Identification of Long Non-coding RNA Candidate Disease Genes Associated with Clinically Reported CNVs in Congenital Heart Disease

Jacqueline S. Penaloza[1,2], Blythe Moreland[2], Jeffrey B. Gaither[1], Benjamin J. Landis[3,4], Stephanie M. Ware[3,4], Kim L. McBride[5], and Peter White[1,2,6]* CCVM Consortium[†]

[1] The Office of Data Sciences, The Abigail Wexner Research Institute, Nationwide Children's Hospital, Columbus, Ohio, United States of America

[2] The Steve and Cindy Rasmussen Institute for Genomic Medicine, The Abigail Wexner Research Institute, Nationwide Children's Hospital, Columbus, Ohio, United States of America

[3] Department of Pediatrics, Indiana University School of Medicine, Indianapolis, Indiana, United States of America

[4] Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, Indiana, United States of America

[5] Department of Medical Genetics, Cumming School of Medicine, University of Calgary, Calgary, Canada

[6] Department of Pediatrics, The Ohio State University College of Medicine, Columbus, Ohio, United States of America

[†] Cytogenomics of Cardiovascular Malformations (CCVM) Consortium – see **Supplemental Methods** for the complete list of Contributing Authors.

Short title: CNVs impacting lncRNAs in congenital heart disease

***Corresponding Author:** Peter White, Ph.D., Battelle Endowed Chair for Quantitative and Computational Medicine, The Office of Data Sciences, The Abigail Wexner Research Institute at Nationwide Children's Hospital, 575 Children's Crossroad, Columbus, OH 43215. USA. Email: peter.white@nationwidechildrens.org

Word Count: 13,124

ABSTRACT

Background: Copy Number Variants (**CNVs**) contribute to 3-10% of isolated Congenital Heart Disease (**CHD**) cases, but their roles in disease pathogenesis are often unclear. Traditionally, diagnostics have focused on protein-coding genes, overlooking the pathogenic potential of non-coding regions constituting 99% of the genome. Long non-coding RNAs (**lncRNAs**) are increasingly recognized for their roles in development and disease.

Methods: In this study, we systematically analyzed candidate lncRNAs overlapping with clinically validated CNVs in 1,363 CHD patients from the Cytogenomics of Cardiovascular Malformations (**CCVM**) Consortium. We identified heart-expressed lncRNAs, constructed a gene regulatory network using Weighted Gene Co-expression Network Analysis (**WGCNA**), and identified gene modules significantly associated with heart development. Functional enrichment analyses and network visualizations were conducted to elucidate the roles of these lncRNAs in cardiac development and disease. The code is stably archived at https://doi.org/10.5281/zenodo.13799847.

Results: We identified 18 lncRNA candidate genes within modules significantly correlated with heart tissue, highlighting their potential involvement in CHD pathogenesis. Notably, lncRNAs such as *lnc-STK32C-3, lnc-TBX20-1*, and *CRMA* demonstrated strong associations with known CHD genes. Strikingly, while only 7.6% of known CHD genes were impacted by a CNV, 68.8% of the CNVs contained a lncRNA expressed in the heart.

Conclusions: Our findings highlight the critical yet underexplored role of lncRNAs in the genomics of CHD. By investigating CNV-associated lncRNAs, this study paves the way for deeper insights into the genetic basis of CHD by incorporating non-coding genomic regions. The research underscores the need for advanced annotation techniques and broader genetic database inclusion to fully capture the potential of lncRNAs in disease mechanisms. Overall, this work emphasizes the importance of the non-coding genome as a pivotal factor in CHD pathogenesis, potentially uncovering novel contributors to disease risk.

INTRODUCTION

Congenital Heart Disease (**CHD**) is one of the most common congenital malformations, with a birth prevalence of 1% worldwide.[1] In 60-75% of CHD cases, the genetic cause of the disease is unknown.[2] Copy Number Variants (**CNVs**), which are subchromosomal variations in the number of copies of DNA segments ranging in size from 1000 base pairs (**bp**) to several megabases (**Mb**), account for 3-10% of the pathogenic causes of isolated CHD.[2] Some recurrent CNVs are associated with a specific type of CHD, including 22q11.2 Deletion Syndrome (also known as DiGeorge syndrome, a ~3Mb microdeletion impacting *TBX1*, a T-box transcription factor required for normal development of the second heart field),[3] Williams-Beuren Syndrome (deletion of 7q11.23 that leads to the haploinsufficiency of nearly 25 genes, including the *ELN* gene that codes for elastin, important for blood vessel elasticity and associated with aortic and pulmonary stenosis)[4] and 1p36 Deletion Syndrome (the most common terminal deletion syndrome with cardiac findings that include congenital heart defects and cardiomyopathy)[5]. In addition to these clinical syndromes, CNVs are linked with a considerably greater risk of non-syndromic CHD.[6] Nonetheless, many distinct CNVs are observed in patients who have the same CHD diagnosis, and often their contribution to CHD pathogenesis remains unclear. The precise mechanisms by which CNVs lead to non-syndromic CHD are still largely unknown, necessitating further investigation into their functional impact and the underlying biological pathways involved.

When considering the molecular genetics of CNVs in CHD, disease etiology can be attributed to changes in the number of gene copies within the CNV. While the established diagnostic approach in CHD primarily emphasizes well-known protein-coding genes, as evidenced by resources such as the CHDgene website (a curated list of 142 high-confidence protein-coding CHD genes),[7] it is becoming increasingly apparent that variation in non-coding regions, which constitute about 99% of the genome, may contribute to the pathogenicity of CNVs.[8] Approximately 60-80% of the human genome can be transcribed into non-coding RNAs (**ncRNAs**), including microRNAs (**miRNAs**), long

non-coding RNAs (**lncRNAs**), small nucleolar RNAs (**snoRNAs**), and several other classes of RNA molecule that provide various regulatory and functional roles in the genome.

lncRNAs are RNA molecules longer than 200 nucleotides that usually do not have a coding sequence.[9–11] They are involved in diverse regulatory roles in gene expression, such as modifying chromatin, regulating transcription and post-transcriptional processes, and regulating RNA processing.[9,10] lncRNAs originate from various genomic regions, including intergenic, intronic, and overlapping regions with other genes.[11,12] At the DNA level, lncRNAs influence chromatin architecture by recruiting chromatin-activating complexes or looping factors, thereby regulating target genes.[9] They also play significant roles in alternative splicing by interacting with splicing factors to control transcription rates and ensure proper gene expression.[12] Post-transcriptionally, lncRNAs interact with RNA-binding proteins through sequence or structural motifs to regulate mRNA splicing. Furthermore, lncRNAs can act as miRNA sponges, binding to miRNAs and preventing their interaction with mRNAs, which is crucial for gene expression.[9] Lastly, lncRNAs regulate protein stability and degradation by altering their structure to interact with proteins involved in key signaling pathways. Together, these roles make lncRNAs vital for differentiation and development, impacting gene regulation at the DNA, RNA, and protein levels.

lncRNAs have been implicated in various biological processes and diseases. Their dysregulation can lead to a variety of conditions, including cardiovascular diseases,[13–15] neurodegenerative disorders,[16–19] and other complex diseases.[20–23] While the role of lncRNAs in CHD is emerging,[13–15] their function as key regulators of gene expression during heart development suggests that they could be critical players in the pathogenesis of CHD. The lncRNA *lnc-STK32C/RP11-432J24.5,* associated with responses to stress, stimuli, and the immune system, is differentially expressed in ischemic hearts.[14] *CRMA* (also known as *CARMA* for *CARdiomyocyte Maturation-Assoicated lncRNA*), previously known as *MIR1-1HG-AS1*, is an antisense lncRNA that regulates NOTCH1 signaling by targeting the miRNAs *MIR1-1*, *MIR1-133a2*, and other ncRNAs.[24] Notably, in

vitro knockdown of CRMA promotes cardiac commitment, suggesting that its downregulation is crucial for directing progenitor cells toward a cardiac fate. The intergenic lncRNA *BANCR* (*BRAF-Activated Non-Protein Coding RNAs*; HSALNG0071734)[25] has been reported to be upregulated in atherosclerotic plaques,[26] and may influence the development of cardiomyopathy.[13] This lncRNA encodes a short open reading frame that produces a microprotein that regulates various cellular functions, including contributing to heart muscle structural integrity and function. *BANCR* has been shown to facilitate cardiomyocyte migration and ventricular enlargement during heart formation in primates, with its expression confirmed to be higher during this critical period of heart development.[27,28] ClinVar, a database of genetic variants, contains several reports of pathogenic CNVs impacting the 9q21.11-21.13 region, which includes the gene *BANCR*.[29,30] However, these CNVs impact several genes in the region, and *BANCR* has not yet been directly associated with CHD.

Several CNVs impacting lncRNAs have been described in CHD patients. In DiGeorge syndrome, the 22q11.2 deletion spans the antisense lncRNA *lnc-TSSK2-8* (HSALNG0134007), which activates canonical Wnt signaling by protecting β-catenin from degradation.[31] Previously, only *TBX1* was thought to be the disease-causing gene in 22q11.2, but evidence suggests that the loss of *lnc-TSSK2-8* contributes to the development of CHD.[32,33] Another intergenic lncRNA associated with CHD is *lnc-NIPA1-4* (HSALNG0104472), located in the pathogenic CNV 15q11.2, where in vitro studies have shown that knockout of this gene significantly impacts cardiomyocyte differentiation.[34]

Despite their significance, the role of most lncRNAs in human health and disease remains largely unknown. This study aims to identify novel lncRNAs that may play a role in the pathogenesis of CHD by examining a large cohort of CHD patients with candidate lncRNAs within clinically validated CNVs (CNV-lncRNAs). By leveraging transcriptomic data from the developing heart and six other organs, we utilized a data mining method known as Weighted Gene Co-expression Network Analysis (**WGCNA**) to assess the expression patterns, interaction networks, and potential mechanisms by which CNV-lncRNAs may influence CHD. We identified several CNV-lncRNAs with

strong associations with key developmental pathways implicated in CHD. Furthermore, to promote transparency and reproducibility, we have made the software code and methodologies used in this analysis publicly available (https://github.com/nch-igm/lncCHDNet/). By providing our codebase as a reusable resource, we aim to contribute to a more comprehensive understanding of lncRNA function in CHD and pave the way for future discoveries in the field.

METHODS

Description and Structure of the CNV Dataset of CHD Patients

Figure 1 gives an overview of the process we followed to identify lncRNAs impacted by CNVs that may play a role in the etiology of CHD. To elucidate the potential role of copy number changes in lncRNA genes in abnormal heart development, we leveraged the recently published CHD cohort from the Cytogenomics of Cardiovascular Malformations (**CCVM**) Consortium.[35] This dataset contains 1,363 CHD patients with both an abnormal echocardiogram and an abnormal finding on chromosomal microarray analysis (**CMA**). CMA is a genome-wide technique that identifies genomic gains or losses (CNVs) and regions of homozygosity (**ROH**). It is recommended as a first-tier test for patients with neurodevelopmental disorders and congenital anomalies[36] and has been integrated into routine practice at many pediatric cardiac centers for infants with severe CHD[37]. Utilizing a centralized echocardiography review, the CCVM registry meticulously categorized each patient's CHD using a hierarchical classification to generalize the cardiac phenotypes.[35] Specific structural heart defects identified through clinical evaluations included, but were not limited to, septal defects (**SDs**), conotruncal heart defects (**CTDs**), and left and right-sided obstructive lesions. All patients had clinically reported pathogenic CNVs as assessed by American College of Medical Genetics and Genomics (**ACMG**) guidelines[38]. Both syndromic and isolated cases of CHD are present in this cohort to allow for comprehensive phenotypic and genotypic correlations.

Ethics approval and consent to participate

The CNV dataset applied in this study is available through the CCVM Consortium. The Consortium's collection of CMA and clinical data was approved by each clinical center's Institutional Review Board and utilized a waiver of informed consent. All data were anonymized before access and we followed general ethical guidelines for data use, including respecting privacy, data integrity, and intellectual property rights.

Preprocessing of the CHD Cohort

The CCVM dataset underwent rigorous data wrangling and filtering to focus the analysis on isolated cases of CHD with CNVs. This filtering considered the CHD classification frequency, CNV length, location, and type (deletion, duplication, or ROH). The original dataset includes CMA data for 1,363 patients with CHD. Of these, 386 patients were classified as having syndromic CHD, while 977 were categorized as non-syndromic (**Supplementary Figure 1**). For this study, 'syndromic' refers to patients with genomic disorders with well-established associations with congenital heart defects, based on consensus opinions within the CCVM Consortium. The list of disorders considered syndromic was curated by the consortium and may not fully encompass all syndromic presentations. Across the non-syndromic patients, a total of 2,416 CNVs were recorded, including heterozygous (n=901) and homozygous (n=16) deletions, duplications (n=770), and runs of homozygosity (ROH, n=729) (**Supplementary Figure 2**). While most patients had a single CNV (n=963), 300 had two, and 100 had three or more CNVs listed on their clinical test report.

The first stage of preprocessing the CHD cohort was to remove all CNV entries capturing ROH. This type of CNV was excluded from our analysis, as discerning the function of a lncRNA contained within these regions would be considerably more complex than events that resulted in a loss or gain of a lncRNA. For the remaining 1,687 CNVs (duplications and deletions), the dataset was next filtered based on the size of the CNV to exclude small CNVs less than 5,000 bp and large CNVs greater than

5,000,000 bp (5 Mb). The lower limit for CNV length is determined by the size detection capabilities of the microarray technology. In contrast, the upper limit of 5 Mb is a commonly used cutoff to exclude aneuploidies and other large chromosomal rearrangements.[35] The highly variable pseudoautosomal regions (PARs) on the X chromosome, which exhibit high rates of recombination and structural variation, were also removed to ensure more consistent and interpretable results in the analysis.

The final steps of preprocessing the data focused on the CHD category and defect classification. After removing ROH, small and large CNVs, and CNVs located in the PARs, 1,126 patients remained with 1,404 CNVs. Of these patients (821 non-syndromic; 305 syndromic), we removed all 305 syndromic patients and those with a diagnosis of cardiomyopathy (n=10) as this is primarily a disease of the heart muscle that may not directly relate to structural congenital defects. Patients with aortopathy (n=15) and arteriopathy (n=14) were removed due to the distinct pathophysiology of these conditions. Five other categories were removed due to their small sample size: CTD and atrioventricular septal defects (**CTD+AVSD**; n=15), Septal defect with Right Ventricular Outflow Tract Obstruction (**Septal RVOTO**; n=11), other (n=9), Patent Ductus Arteriosus (**PDA**; n=2), and Single Ventricle, other subtype (**SV,os**; n=2). The exclusion of these rare phenotypes enabled us to focus the analysis on the recurrent cardiac malformations that constitute the majority of phenotypes observed in this CHD cohort. The final dataset comprised 743 patients with 938 CNVs.

RNA-seq Dataset and Processing for CNV-lncRNA analysis

We utilized the human organ RNA-seq time series of the development of seven major organs (ArrayExpress Accession Number E-MTAB-6814).[39] The dataset covers 23 developmental stages across 7 organs (n=313: brain/forebrain 55, hindbrain/cerebellum 59, heart 50, kidney 40, liver 50, ovary 18, and testis 41), starting at four post-conception weeks (**PCW**) until 58-63 years of age. Normalized expression levels of lncRNAs calculated from this time series, expressed as transcripts

per million (**TPM**), were obtained from the publicly accessible LncExpDB database (available at lncRNA Expression Database).[40] For a description of how LncExpDB processed the transcriptomic data, see **Supplemental Methods**.

The RNA-seq data include gene expression levels for 101,293 lncRNAs and 19,957 protein-coding genes. When performing WGCNA, applying a minimum expression threshold of at least one transcript per million (**TPM**) is recommended, ensuring that only genes expressed in at least one sample are included in the analysis[34]. Applying this threshold to all samples in the developmental time series, 18,317 protein-coding genes had a minimum expression level of ≥1 TPM. For lncRNAs, we required a minimum expression level of ≥1 TPM in any of the 50 heart samples, as recommended for building the co-expression network. This filtering criterion identified 24,009 lncRNAs expressed in the heart at any developmental stage, along with corresponding expression levels for these genes across the other six organs.

Identifying the CNV-lncRNA

Next, we wanted to identify all lncRNAs that fell within the 938 CNVs in the 743 isolated CHD patients in the processed CCVM cohort. Given the CNV coordinates of the original data were mapped to GRCh37, a liftover was performed to map regions to reference genome build GRCh38. The lncExpDB RNA-seq data were generated using a custom LncBook gene transfer format (**GTF**) file (LncBookv1.9_GENCODEv33_GRCh38.gtf). As such, this same file was used to identify CNV-lncRNAs. Although protein-coding genes are present in the GTF file, only lncRNAs were mapped to the CNV coordinates. Coordinates of CNVs were intersected with the LncBook lncRNAs utilizing the R package GenomicsRanges (v1.56.1). A total of 15,261 unique lncRNAs fell within clinically reported CNVs in the CCVM cohort. The median number of lncRNAs per CNV was 10, with an IQR of 24, and a maximum of 356 lncRNAs in a single CNV (**Supplementary Figure 3** shows the distribution of lncRNAs within the CNVs). Of the original 743 patients, 720 had a CNV overlapping 1 or more lncRNAs.

The list of 24,009 lncRNAs expressed in the heart at any developmental stage was intersected with the 15,262 unique lncRNAs falling within a CNV in the CCVM cohort. This resulted in a final list of 3,608 CNV-lncRNAs with evidence of heart expression, along with 18,317 unique protein-coding genes expressed in any tissue at any development stage, that together can be further analyzed in the co-expression network (**Figure 2E**). The median number of expressed lncRNAs per CNV was 2, with an IQR of 6, and a maximum of 110 expressed lncRNAs in a single CNV (**Supplementary Figure 4** shows the distribution of expressed lncRNAs within the CNVs). Of the original 743 patients, 557 had a CNV overlapping 1 or more expressed lncRNAs.

Network Construction

Protein-coding genes are well studied, and their functions are better understood than lncRNAs. Therefore, we can leverage the expression pattern of protein-coding genes to infer the function of lncRNAs. The final gene set consisted of 21,925 total genes, of which 18,317 are protein-coding genes expressed in any tissue, and 3,608 unique lncRNAs expressed in the human heart and falling within clinically reported CNVs in the CCVM cohort. We next constructed gene regulatory networks using lncRNAs and protein-coding genes identified through Weighted Gene Co-expression Network Analysis (**WGCNA**).[41] WGCNA is a systems biology method that identifies clusters, or modules, of genes that show high levels of correlation in their expression patterns. These modules are then summarized using module eigengenes. A module eigengene captures the average expression profile of a gene module.[41] This allows the module to be associated with different tissues, and hub genes specifically related to the heart to be identified. Building on this framework, we implemented stepwise module detection to refine our network analysis further, focusing on CNV-lncRNA co-expression networks to identify novel lncRNA hub genes.

In contrast to the block-wise module detection approach taken by Lu *et al.*,[34] our study employs stepwise module detection for annotating CNV-lncRNAs. This approach offers greater

control and precision in network construction by allowing the merging of similar modules. In contrast, the block-wise method does not and places genes into independent blocks, even when there may be potential similarities.[41] For module detection, an appropriate soft-threshold power must be specified. WGCNA aims to create a network that follows a scale-free topology, characterized by a structure in which a few nodes are highly connected, and most nodes have low connectivity. We evaluated a range of powers, selecting a power value of 9 (R^2=0.76), to ensure an optimal balance between achieving a scale-free topology and maintaining network connectivity (**Supplemental Figure 5**).

First, the topological overlap matrix (**TOM**) similarity was calculated with an unsigned network (see **Supplemental Methods** for details). The unsigned network considers positive and negative gene correlations but treats them equivalently without distinguishing their specific contribution.[41,42] We chose an unsigned network to capture both activators and repressors of the heart developmental process. Second, network modules were detected by hierarchical clustering on the distance matrix derived from the TOM, specifically using $1 - TOM$ to measure gene-gene dissimilarity (**Supplemental Figure 6**), enabling genes to be grouped into co-expression modules based on their correlation patterns.[41] Each module is assigned a unique identifier in the form of a color, which provides a visual and intuitive way to distinguish between different modules. These colors are arbitrary and serve as labels, but they remain consistent throughout the analysis, making it easy to track modules across different plots and analyses. For example, genes assigned to the "magenta" module exhibit similar expression patterns, while those in the "green" module form a distinct co-expression group. This color-based labeling system facilitates the interpretation of network topology and gene clustering results. Third, each module was summarized by its eigengene. Hierarchical clustering of these eigengenes, using dissimilarity measures (1 – correlation), was used to refine module granularity and merge similar modules at a 0.1 dissimilarity threshold (**Supplemental Figure 7**).

Relate Modules to Organs and Developmental Stages

The final steps in our analysis focused on identifying which modules correlated most closely with the developing heart and identifying hub genes. First, the module-trait relationships were measured by calculating the Pearson correlation of the module eigengenes with the tissue type. To ensure robustness against Type I errors due to multiple comparisons across the seven organs, we applied a stringent p-value threshold of ≤ 0.001.[41] Modules with high correlations to particular tissues, such as the heart, were highlighted as tissue-specific.

In addition to using the Pearson correlations, we visualized these relationships through a dendrogram, traditionally used to cluster co-expressed modules based on their eigengenes. In this case, we treated tissue types as an additional external trait and extended the dendrogram to reflect the similarity of module eigengenes to these tissue traits (**Supplemental Figure 8**). To complement the dendrogram, we also used an adjacency heatmap, which provides a visual representation of the correlation strengths between modules and tissues (**Supplemental Figure 9**). We focused on identifying modules most strongly correlated with the heart (**Supplemental Figures 10-12**), focusing on the 50 heart samples to developmental stages (**Supplemental Figures 13-15**).

An enrichment analysis was conducted on each WGCNA module to assess the representation of lncRNAs, protein-coding genes, and CHD genes. Hypergeometric and Fisher's exact tests were employed to determine whether these gene types were significantly enriched or depleted within the modules. Fisher's exact test p-values were adjusted for multiple testing, and the results were classified as "Enriched," "Depleted," or "Not Significant" (see **Supplemental Methods** for details.)

Next, we performed gene ontology (**GO**) and pathway enrichment analysis on the protein-coding genes within each module to identify overrepresented gene sets. This analysis focused on three GO categories: biological processes (**BP**), molecular functions (**MF**), and cellular components (**CC**), adjusting p values for multiple testing (see **Supplemental Methods** for details).

Identification of Hub Genes

In WGCNA, a hub gene is a highly connected gene within a module, strongly correlated with the module eigengene, and often pivotal in the biological processes the module represents. Hub genes for modules of interest are identified by calculating the module memberships (**MM**) for all genes and selecting those with $R^2 \geq 0.80$.[43] High MM values signify a strong and specific connection between a gene and a particular module. Low MM values indicate a weaker connection, showing that the gene's expression pattern does not align closely with the module's expression. In addition to MM, we calculated the correlation between each gene's expression and the heart trait to determine gene-trait significance (**GS**) values. Here, we refer to the "heart trait" as a variable representing the expression data specific to heart samples across the dataset. GS represents how strongly each gene is associated with the heart trait, compared to other organs. Scatter plots were generated to visualize the relationship between MM and GS for each module, helping to identify hub genes with high MM and GS values. Gene aliases of the hub genes are identified by the LncExpDB database.[40] The classification of lncRNAs among the hub genes was obtained from LncBook (Version 2.00).[44] Additionally, LncExpDB was utilized to overview lncRNA tissue expression in organs not included in the developmental dataset.

To visualize the networks of correlated genes, the modules with significant association with the heart were exported to Cytoscape,[45] where nodes represent genes and edges represent gene co-expression derived from the TOM matrix. We applied an edge weight minimum of 0.3 to refine the network, removing isolated nodes and small sub-networks for clarity. We focused on networks containing known CHD genes within the module of interest. To identify whether a protein-coding gene was a CHD gene, we cross-referenced our gene list with 142 high-confidence CHD genes from the CHDgene database. The CHDgene website (https://chdgene.victorchang.edu.au) is a valuable resource for clinicians and researchers, offering a curated list of genes with variants consistently shown to cause CHD in humans.

Embryonic heart Gini coefficient scores were used to quantify tissue-specific gene expression during human heart development (see **Supplemental Methods** for details).[28] VanOudenhove *et al.* proposed that a higher Gini coefficient (e.g., >0.5) indicated a gene's expression was more specific to the embryonic heart, identifying genes enriched for heart-related functions like TBX5, IRX4, and HAND1. To integrate this approach into the current analysis, LncBook identifiers were mapped to Ensembl gene accessions to retrieve Gini coefficients for genes with maximal expression in the embryonic heart.

RESULTS

The CCVM cohort comprises 1,363 patients and includes both syndromic and non-syndromic CHD (**Supplemental Figure 1**).[35] We focused our analysis on patients with non-syndromic CHD, applying a filtering scheme that removed ROH, deletions less than 5 thousand base pairs (**Kb**), aneuploidies, and CNVs falling within the pseudoautosomal region (**Figure 2A**). The filtered dataset has 743 non-syndromic patients with a total of 938 CNVs larger than 5 Kb, but less than 5 million base pairs (**Mb**) (**Supplementary Figure 2**). Eight different subtypes of CHD were represented in the cohort: CTD, SD, left ventricular obstructive lesion (**LVOTO**), right ventricular obstructive lesion (**RVOTO**), SD + LVOTO, heterotaxy (**HTX**), atrioventricular septal defect (**AVSD**), and anomalous pulmonary venous return (**APVR**) (**Figure 2B**).

Identification of CNV-associated lncRNA genes (CNV-lncRNAs)

The CCVM CMA data was successfully lifted from GRCh37 to GRCh38 to annotate lncRNA and protein-coding genes within each CNV using a comprehensive set of over 100,000 putative lncRNA genes from LncExpDB.[40] Following the liftover process, some CNVs were split into multiple regions, resulting in a final dataset of 953 CNVs. The frequency of CNVs for the eight cardiac phenotypes is ranked from most to least in **Figure 2C**. There are 494 CNV gains (blue) and 459 losses (red). 573

patients had a single CNV, 140 had two CNVs, and 30 had three or more CNVs. The chromosomal distribution for gains and losses shows clinically reported CNVs in the CCVM non-syndromic CHD patients impact all chromosomes (**Figure 2D**). Overall, chromosome X had the most CNVs with 88 (60 gains and 28 losses). Chromosome Y had the least with 7 CNVs (5 gains and 2 losses). Only 72 of the 953 CNVs (7.6%) affected a known CHD gene. Of the 142 known CHD genes in the current CHDgene list, 33 (23%) were impacted by a CNV in our dataset. Notably, some of these known genes were recurrently affected, including *MYH11* (13 CNVs), *LZTR1* (8 CNVs), and *NOTCH1* (8 CNVs).

To identify CNV-associated lncRNAs (CNV-lncRNAs), we used RNA-seq data from a developmental time series of seven major human organs. For the WGCNA analysis, we first identified the subset of protein-coding genes expressed in any tissue across developmental stages. To focus on lncRNAs potentially involved in heart development and thus relevant to CHD, we restricted our analysis to lncRNAs expressed in the heart. We applied a minimum expression threshold of at least one TPM, resulting in 18,317 protein-coding genes and 57,627 lncRNAs that were expressed at any developmental time point across the seven different organs.

Among the expressed lncRNAs, 24,009 (42%) were expressed in the heart at some developmental stage. Of these heart-expressed lncRNAs, 3,608 (15%) were located within clinically reported CNVs from the CCVM cohort (**Figure 2E**). Out of the original 743 patients, 557 had CNVs overlapping one or more of these 3,608 heart-expressed lncRNAs, which we refer to as CNV-lncRNAs. These CNV-lncRNAs were used as input for the WGCNA analysis.

WGCNA-Based Identification and Functional Analysis of Heart-Specific Gene Modules

For the WGCNA network construction, the TOM adjacency matrix contained a pairwise comparison for 21,925 total genes: 18,317 protein-coding genes expressed in any tissue and 3,608 CNV-lncRNAs derived from the intersection of lncRNAs expressed in the heart and CNVs in the CCVM cohort. The dendrogram constructed from hierarchical clustering of the distance TOM matrix

contains information on the structure of gene expression correlations (**Supplementary Figure 6**). The hierarchical clustering constructs the modules for all genes. Each gene is tagged underneath by the color according to the module. For each module identified using the TOM-based clustering, eigengenes were calculated as the first principal component of the gene expression profiles, summarizing the overall expression pattern within each module. The hierarchical clustering initially resulted in 53 modules (**Supplementary Figure 7A**). Merging similar modules below the dissimilarity threshold of 0.1 led to 40 modules (**Supplementary Figure 7B**). All 40 modules and their associated counts of protein-coding genes, lncRNA genes, and known CHD genes are shown in **Supplemental Data Table S1**. Statistical analysis for gene enrichment of lncRNA genes and known CHD genes was performed using the hypergeometric distribution. Six modules (grey, magenta, paleturquoise, green, lightcyan1, and lightyellow) were enriched for lncRNAs. Four modules demonstrated enrichment of known CHD genes (darkturquoise, darkolivegreen, purple, and magenta). The magenta module (ME03) was the only module to display significant enrichment for both lncRNAs ($p = 8.09 \times 10^{-08}$) and known CHD genes ($p=0.0112$).

Pearson correlation values for the module eigengene-organ association across the 7 tissues for the 40 modules are depicted in the heatmap in **Figure 3** (modules and their associated correlations and p-values are available in **Supplemental Data Table S2**). The stronger the correlation, the higher the intensity of the color, with red representing positive correlations, blue representing negative correlations, and white representing no correlation. Modules with high positive or negative correlations ($|r| > 0.6$) and p-values below 0.001 are considered tissue-specific. For example, module ME40 (pink) is highly specific to the liver ($r = 0.96$, $p = 5 \times 10^{-174}$), while module ME28 (royalblue) is highly specific to the kidney ($r = 0.83$, $p = 3 \times 10^{-82}$).

Hierarchical clustering of the modules demonstrated that the magenta module clustered most closely to the heart samples (**Supplementary Figure 8**). The magenta module (ME03), comprising 567 genes, demonstrated significant heart tissue specificity, with a positive correlation

value of 0.95 and a p-value of 4×10^{-156}. This module showed no significant positive correlations in the other organs (**Supplemental Data Table S3**). Focusing on the heart samples, we compared each module's eigengene and developmental stage, using the age of the samples in post-conception (**Supplemental Data Table S4**). Modules with high positive correlations (in red) indicate increasing expression with developmental stage, while negative correlations (in blue) suggest decreasing expression over time (**Supplementary Figure 13**). The magenta module had a significant negative correlation ($r = -0.50$, $p = 0.0002$), clustering most closely with the mid-development (11–20 PCW) stage (**Supplementary Figure 14** and **Supplementary Figure 15**). In the magenta module, eigengene expression increased from conception, peaking around fetal mid-development (10-20 PCW) (**Figure 4**). Following this peak, expression levels gradually declined, reaching their lowest levels in the mature heart. CNVs impacting one or more of the 567 genes (425 protein-coding and 142 lncRNA genes) in the magenta module are found across all eight CHD phenotypes (**Supplemental Data Table S5**). In total, 154 (21%) of non-syndromic CHD patients had a CNV impacting the module (**Supplemental Data Table S6**).

The functional associations of the magenta module were identified using gene ontology terms to reinforce its connection to the heart. Functional enrichment of GO ontology terms was performed using the 425 protein-coding genes within the magenta module. The analysis focused solely on protein-coding genes because lncRNA genes generally lack well-defined functional annotations in existing gene ontology databases, making them less suitable for GO enrichment analysis. After combining semantically similar terms, 159 GO terms were functionally enriched (**Supplemental Data Table S7**), with the top ten for each category illustrated in **Supplementary Figure 16**. These significant GO terms predominantly relate to muscle structure and function, highlighting critical processes such as muscle cell development, differentiation, and assembly. The identified cellular components include key structural elements like the sarcomere, myofibril, and contractile fibers. Additionally, molecular functions such as actin-binding, structural constituents of muscle, and

electron transfer activity further emphasize the role of these genes in muscle contraction and energy transduction. Of the 159 enriched GO terms, 26 were related to the heart and the cardiovascular organ system (**Figure 5**). These terms included biological processes related to cardiac tissue muscle development ($p = 4.41 \times 10^{-29}$), heart contraction ($p = 1.44 \times 10^{-21}$), heart morphogenesis ($p = 2.20 \times 10^{-16}$), cardiac chamber development ($p = 2.79 \times 10^{-13}$), vasculogenesis ($p = 6.44 \times 10^{-05}$) and embryonic heart tube development ($p = 0.0037$).

Identification and Analysis of Hub Genes: Central Regulators of Heart-Specific Gene Networks

Genes with high module membership (**MM**), indicating strong connectivity within the module, and high gene significance (**GS**), indicating a strong association with the heart trait, were identified as hub genes (**Figure 6**). These genes are likely key drivers of the biological processes relevant to heart function and development. The correlation coefficients for the protein-coding genes (turquoise, r=0.968) and lncRNA genes (purple, r=0.972) both indicate strong positive correlation with p-value < 0.001. The correlation cutoff for module membership of $r \geq 0.80$ led to the identification of 146 hub genes (25.7% of the total 567 genes in the module) (**Supplemental Data Table S8**). Within the hub genes, 18 were lncRNAs (12.3%, p=0.9326). Strikingly, eight of the nine known CHD genes in the module were within the hub, representing a highly significant enrichment ($p=4.787 \times 10^{-06}$). The lncRNA and known CHD hub genes and their corresponding MM and GS values are shown in **Table 1**. Our analysis identified 18 lncRNAs as potential clinical targets from the CCVM cohort of clinically validated CNVs associated with non-syndromic CHD (**Table 2**).

The relationships within the magenta module were further explored through co-expression network visualization (**Figure 7**). To enhance the clarity of the network, we applied a stringent filter to the edges, retaining only those with a TOM value greater than 0.30. This filtering resulted in a network comprising 67 nodes with high module membership (MM.Cor ≥ 0.80) and strong co-expression interaction. Known CHD genes in the module were included in the network regardless of

their edge threshold interaction value. As a result, the network included 8 CHD genes, 7 lncRNA genes, and 52 protein-coding genes. Notably, all genes in this network displayed a co-expression interaction with *NKX2-5*, a critical regulator of heart development.

The Gini index of over 36,000 genes from 31 tissues, developed using independent human heart RNA-seq and GTEx datasets, was mapped to the hub genes.[28] The Gini coefficient with respect to the distribution of expression levels across samples identifies genes likely important in heart development by highlighting those with expression patterns specific to the heart during embryogenesis. VanOudenhove *et al.* proposed that candidate CHD genes have a Gini coefficient greater than 0.5 and, of the tissues studied, the gene must be maximally expressed in the embryonic heart. Within the hub genes identified in the magenta module, all eight known CHD genes and seven of the 18 lncRNAs genes had Gini scores from their study (**Table 3**). Four of the seven lncRNA candidates met the CHD gene candidate criteria: *lnc-TBX20-1* (Gini = 0.959), *lnc-STK32C-3* (Gini = 0.867), *CRMA* (Gini = 0.662), and *lnc-DOCK8-2* (Gini = 0.631). As was observed in the known CHD genes MYH6 and MYH7, the other candidate lncRNA genes had high expression in adult heart and muscle, or had a Gini score below 0.5, thus not meeting VanOudenhove *et al.'s* suggested criteria for a CHD candidate gene.

DISCUSSION

We applied WGCNA to cluster CNV-lncRNAs using a similar approach to that described for the heart and kidney.[34,46] CNV-associated lncRNAs were identified in clinically reported CNVs from a cohort of patients with isolated CHD. Utilizing human organ transcriptomic time-series data for the development of seven major organs, we clustered gene expression of both protein-coding and CNV-lncRNA genes into modules. These modules were then correlated with specific organs, allowing us to identify a module of co-expressed genes highly specific to the developing heart. We identified

eighteen CHD-candidate lncRNA genes within this heart-specific module (the magenta module, ME03). Seven of these eighteen lncRNAs had significant co-expression relationships with known CHD genes, and four were confirmed as candidate CHD genes in an orthogonal data set examining gene expression patterns specific to the heart during embryogenesis. Our study offers a well-documented and reproducible programming pipeline that applies a widely used method in a novel way to identify candidate lncRNAs within a disease-specific context and predict their functions. This approach empowers researchers to systematically study CNV-lncRNAs, providing a comprehensive toolset for future investigations. Moreover, our work expands the analysis of CNV-lncRNAs by utilizing a larger and more recent dataset of clinically confirmed CNVs from CHD patients in the CCVM dataset, further enhancing the scope and applicability of the research.

Clinical genetics in CHD primarily emphasizes well-known protein-coding genes, as evidenced by resources such as the high-confidence CHDgene list.[7] However, our analysis of the CCVM cohort revealed that only 72 of the 953 clinically reported CNVs (7.6%) affected one of these known CHD genes, suggesting that many potential pathogenic factors remain unidentified. By contrast, 656 of the 953 clinically reported CNVs (68.8%) contain one or more lncRNAs expressed in the heart, i.e., CNV-lncRNAs, underscoring the need to consider the non-coding genome as a crucial factor in the genetic etiology of CHD.

Through WGCNA, we identified a highly heart-correlated module (the magenta module, ME03, **Figure 3**) consisting of 567 genes with significant heart tissue specificity, clustering most closely with the mid-development stage (11–20 PCW). No significant positive correlations for this module were observed in other organs. CNVs affecting these genes were present across all eight CHD phenotypes, with 21% of non-syndromic CHD patients carrying a CNV impacting the module. In this module, eigengene expression increased steadily from conception, peaking during fetal mid-development (10–20 PCW) (**Figure 4**). After this peak, expression levels gradually declined, reaching their lowest point in the mature heart. This expression pattern suggests that the genes within the

magenta module are most active during early to mid-heart development, a critical period for structural refinement, including chamber septation, conduction system development, and the initiation of blood circulation. Functional enrichment analysis of the protein-coding genes in this module identified multiple enriched GO terms directly linked to heart and cardiovascular development, highlighting processes such as cardiac muscle development, heart contraction, and morphogenesis (**Figure 5**).

Additionally, the analysis revealed significant enrichment of genes involved in mitochondrial function, energy production, and metabolic processes, underscoring the critical role of cellular energy dynamics in heart development (**Supplemental Data Table 7** and **Supplemental Figure 16**). Given the essential role of mitochondria in providing the energy required for cardiac muscle contraction and overall cardiovascular function, this enrichment of mitochondrial and metabolic genes suggests that dysregulation in energy metabolism may contribute to CHD pathogenesis. Overall, the developmental gene expression pattern and the identified gene ontology terms suggest that genes in this module likely play pivotal roles in heart development and may serve as candidate genes for CHD.

Within the module of genes highly correlated with the developing heart, we identified eighteen highly connected lncRNA hub genes strongly correlated with the module eigengene (**Figure 6**). These genes are likely key drivers of the biological processes relevant to heart function and development. However, the lack of Ensembl accession numbers for many of our candidate lncRNAs limited our search for external information. Notably, only nine of these lncRNAs possess recognized Ensembl identifiers; the other nine were putative lncRNAs identified by LncBook. This highlights significant constraints in current gene annotation methods, which are often too stringent for ncRNA. For instance, Ensembl's criteria exclude intergenic lncRNAs with open reading frames that exceed 35% of their length or contain protein domains.[47] Consequently, our analysis and subsequent discussions were confined to those lncRNA candidates with Ensembl identifiers, limiting the full

exploration of the potential roles of the remaining putative lncRNAs in heart development and emphasizing the need for more inclusive gene annotation frameworks for non-coding RNAs.

Seven of the eighteen CNV-lncRNA hub genes were co-expressed with *NKX2-5*, a well-known transcription factor critical for early heart development: HSALNG0057201 (Lnc-TBX20-1), HSALNG0057200 (ENSG00000289335), HSALNG0055689, HSALNG0092764 (Lnc-CCDC59-5), HSALNG0096048 (ENSG00000256672), HSALNG0047724 (Lnc-ECI2-2), and HSALNG0131540 (MIR1-1HG) (**Figure 7A**). *NKX2-5* is expressed in the first and second heart fields to maintain chamber-specific identities.[48] Pathogenic variants in *NKX2-5* are associated with a variety of CHDs, including atrial[49] and ventricular SDs[50], AVSDs[51], and others[52,53]. Notably, HSALNG0131540 or *MIR1-1HG* was also significantly co-expressed with well-established CHD genes, *TBX5*[54,55], *MYH6*[56,57], and *MYH7*[58,59] (**Figure 7B**). The co-expression of these lncRNAs with *NKX2-5* and other known CHD genes underscores their potential roles in heart development and cardiac function. This suggests that they may be involved in similar developmental pathways and could contribute to the pathogenesis of congenital heart defects.

Four of our CNV-lncRNA hub genes—*lnc-TBX20-1*, *lnc-STK32C-3*, *CRMA*, and *lnc-DOCK8-2*—were also highly ranked as candidate CHD genes in an external validation (**Table 3**). These genes met the criteria of having the highest expression in the embryonic heart across 32 different organs and tissues, along with a significant Gini index, indicating strong heart-specificity.[28] Established CHD genes that meet the Gini index criteria include *TBX20* (Gini=0.948) and *NKX2-5* (Gini=0.935). Unexpectedly, the three known CHD genes—*MYBPC3*, *MYH6*, and *MYH7*—were excluded from VanOudenhove's candidate gene set because their expression extends beyond the embryonic heart into the adult heart and muscle. Despite the Gini index limitations to score all known CHD genes, this index did confirm *lnc-TBX20-1*, *lnc-STK32-C*, *CRMA*, and *lnc-DOCK8-2*—as potential contributors to heart-specific gene regulation and strengthens their candidacy for further investigation in CHD research.

The intergenic lncRNA *lnc-TBX20-1* (HSALNG0057201) was particularly interesting due to its high Gini index and proximity to the known CHD gene *TBX20*[60]. Of the known CHD genes and CNV-lncRNAs, *lnc-TBX20-1* had the highest Gini index (0.959) and peak expression in the embryonic heart, making it a strong CHD candidate lncRNA. The CNV associated with *lnc-TBX20-1* is a 350 Kb loss of 7p14.2 in a patient with CTD. This region contains the known CHD gene, *TBX20*, and two other lncRNA candidates, HSALNG0057200 and HSALNG0057214. The location of lncRNAs is important as lncRNAs are known to regulate nearby genes.[10] While it is not possible to determine whether the effects of this pathogenic CNV are due to the loss of *TBX20* or the lncRNA genes, the co-expression of *lnc-TBX20-1* and HSALNG0057200 with *NKX2-5* is particularly intriguing. *NKX2-5* and *TBX20* are transcription factors that interact in heart development regulation.[61] This raises the possibility that these lncRNAs may have regulatory roles in connection with *TBX20* and potentially *NKX2-5*, though experimental evidence is still needed to support this hypothesis.

lncRNAs function in processes ranging from chromatin modification to nuclear organization and are emerging as critical players in cardiovascular development and disease. However, their complexity often extends beyond the traditional definitions of non-coding genes, as exemplified by *lnc-STK32C-3*'s (HSALNG0081800) coding potential.[13] The CNVs associated with *lnc-STK32C-3* are two separate 2.9 MB deletions of 10q26.3 observed in two CCVM patients with RVOTO. We identified this CNV-lncRNA as a candidate CHD gene with the highest module membership in the magenta module, indicating specificity in cardiac development. This gene was also indicated as a candidate CHD gene by the Gini index with highly significant Gini index of 0.867, and is most highly expressed in the embryonic heart across the 32 organs and tissues transcriptionally profiled.[28] We are the first to suggest that this lncRNA is involved in the initial formation of the heart. Previous research suggests that *lnc-STK32C-3* is abundantly expressed in ischemic hearts and it is suggested to be involved in myocardial injury.[14] This aligns with the broader understanding that genes can turn on and off in response to developmental cues and cardiac injury, as shown by the protein-coding gene *VEGF-B*.[62]

Therefore, if protein-coding genes can have dual roles under different conditions expressed by the heart, it is plausible that lncRNAs can, too. Lastly, while it is known that *lnc-STK32C-3* encodes microproteins and secretes them in the adult heart,[27] whether it can also perform this function during cardiac development remains to be explored.

CRMA is a known regulator of cardiac differentiation and has been experimentally validated as important for heart development.[24,63] Variants affecting *CRMA* are therefore likely pathogenic and could help explain the CHD phenotype. The CNV associated with *CRMA* is a 2.9 MB gain in the 20q13.3 region in a patient with SD and LVOTO. In addition to *CRMA*, this CNV impacts the known CHD gene *GATA5*[64,65] and the lncRNA *MIR1-1HG*[24]. *MIR1-1HG* is the precursor of two cardiogenic miRNAs, known as *MIR1-1* and *MIR-133a2*. *GATA5* has only been implicated in CHD as a loss-of-function mutation in humans,[64] whereas in our cohort patients with 20q13.3 is a gain-of-function mutation. When CRMA is expressed, it downregulates *MIR1-1* and *MIR133a2,* pushing embryonic stem cells into a mature cardiac fate.[24,63] While the knockdown of *CRMA* promotes cardiac differentiation through increased expression of MIR-133a2, MIR-133a1 targets RBPJ to inhibit the NOTCH pathway.[24] We hypothesize that in a gain-of-function mutation CRMA may suppress the cardiac miRNAs expression beyond normal levels, leading to dysregulation of the NOTCH pathway causing cardiac malformations. Together, these networks of genes are required for the coordinated regulation of cardiogenic differentiation in embryonic stem cells.

Three CNVs impacting 9p24.3 were associated with *lnc-DOCK8-2:* a 305 Kb duplication in a patient with RVOTO, a 354 Kb duplication in a patient with CTD, and a 100 Kb deletion in a patient with LVOTO. While the CNVs do not impact any known CHD genes, in addition to the lncRNA, these smaller CNVs impact two protein-coding genes, *DOCK8* and *KANK1*. *DOCK8* is a known Mendelian disease gene, resulting in a rare autosomal recessive immunologic disorder characterized by recurrent staphylococcal infections of the skin and respiratory tract.[66] *KANK1* is an essential regulator of the actin cytoskeleton, similar to Rho GTPases, which also regulate actin dynamics

through Rho-kinase (ROCK), and is required for myogenic differentiation.[67] Rho GTPase signaling is involved in cardiac looping and early development of the heart, and dysregulation of this pathway may lead to structural heart defects.[68] Whether *lnc-DOCK8-2* regulates either of these genes is unknown, and like the majority of CNV-lncRNAs that we identified, further studies are required.

Reproducible bioinformatics methods and well-documented software are still significant barriers to advancing biomedical research, often hindering the ability to replicate and build upon existing studies. To address this challenge, we developed comprehensive and well-documented R notebooks that streamline our analytical workflow and fully automate the generation of all tables and figures, enabling easy replication and modification of the entire pipeline. This approach provides a robust and adaptable framework standardizing co-expression analysis to CNV-lncRNAs that can be applied to other genomic datasets, facilitating broader investigations into the roles of lncRNAs in tissue-specific and disease-specific contexts. As annotations of non-coding RNAs continue to improve and new genomic studies of CHD patients emerge, our methods can be reused and extended to uncover additional candidate genes and pathways involved in CHD.

Conclusion

In this study, we identified and characterized previously unrecognized lncRNAs associated with well-established protein-coding CHD genes, confirming prior findings while revealing new potential links, such as the involvement of lnc-DOCK8-2 and HSALNG0096048 in our CHD cohort. By applying the advanced bioinformatics approach of WGCNA, we uncovered insights into the complexities of CNV-associated lncRNAs within a CHD context. These findings emphasize the potential pathogenic roles of non-coding regions, which are often overlooked by traditional genetic studies. Our work highlights the significant, yet underexplored, contribution of lncRNAs to the genomics of CHD, opening new pathways for diagnosing and understanding the genetic basis of CHD.

In addition to these discoveries, we have made our fully documented R code publicly accessible, enabling other researchers to replicate and build upon our findings. This transparency facilitates further investigation into the uncharted territory of lncRNAs, their critical roles in heart development, and their probable link to CHD. As we continue to advance our understanding of non-coding RNAs and their role in CHD, we anticipate that this research will ultimately lead to better diagnostic tools and improved outcomes for individuals affected by CHD. As genome sequencing is incorporated into the standard of care for patients with critical CHDs, we encourage clinical labs to consider reporting CNVs and single nucleotide variants impacting lncRNAs.

Acknowledgments

We thank the Nationwide Children's Hospital Foundation and The Abigail Wexner Research Institute at Nationwide Children's Hospital for generously supporting this work. These funding bodies had no role in the study's design, collection, analysis, and interpretation of data, nor in writing the manuscript. Thank you to Thomas Hohmann, who contributed to data analysis early on in the project, and Dhruv Prasad, who tested the reproducibility of the CNV-lncRNA pipeline.

Source of Funding

Research reported in this publication was supported by the American Heart Association Transformational Award AHA 19TPA34850054 (SMW) and the National Heart, Lung, And Blood Institute of the National Institutes of Health under Award Number F31HL168950 (JSP). The content is solely the authors' responsibility and does not necessarily represent the official views of these funding agencies.

Disclosures

The authors declare no conflicts of interest.

Ethics Approval and Consent to Participate

This study exclusively employed publicly available deidentified datasets and thus did not involve direct human subjects research. For the CCVM data, collection of the clinical CMA reports utilized a waiver of informed consent approved by each clinical center's Institutional Review Board. In conducting this analysis, we strictly adhered to the datasets' terms of use, access, and distribution as outlined by their respective sources. We ensured that our research methods and objectives were aligned with the ethical guidelines for research and data use, including respecting privacy, intellectual property rights, and the integrity of the data.

Data Availability

The CCVM dataset, used for identifying CNV-lncRNA interactions, is accessible through the CCVM Consortium. The RNA-seq data, including expression levels of lncRNAs, is available from LncExpDB at the lncRNA Expression Database (https://ngdc.cncb.ac.cn/lncexpdb/). The sample metadata was downloaded from Array Express (https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-6814). The LncBook GTF file, used to map CNV-lncRNAs to the GRCh38 genome build and quantify gene expression, is available at LncBook (http://bigd.big.ac.cn/lncbook).

The CHDgene database, which provided a curated list of genes associated with CHD, is available at https://chdgene.victorchang.edu.au. The Gini Index data used for tissue-specificity analysis, published by the Cotney Lab, is available at https://cotneylab.cam.uchc.edu.

The source code used for CNV-lncRNA mapping, preprocessing, and network construction is available under an open-source license on GitHub (https://github.com/nch-igm/lncCHDNet/). This ensures that other researchers can freely access, modify, and implement the code in their respective analyses. The code is also stably archived on Zenodo at https://doi.org/10.5281/zenodo.13799847 to ensure long-term availability and reproducibility.

SUPPLEMENTAL MATERIAL

Supplemental Methods

Supplemental figures:

- **Supplemental Figure 1:** Counts of Syndromic and Non-syndromic Patients by Diagnosis.
- **Supplemental Figure 2:** CNV Size Distributions Across Isolated CHD Patients.
- **Supplemental Figure 3:** Distribution of lncRNA Counts Within CNVs.
- **Supplemental Figure 4:** Distribution of Expressed lncRNA Counts Within CNVs.
- **Supplemental Figure 5:** Selection of Soft-Thresholding Power for WGCNA Network Construction.
- **Supplemental Figure 6:** Gene Dendrogram and Module Assignment Based on Dynamic Tree Cut.
- **Supplemental Figure 7:** Hierarchical Clustering of Module Eigengenes Before and After Merging.
- **Supplemental Figure 8:** Module Eigengene Dendrogram for Organ-Specific Modules.
- **Supplemental Figure 9:** Module Eigengene Adjacency Heatmap for Organ-Specific Modules.
- **Supplemental Figure 10:** Correlation Between Module Eigengenes and Heart Trait.
- **Supplemental Figure 11:** Module Eigengene Dendrogram for Heart-Specific Correlations.
- **Supplemental Figure 12:** Module Eigengene Adjacency Heatmap for Heart-Specific Correlations.

- **Supplemental Figure 13:** Correlation Between Module Eigengenes and Age in Heart Samples.
- **Supplemental Figure 14:** Module Eigengene Dendrogram for Heart Developmental Stages.
- **Supplemental Figure 15:** Module Eigengene Adjacency Heatmap for Heart Developmental Stages.
- **Supplemental Figure 16:** GO Functional Enrichment Analysis for Protein-Coding Genes in the Magenta Module.

Supplemental data tables:

- **Supplemental Data Table 1:** WGCNA modules and their associated counts of protein-coding genes, lncRNA genes, and known CHD genes.
- **Supplemental Data Table 2:** Pearson correlation values for the module eigengene-organ association across the 7 tissues for the 40 modules.
- **Supplemental Data Table 3:** Pearson correlation of modules correlated with the heart trait.
- **Supplemental Data Table 4:** Pearson correlation of modules correlated with heart developmental stage.
- **Supplemental Data Table 5:** Gene-level data, highlighting module membership (MM) and gene significance (GS).
- **Supplemental Data Table 6:** Representation of CNVs by patient and diagnosis within each module.
- **Supplemental Data Table 7:** Gene Ontology (GO) enrichment results for the magenta module (ME03).
- **Supplemental Data Table 8:** Identification of hub genes impacted by CNVs in the magenta module (ME03).

REFERENCES

1. Bouma BJ, Mulder BJM. Changing Landscape of Congenital Heart Disease. *Circulation Research*. 2017;120(6):908–922.
2. Pierpont ME, Brueckner M, Chung WK, Garg V, Lacro RV, McGuire AL, Mital S, Priest JR, Pu WT, Roberts A, Ware SM, Gelb BD, Russell MW, American Heart Association Council on Cardiovascular Disease in the Young; Council on Cardiovascular and Stroke Nursing; and Council on Genomic and Precision Medicine. Genetic Basis for Congenital Heart Disease: Revisited: A Scientific Statement From the American Heart Association. *Circulation*. 2018;138(21):e653–e711.
3. Goldmuntz E. 22q11.2 deletion syndrome and congenital heart disease. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*. 2020;184(1):64–72.
4. Yuan S-M. Congenital heart defects in Williams syndrome. *The Turkish Journal of Pediatrics*. 2017;59(3):225–232.
5. Kaur T, Sriram CS, Prasanna P, Kohli U. Cardiovascular Phenotypic Spectrum of 1p36 Deletion Syndrome. *Journal of Pediatric Genetics*. 2023;12(4):329–334.
6. Southard AE, Edelmann LJ, Gelb BD. Role of copy number variants in structural birth defects. *Pediatrics*. 2012;129(4):755–763.
7. Yang A, Alankarage D, Cuny H, Ip EKK, Almog M, Lu J, Das D, Enriquez A, Szot JO, Humphreys DT, Blue GM, Ho JWK, Winlaw DS, Dunwoodie SL, Giannoulatou E. CHDgene: A Curated Database for Congenital Heart Disease Genes. *Circulation. Genomic and Precision Medicine*. 2022;15(3):e003539.
8. ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*. 2007;447(7146):799–816.
9. Statello L, Guo C-J, Chen L-L, Huarte M. Gene regulation by long non-coding RNAs and its biological functions. *Nature Reviews. Molecular Cell Biology*. 2021;22(2):96–118.

10. Mattick JS, Amaral PP, Carninci P, Carpenter S, Chang HY, Chen L-L, Chen R, Dean C, Dinger ME, Fitzgerald KA, Gingeras TR, Guttman M, Hirose T, Huarte M, Johnson R, et al. Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nature Reviews Molecular Cell Biology*. 2023;24(6):430–447.

11. Volders P-J, Anckaert J, Verheggen K, Nuytens J, Martens L, Mestdagh P, Vandesompele J. LNCipedia 5: towards a reference set of human long non-coding RNAs. *Nucleic Acids Research*. 2019;47(D1):D135–D139.

12. Fernandes JCR, Acuña SM, Aoki JI, Floeter-Winter LM, Muxel SM. Long Non-Coding RNAs in the Regulation of Gene Expression: Physiology and Disease. *Non-coding RNA*. 2019;5(1):17.

13. van Heesch S, Witte F, Schneider-Lunitz V, Schulz JF, Adami E, Faber AB, Kirchner M, Maatz H, Blachut S, Sandmann C-L, Kanda M, Worth CL, Schafer S, Calviello L, Merriott R, et al. The Translational Landscape of the Human Heart. *Cell*. 2019;178(1):242-260.e29.

14. Saddic LA, Sigurdsson MI, Chang T-W, Mazaika E, Heydarpour M, Shernan SK, Seidman CE, Seidman JG, Aranki SF, Body SC, Muehlschlegel JD. The Long Noncoding RNA Landscape of the Ischemic Human Left Ventricle. *Circulation. Cardiovascular Genetics*. 2017;10(1):e001534.

15. Kohlmaier A, Holdt LM, Teupser D. Long noncoding RNAs in cardiovascular disease. *Current Opinion in Cardiology*. 2023;38(3):179–192.

16. Sierra C, Sabariego-Navarro M, Fernández-Blanco Á, Cruciani S, Zamora-Moratalla A, Novoa EM, Dierssen M. The lncRNA Snhg11, a new candidate contributing to neurogenesis, plasticity, and memory deficits in Down syndrome. *Molecular Psychiatry*. 2024.

17. Zhao H, Lou G, Shao Y, Wang T, Wang H, Guo Q, Yang W, Liu H, Liao S. Competing Endogenous RNAs Crosstalk in Hippocampus: A Potential Mechanism for Neuronal Developing Defects in Down Syndrome. *Journal of molecular neuroscience: MN*. 2024;74(2):32.

18. Chaudhari S, Acharya LP, Jasti DB, Ware AP, Gorthi SP, Satyamoorthy K. Discovery of a Novel Shared Variant Among RTEL1 Gene and RTEL1-TNFRSF6B lncRNA at Chromosome 20q13.33 in Familial Progressive Myoclonus Epilepsy. *International Journal of Genomics*. 2024;2024:7518528.

19. Wu Q, Yi X. Down-regulation of Long Noncoding RNA MALAT1 Protects Hippocampal Neurons Against Excessive Autophagy and Apoptosis via the PI3K/Akt Signaling Pathway in Rats with Epilepsy. *Journal of molecular neuroscience: MN*. 2018;65(2):234–245.

20. Hu D-L, Wang G, Yu J, Zhang L-H, Huang Y-F, Wang D, Zhou H-H. Epigallocatechin-3-gallate modulates long non-coding RNA and mRNA expression profiles in lung cancer cells. *Molecular Medicine Reports*. 2019;19(3):1509–1520.

21. Li Y, Wang Z, Shi H, Li H, Li L, Fang R, Cai X, Liu B, Zhang X, Ye L. HBXIP and LSD1 Scaffolded by lncRNA Hotair Mediate Transcriptional Activation by c-Myc. *Cancer Research*. 2016;76(2):293–304.

22. Zhu X, Liu Y, Cui J, Lv J, Li C, Lu J, Huo X, Dou J, Bai Z, Chen Z, Du X. LncRNA LYPLAL1-DT screening from type 2 diabetes with macrovascular complication contributes protective effects on human umbilical vein endothelial cells via regulating the miR-204-5p/SIRT1 axis. *Cell Death Discovery*. 2022;8(1):245.

23. Li Y, Sun W, Li J, Du R, Xing W, Yuan X, Zhong G, Zhao D, Liu Z, Jin X, Pan J, Li Y, Li Q, Kan G, Han X, et al. HuR-mediated nucleocytoplasmic translocation of HOTAIR relieves its inhibition of osteogenic differentiation and promotes bone formation. *Bone Research*. 2023;11(1):53.

24. Kay M, Soltani BM, Nemir M, Aghagolzadeh P, Pezzuto I, Chouvardas P, Ruberto F, Movahedi F, Ansari H, Baharvand H, Pedrazzini T. The conserved long non-coding RNA *CARMA* regulates cardiomyocyte differentiation. *Cardiovascular Research*. 2022;118(10):2339–2353.

25. Flockhart RJ, Webster DE, Qu K, Mascarenhas N, Kovalski J, Kretz M, Khavari PA. BRAFV600E remodels the melanocyte transcriptome and induces BANCR to regulate melanoma cell migration. *Genome Research*. 2012;22(6):1006–1014.

26. Li H, Liu X, Zhang L, Li X. LncRNA BANCR facilitates vascular smooth muscle cell proliferation and migration through JNK pathway. *Oncotarget*. 2017;8(70):114568–114575.

27. Wilson KD, Ameen M, Guo H, Abilez OJ, Tian L, Mumbach MR, Diecke S, Qin X, Liu Y, Yang H, Ma N, Gaddam S, Cunningham NJ, Gu M, Neofytou E, et al. Endogenous Retrovirus-Derived lncRNA BANCR Promotes Cardiomyocyte Migration in Humans and Non-human Primates. *Developmental Cell*. 2020;54(6):694-709.e9.

28. VanOudenhove J, Yankee TN, Wilderman A, Cotney J. Epigenomic and Transcriptomic Dynamics During Human Heart Organogenesis. *Circulation Research*. 2020;127(9):e184–e209.

29. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, Karapetyan K, Katz K, Liu C, Maddipatla Z, Malheiro A, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*. 2018;46(D1):D1062–D1067.

30. Kaminsky EB, Kaul V, Paschall J, Church DM, Bunke B, Kunig D, Moreno-De-Luca D, Moreno-De-Luca A, Mulle JG, Warren ST, Richard G, Compton JG, Fuller AE, Gliem TJ, Huang S, et al. An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*. 2011;13(9):777–784.

31. Fa J, Zhang X, Zhang X, Qi M, Zhang X, Fu Q, Xu Z, Gao Y, Wang B. Long Noncoding RNA lnc-TSSK2-8 Activates Canonical Wnt/β-Catenin Signaling Through Small Heat Shock Proteins HSPA6 and CRYAB. *Frontiers in Cell and Developmental Biology*. 2021;9:660576.

32. Lindsay EA, Vitelli F, Su H, Morishima M, Huynh T, Pramparo T, Jurecic V, Ogunrinu G, Sutherland HF, Scambler PJ, Bradley A, Baldini A. Tbx1 haploinsufficiency in the DiGeorge syndrome region causes aortic arch defects in mice. *Nature*. 2001;410(6824):97–101.

33. Chapnik E, Sasson V, Blelloch R, Hornstein E. Dgcr8 controls neural crest cells survival in cardiovascular development. *Developmental Biology*. 2012;362(1):50–56.

34. Lu Y, Fang Q, Qi M, Li X, Zhang X, Lin Y, Xiang Y, Fu Q, Wang B. Copy number variation-associated lncRNAs may contribute to the etiologies of congenital heart disease. *Communications Biology*. 2023;6(1):189.

35. Landis BJ, Helvaty LR, Geddes GC, Lin J-HI, Yatsenko SA, Lo CW, Border WL, Wechsler SB, Murali CN, Azamian MS, Lalani SR, Hinton RB, Garg V, McBride KL, Hodge JC, et al. A Multicenter Analysis of Abnormal Chromosomal Microarray Findings in Congenital Heart Disease. *Journal of the American Heart Association*. 2023;12(18):e029340.

36. Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, Church DM, Crolla JA, Eichler EE, Epstein CJ, Faucett WA, Feuk L, Friedman JM, Hamosh A, Jackson L, et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *American Journal of Human Genetics*. 2010;86(5):749–764.

37. Geddes GC, Basel D, Frommelt P, Kinney A, Earing M. Genetic Testing Protocol Reduces Costs and Increases Rate of Genetic Diagnosis in Infants with Congenital Heart Disease. *Pediatric Cardiology*. 2017;38(7):1465–1470.

38. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL, ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*. 2015;17(5):405–424.

39. Cardoso-Moreira M, Halbert J, Valloton D, Velten B, Chen C, Shao Y, Liechti A, Ascenção K, Rummel C, Ovchinnikova S, Mazin PV, Xenarios I, Harshman K, Mort M, Cooper DN, et al. Gene expression across mammalian organ development. *Nature*. 2019;571(7766):505–509.

40. Li Z, Liu L, Jiang S, Li Q, Feng C, Du Q, Zou D, Xiao J, Zhang Z, Ma L. LncExpDB: an expression database of human long non-coding RNAs. *Nucleic Acids Research*. 2021;49(D1):D962–D968.

41. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*. 2008;9:559.

42. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*. 2005;4:Article17.

43. Meng J, Chen S, Han J-X, Tan Q, Wang X-R, Wang H-Z, Zhong W-L, Qin Y, Qiao K-L, Zhang C, Gao W-F, Lei Y-Y, Liu H-J, Liu Y-R, Zhou H-G, et al. Derepression of co-silenced tumor suppressor genes by nanoparticle-loaded circular ssDNA reduces tumor malignancy. *Science Translational Medicine*. 2018;10(442):eaao6321.

44. Li Z, Liu L, Feng C, Qin Y, Xiao J, Zhang Z, Ma L. LncBook 2.0: integrating human long non-coding RNAs with multi-omics annotations. *Nucleic Acids Research*. 2023;51(D1):D186–D191.

45. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*. 2003;13(11):2498–2504.

46. Lu Y, Zhou Y, Guo J, Qi M, Lin Y, Zhang X, Xiang Y, Fu Q, Wang B. Integrated analysis of copy number variation-associated lncRNAs identifies candidates contributing to the etiologies of congenital kidney anomalies. *Communications Biology*. 2023;6(1):735.

47. Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, Becker A, Bennett R, Berry A, Bhai J, Bhurji SK, Bignell A, Boddu S, Branco Lins PR, Brooks L, et al. Ensembl 2023. *Nucleic Acids Research*. 2023;51(D1):D933–D941.

48. de Sena-Tomás C, Aleman AG, Ford C, Varshney A, Yao D, Harrington JK, Saúde L, Ramialison M, Targoff KL. Activation of Nkx2.5 transcriptional program is required for adult myocardial repair. *Nature Communications*. 2022;13(1):2970.

49. Schott J-J, Benson DW, Basson CT, Pease W, Silberbach GM, Moak JP, Maron BJ, Seidman CE, Seidman JG. Congenital Heart Disease Caused by Mutations in the Transcription Factor NKX2-5. *Science*. 1998;281(5373):108–111.

50. Benson DW, Silberbach GM, Kavanaugh-McHugh A, Cottrill C, Zhang Y, Riggs S, Smalls O, Johnson MC, Watson MS, Seidman JG, Seidman CE, Plowden J, Kugler JD. Mutations in the cardiac transcription factor NKX2.5 affect diverse cardiac developmental pathways. *The Journal of Clinical Investigation*. 1999;104(11):1567–1573.

51. Reamon-Buettner SM, Hecker H, Spanel-Borowski K, Craatz S, Kuenzel E, Borlak J. Novel NKX2-5 mutations in diseased heart tissues of patients with cardiac malformations. *The American Journal of Pathology*. 2004;164(6):2117–2125.

52. Izumi K, Noon S, Wilkens A, Krantz ID. NKX2.5 mutation identification on exome sequencing in a patient with heterotaxy. *European Journal of Medical Genetics*. 2014;57(10):558–561.

53. Elliott DA, Kirk EP, Yeoh T, Chandar S, McKenzie F, Taylor P, Grossfeld P, Fatkin D, Jones O, Hayes P, Feneley M, Harvey RP. Cardiac homeobox gene NKX2-5mutations and congenital heart disease. *Journal of the American College of Cardiology*. 2003;41(11):2072–2076.

54. Basson CT, Bachinsky DR, Lin RC, Levi T, Elkins JA, Soults J, Grayzel D, Kroumpouzou E, Traill TA, Leblanc-Straceski J, Renault B, Kucherlapati R, Seidman JG, Seidman CE. Mutations in human TBX5 [corrected] cause limb and cardiac malformation in Holt-Oram syndrome. *Nature Genetics*. 1997;15(1):30–35.

55. Li QY, Newbury-Ecob RA, Terrett JA, Wilson DI, Curtis AR, Yi CH, Gebuhr T, Bullen PJ, Robson SC, Strachan T, Bonnet D, Lyonnet S, Young ID, Raeburn JA, Buckler AJ, et al. Holt-Oram syndrome is caused by mutations in TBX5, a member of the Brachyury (T) gene family. *Nature Genetics*. 1997;15(1):21–29.

56. Ching Y-H, Ghosh TK, Cross SJ, Packham EA, Honeyman L, Loughna S, Robinson TE, Dearlove AM, Ribas G, Bonser AJ, Thomas NR, Scotter AJ, Caves LSD, Tyrrell GP, Newbury-Ecob RA, et al. Mutation in myosin heavy chain 6 causes atrial septal defect. *Nature Genetics*. 2005;37(4):423–428.

57. Granados-Riveron JT, Ghosh TK, Pope M, Bu'Lock F, Thornborough C, Eason J, Kirk EP, Fatkin D, Feneley MP, Harvey RP, Armour JAL, David Brook J. Alpha-cardiac myosin heavy chain (MYH6)

mutations affecting myofibril formation are associated with congenital heart defects. *Human Molecular Genetics*. 2010;19(20):4007–4016.

58. Ritter A, Leonard J, Gray C, Izumi K, Levinson K, Nair DR, O'Connor M, Rossano J, Shankar V, Chowns J, Marzolf A, Owens A, Ahrens-Nicklas RC. *MYH7* variants cause complex congenital heart disease. *American Journal of Medical Genetics Part A*. 2022;188(9):2772–2776.

59. Polyakova E, Van Gils JM, Stöger JL, Kiès P, Egorova AD, Koopmann TT, Van Dijk T, DeRuiter MC, Barge-Schaapveld DQCM, Jongbloed MRM. New Genetic Variant in the *MYH7* Gene Associated With Hypoplastic Right Heart Syndrome and Hypertrophic Cardiomyopathy in the Same Family. *Circulation: Genomic and Precision Medicine*. 2023;16(6).

60. Kirk EP, Sunde M, Costa MW, Rankin SA, Wolstein O, Castro ML, Butler TL, Hyun C, Guo G, Otway R, Mackay JP, Waddell LB, Cole AD, Hayward C, Keogh A, et al. Mutations in cardiac T-box factor gene TBX20 are associated with diverse cardiac pathologies, including defects of septation and valvulogenesis and cardiomyopathy. *American Journal of Human Genetics*. 2007;81(2):280–291.

61. Stennard FA, Costa MW, Elliott DA, Rankin S, Haast SJP, Lai D, McDonald LPA, Niederreither K, Dolle P, Bruneau BG, Zorn AM, Harvey RP. Cardiac T-box factor Tbx20 directly interacts with Nkx2-5, GATA4, and GATA5 in regulation of gene expression in the developing heart. *Developmental Biology*. 2003;262(2):206–224.

62. Räsänen M, Sultan I, Paech J, Hemanthakumar KA, Yu W, He L, Tang J, Sun Y, Hlushchuk R, Huan X, Armstrong E, Khoma O-Z, Mervaala E, Djonov V, Betsholtz C, et al. VEGF-B Promotes Endocardium-Derived Coronary Vessel Development and Cardiac Regeneration. *Circulation*. 2021;143(1):65–77.

63. Juni RP, Boon RA. CARMA: what goes around, comes around for the heart. *Cardiovascular Research*. 2022;118(10):2227–2228.

64. Wei D, Bao H, Liu X-Y, Zhou N, Wang Q, Li R-G, Xu Y-J, Yang Y-Q. GATA5 loss-of-function mutations underlie tetralogy of fallot. *International Journal of Medical Sciences*. 2013;10(1):34–42.

65. Padang R, Bagnall RD, Richmond DR, Bannon PG, Semsarian C. Rare non-synonymous variations in the transcriptional activation domains of GATA5 in bicuspid aortic valve disease. *Journal of Molecular and Cellular Cardiology*. 2012;53(2):277–281.

66. Zhang Q, Davis JC, Lamborn IT, Freeman AF, Jing H, Favreau AJ, Matthews HF, Davis J, Turner ML, Uzel G, Holland SM, Su HC. Combined immunodeficiency associated with DOCK8 mutations. *The New England Journal of Medicine*. 2009;361(21):2046–2055.

67. Nguyen MT, Lee W. Kank1 Is Essential for Myogenic Differentiation by Regulating Actin Remodeling and Cell Proliferation in C2C12 Progenitor Cells. *Cells*. 2022;11(13):2030.

68. Dickover M, Hegarty JM, Ly K, Lopez D, Yang H, Zhang R, Tedeschi N, Hsiai TK, Chi NC. The atypical Rho GTPase, RhoU, regulates cell-adhesion molecules during cardiac morphogenesis. *Developmental Biology*. 2014;389(2):182–191.

FIGURE AND TABLE LEGENDS

Figure 1. Workflow for Identifying lncRNA Candidates in CHD. This figure provides an overview of the multi-step methodological approach used in our analysis to investigate the roles of lncRNAs in CHD. In the first step, we integrate genomic information about lncRNAs from our cohort with expression data. In the second step, WGCNA is applied to identify modules that exhibit correlated expression across various tissues. The third step involves identifying hub genes, focusing on lncRNA candidates in the heart-specific modules. Finally, the network is visualized to highlight the interrelationships among CHD-related genes, lncRNAs, and protein-coding genes, showcasing the comprehensive methods applied in our analysis.

Figure 2. Overview of CHD Categories and CNVs in the CCVM Cohort. This figure details the selection of the CNV-LncRNAs reported in the CCVM cohort of isolated CHD patients. (A) Schematic representation of the CNV filtering process that refined the dataset for more targeted analysis. (B) The bar plot shows CHD category distribution in isolated (non-syndromic) CHD patients across the CCVM study cohort. Diagnoses are ranked by patient count, with bars indicating totals. Bars are color-coded: sky blue for patients included in the analysis and white for those excluded. (C) A Sankey diagram shows the distribution of isolated CHD cases in the cohort, detailing associated cardiac phenotypes and the total number of CNVs across all patients. Each node represents a category, diagnosis, or CNV type labeled with the total CNV count. The flow widths between nodes reflect the number of CNVs, highlighting the relative distribution of gains and losses across different CHD diagnoses in the cohort. (D) The circos plot visualizes CNV gains and losses across the human genome (hg38). The outer ideogram track displays chromosome bands, with black lines indicating gene-poor G-bands and red lines marking centromeres. The blue histogram track represents CNV gains, and the red track represents CNV losses. Taller bars highlight regions with higher frequencies of CNVs, pinpointing the genomic areas most impacted by gains and losses in the study cohort. (E) Venn diagram of expressed lncRNAs and CNV-associated lncRNAs illustrating the overlap between

lncRNAs expressed in any tissue (blue), heart-expressed lncRNAs (red), and CNV-associated lncRNAs (green). The overlap regions indicate the number of lncRNAs shared between these categories, with the 3,608 heart-expressed lncRNAs used in our analysis being a subset of the total expressed lncRNAs.

Figure 3. Module-Tissue Association Heatmap Across Seven Organ Systems. This heatmap helps identify gene modules critical for the development and function of specific tissues. It illustrates the correlation between the modules identified by WGCNA and the different tissue types across the developmental time series. Each row represents a module eigengene, color-coded according to the original module colors prefixed by "ME," while each column corresponds to a specific tissue. Positive correlations are depicted in red, negative correlations in blue, and white indicates no correlation. The color intensity reflects the strength of the correlation, with numerical values in each cell displaying the correlation coefficient, followed by the p-value in parentheses for statistical significance. Only significant correlations ($p < 0.001$) are shown. Modules with high positive or negative correlations ($> |0.6|$) are considered tissue-specific, highlighting potential tissue-specific gene expression patterns. Module ME03 (magenta) notably shows the highest positive correlation ($R = 0.95$) with heart tissue, suggesting a strong link to heart development and function.

Figure 4. Magenta Module Eigengene Expression Across Developmental Stages in the Heart. The eigengene expression for the magenta module (ME03) is shown across various developmental stages in the heart, plotted by age (PCW). The red data points represent the average eigengene expression at each stage, with a LOESS-smoothed line (blue) fitted to highlight the trend over time. The gray shading indicates the 95% confidence interval, showing the range within which the trend line is expected to fall. Vertical dashed gray lines indicate key developmental milestones: 10 weeks, 20 weeks, birth (38 weeks), 1 year, and 5 years post-conception. These milestones mark significant transitions in heart development, providing context for the observed changes in eigengene

expression. The expression level increases during early development, peaking between 10–20 PCW and gradually decreasing.

Figure 5. Functional Enrichment of Heart-Related Biological Processes in the Magenta Module. This bar chart presents 26 significantly enriched GO biological process (BP) terms related to the heart and cardiovascular system, derived from the protein-coding genes in the magenta module. Semantically similar terms were collapsed, and the most significant terms were selected based on their lowest adjusted p-values (Benjamini-Hochberg method), all below 0.05. Each bar represents the -log10 transformation of the adjusted p-values, indicating the relative statistical significance. The analysis highlights the magenta module's enrichment for genes with key roles in heart development and function.

Figure 6. Identification of Hub Genes in the Magenta Module. This scatter plot illustrates the relationship between Module Membership (MM) and Gene Significance (GS) for genes within the magenta module (ME03). Module Membership reflects how strongly a gene is associated with the module eigengene, while Gene Significance measures the correlation between a gene's expression and the heart trait. Genes are categorized into lncRNAs (purple circles) and protein-coding genes (teal squares). The top hub genes, which are highly connected within the module and strongly associated with the heart trait, are emphasized by a dashed vertical line at MM = 0.8. The plot also includes correlation (R) and p-values for both lncRNAs and protein-coding genes, calculated separately and displayed on the graph to indicate the strength and significance of the correlation between Module Membership and Gene Significance. This visualization is crucial for identifying lncRNA hub genes within the module, which may play essential roles in the heart-related biological processes associated with the module.

Figure 7. Gene Network Visualization of Hub Genes in the Magenta Module. This figure visualizes the co-expression network of hub genes within the magenta module (ME03), generated using Cytoscape. Nodes represent hub genes, color-coded by gene type: gold for known CHD genes,

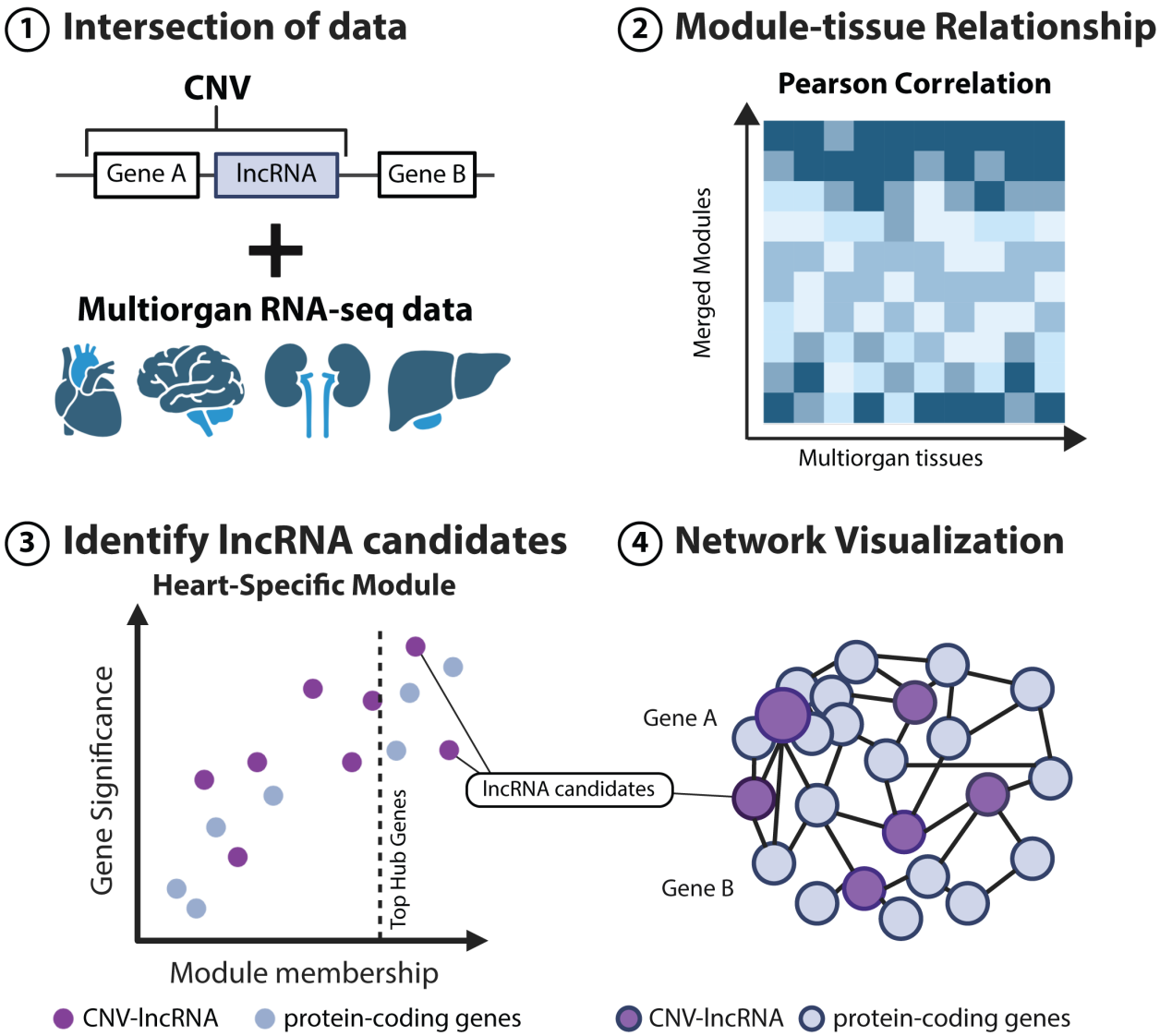
purple for lncRNAs, and light green for protein-coding genes. Edges indicate co-expression interactions between genes based on the TOM matrix, with a threshold of TOM > 0.30. Hub genes are defined as those with high module membership (R ≥ 0.80). Co-expression edges connecting to lncRNA genes are highlighted in red. The network is divided into two panels for clarity: (**A**) Seven of the eighteen CNV-lncRNA hub genes are co-expressed with *NKX2-5,* a critical transcription factor for early heart development. (**B**) A highly interconnected network of known CHD genes and other protein-coding genes, all co-expressed with *NKX2-5*. For clarity, redundant edges were removed and indicated with an arrow and curly brace.

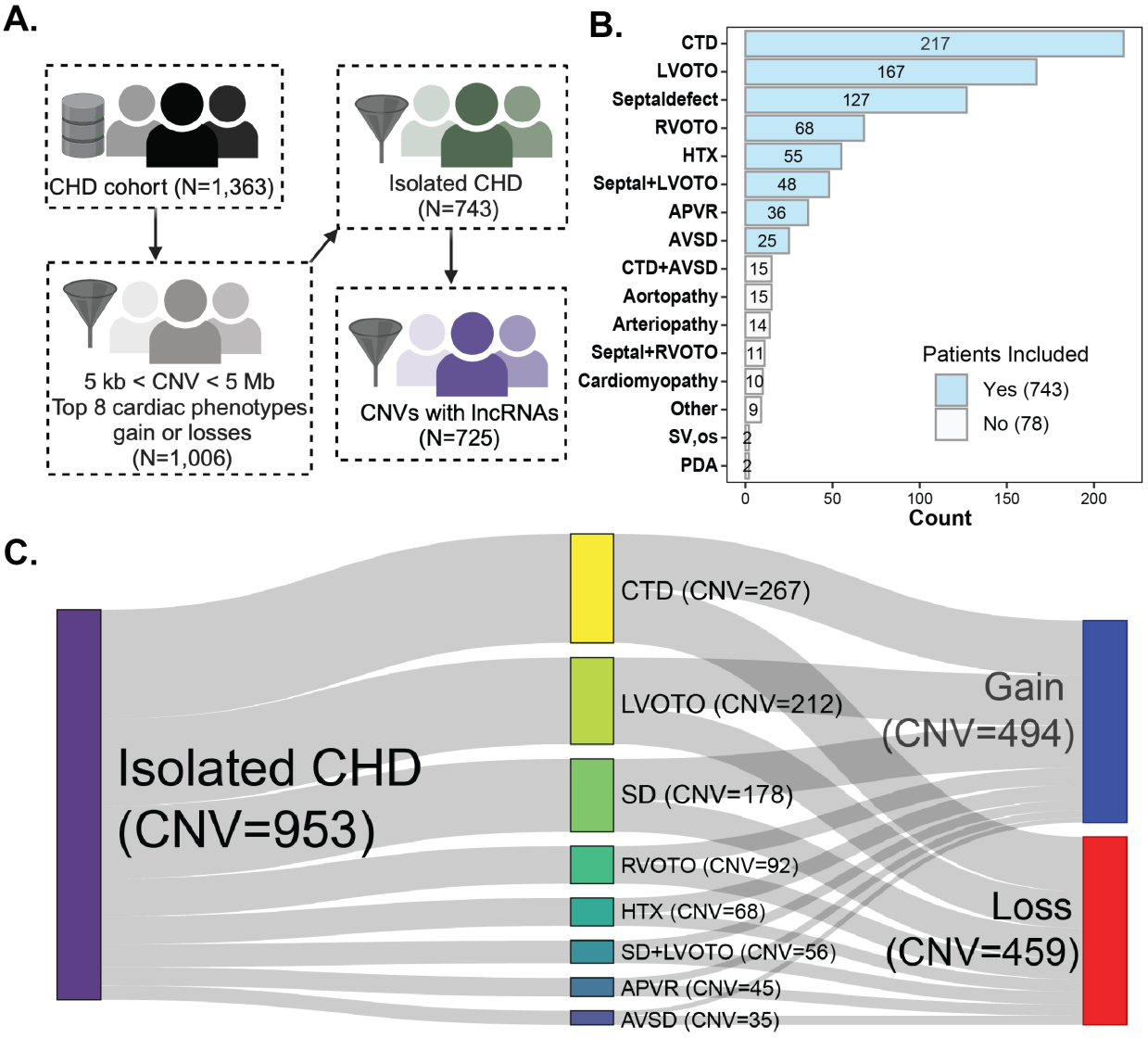
Table 1. Known CHD and lncRNA Hub Genes from the Magenta Module (ME03). For each gene, the table provides the module membership correlation (MM.Cor) and its associated p-value, which indicates the strength of each gene's association with the magenta module. Additionally, the table shows the gene significance correlation (GS.Cor) and its p-value, reflecting the correlation between each gene and heart-related traits. This summary highlights hub genes highly correlated with the module, including both protein-coding CHD genes and lncRNAs.

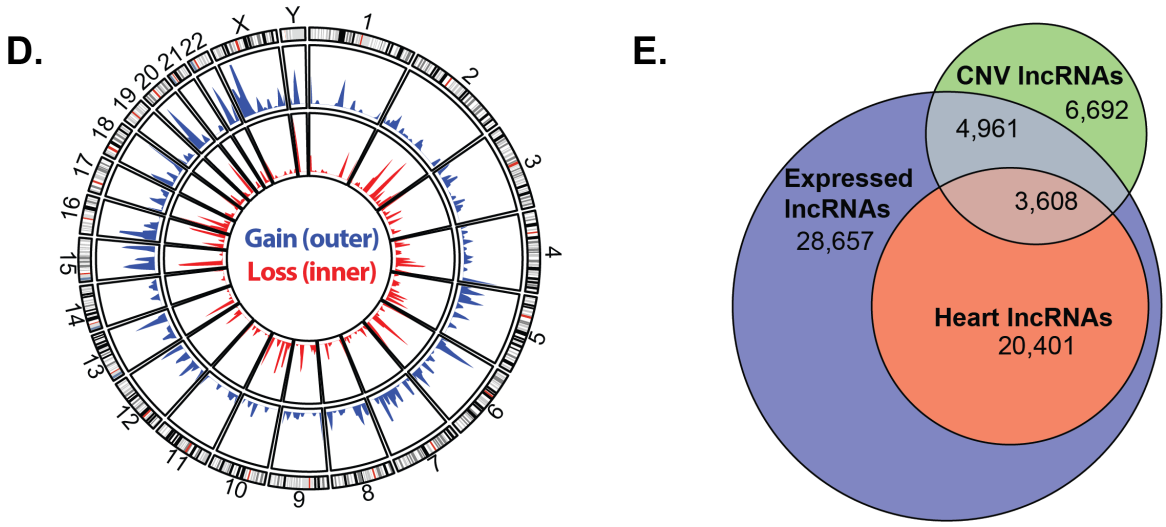
Table 2: lncRNA Hub Genes and Associated CNVs and Cardiac Phenotypes. Each lncRNA from the magenta module (ME03) is listed alongside the corresponding chromosomal region of the overlapping CNV (e.g., gain or loss) and the associated cardiac phenotype. The table also includes patient identifiers and CHD classification from the CCVM cohort who exhibit the relevant CNV, directly linking the genetic variations and observed heart conditions.

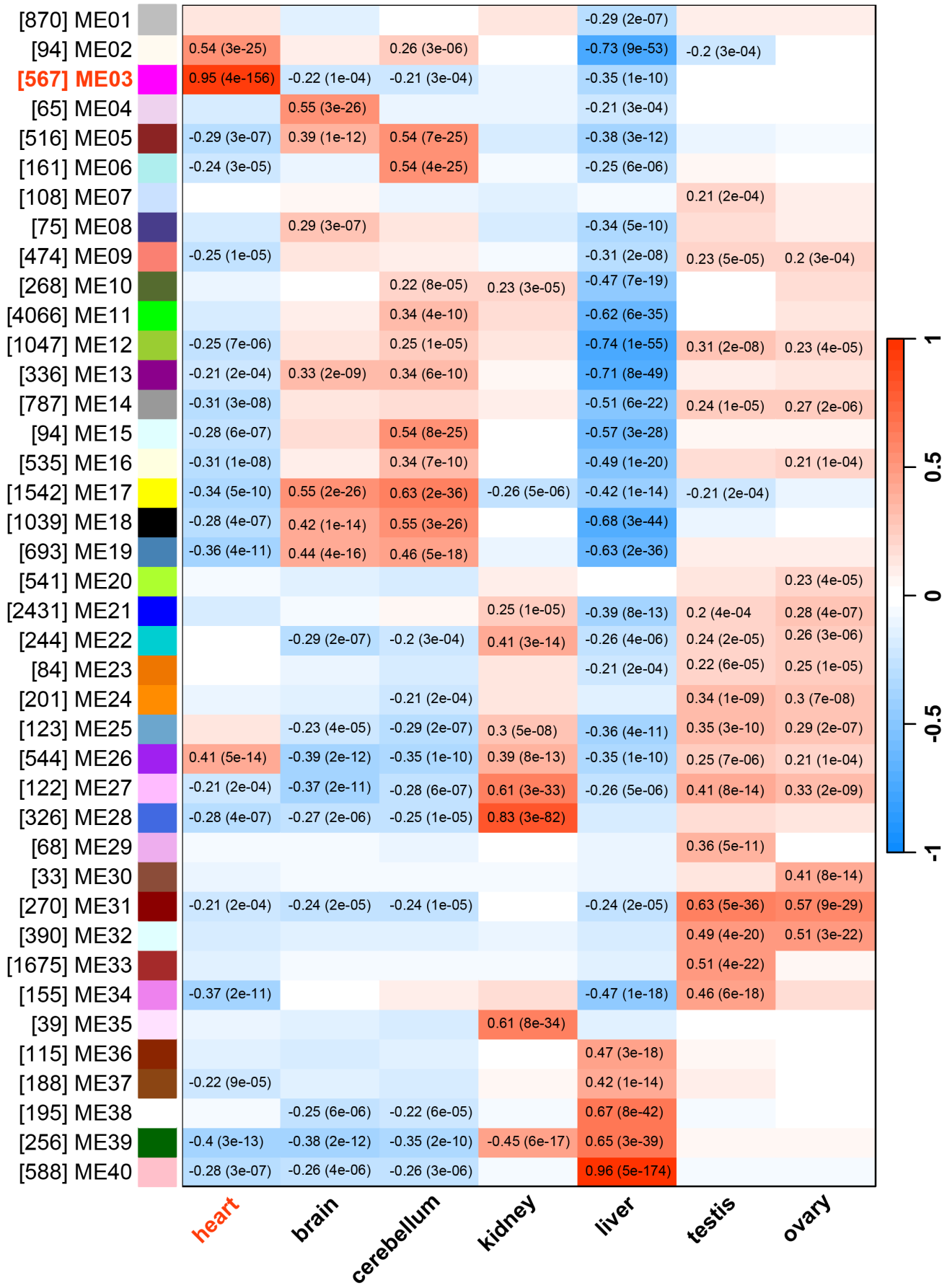
Table 3: Gini Coefficient Analysis of Hub Genes. This table presents the Gini coefficients for hub genes identified in the magenta module (ME03). The table includes hub genes that are either known CHD genes or lncRNAs. The Gini coefficient, a measure of tissue-specific gene expression, is calculated for each gene, with higher values indicating tissue-specific expression. Genes with a Gini coefficient > 0.5 and maximal expression in the embryonic heart are considered to be candidate CHD genes. Out of 8 known CHD hub genes, five were found to have significant embryonic heart-specific

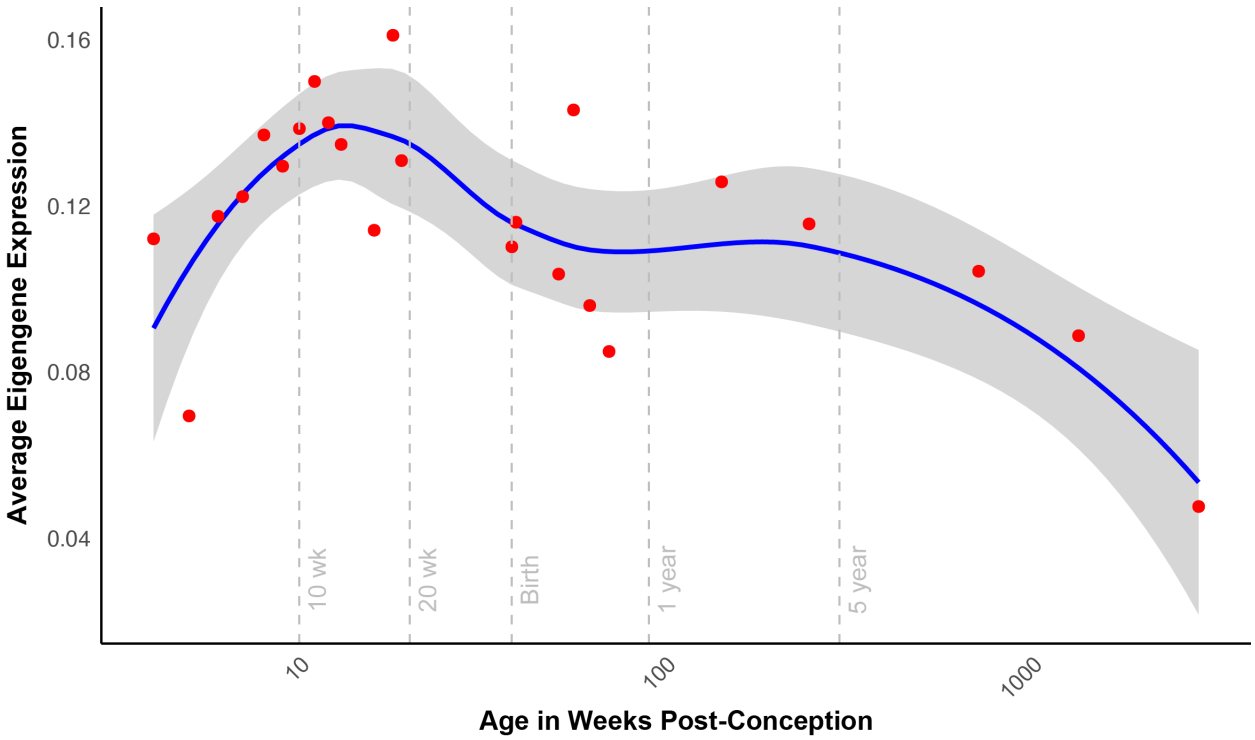
expression (Gini coefficient > 0.5). Among the 18 lncRNA hub genes, 7 of the 9 with Ensembl accessions have a Gini coefficient, and five are maximally expressed in the embryonic heart. Of these, 4 lncRNAs exhibit significant heart-specific expression with a Gini coefficient greater than 0.5.

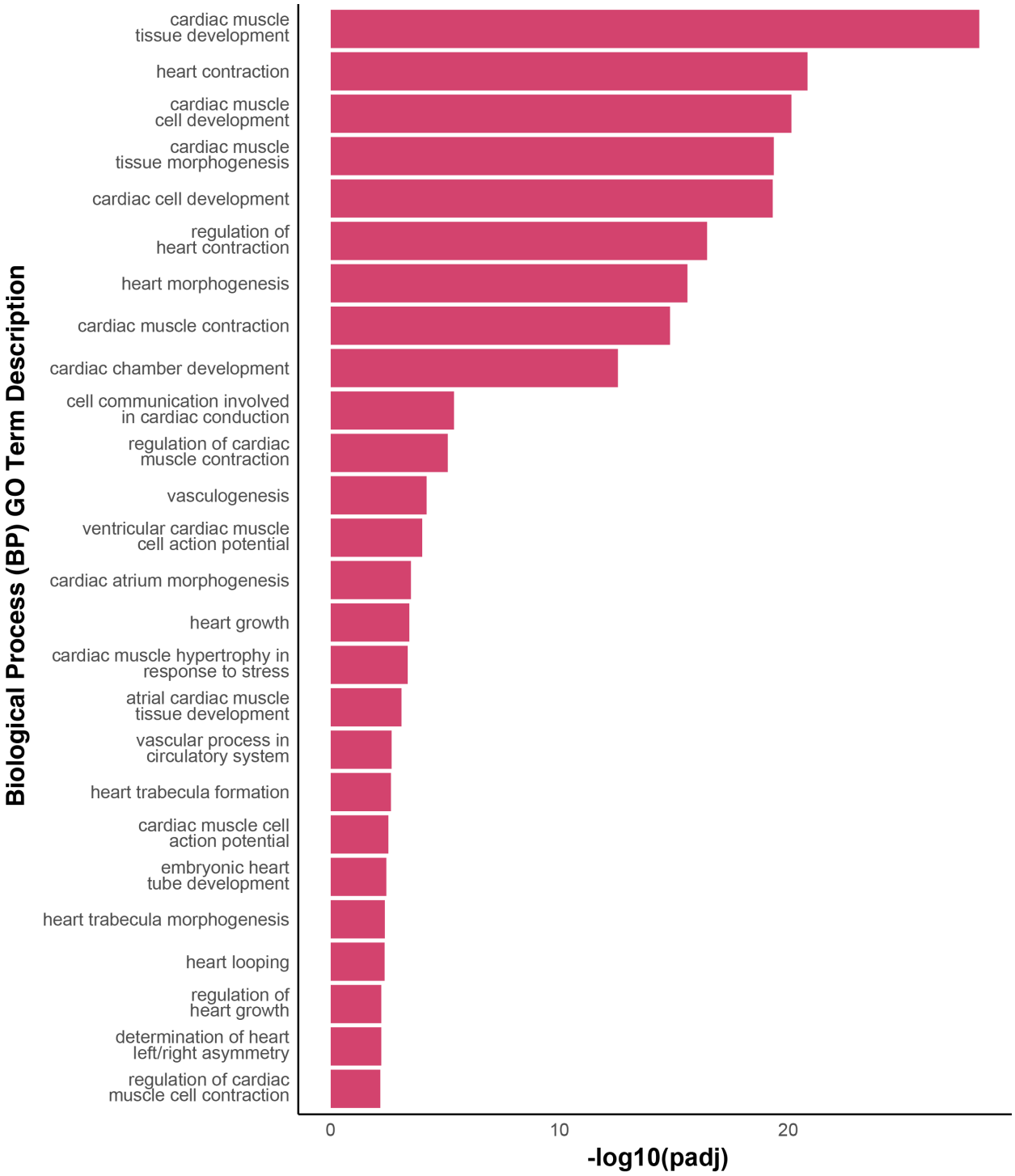

1 Intersection of data
CNV
Gene A
lncRNA
Gene B
Multiorgan RNA-seq data
2 Module-tissue Relationship
Pearson Correlation
Merged Modules
Multiorgan tissues
3 Identify lncRNA candidates
Heart-Specific Module
Gene Significance
Top Hub Genes
Module membership
lncRNA candidates
CNV-lncRNA
protein-coding genes
4 Network Visualization
Gene A
Gene B
CNV-lncRNA
protein-coding genes

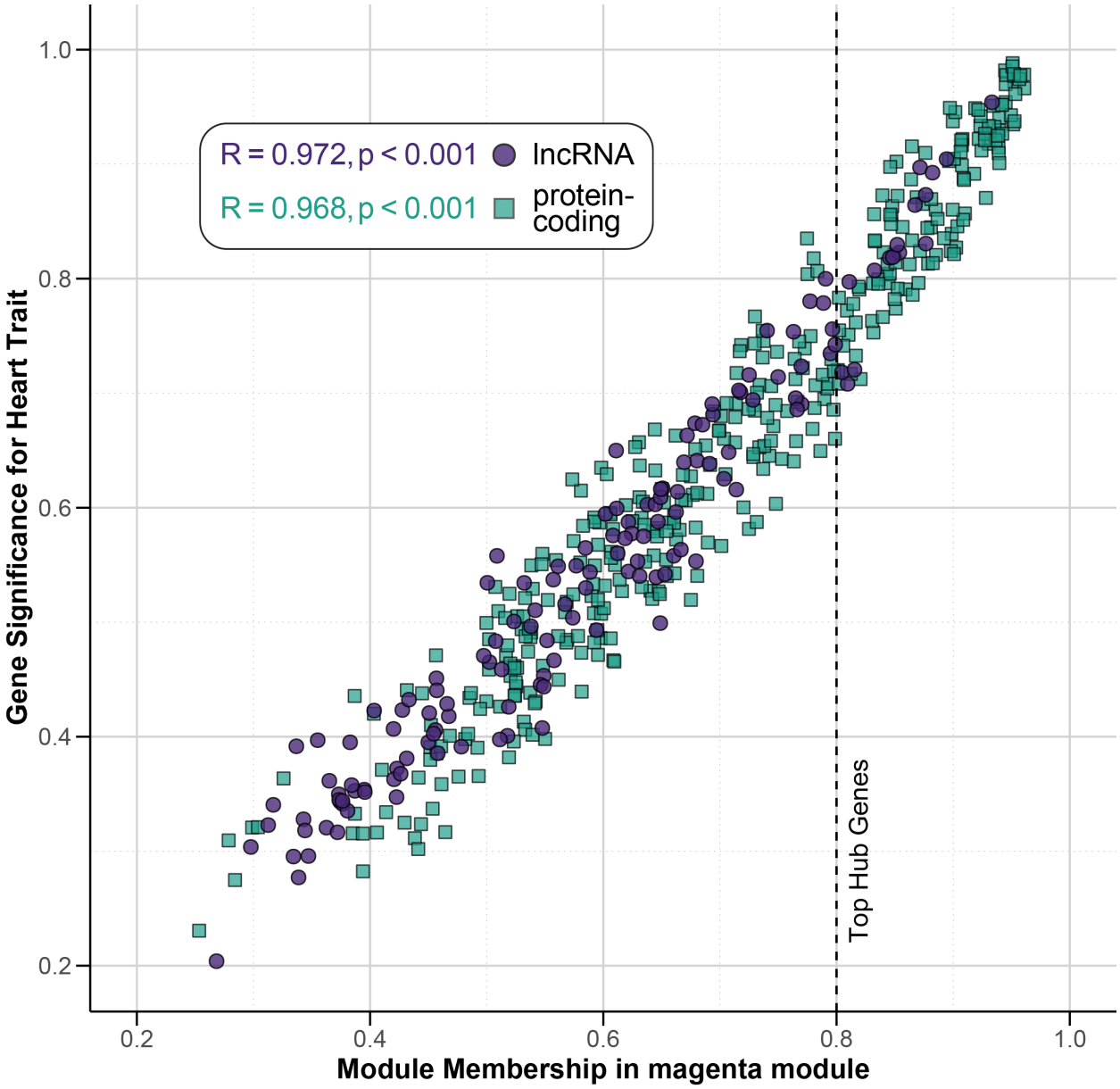

A.
CHD cohort (N=1,363)
Isolated CHD (N=743)
5 kb < CNV < 5 Mb
Top 8 cardiac phenotypes gain or losses
(N=1,006)
CNVs with lncRNAs
(N=725)
B.
CTD 217
LVOTO 167
Septaldefect 127
RVOTO 68
HTX 55
Septal+LVOTO 48
APVR 36
AVSD 25
CTD+AVSD 15
Aortopathy 15
Arteriopathy 14
Septal+RVOTO 11
Cardiomyopathy 10
Other 9
SV,os 2
PDA 2
Patients Included
Yes (743)
No (78)
0
50
100
150
200
Count
C.
Isolated CHD
(CNV=953)
CTD (CNV=267)
LVOTO (CNV=212)
SD (CNV=178)
RVOTO (CNV=92)
HTX (CNV=68)
SD+LVOTO (CNV=56)
APVR (CNV=45)
AVSD (CNV=35)
Gain
(CNV=494)
Loss
(CNV=459)

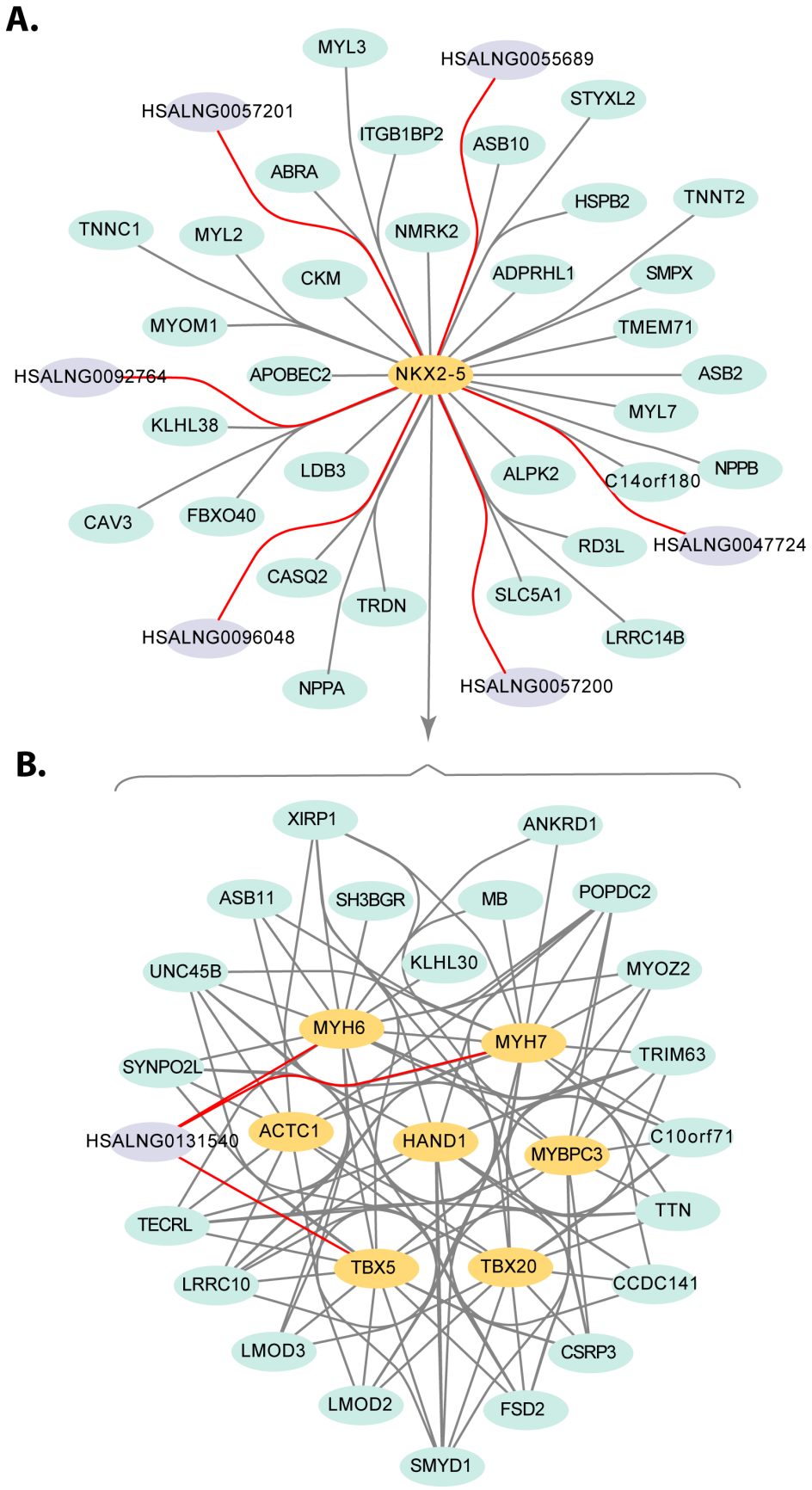

D.
Gain (outer)
Loss (inner)
E.
CNV lncRNAs
6,692
4,961
Expressed
lncRNAs
28,657
3,608
Heart lncRNAs
20,401


[870] ME01
[94] ME02
[567] ME03
[65] ME04
[516] ME05
[161] ME06
[108] ME07
[75] ME08
[474] ME09
[268] ME10
[4066] ME11
[1047] ME12
[336] ME13
[787] ME14
[94] ME15
[535] ME16
[1542] ME17
[1039] ME18
[693] ME19
[541] ME20
[2431] ME21
[244] ME22
[84] ME23
[201] ME24
[123] ME25
[544] ME26
[122] ME27
[326] ME28
[68] ME29
[33] ME30
[270] ME31
[390] ME32
[1675] ME33
[155] ME34
[39] ME35
[115] ME36
[188] ME37
[195] ME38
[256] ME39
[588] ME40
-0.29 (2e-07)
0.54 (3e-25) 0.26 (3e-06) -0.73 (9e-53) -0.2 (3e-04)
0.95 (4e-156) -0.22 (1e-04) -0.21 (3e-04) -0.35 (1e-10)
0.55 (3e-26) -0.21 (3e-04)
-0.29 (3e-07) 0.39 (1e-12) 0.54 (7e-25) -0.38 (3e-12)
-0.24 (3e-05) 0.54 (4e-25) -0.25 (6e-06)
0.21 (2e-04)
0.29 (3e-07) -0.34 (5e-10)
-0.25 (1e-05) -0.31 (2e-08) 0.23 (5e-05) 0.2 (3e-04)
0.22 (8e-05) 0.23 (3e-05) -0.47 (7e-19)
0.34 (4e-10) -0.62 (6e-35)
-0.25 (7e-06) 0.25 (1e-05) -0.74 (1e-55) 0.31 (2e-08) 0.23 (4e-05)
-0.21 (2e-04) 0.33 (2e-09) 0.34 (6e-10) -0.71 (8e-49)
-0.31 (3e-08) -0.51 (6e-22) 0.24 (1e-05) 0.27 (2e-06)
-0.28 (6e-07) 0.54 (8e-25) -0.57 (3e-28)
-0.31 (1e-08) 0.34 (7e-10) -0.49 (1e-20) 0.21 (1e-04)
-0.34 (5e-10) 0.55 (2e-26) 0.63 (2e-36) -0.26 (5e-06) -0.42 (1e-14) -0.21 (2e-04)
-0.28 (4e-07) 0.42 (1e-14) 0.55 (3e-26) -0.68 (3e-44)
-0.36 (4e-11) 0.44 (4e-16) 0.46 (5e-18) -0.63 (2e-36)
0.23 (4e-05)
0.25 (1e-05) -0.39 (8e-13) 0.2 (4e-04 0.28 (4e-07)
-0.29 (2e-07) -0.2 (3e-04) 0.41 (3e-14) -0.26 (4e-06) 0.24 (2e-05) 0.26 (3e-06)
-0.21 (2e-04) 0.22 (6e-05) 0.25 (1e-05)
-0.21 (2e-04) 0.34 (1e-09) 0.3 (7e-08)
-0.23 (4e-05) -0.29 (2e-07) 0.3 (5e-08) -0.36 (4e-11) 0.35 (3e-10) 0.29 (2e-07)
0.41 (5e-14) -0.39 (2e-12) -0.35 (1e-10) 0.39 (8e-13) -0.35 (1e-10) 0.25 (7e-06) 0.21 (1e-04)
-0.21 (2e-04) -0.37 (2e-11) -0.28 (6e-07) 0.61 (3e-33) -0.26 (5e-06) 0.41 (8e-14) 0.33 (2e-09)
-0.28 (4e-07) -0.27 (2e-06) -0.25 (1e-05) 0.83 (3e-82)
0.36 (5e-11)
0.41 (8e-14)
-0.21 (2e-04) -0.24 (2e-05) -0.24 (1e-05) -0.24 (2e-05) 0.63 (5e-36) 0.57 (9e-29)
0.49 (4e-20) 0.51 (3e-22)
0.51 (4e-22)
-0.37 (2e-11) -0.47 (1e-18) 0.46 (6e-18)
0.61 (8e-34)
0.47 (3e-18)
-0.22 (9e-05) 0.42 (1e-14)
-0.25 (6e-06) -0.22 (6e-05) 0.67 (8e-42)
-0.4 (3e-13) -0.38 (2e-12) -0.35 (2e-10) -0.45 (6e-17) 0.65 (3e-39)
-0.28 (3e-07) -0.26 (4e-06) -0.26 (3e-06) 0.96 (5e-174)
heart
brain
cerebellum
kidney
liver
testis
ovary
1
0.5
0
-0.5
-1


Average Eigengene Expression
0.16
0.12
0.08
0.04
10 wk
20 wk
Birth
1 year
5 year
10
100
1000
Age in Weeks Post-Conception


Biological Process (BP) GO Term Description
cardiac muscle tissue development
heart contraction
cardiac muscle cell development
cardiac muscle tissue morphogenesis
cardiac cell development
regulation of heart contraction
heart morphogenesis
cardiac muscle contraction
cardiac chamber development
cell communication involved in cardiac conduction
regulation of cardiac muscle contraction
vasculogenesis
ventricular cardiac muscle cell action potential
cardiac atrium morphogenesis
heart growth
cardiac muscle hypertrophy in response to stress
atrial cardiac muscle tissue development
vascular process in circulatory system
heart trabecula formation
cardiac muscle cell action potential
embryonic heart tube development
heart trabecula morphogenesis
heart looping
regulation of heart growth
determination of heart left/right asymmetry
regulation of cardiac muscle cell contraction
0
10
20
-log10(padj)


R = 0.972, p < 0.001
lncRNA
R = 0.968, p < 0.001
protein-coding
Top Hub Genes
Gene Significance for Heart Trait
Module Membership in magenta module
0.2
0.4
0.6
0.8
1.0


A.
MYL3
HSALNG0055689
STYXL2
HSALNG0057201
ITGB1BP2
ASB10
ABRA
HSPB2
TNNT2
TNNC1
MYL2
NMRK2
CKM
ADPRHL1
SMPX
MYOM1
TMEM71
HSALNG0092764
APOBEC2
NKX2-5
ASB2
MYL7
KLHL38
LDB3
ALPK2
C14orf180
NPPB
CAV3
FBXO40
RD3L
HSALNG0047724
CASQ2
SLC5A1
TRDN
HSALNG0096048
LRRC14B
NPPA
HSALNG0057200
B.
XIRP1
ANKRD1
ASB11
SH3BGR
MB
POPDC2
KLHL30
UNC45B
MYOZ2
MYH6
MYH7
TRIM63
SYNPO2L
HSALNG0131540
ACTC1
HAND1
MYBPC3
C10orf71
TTN
TECRL
TBX5
TBX20
LRRC10
CCDC141
LMOD3
CSRP3
LMOD2
FSD2
SMYD1

TABLES

Table 1. Known CHD and lncRNA hub genes from the magenta module (ME03).

Accession	Gene Symbols / Synonymns	MM.Cor	MM.Pvalue	GS.Cor	GS.Pvalue	Classification
Know CHD Protein-Coding Genes						
ENSG00000183072	NKX2-5	0.95	1.215E-153	0.98	2.980E-213	Protein-coding
ENSG00000134571	MYBPC3	0.93	9.593E-138	0.93	1.460E-140	Protein-coding
ENSG00000089225	TBX5	0.92	1.236E-131	0.94	2.262E-149	Protein-coding
ENSG00000092054	MYH7	0.92	7.219E-130	0.95	5.668E-156	Protein-coding
ENSG00000159251	ACTC1	0.91	3.250E-119	0.90	1.605E-113	Protein-coding
ENSG00000197616	MYH6	0.90	2.217E-115	0.95	1.557E-153	Protein-coding
ENSG00000113196	HAND1	0.85	2.988E-87	0.85	1.218E-87	Protein-coding
ENSG00000164532	TBX20	0.85	1.172E-87	0.86	2.123E-94	Protein-coding
lncRNA Genes						
HSALNG0131540	MIR1-1HG MIR133A2HG Lnc-SLCO4A1-8	0.93	2.219E-140	0.95	9.435E-165	Intergenic
HSALNG0057201	Lnc-TBX20-1 ENSG00000226063	0.89	1.122E-110	0.90	4.646E-117	Intergenic
HSALNG0055689	HSALNG0055689	0.88	8.389E-104	0.89	1.328E-109	Antisense
HSALNG0057200	ENSG00000289335	0.88	9.246E-101	0.87	4.126E-99	Intergenic
HSALNG0081800	Lnc-STK32C-3 ENSG00000226900	0.88	8.062E-101	0.83	4.435E-81	Intergenic
HSALNG0047724	Lnc-ECI2-2; LOC101927888 ENSG00000231811	0.87	2.538E-98	0.90	2.27E-112	Intergenic
HSALNG0092764	Lnc-CCDC59-5 ENSG00000288941	0.87	3.134E-96	0.86	7.883E-95	Intergenic
HSALNG0057214	HSALNG0057214	0.85	9.024E-87	0.82	9.348E-77	Intergenic
HSALNG0088492	lnc-B4GALNT3-4	0.85	3.362E-90	0.82	2.549E-78	Antisense, Sense
HSALNG0088499	LINC02455	0.85	7.368E-88	0.82	6.253E-77	Intergenic
HSALNG0096048	LINC02455 ENSG00000256672	0.85	1.990E-89	0.83	1.008E-80	Antisense
HSALNG0062838	HSALNG0062838	0.83	8.777E-82	0.81	3.391E-73	Sense
HSALNG0119665	ENSG00000272625 antisense to LPIN2	0.82	8.212E-76	0.72	2.238E-51	Antisense, Sense
HSALNG0062865	lnc-RP11-63E5.6.1-3	0.81	6.605E-74	0.71	6.227E-49	Antisense
HSALNG0114768	Lnc-PMP22-2 ENSG00000237377	0.81	2.341E-74	0.80	3.837E-70	Intergenic
HSALNG0054330	HSALNG0054330	0.80	6.152E-70	0.76	3.558E-59	Antisense
HSALNG0069545	lnc-DOCK8-2	0.80	2.181E-72	0.72	6.540E-51	Sense
HSALNG0131539	CRMA ENSG00000174403 C20orf166-AS1 MIR1-1HG-AS1	0.80	1.225E-70	0.74	4.564E-56	Intergenic

Table 2. lncRNA hub genes and associated CNVs and cardiac phenotype

LncBook Identifier	Overlapping CNV	Cardiac Phenotype	CCVM Patient(s)
HSALNG0047724	6p25.1 loss	CTD	003-0293
HSALNG0054330	10q23.1 gain 6q25.1 gain	CTD	000-0369
HSALNG0055689	7p22.3p22.2 loss	CTD	002-0019
HSALNG0057200 HSALNG0057201 HSALNG0057214	7p14.2 loss*	CTD	003-0191
HSALNG0062838	7q36.3 loss	LVOTO	001-0182
HSALNG0062865	8p23.3 loss 8p23.3 gain	CTD LVOTO	004-0065 003-0116
HSALNG0069545	9p24.3 gain (2) 9p24.3 loss	CTD RVOTO LVOTO	008-0053 001-0078 003-0375
HSALNG0081800	10q26.3 loss (2)	RVOTO (2)	002-0008 009-0009
HSALNG0088492 HSALNG0088499	12p13.33 gain 2p16.3 loss	SD+LVOTO	007-0016
HSALNG0092764	12q21.31 loss 16p13.11 gain	SD	003-0440
HSALNG0096048	13q12.3 gain	CTD	007-0002
HSALNG0114768	2p21 gain 5p15.2 loss 17p12 gain	AVSD	002-0007
HSALNG0119665	18p11.32p11.31 loss 18p11.32p11.31 gain	APVR SD	005-0036 000-0224
HSALNG0131539 HSALNG0131540	20q13.33 gain	SD+LVOTO	001-0285

*This CNV also contains the known CHD gene TBX20

Table 3: Gini Coefficient Analysis of Hub Genes in the Magenta Module.

Accession	Symbol	Type	Gini Coefficient	Tissue with max expression	Max Expression
ENSG00000226063	Lnc-TBX20-1 (HSALNG0057201)	lncRNA	0.959	Embryonic Heart	167
ENSG00000164532	TBX20	CHD	0.948	Embryonic Heart	5117
ENSG00000197616	MYH6	CHD	0.936	Heart	323262
ENSG00000183072	NKX2-5	CHD	0.935	Embryonic Heart	9467
ENSG00000134571	MYBPC3	CHD	0.934	Heart	91682
ENSG00000231811	Lnc-ECI2-2 (HSALNG0047724)	lncRNA	0.93	Heart	85
ENSG00000174407	MIR1-1HG (HSALNG0131540)	lncRNA	0.918	Muscle	5447
ENSG00000159251	ACTC1	CHD	0.915	Embryonic Heart	290948
ENSG00000092054	MYH7	CHD	0.905	Muscle	457406
ENSG00000113196	HAND1	CHD	0.889	Embryonic Heart	2246
ENSG00000089225	TBX5	CHD	0.873	Embryonic Heart	5400
ENSG00000226900	Lnc-STK32C-3 (HSALNG0081800)	lncRNA	0.867	Embryonic Heart	567
ENSG00000174403	CRMA / MIR1-1HG-AS1 (HSALNG0131539)	lncRNA	0.662	Embryonic Heart	2167
ENSG00000226403	lnc-DOCK8-2 (HSALNG0069545)	lncRNA	0.631	Embryonic Heart	398
ENSG00000272625	antisense to LPIN2 (HSALNG0119665)	lncRNA	0.251	Embryonic Heart	81