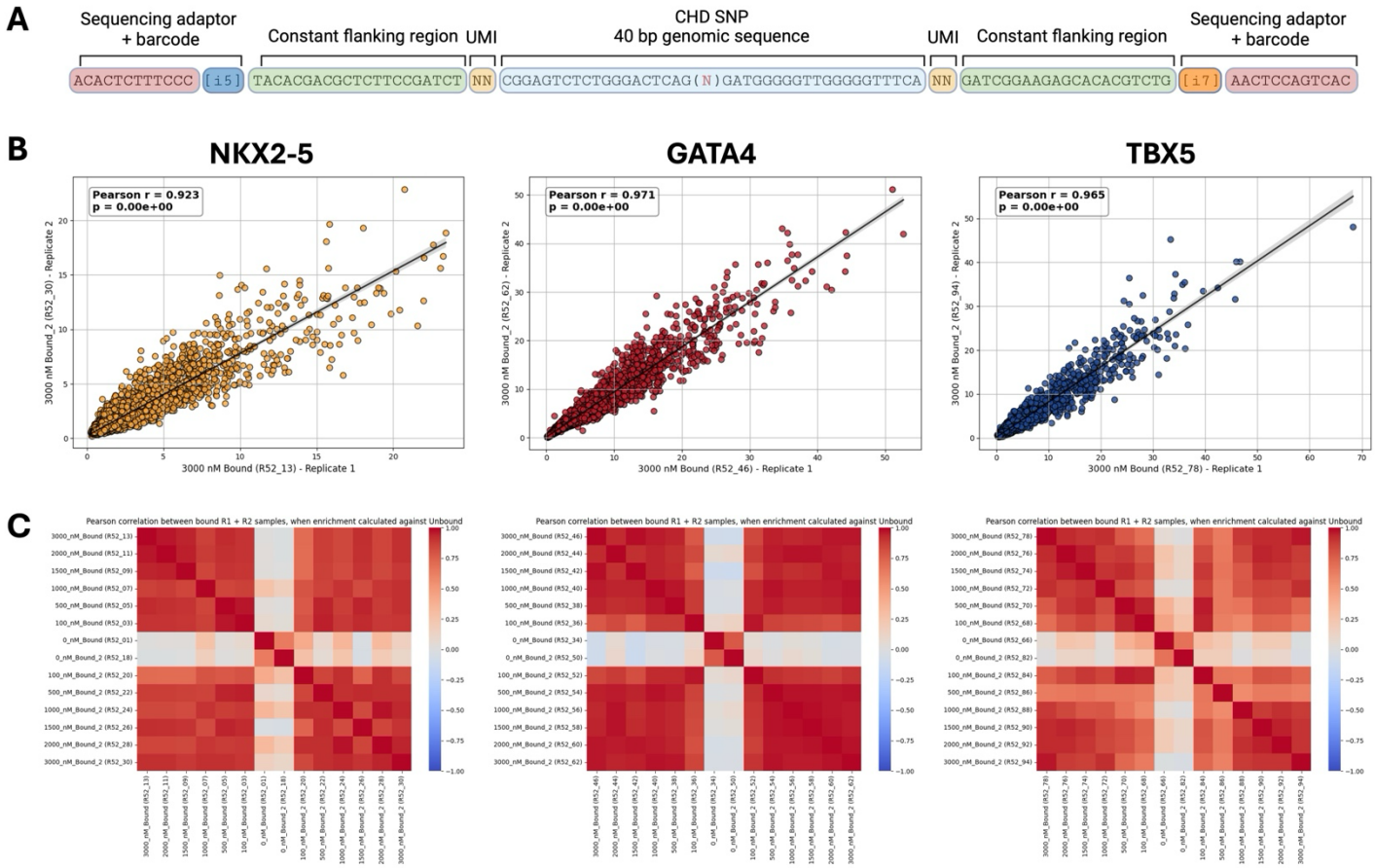
125. Hu, H., Lin, S., Wang, S., and Chen, X. (2020). The Role of Transcription Factor 21 in Epicardial Cell Differentiation and the Development of Coronary Heart Disease. Front Cell Dev Biol *8*. https://doi.org/10.3389/fcell.2020.00457.
126. Miller, C.L., Anderson, D.R., Kundu, R.K., Raiesdana, A., Nürnberg, S.T., Diaz, R., Cheng, K., Leeper, N.J., Chen, C.H., Chang, I.S., et al. (2013). Disease-Related Growth Factor and Embryonic Signaling Pathways Modulate an Enhancer of TCF21 Expression at the 6q23.2 Coronary Heart Disease Locus. PLoS Genet *9*. https://doi.org/10.1371/journal.pgen.1003652.
127. Miller, C.L., Haas, U., Diaz, R., Leeper, N.J., Kundu, R.K., Patlolla, B., Assimes, T.L., Kaiser, F.J., Perisic, L., Hedin, U., et al. (2014). Coronary Heart Disease-Associated Variation in TCF21 Disrupts a miR-224 Binding Site and miRNA-Mediated Regulation. PLoS Genet *10*. https://doi.org/10.1371/journal.pgen.1004263.
128. Tandon, P., Miteva, Y. V., Kuchenbrod, L.M., Cristea, I.M., and Conlon, F.L. (2012). Tcf21 regulates the specification and maturation of proepicardial cells. Development (Cambridge) *140*. https://doi.org/10.1242/dev.093385.
129. Yi, Y., Zhang, H., Chen, M., Chen, B., Chen, Y., Li, P., Zhou, H., Ma, Z., and Jiang, H. (2023). Inhibition of multiple uptake transporters in cardiomyocytes/mitochondria alleviates doxorubicin-induced cardiotoxicity. Chem Biol Interact *382*. https://doi.org/10.1016/j.cbi.2023.110627.
130. Huang, K.M., Thomas, M.Z., Magdy, T., Eisenmann, E.D., Uddin, M.E., DiGiacomo, D.F., Pan, A., Keiser, M., Otter, M., Xia, S.H., et al. (2021). Targeting OCT3 attenuates doxorubicin-induced cardiac injury. Proc Natl Acad Sci U S A *118*. https://doi.org/10.1073/pnas.2020168118.
131. Zhang, J., Song, Y., Wang, X., Wang, X., Li, S., Song, X., Zhao, C., Qi, J., Tian, Y., Zhao, B., et al. (2025). The transcription factor PITX1 cooperates with super-enhancers to regulate the expression of DUSP4 and inhibit pyroptosis in pulmonary artery smooth muscle cells. Respir Res *26*. https://doi.org/10.1186/s12931-025-03222-9.
132. Sohail, A., Nicoll, O., and Bendall, A.J. (2025). Assessing candidate DLX-regulated genes in the first pharyngeal arch of chick embryos. Developmental Dynamics. https://doi.org/10.1002/dvdy.765.
133. Yu, Q., Cai, B., Zhang, Y., Xu, J., Liu, D., Zhang, X., Han, Z., Ma, Y., Jiao, L., Gong, M., et al. (2023). Long non-coding RNA LHX1-DT regulates cardiomyocyte differentiation through H2A.Z-mediated LHX1 transcriptional activation. iScience *26*. https://doi.org/10.1016/j.isci.2023.108051.
134. Pickett, C.J., Gruner, H.N., and Davidson, B. (2024). Lhx3/4 initiates a cardiopharyngeal-specific transcriptional program in response to widespread FGF signaling. PLoS Biol *22*. https://doi.org/10.1371/journal.pbio.3002169.
135. Smagulova, F.O., Manuylov, N.L., Leach, L.L., and Tevosian, S.G. (2008). GATA4/FOG2 transcriptional complex regulates Lhx9 gene expression in murine heart development. BMC Dev Biol *8*. https://doi.org/10.1186/1471-213X-8-67.
136. Habets, P.E.M.H., Moorman, A.F.M., Clout, D.E.W., Van Roon, M.A., Lingbeek, M., Van Lohuizen, M., Campione, M., and Christoffels, V.M. (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. https://doi.org/10.1101/gad.222902.
137. Gitter, A., Lu, Y., and Bar-joseph, Z. (2010). Regulatory Regions in DNA: Promoters, Enhancers, Silencers, and Insulators. Computational Biology of Transcription Factor Binding, Methods in Molecular Biology *674*. https://doi.org/10.1007/978-1-60761-854-6.
138. Pang, B., van Weerd, J.H., Hamoen, F.L., and Snyder, M.P. (2023). Identification of non-coding silencer elements and their regulation of gene expression. Nat Rev Mol Cell Biol *24*, 383–395. https://doi.org/10.1038/s41580-022-00549-9.
139. Cornejo-Páramo, P., Roper, K., Degnan, S.M., Degnan, B.M., and Wong, E.S. (2022). Distal regulation, silencers, and a shared combinatorial syntax are hallmarks of animal embryogenesis. Genome Res *32*, 474–487. https://doi.org/10.1101/gr.275864.121.

140. Lin, A.C., Roche, A.E., Wilk, J., and Svensson, E.C. (2004). The N termini of Friend of GATA (FOG) proteins define a novel transcriptional repression motif and a superfamily of transcriptional repressors. Journal of Biological Chemistry *279*. https://doi.org/10.1074/jbc.M411240200.
141. Robbe, Z.L., Shi, W., Wasson, L.K., Scialdone, A.P., Wilczewski, C.M., Sheng, X., Hepperla, A.J., Akerberg, B.N., Pu, W.T., Cristea, I.M., 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. https://doi.org/10.1101/gad.349154.121.
142. Han, X., Tang, J., Chen, T., and Ren, G. (2019). Restoration of GATA4 expression impedes breast cancer progression by transcriptional repression of ReLA and inhibition of NF-κB signaling. J Cell Biochem *120*. https://doi.org/10.1002/jcb.27455.
143. Leatherbury, L., and Berul, C.I. (2017). Genetics of Congenital Heart Disease: Is the Glass Now Half-Full? Circ Cardiovasc Genet *10*. https://doi.org/10.1161/CIRCGENETICS.117.001746.
144. Yin, Y., Ye, L., Chen, M., Liu, H., and Miao, J. (2024). Unraveling cardiomyocyte responses and intercellular communication alterations in primary carnitine deficiency cardiomyopathy via single-nucleus RNA sequencing. Heliyon *10*. https://doi.org/10.1016/j.heliyon.2024.e33581.
145. Lee, J.H., Cho, M.H., Hersh, C.P., McDonald, M.L.N., Wells, J.M., Dransfield, M.T., Bowler, R.P., Lynch, D.A., Lomas, D.A., Crapo, J.D., et al. (2015). IREB2 and GALC are associated with pulmonary artery enlargement in chronic obstructive pulmonary disease. Am J Respir Cell Mol Biol *52*. https://doi.org/10.1165/rcmb.2014-0210OC.
146. Rafi, M.A., Rao, H.Z., Luzi, P., Luddi, A., Curtis, M.T., and Wenger, D.A. (2015). Intravenous injection of AAVrh10-GALC after the neonatal period in twitcher mice results in significant expression in the central and peripheral nervous systems and improvement of clinical features. Mol Genet Metab *114*. https://doi.org/10.1016/j.ymgme.2014.12.300.
147. Le Bras, A. (2018). Basic research: RNA deadenylation by CCR4-NOT controls heart function. Nat Rev Cardiol *15*. https://doi.org/10.1038/nrcardio.2018.14.
148. Elmén, L., Volpato, C.B., Kervadec, A., Pineda, S., Kalvakuri, S., Alayari, N.N., Foco, L., Pramstaller, P.P., Ocorr, K., Rossini, A., et al. (2020). Silencing of CCR4-NOT complex subunits affects heart structure and function. DMM Disease Models and Mechanisms *13*. https://doi.org/10.1242/dmm.044727.
149. Bukas, C., Galter, I., da Silva-Buttkus, P., Fuchs, H., Maier, H., Gailus-Durner, V., Müller, C.L., Hrabě de Angelis, M., Piraud, M., and Spielmann, N. (2023). Echo2Pheno: a deep-learning application to uncover echocardiographic phenotypes in conscious mice. Mammalian Genome *34*. https://doi.org/10.1007/s00335-023-09996-x.
150. Yamaguchi, T., Suzuki, T., Sato, T., Takahashi, A., Watanabe, H., Kadowaki, A., Natsui, M., Inagaki, H., Arakawa, S., Nakaoka, S., et al. (2018). The CCR4-NOT deadenylase complex controls atg7-dependent cell death and heart function. Sci Signal *11*. https://doi.org/10.1126/scisignal.aan3638.
151. Wang, Y., and Xian, H. (2022). Identifying Genes Related to Acute Myocardial Infarction Based on Network Control Capability. Genes (Basel) *13*. https://doi.org/10.3390/genes13071238.
152. Ngai, D., Lino, M., Rothenberg, K.E., Simmons, C.A., Fernandez-Gonzalez, R., and Bendeck, M.P. (2020). DDR1 (Discoidin Domain Receptor-1)-RhoA (Ras Homolog Family Member A) Axis Senses Matrix Stiffness to Promote Vascular Calcification. Arterioscler Thromb Vasc Biol *40*. https://doi.org/10.1161/ATVBAHA.120.314697.
153. Ma, K., Xie, M., He, X., Liu, G., Lu, X., Peng, Q., Zhong, B., and Li, N. (2018). A novel compound heterozygous mutation in VARS2 in a newborn with mitochondrial cardiomyopathy: A case report of a Chinese family. BMC Med Genet *19*. https://doi.org/10.1186/s12881-018-0689-3.
154. Bruni, F., Di Meo, I., Bellacchio, E., Webb, B.D., McFarland, R., Chrzanowska-Lightowlers, Z.M.A., He, L., Skorupa, E., Moroni, I., Ardissone, A., et al. (2018). Clinical, biochemical, and genetic features associated with VARS2-related mitochondrial disease. Hum Mutat *39*. https://doi.org/10.1002/humu.23398.
155. Kayvanpour, E., Wisdom, M., Lackner, M.K., Sedaghat-Hamedani, F., Boeckel, J.N., Müller, M., Eghbalian, R., Dudek, J., Doroudgar, S., Maack, C., et al. (2022). VARS2 Depletion Leads to

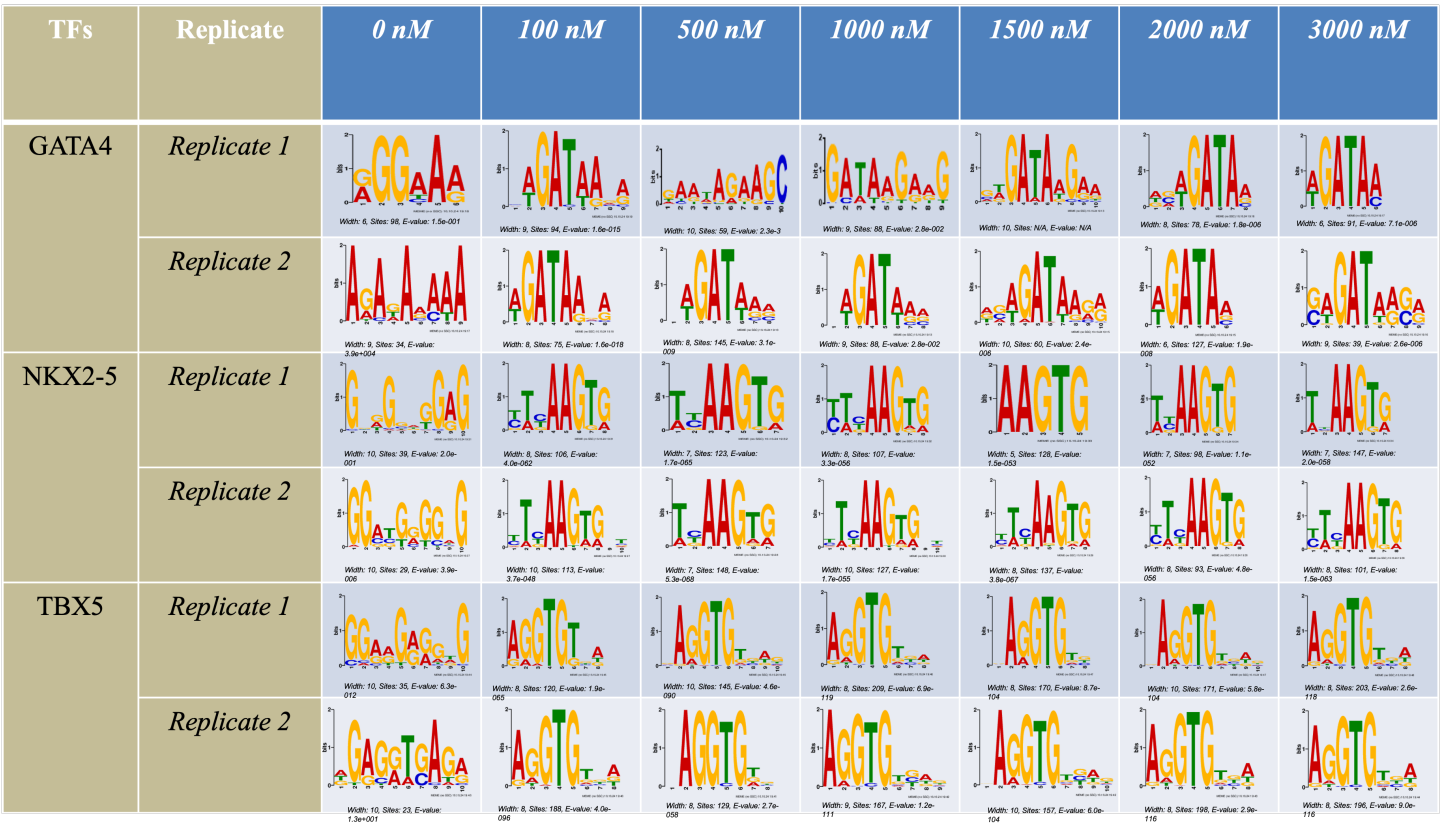
Activation of the Integrated Stress Response and Disruptions in Mitochondrial Fatty Acid Oxidation. Int J Mol Sci *23*. https://doi.org/10.3390/ijms23137327.

156. Cordell, H.J., Töpf, A., Mamasoula, C., Postma, A. V., Bentham, J., Zelenika, D., Heath, S., Blue, G., Cosgrove, C., Granados riveron, J., et al. (2013). Genome-wide association study identifies loci on 12q24 and 13q32 associated with Tetralogy of Fallot. Hum Mol Genet *22*, 1473–1481. https://doi.org/10.1093/hmg/dds552.
157. Cordell, H.J., Bentham, J., Topf, A., Zelenika, D., Heath, S., Mamasoula, C., Cosgrove, C., Blue, G., Granados-Riveron, J., Setchfield, K., et al. (2013). Genome-wide association study of multiple congenital heart disease phenotypes identifies a susceptibility locus for atrial septal defect at chromosome 4p16. Nat Genet *45*, 822–824. https://doi.org/10.1038/ng.2637.
158. Cordell, H.J., Bentham, J., Topf, A., Zelenika, D., Heath, S., Mamasoula, C., Cosgrove, C., Blue, G., Granados-Riveron, J., Setchfield, K., et al. (2013). Genome-wide association study of multiple congenital heart disease phenotypes identifies a susceptibility locus for atrial septal defect at chromosome 4p16. Nat Genet *45*, 822–824. https://doi.org/10.1038/ng.2637.
159. Mitchell, L.E., Agopian, A.J., Bhalla, A., Glessner, J.T., Kim, C.E., Swartz, M.D., Hakonarson, H., and Goldmuntz, E. (2015). Genome-wide association study of maternal and inherited effects on left-sided cardiac malformations. Hum Mol Genet *24*, 265–273. https://doi.org/10.1093/hmg/ddu420.
160. 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. Developmental Dynamics *237*, 3809–3819. https://doi.org/10.1002/dvdy.21803.
161. Ardlie, K.G., DeLuca, D.S., Segrè, A. V., Sullivan, T.J., Young, T.R., Gelfand, E.T., Trowbridge, C.A., Maller, J.B., Tukiainen, T., Lek, M., et al. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (1979) *348*. https://doi.org/10.1126/science.1262110.
162. Xiao, Y., Wang, J., Li, J., Zhang, P., Li, J., Zhou, Y., Zhou, Q., Chen, M., Sheng, X., Liu, Z., et al. (2023). An analytical framework for decoding cell type-specific genetic variation of gene regulation. Nat Commun *14*. https://doi.org/10.1038/s41467-023-39538-7.
163. Erratum: Genetic effects on gene expression across human tissues (Nature (2017) 550 (204-213) DOI: 10.1038/nature24277) (2018). Nature *553*. https://doi.org/10.1038/nature25160.
164. Siraj, L., Castro, R.I., Dewey, H., Kales, S., Nguyen, T.T.L., Kanai, M., Berenzy, D., Mouri, K., Wang, Q., McCaw, Z.R., et al. (2024). Functional dissection of complex and molecular trait variants at single nucleotide resolution. Biorxiv. https://doi.org/10.1101/2024.05.05.592437.
165. Huang, L., Rosen, J.D., Sun, Q., Chen, J., Wheeler, M.M., Zhou, Y., Min, Y.I., Kooperberg, C., Conomos, M.P., Stilp, A.M., et al. (2022). TOP-LD: A tool to explore linkage disequilibrium with TOPMed whole-genome sequence data. Am J Hum Genet *109*, 1175–1181. https://doi.org/10.1016/j.ajhg.2022.04.006.
166. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature *590*, 290–299. https://doi.org/10.1038/s41586-021-03205-y.
167. Shook, M.S., Lu, X., Chen, X., Parameswaran, S., Edsall, L., Trimarchi, M.P., Ernst, K., Granitto, M., Forney, C., Donmez, O.A., et al. (2024). Systematic identification of genotype-dependent enhancer variants in eosinophilic esophagitis. Am J Hum Genet *111*, 280–294. https://doi.org/10.1016/j.ajhg.2023.12.008.
168. Maury, E.A., Jones, A., Seplyarskiy, V., Thanh, T., Nguyen, L., Rosenbluh, C., Bae, T., Wang, Y., Abyzov, A., Khoshkhoo, S., et al. (2024). Somatic mosaicism in schizophrenia brains reveals prenatal mutational processes. Science (1979) *386*, 217–224.

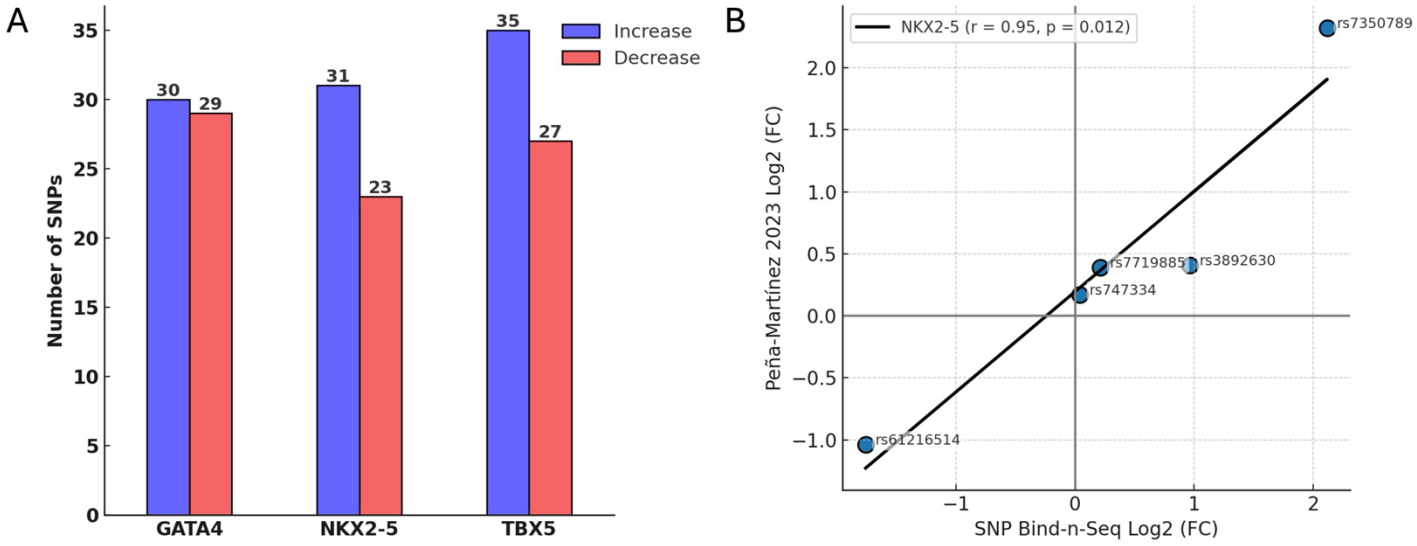
Supplementary Figures



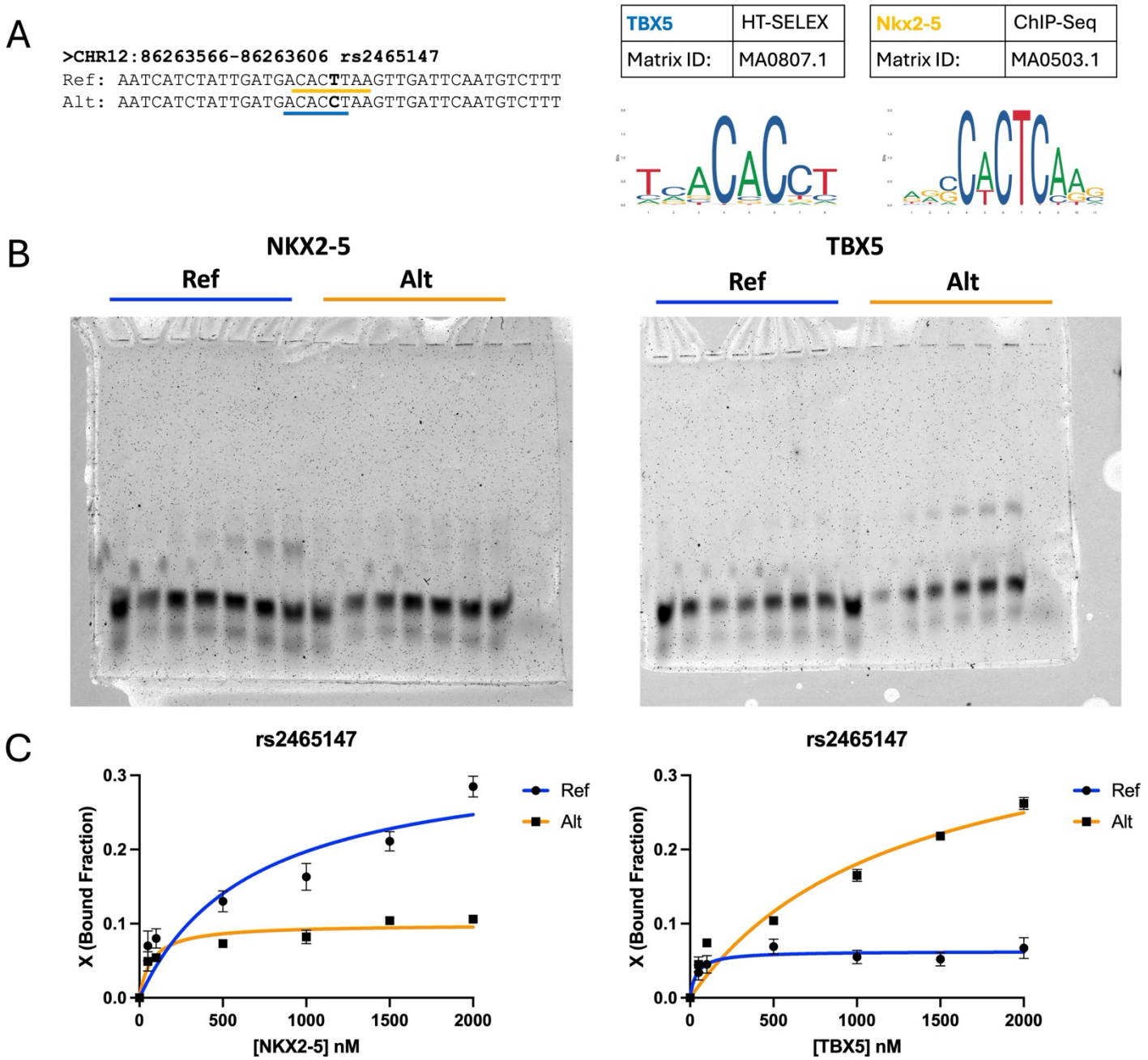
Supplementary Figure 1: SNP Bind-n-Seq library anatomy and replicate correlation. **A)** SNP Bind-n-Seq library sequence features, structures, and constant regions. **B)** Experimental correlation between replicates of NKX2-5 (left), GATA4 (middle), and TBX5 (right) at 3,000 nM. **C)** Correlation Matrix of Replicate Enrichment Values Relative to Unbound Conditions. Heatmap of the correlation matrix of NKX2-5 (left), GATA4 (middle), and TBX5 (right) enrichment across concentrations, relative to the corresponding Unbound condition in both replicates. Each cell represents the Pearson correlation coefficient between two concentrations, indicating the similarity in enrichment patterns.



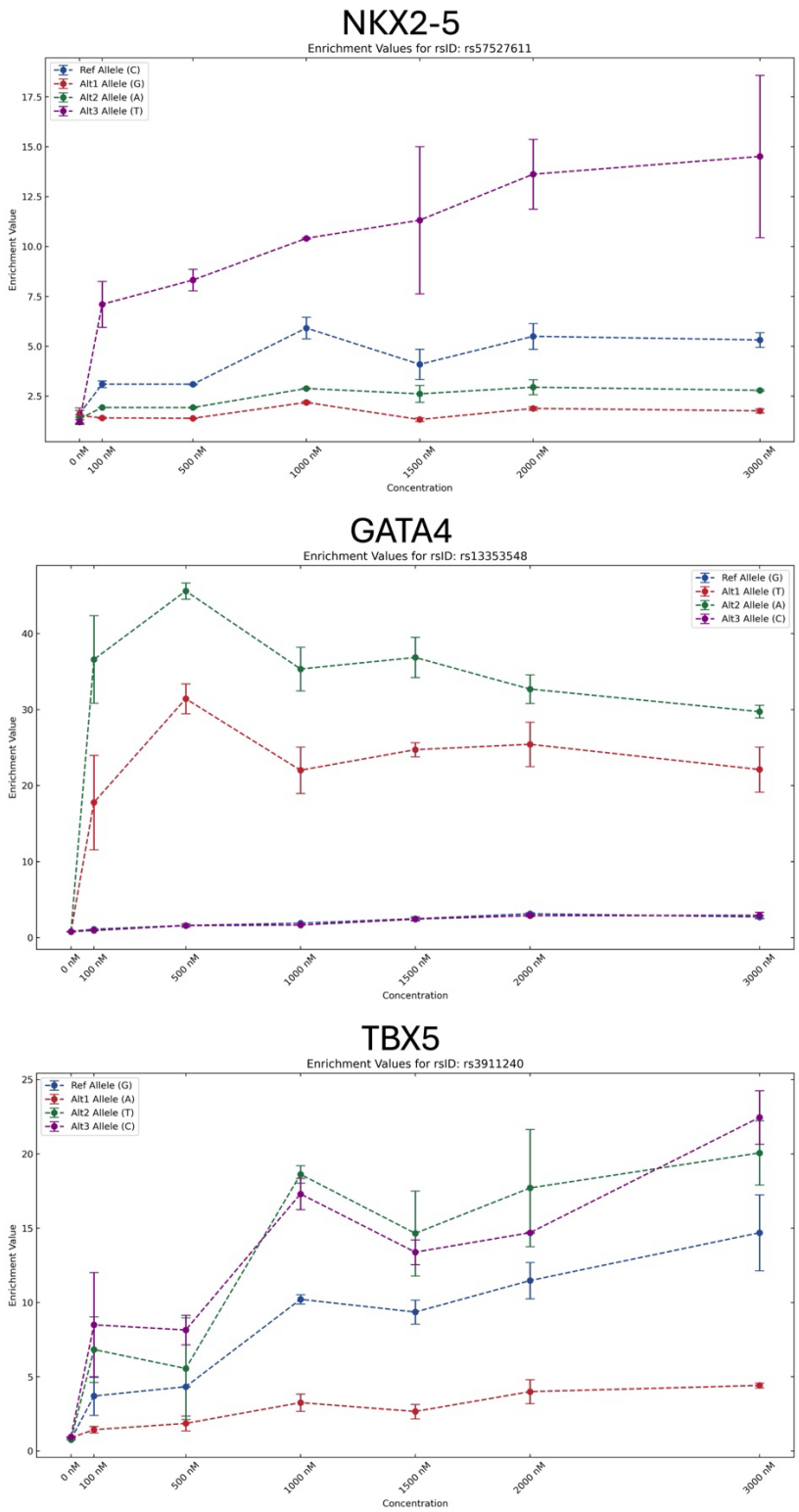
Supplementary Figure 2: TF motif enrichment across SNP Bind-n-Seq replicates and concentration points. Motifs were generated using MEME with the top 500 sequences with the highest K_A values.



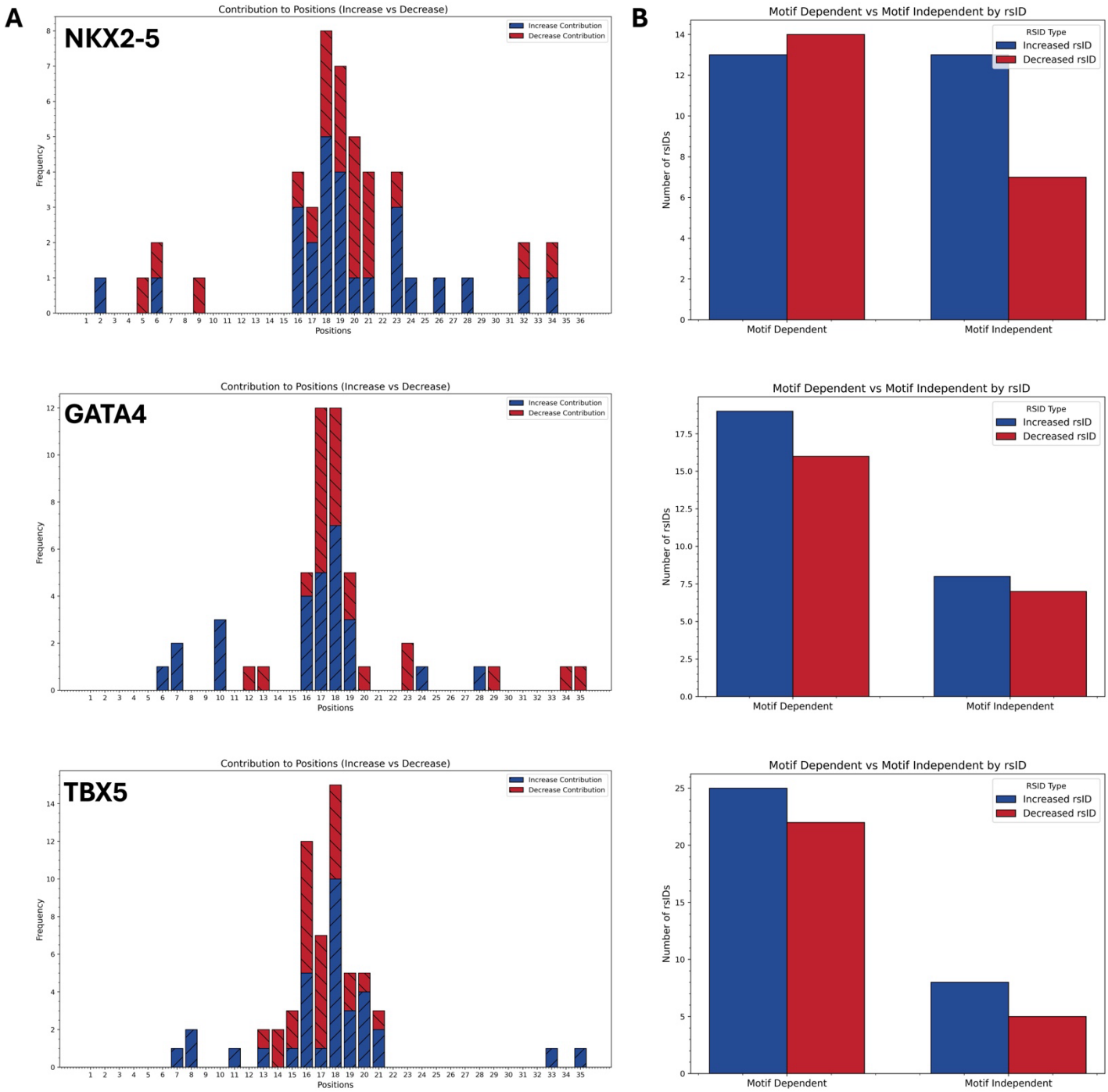
Supplementary Figure 3: Variants with differential allelic binding identified through SNP Bind-n-Seq. **A)** Number of variants with allelic binding for NKX2-5, GATA4, and TBX5. **B)** Fold change correlations of previously described variants for NKX2-5 binding affinity.



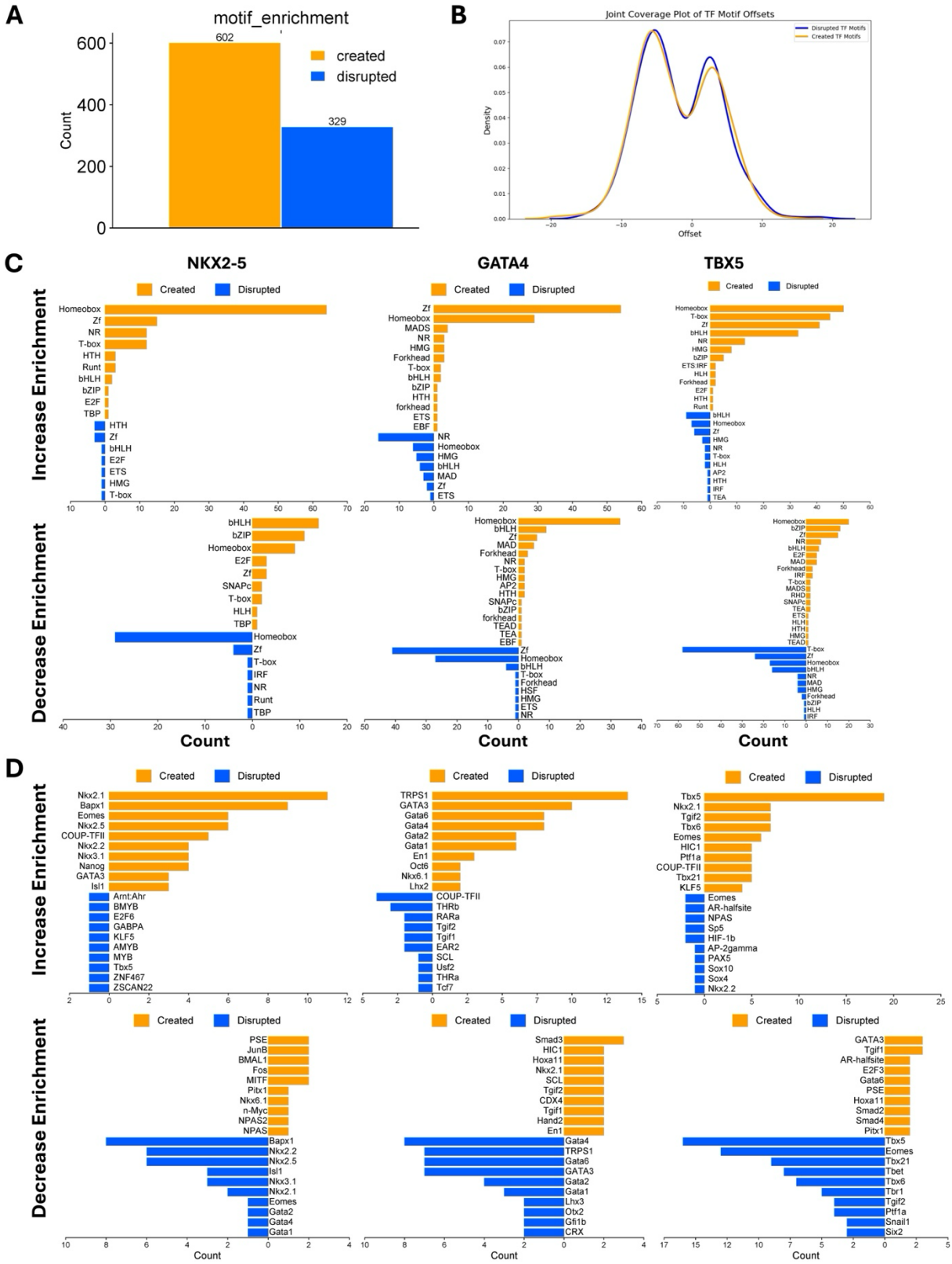
Supplementary Figure 4: In vitro validation of SNP Bin-n-Seq variant rs2465147. **A)** Reference and alternate sequences. TF binding motifs are highlighted in yellow for NKX2-5 and blue for TBX5. JASPAR motifs from NKX2-5 and TBX5 are displayed on the right. **B)** Electrophoretic mobility shift assay (EMSA) of rs2465147 for NXK2-5 (left) and TBX5 (right). **C)** Binding curves generated from EMSA of rs2465147 in triplicate.



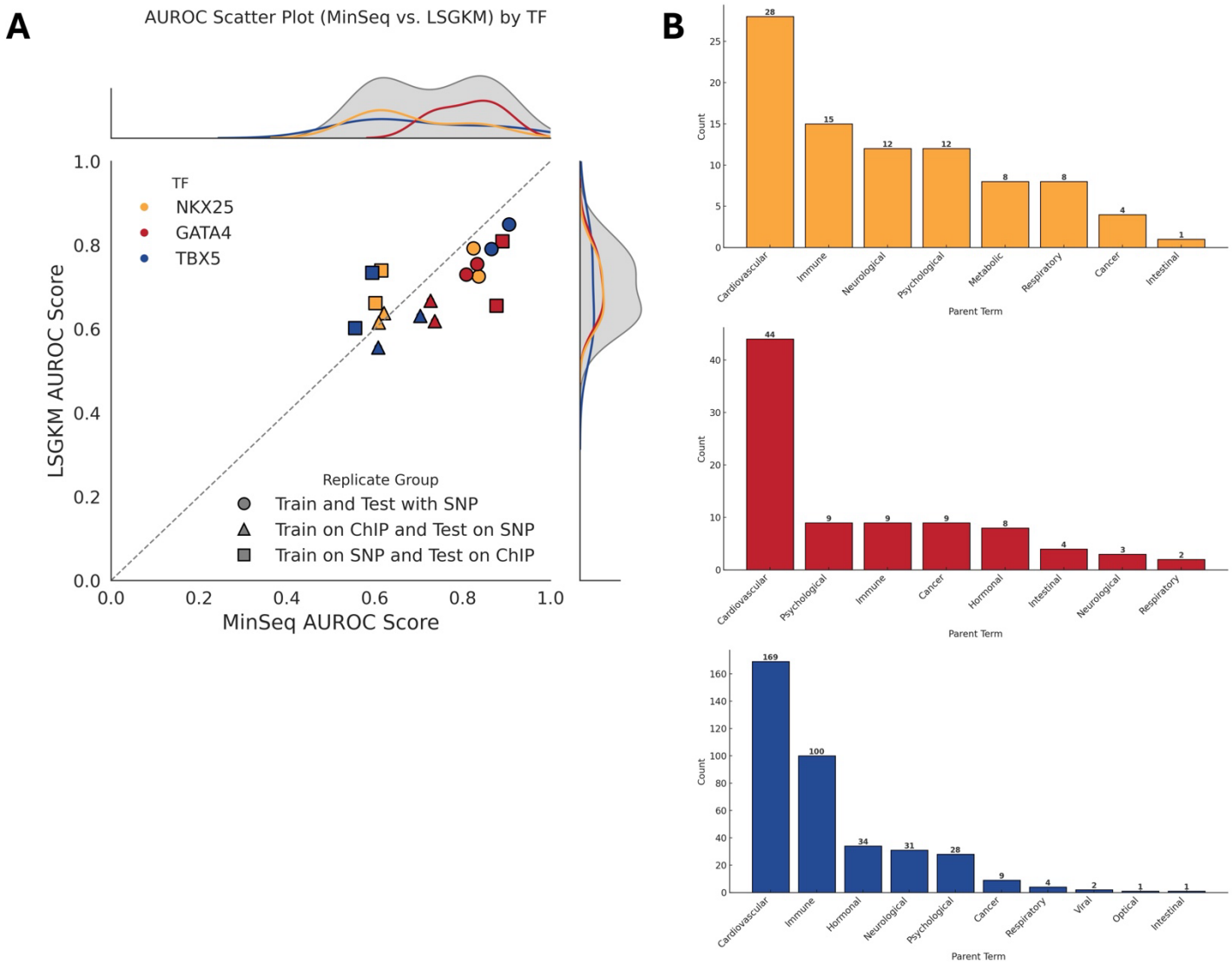
Supplementary Figure 5: Allelic enrichment curve of non-CHD-risk alternate alleles. Variants rs57527611, rs13353548, and rs3911240 are plotted for NKX2-5 (top), GATA4 (middle), and TBX5 (bottom) binding, respectively. Reference alleles (Ref) are represented in blue, and tag-SNP alleles from the GWAS catalog (Alt 2) are represented in red. Permutated alleles (alternate non-risk; Alt 2 and Alt 3) are represented in green and purple, respectively.



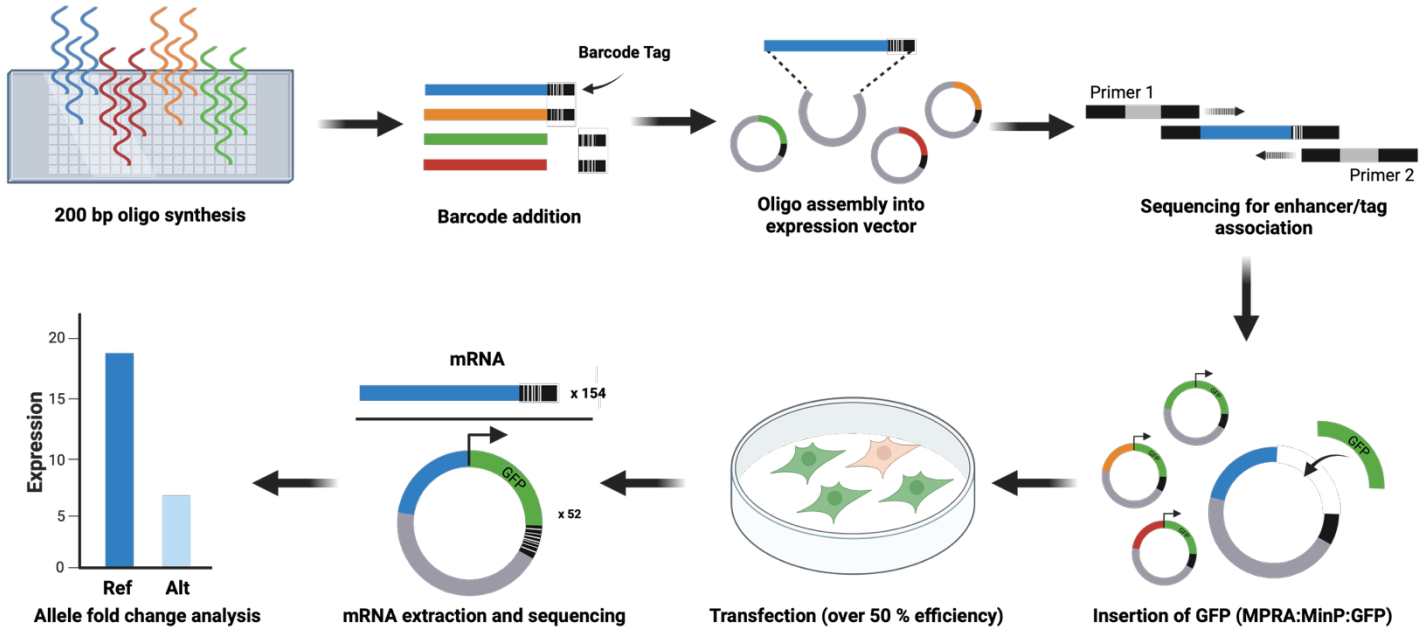
Supplementary Figure 6: Motif disruption analysis of variants with allele-specific binding. **A)** Binding motifs distribution of variants with allele-specific binding. X-axis represents the 40 bp window centered on the variant. **B)** Count of variants that directly create or disrupt TF binding motifs (motif dependent), adjacent to TF motif (motif independent). Variants creating binding motifs are represented in blue, whereas disrupting variants are represented in red.



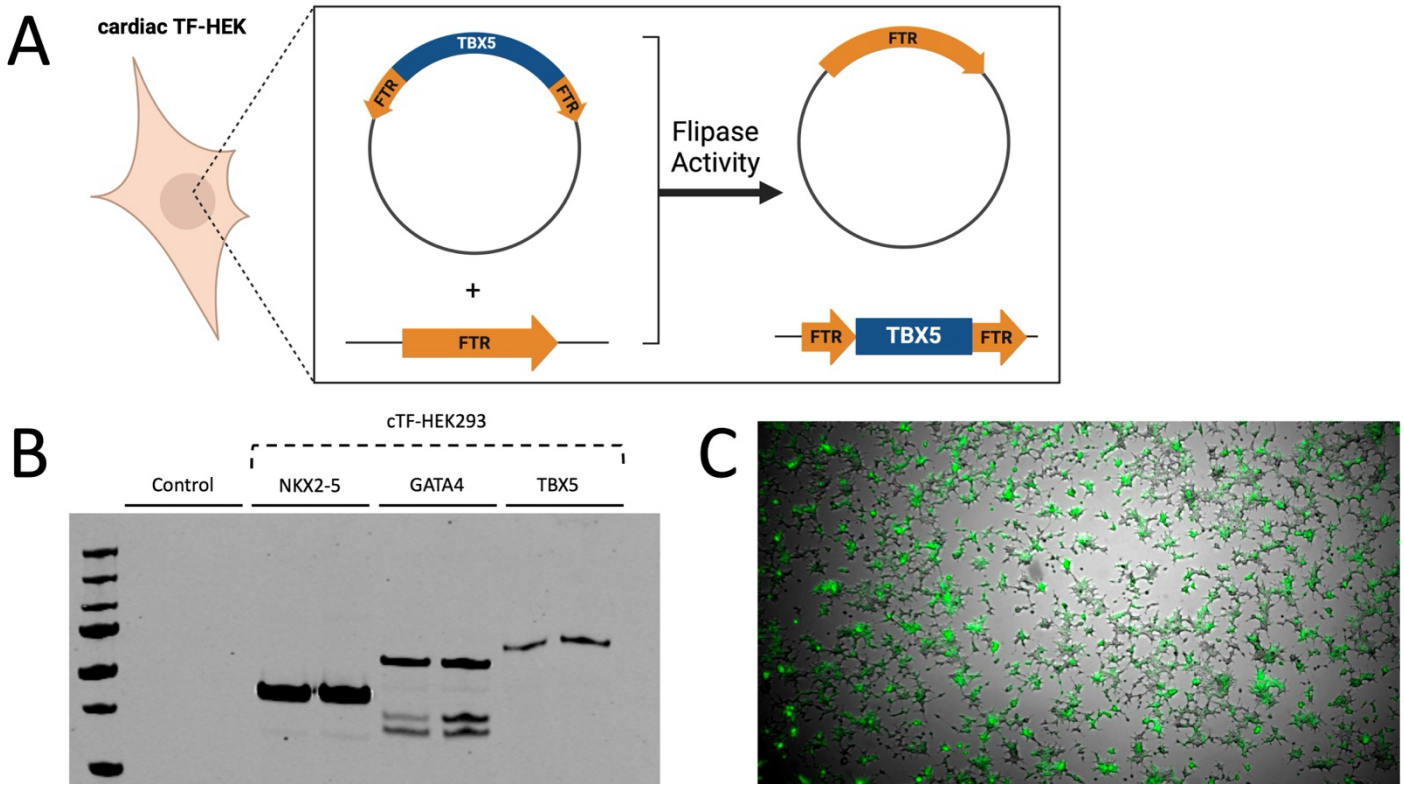
Supplementary Figure 7: Homer motif enrichment analysis of variants with allele-specific binding. **A)** Number of variants that created (orange) or disrupted (blue) TF binding motifs. **B)** Frequency of created and disrupted TF binding motifs relative to variant position (X = 0). **C-D**) Number of motif created or disrupted for **C)** TF families and **D)** specific TFs. Motif enrichment analysis are displayed separately for variants that increased or decreased binding affinity for NXK2-5, GATA4, and TBX5.



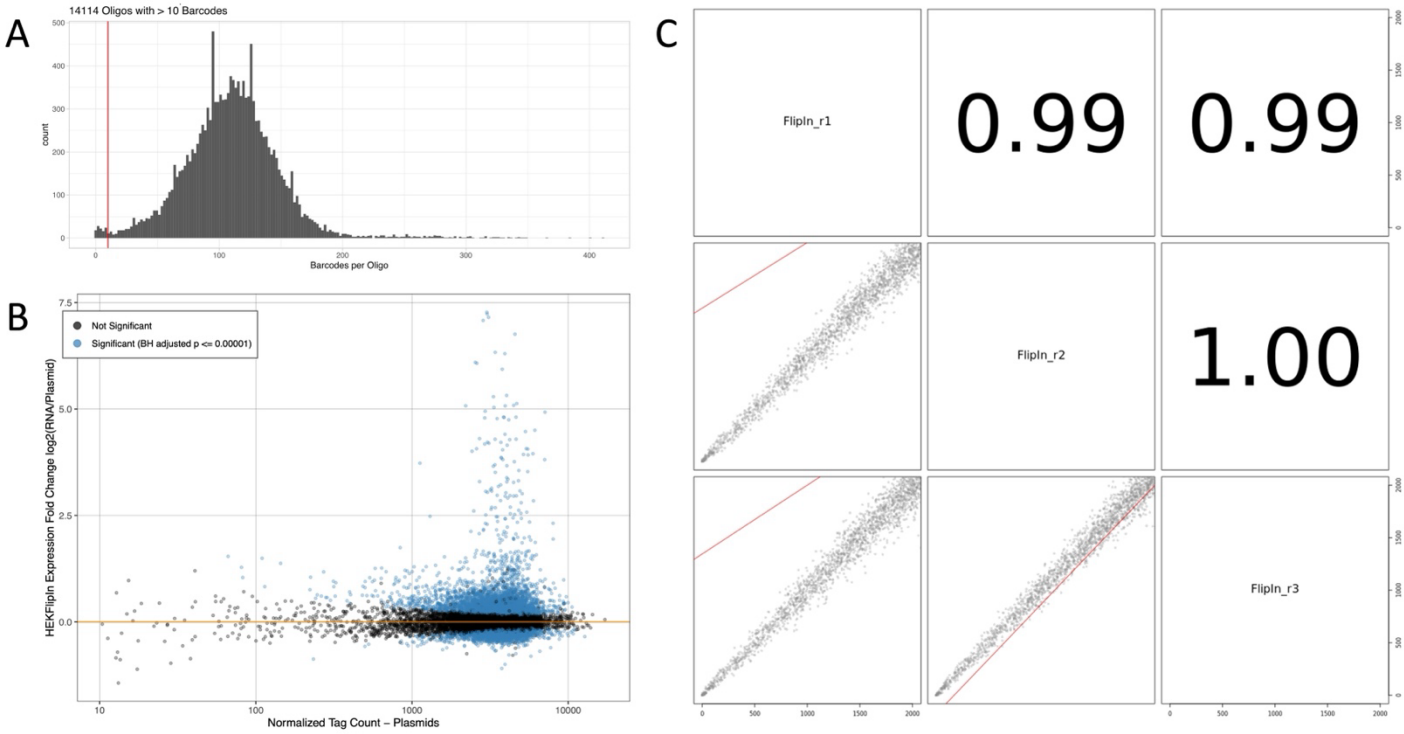
Supplementary Figure 8: Training predictive models with SNP Bind-n-Seq experimental data. **A)** Scatter plot comparing MinSeqChIP and LSGKM classifier performance for three transcription factors (NKX2-5, GATA4, TBX5), with AUROC scores shown on the x-axis (MinSeqChIP) and y-axis (LSGKM). Marker shapes indicate evaluation strategies: circles represent training and testing on SNPs, triangles indicate training on ChIP-seq and testing on SNPs, and squares indicate training on SNPs and testing on ChIP-seq. For each condition, two points are plotted because the same positive set was evaluated against two different negative sets: LSGKM-generated negatives and genomic negatives located 6,000 bp away from positives. In SNP→SNP evaluation, a 60–40 split is used for both classifiers, with 60% of the data (300 positives and 300 negatives) used for training and 40% (200 positives and 200 negatives) used for testing. In ChIP→SNP evaluation, training is performed on the top 1,000 ChIP-seq peaks and testing on 500 SNP positives (controls excluded). In SNP→ChIP evaluation, training is on 500 SNP positives (controls excluded) and testing on 1,000 ChIP-seq peaks. The dashed diagonal line indicates equal classifier performance, while marginal kernel density estimates summarize the overall distribution of AUROC scores. Points above the diagonal indicate superior LSGKM performance, whereas points below indicate better MinSeqChIP performance. **B)** Number of variants predicted to alter TF binding per disease parent term from the GWAS catalog. NKX2-5 is represented in yellow, GATA4 in red, and TBX5 in blue.



Supplementary Figure 9: MPRA workflow.



Supplementary Figure 10: Generating a cardiac TF stably-expressing HEK293 cell line. **A)** Diagram of the HEK293 FlpIn system to integrate the cardiac TF gene into the genomic landing pad. **B)** Confirmation of NXK2-5, GATA4, and TBX5 expression in HEK FlpIn cell line through Western Blot. **C)** GFP library mock transfection of HEK FlpIn.



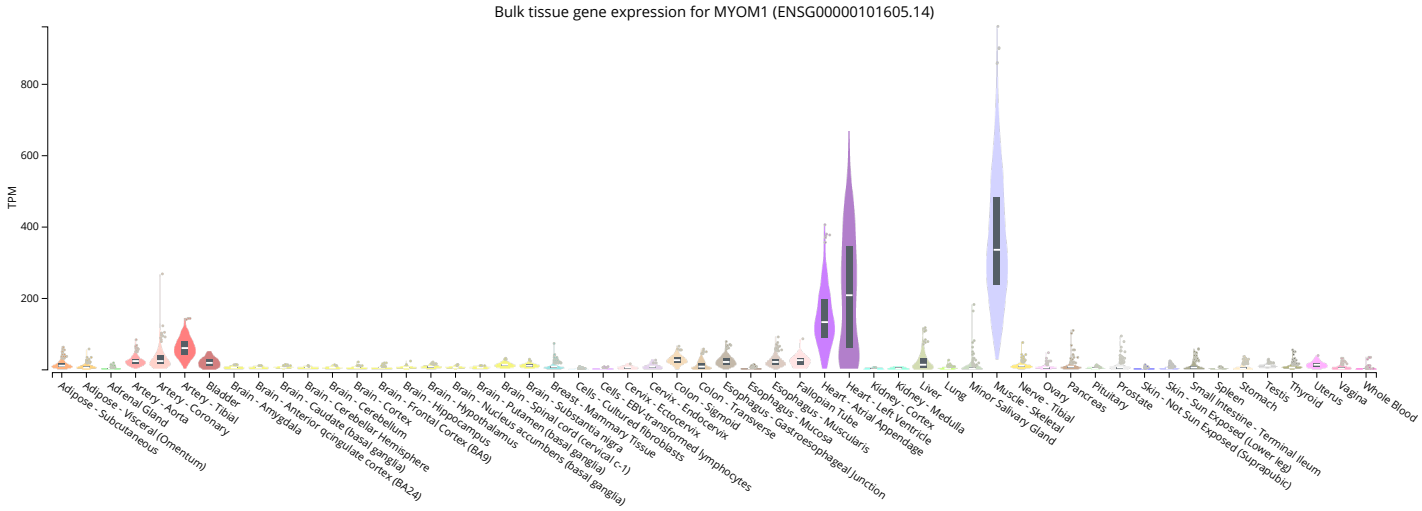
Supplementary Figure 11: MPRA quality and reproducibility analysis. **A)** Number of variants with >10 unique barcodes per oligo. **B)** Fold change of oligos compared to barcode tag counts. **C)** Correlation between biological triplicates.

	Overlap with cardiac enhancer	No overlap with cardiac enhancer	Total
enAllele	92	817	909
Non enAlleles (remaining MPRA library oligos)	295	13,057	13, 352
Total	387	13,874	14,261 (MPRA library)
Odds Ratio = 4.98		P-value = 2.10×10^{-29}	

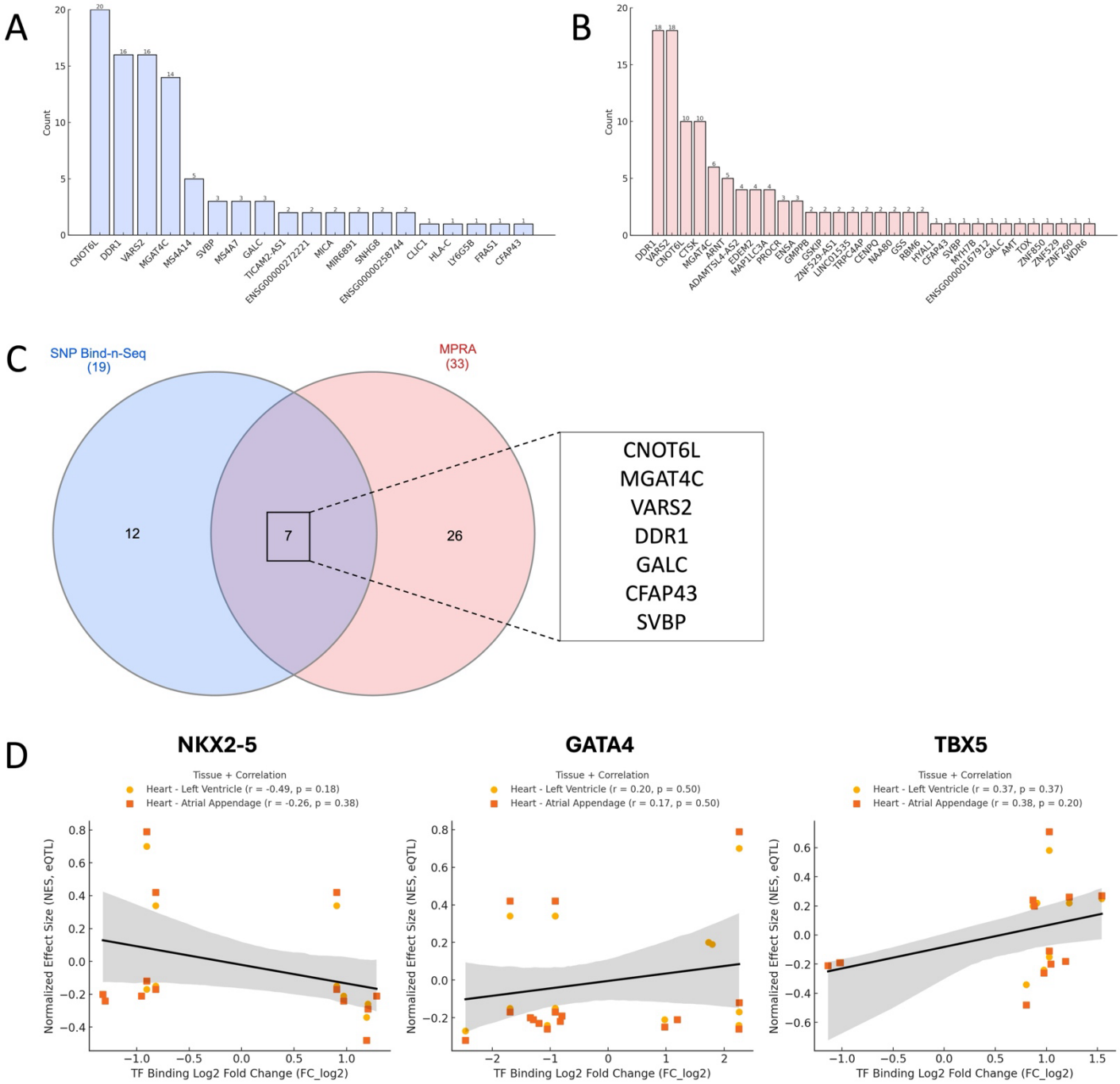
	Overlap with heart DGF	No overlap with heart DGF	Total
enAllele	90	819	909
Non enAlleles (remaining MPRA library oligos)	382	12,970	13, 352
Total	472	13,789	14,261 (MPRA library)
Odds Ratio = 3.73		P-value = 1.57×10^{-21}	

	Overlap with both cardiac CREs	No overlap with both cardiac CREs	Total
enAllele	38	871	909
Non enAlleles (remaining MPRA library oligos)	0	13,352	13, 352
Total	38	14,223	14,261 (MPRA library)
Odds Ratio = infinite		P-value = 1.61×10^{-23}	

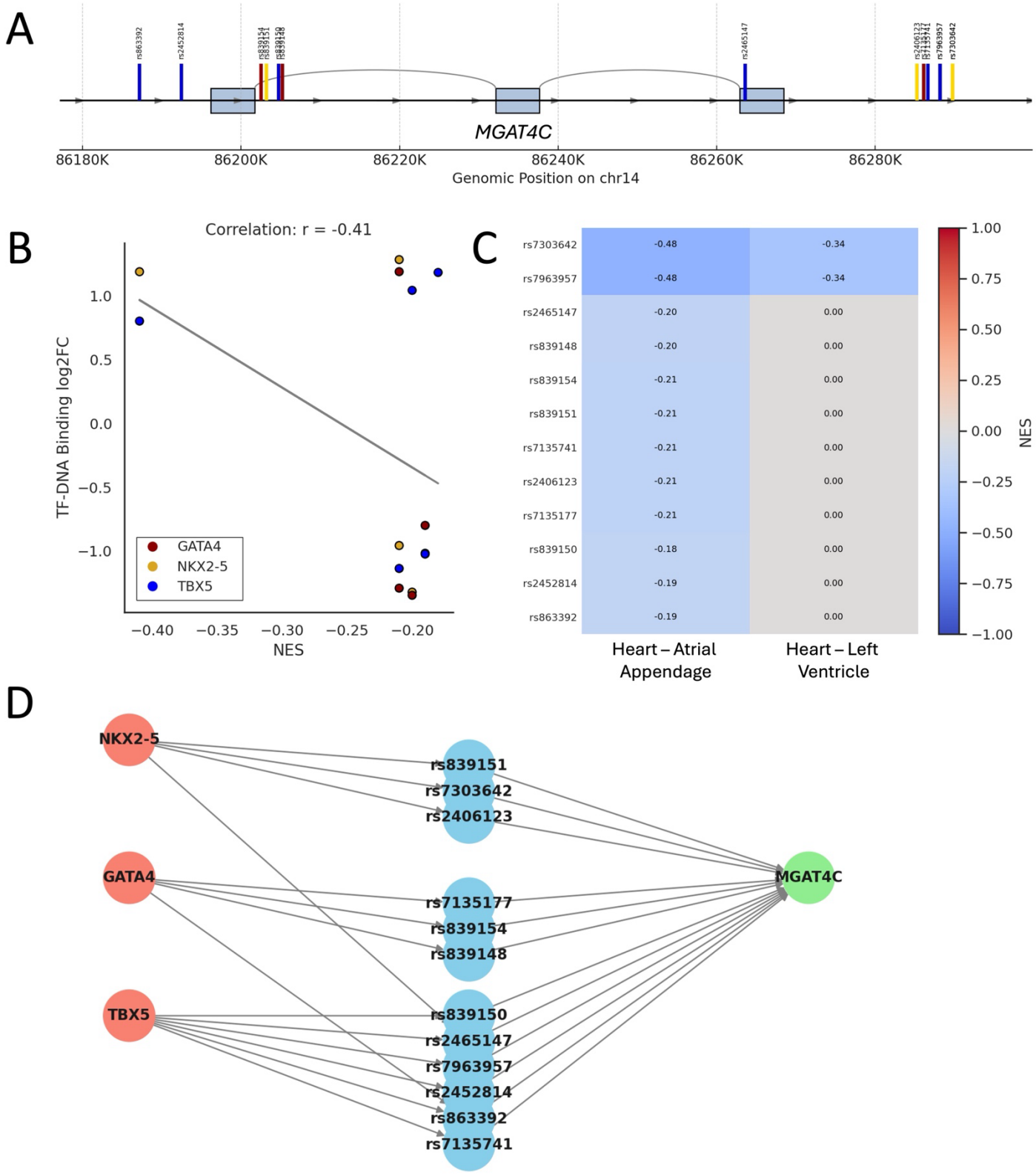
Supplementary Figure 12: Contingency tables to determine the significance of overlap between enhancer alleles (enAlleles) with cardiac regulatory elements.



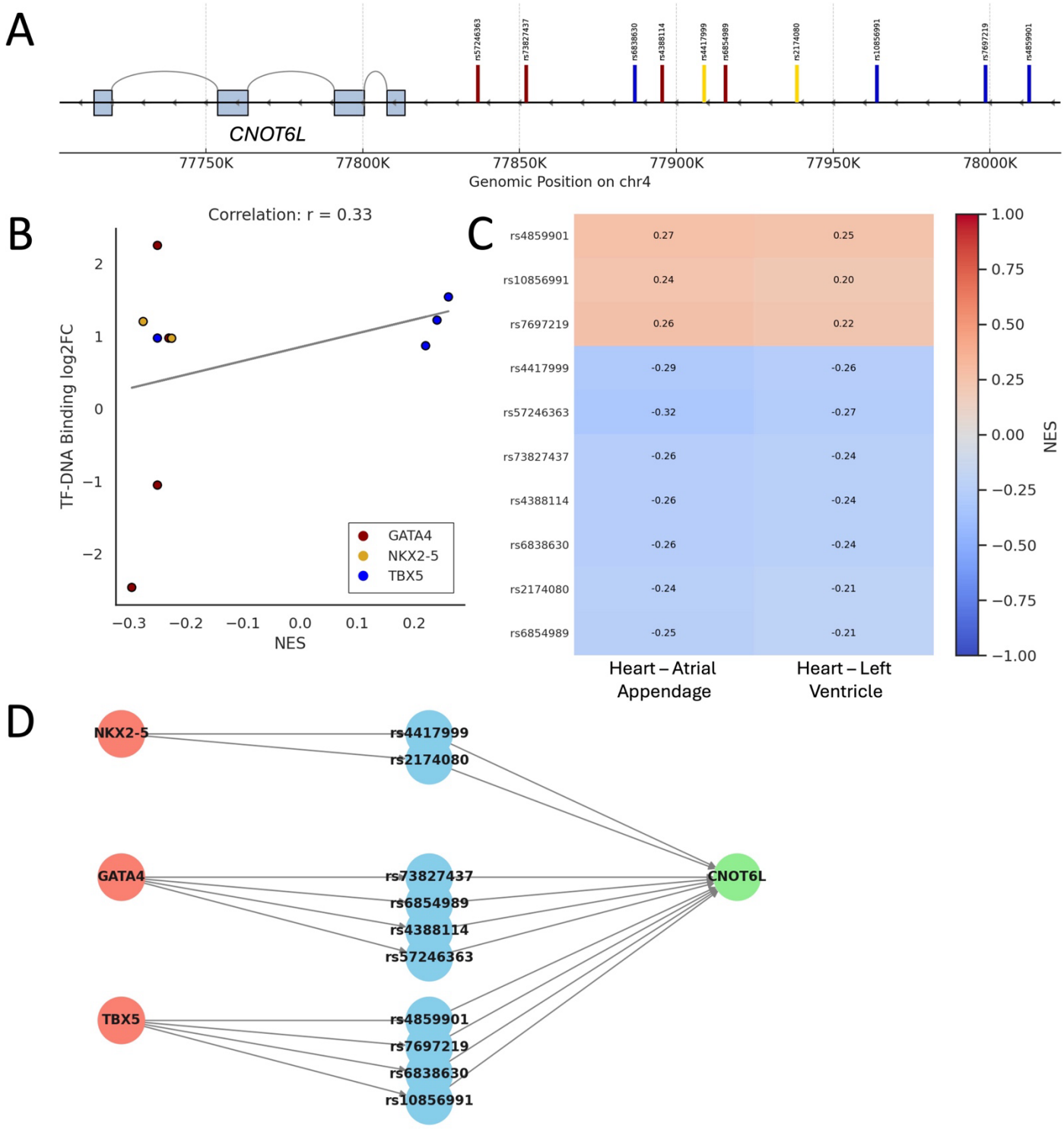
Supplementary Figure 13: MYOM1 expression profiles in multiple tissues. Cardiac tissues (the heart atrial appendage and left ventricle) are displayed in purple.



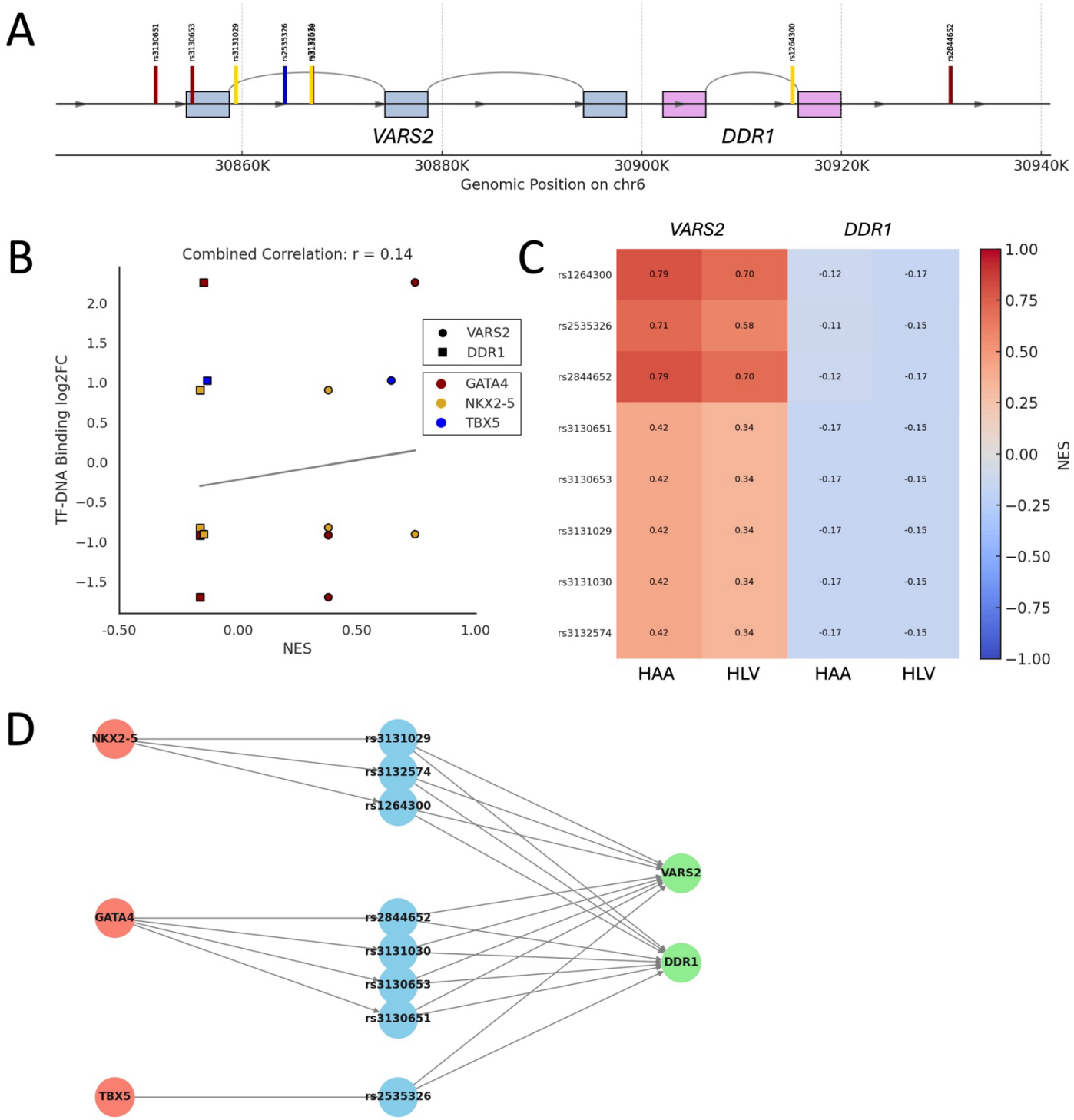
Supplementary Figure 15: Cardiac eQTL analysis of variants with genotype-dependent regulatory activity in SNP Bind-n-Seq and MPRA. **A-B)** List of genes (x-axis) and count of variants (y-axis) with cardiac eQTLS from the **A)** SNP Bind-n-Seq and **B)** MPRA experiments. **C)** Venn diagram of genes in cardiac eQTL genes from both experiments. **D**) Correlation analysis of TF binding fold change and eQTL normalized effect size.



Supplementary Figure 15: Cardiac eQTL analysis of MGAT4C. **A)** Genomic map of MGAT4C with variants in cardiac eQTL. Variants are displayed as colored lines if they altered NKX2-5 (yellow), GATA4 (red), and TBX5 (blue) binding. **B)** Correlation analysis of cardiac eQTL NES and TF binding fold change. **C)** Heatmap of NES of each variant per tissue. **D**) Interaction network of cardiac eQTL genes, variants, TF with altered binding.



Supplementary Figure 16: Cardiac eQTL analysis of CNOT6L. **A)** Genomic map of MGAT4C with variants in cardiac eQTL. Variants are displayed as colored lines if they altered NKX2-5 (yellow), GATA4 (red), and TBX5 (blue) binding. **B)** Correlation analysis of cardiac eQTL NES and TF binding fold change. **C)** Heatmap of NES of each variant per tissue. **D**) Interaction network of cardiac eQTL genes, variants, TF with altered binding.



Supplementary Figure 17: Cardiac eQTL analysis of VARS2 and DDR1. **A)** Genomic map of MGAT4C with variants in cardiac eQTL. Variants are displayed as colored lines if they altered NKX2-5 (yellow), GATA4 (red), and TBX5 (blue) binding. **B)** Correlation analysis of cardiac eQTL NES and TF binding fold change. **C)** Heatmap of NES of each variant per tissue. **D**) Interaction network of cardiac eQTL genes, variants, TF with altered binding.