(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
20 January 2022 (20.01.2022)

WIPO | PCT

(10) International Publication Number
WO 2022/015998 A1

(51) International Patent Classification:
C12Q 1/68 (2018.01) *G06F 19/20* (2011.01)
C12Q 1/6879 (2018.01) *G06F 19/24* (2011.01)
G06F 19/18 (2011.01) *G16B 25/00* (2019.01)

(21) International Application Number:
PCT/US2021/041853

(22) International Filing Date:
15 July 2021 (15.07.2021)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/051,983 15 July 2020 (15.07.2020) US

(71) Applicant: **UNIVERSITY OF CONNECTICUT** [US/US]; Technology Commercialization Services, 400 Farmington Avenue-MC6400, Farmington, Connecticut 06032 (US).

(72) Inventors; and
(71) Applicants: **COTNEY, Justin L.** [US/US]; c/o University of Connecticut, 400 Farmington Avenue-MC6400, Farmington, Connecticut 06032 (US). **VANOUDENHOVE, Jennifer J.** [US/US]; c/o University of Connecticut, 400 Farmington Avenue-MC6400, Farmington, Connecticut 06032 (US). **YANKEE, Tara** [US/US]; c/o University of Connecticut, 400 Farmington Avenue-MC6400, Farmington, Connecticut 06032 (US).

(74) Agent: **MUELLER, Lisa V.**; Casimir Jones, S.C., 2275 Deming Way, Suite 310, Middleton, Wisconsin 53562 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK,

(54) Title: GENE PANELS AND METHODS OF USE THEREOF FOR SCREENING AND DIAGNOSIS OF CONGENITAL HEART DEFECTS AND DISEASES

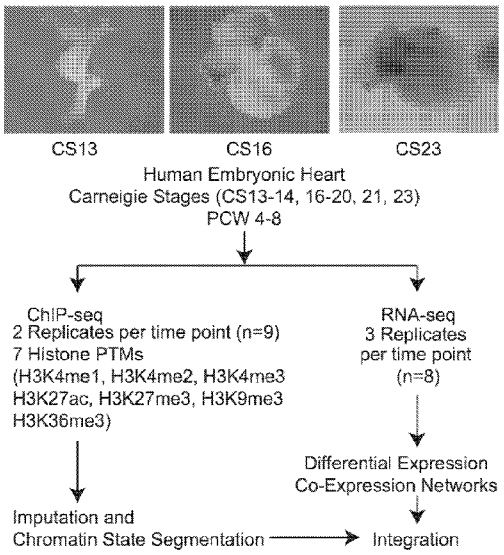


FIG. 1A

(57) Abstract: The present disclosure relates to gene panels and methods of use thereof for the assessment of risk of congenital heart defect or disease in a subject.

[Continued on next page]

WO 2022/015998 A1

EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

GENE PANELS AND METHODS OF USE THEREOF FOR SCREENING AND DIAGNOSIS OF CONGENITAL HEART DEFECTS AND DISEASES

STATEMENT OF GOVERNMENT SUPPORT

This invention was made with government support under Grant No GM119465 from the National Institutes of Health. The government has certain rights in the invention.

STATEMENT OF RELATED APPLICATIONS

This application claims priority to U.S. Provisional Patent Application No. 63/051,983, filed July 15, 2020, the entire contents of which are incorporated herein by reference.

BACKGROUND

Congenital heart defects and diseases (CHDs) are among the most common birth defects, affecting about 1% of live births worldwide. CHDs remain the leading cause of infant mortality in developed nations (van der Linde *et al.* Birth prevalence of congenital heart disease worldwide: a systematic review and meta-analysis. *J Am Coll Cardiol.* 2011;58:2241–47; Liu *et al.* Global birth prevalence of congenital heart defects 1970-2017: updated systematic review and meta-analysis of 260 studies. Int J Epidemiol. 2019;48:455–63; Centers for Disease Control and Prevention (CDC). Racial differences by gestational age in neonatal deaths attributable to congenital heart defects—United States, 2003–2006. *MMWR Morb Mortal Wkly Rep.* 2010;59:1208–11). Significant advancements in medical and surgical interventions, particularly those during the first year of life, have dramatically improved survivability of those with CHD to adulthood (Marelli *et al.* Lifetime prevalence of congenital heart disease in the general population from 2000 to 2010. *Circulation* 2014;*130*: 749–56). With these improvements come significant costs. The estimated annual cost of hospitalization related to CHD is over $5 billion dollars in the US alone. Despite these improved treatments CHD is still the leading cause of infant mortality in developed nations, contributing to over 4% of all neonatal deaths in the US (Centers for Disease Control and Prevention (CDC), 2010). Understanding the cause of CHD is therefore of extreme importance not only to reduce its incidence, but to provide better long-term care and education for the ever-increasing number of adults living with CHD.

SUMMARY

The present disclosure generally describes a well-connected hub genes with heart-specific expression targeted by embryonic heart-specific enhancers are likely disease candidates. These functional annotations will allow for better interpretation of whole genome sequencing data in the large number of patients affected by congenital heart defects.

Accordingly, provided herein are methods of identifying a subject as at risk of having a congenital heart defect. In some embodiments, the method of identifying a subject as at risk of having a congenital heart defect comprise assessing a plurality of genes in a sample obtained from the subject. In some embodiments, the plurality of genes are selected from the genes listed in Table 1. In some embodiments, the plurality of genes comprise at least 5 genes listed in Table 1. In some embodiments, the plurality of genes comprises at least 10 genes listed in Table 1. In some embodiments, the plurality of genes comprises at least 25 genes listed in Table 1. In some embodiments, the panel of genes comprises at least 50 genes listed in Table 1. In some embodiments, the panel of genes comprises at least 100 genes listed in Table 1. In some embodiments, the plurality of genes comprises at least 150 genes listed in Table 1. In some embodiments, the plurality of genes comprises at least 200 genes listed in Table 1. In some embodiments, the plurality of genes comprises at least 250 genes listed in Table 1. In some embodiments, the plurality of genes comprises all of the genes listed in Table 1.

In some embodiments, wherein assessing the plurality of genes comprises determining a copy number for at least one of the plurality of genes. In some embodiments, assessing the plurality of genes comprises detecting the presence of one or more single nucleotide polymorphisms (SNPs) for at least one of the plurality of genes. In some embodiments, assessing the plurality of genes comprises measuring expression of a product of at least one of the plurality of genes.

In some embodiments, provided herein is a panel of genes for assessing risk of congenital heart defects. In some embodiments, the panel comprises at least 10 genes listed in Table 1. In some embodiments, the panel comprises at least 25 genes listed in Table 1. In some embodiments, the panel comprises at least 50 genes listed in Table 1. In some embodiments, the panel comprises at least 100 genes listed in Table 1. In some embodiments, the panel comprises

at least 150 genes listed in Table 1. In some embodiments, the panel comprises at least 200 genes listed in Table 1. In some embodiments, the panel comprises at least 250 genes listed in Table 1. In some embodiments, the panel comprises all of the genes listed in Table 1.

DESCRIPTION OF THE DRAWINGS

FIG. 1A (top) shows representative images of primary human embryonic heart tissue at indicated Carnegie stages (CS). **FIG. 1A** (bottom) shows the data types collected and downstream analyses performed in this study.

FIG. 1B shows the principal component analysis of genome-wide primary and imputed pchromatin immunoprecipitation followed by next-generation sequencing (ChIP-seq) signals. Each mark is indicated by separate colors. Primary samples are shown as triangles and imputed data as circles. Grouping of marks and overall function are indicated in normal and bold text, respectively.

FIG. 1C shows the total numbers of each ChromHMM (chromatin state) identified in segmentation of each individual embryonic tissue sample. Samples are ordered from left to right as the earliest to latest time points. Legend of colors is located below using conventions defined by Roadmap Epigenome.

FIG. 1D shows the average numbers of each chromatin state for all heart samples (red) and all Roadmap Epigenome samples (gray) are shown. Error bars represent SDs for each chromatin state and tissue group PC indicates principal component; PCW, postconception weeks; and PTM, post translational modification.

FIG. 2A shows the tSNE (t-distributed stochastic neighborhood embedding) projection of imputed H3K27ac P signals at 444 413 enhancer segments from tissues profiled by Roadmap Epigenome and in this study. Dots are color coded by tissue as indicated and labeled as each individual tissue samples as profiled by Roadmap Epigenome or in this study.

FIG. 2B shows the fraction of each of the 25 ChromHMM States, EMERGE, and Dickel datasets that overlap with either active heart enhancers (unshaded) or enhancers active in tissues other than heart (shaded) as tested by the Vista Enhancer Browser (enhancer.lbl.gov). Significance of difference of overlap between heart and other tissue was calculated using the Mann-Whitney U test and is shown at top (*P≤0.05, **P≤0.01, ***P≤0.001, ****P≤0.0001).

FIG. 2C shows the gene ontology enrichments for indicated functional categories for putative novel strong enhancer segments identified in human embryonic heart vs Roadmap Epigenome (n=12 395). Putative enhancers were assigned to genes and significance determined by the Genomic Regions Enrichment of Annotations Tool (GREAT). Position of each dot is based on −log10(binomial FDR) and colored by binomial fold enrichment calculated by GREAT.

FIG. 2D shows the top most significantly enriched motifs in putative EHEs calculated by Hypergeometric Optimization of Motif Enrichment (HOMER). Shown are the position weight matrix for each motif, transcription factor predicted to bind that motif, and HOMER P: HOMER known motifs (top) and de novo motifs (bottom). AF-2 indicates activation function 2; AV, atrioventricular; ESC, embryonic stem cell; ESD, ESC-derived; FDR, false discovery Rate; GI, gastrointestinal; and GO, gene ontology.

FIG. 3A shows the delineation of 3 major stages of heart development during the embryonic period based on Carnegie staging (CS).

FIG. 3B is a heat map of signal at putative enhancers differentially marked with H3K27ac.

FIG. 3C is the same as FIG. 3B but with H3K4me2.

FIG. 3D is a heat map of Z scores for level of significance of motifs enriched in each class of differentially regulated enhancers based on pairwise comparisons of replicates of H3K27ac signal at all embryonic heart enhancer segments using DiffBind. Comparisons are indicated as follows: early up vs mid (EVM), early up vs late (EVL), mid up vs early (MVE), mid up vs late (MVL), late up vs mid (LVM), and late up vs early (LVE). The more significantly enriched motifs are colored yellow.

FIG. 3E is the same as in FIG. 3D but using H3K4me2 signals.

FIG. 3F is a heat map of most variable Z scores for significance of enrichment of gene ontology categories for genes assigned a differentially activated enhancer by the Genomic Regions Enrichment of Annotations Tool. Comm indicates communication; and TGFβ, transforming growth factor-beta.

FIG. 4A is a University of California Santa Cruz (UCSC) Browser shot of *NKX2.5* gene locus showing individual embryo ChromHMM (chromatin state) annotations from this study and Roadmap Epigenome. Samples are ordered from top to bottom based on developmental age,

earliest to latest. Chromatin states are indicated by color segments using color convention from **FIG. 1C**. Strong human embryonic heart (HEH) enhancers are shown in black, and super enhancers and super enhancers unique to HEH are shown in orange.

FIG. 4B is a UCSC Browser shot of locus near the *TBX20* gene using the same conventions as in **FIG. 4A**. The region upstream of the *TBX20* gene is a human embryonic heart–specific super enhancer (orange bar). Of note are the strong HEH-specific enhancer states track, as well as the experimentally validated enhancer elements with images to the right. In the **bottom**, all the Roadmap Epigenome ChromHMM segmentations are stacked showing the region is not similarly active in any other profiled tissue.

FIG. 4C are box plots of fold enrichment of overlap of each indicated chromatin state in the human embryonic heart or brain with anchor points identified by capture Hi-C interactions in iPSC-derived cardiomyocytes over matched randomly selected segments. Solid boxes represent embryonic heart chromatin segments while dotted boxes represent adult brain chromatin segments. Significance of difference between embryonic heart and adult brain fold enrichments was calculated using the Mann-Whitney *U* test and is shown at **top** (*$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$, ****$P \leq 0.0001$). The largest increases in fold enrichments for embryonic heart were identified for strong enhancer states 13 and 14.

FIG. 5A is a scatterplot of the log2 fold enrichment and log10 Bonferroni-adjusted significance level of genome-wide association study (GWAS) variants associated with systolic blood pressure in all enhancer segments identified in the strong enhancer states for each embryonic heart sample (bright red), the total reproducible strong enhancers from the whole dataset (dark red), or other tissues in Roadmap Epigenome (blue). All values calculated using only variants with $P<5\times10^{-8}$ from GWAS catalog using GREGOR (genomic regulatory elements and Gwas overlap algoRithm).

FIG. 5B is the same as in **FIG. 5A** using GWAS variants associated with electrocardiograph traits and measures.

FIG. 5C is the same as in **FIG. 5A** using GWAS variants associated with resting heart rate.

FIG. 5D is the same as in **FIG. 5A** using GWAS variants associated with QRS complex traits.

FIG. 5E shows the enrichment of GWAS analysis *P* for atrial fibrillation in all ChromHMM (chromatin state) annotations as determined by GARFIELD (GWAS analysis of regulatory and functional information enrichment with LD correction). Scatterplot of the odds ratio of atrial fibrillation GWAS SNPS (single nucleotide polymorphism) using the 1×10^{-8} threshold by the log10 GARFIELD Bonferroni-adjusted *P*. Samples from this study (triangle symbol) and Roadmap Epigenome (star symbol) are colored by chromatin state as indicated by the color key. Atrial fibrillation shows the greatest enrichment in strong enhancers identified in embryonic heart tissues.

FIG. 5F is the same as in **FIG. 5E** using GWAS summary statistics for systemic lupus erythematosus. Lupus shows the greatest enrichment in strong enhancers identified in immune cell types sorted from blood. Lupus also shows enrichment in repressed and bivalent states in human embryonic heart.

FIG. 6A shows a heat map showing specificity of expression for 5167 genes identified with elevated Gini scores (>0.5) for 25 tissues from GTEx and embryonic heart. Brain, spleen, and embryonic heart–specific genes are identified as colored leaves on the dendrogram along the left of the plot.

FIG. 6B shows the gene ontology enrichments for genes identified as specific for heart, spleen, and embryonic heart, respectively, based on genes from indicated color-coded clusters in **FIG. 6A**.

FIG. 6C shows a heat map of Z scores of normalized gene expression for genes identified as differentially expressed in pairwise comparisons of replicates from each of Carnegie stage in our developmental series. Dendrogram on the left is hierarchical clustering of genes across a developmental series. The genes were color coded by cutting the dendrogram at a height, which would result in 4 groups. Purple most highly expressed early. Pink and green expressed most strongly in intermediate stages of the series. Blue genes are most strongly expressed at the end of the developmental series.

FIG. 6D shows the gene ontology enrichment maps from the purple (left) and blue (right) gene sets identified in **FIG. 6C**. The size of each dot represents the number of genes, and the color scale represents the –log2 transformed Benjamini and Hochberg–adjusted *P* of each ontology. Darker colors indicate higher significance. The edges connect overlapping gene sets. The location of each dot is determined by the overlap ratio (OvR) calculated by enrichplot.

Genes active early are enriched for functions related to embryonic patterning and morphogenesis, whereas genes active late in embryonic heart development are enriched for vasculature development and ion channel function.

FIG. 7A is a plot of gene expression values from embryonic heart (red), adult heart (purple), brain (green), or all other tissues (gray) for genes assigned indicated number of EHEs as determined by Genomic Regions Enrichment of Annotations Tool. Genes assigned multiple EHEs are more strongly expressed in the embryonic heart than in other tissues. Significant differences in distributions of gene expression values in each comparison were determined based on Mann-Whitney *U* test.

FIG. 7B is a histogram of distances of EHEs (red) or randomly selected sets of enhancers (gray) to the nearest heart-specific gene (Gini, >0.75) in 10-kb bins up to 100 kb. Overall EHEs are enriched near heart-specific genes over all distances up to 100 kb. Error bars indicate SD of 1000 random permutations of enhancers.

FIG. 7C is a network plot of gene modules identified by WGCNA using embryonic heart gene expression data. A Pearson correlation of the module eigenvectors was calculated for the edges. Positive correlations of ≥0.5 were included. The location of each module is determined by multidimensional scaling (MDS) of the module eigengene vectors. Modules are color coded based on names assigned by WGCNA. Size of dots indicates the number of genes in each module. Each module is labeled based on the most significant biological process category gene ontology enrichment determined by Database for Annotation, Visualization, and Integrated Discovery (DAVID). Modules are grouped based on related functional category enrichments and distance in MDS space.

FIG. 7D shows the trajectories of expression based on eigenvectors reported by WGCNA for each module across the developmental series. Groups and color coding are the same as in **FIG. 7C**. Group 1 modules have generally declining expression and include many genes involved in developmental patterning. Group 3 modules generally have increasing expression. Groups 2 and 4 have multiphasic but offset expression and contain genes involved in chromatin regulation and muscle cell differentiation and function. GTEx indicates Genotype-Tissue Expression Project.

FIG. 8A shows dot plots of gene enrichment within the WGCNA modules. The lists of genes used are curated from multiple sources, while embryonic heart–specific enhancer segments (EHEs) and Gini are from this article. The groups correspond to **FIG. 5**.

FIG. 8B shows network of multidimensional scaling coordinates and pairwise correlation scores for the violet module in group 4 in **FIG. 8D**, which is enlarged to show detail. All genes with correlation value >0.88 with any other gene are plotted. Size of shape indicates highly connected hub genes. Diamonds represent genes assigned EHEs. Purple filled shapes indicate heart-specific gene expression (Gini, ≥0.5). Hub genes are labeled with gene symbol. Genes directly positively regulated by NKX2-5 (NK2 homeobox 5) binding are indicated with yellow. Several hub genes that have all these criteria are listed in larger yellow text.

FIG. 8C is a histogram of loss-of-function observed/expected upper bound fraction (LOEUF) deciles of hub genes or randomly selected nonhub genes from all modules in the WGCNA network. Deciles range from decile 1 (d1), which represent the most constrained genes to d10, genes that are the most tolerant to putative loss-of-function (pLoF) variation. Error bars indicate SD of 1000 random permutations of nonhub genes.

FIG. 8D is a histogram of the number of gene-scrambled modules that have protein-protein interaction (ppi) enrichment at a Bonferroni-adjusted *P* of <0.05. The vertical orange line marks the 15 modules that have significant ppi in the actual WGCNA network. CHD indicates congenital heart defects; MYOZ2, myozenin 2; NCC, neural crest cell; NHE, novel heart enhancer; OR, odds ratio; and pLI, probability of loss of function intolerance.

FIG. 9 shows example images of embryos at each carnegie stage.

FIG. 10A shows the Pearson correlation of primary ChIP-Seq signals by 10kb bins across the genome showing general correlation by mark.

FIG. 10B shows the Pearson correlation of imputed ChIP-Seq signals by 10kb bins across the genome showing better correlation between marks than the primary signal.

FIG. 11 shows the individual state names/classifications and color coding for each model of the 15, 18, and 25 state models.

FIG. 12A shows the distance of segments from protein coding TSS in base pairs per state with Dickel and EMERGE.

FIG. 12B shows the length of segments per state with Dickel and EMERGE in base pairs.

FIG. 12C shows the conservation based on phyloP (phylogenetic p-values) scores per state with Dickel and EMERGE from the PHAST package for multiple alignments of 99 vertebrate genomes to the human genome (100 way). Significance of difference of scores between strong enhancer state 13 and other states along with Dickel and EMERGE were calculated using the Mann-Whitney test and is shown at top (p-value ≤ 0.05 = *, ≤ 0.01= **, ≤ 0.001= ***, ≤ 0.0001 = ****).

FIG. 12D shows the CADD scores per state with Dickel and EMERGE.

FIG. 12E shows the LINSIGHT scores per state with Dickel and EMERGE.

FIG. 12F shows the activity of overlap with MPRA for in vivo mouse cardiomyocytes. Significance of difference of activity score between repressed state 24 and other states including the Dickel and EMERGE datasets were calculated using the Mann-Whitney test and is shown at top (p-value ≤ 0.05 = *, ≤ 0.01= **, ≤ 0.001= ***, ≤ 0.0001 = ****).

FIG. 13A shows the Pearson correlation heatmap of H3K27ac signal at putative enhancer segments showing general correlation by tissue type. Red indicates greater correlation, blue lower correlation.

FIG. 13B shows the tSNE projection of imputed H3K27ac p-value signals at 444,413 enhancer segments from tissues profiled by Roadmap Epigenome, in embryonic craniofacial tissue, in embryonic heart tissue from this study, and from a study looking at sorted nuclei from heart tissue from multiple developmental stages.

FIG. 14A shows the 15 state ChromHMM model analysis of the embryonic heart data. Boxplot where solid bar indicates fraction of overlap of VISTA heart positive enhancers with each state, shaded bar indicates overlap with enhancers from VISTA positive in tissues other than heart.

FIG. 14B shows the 18-state model with the same conventions as **FIG. 14A**.

FIG. 15A is a boxplot representing the fraction of each state in our embryonic heart samples that overlaps with peaks called from EMERGE bedgraph showing the greatest overlap between EMERGE and our TSS and Promoter States (specifically States 1-3), followed by strong enhancer states (States 13-15), and transcribed and regulatory states (States 9-10). In the upper right corner is an inset of a meta-gene plot of the distribution of EMERGE peaks, which also shows that a majority of peaks can be found very close to the TSS.

FIG. 15B is a violin plot of the distribution of the scores calculated over EMERGE peaks from bedgraph signal, with again the highest concentration of higher scores being seen in the TSS and promoter states (States 1-4).

FIG. 15C are ROC curves calculated for each of our embryonic heart samples, as well as the Dickel and EMERGE datasets, showing similar or higher AUCs for our heart samples when compared to the other datasets.

FIG. 16A shows the overlap of enhancers from Dickel *et al.* by state as represented by percentage of overlap.

FIG. 16B (Left) shows the overlap of the reproducible human embryonic heart enhancers (177,412) with the Dickel compendium of heart enhancers (82,119); (Right) shows the significance of overlap, shown by 1000 iterations of overlap with shuffled coordination of the larger dataset, $p<0.001$.

FIG. 16C (Left) shows the overlap of the strong novel human embryonic heart enhancers (12,395) with the Dickel compendium of heart enhancers (82,119); (Right) shows the significance of overlap, shown by 1000 iterations of overlap with shuffled coordination of the larger dataset, $p<0.001$.

FIG. 16D (Left) shows the overlap of the strong novel human embryonic heart enhancers (12,395) with the enhancers from the Dickel compendium that show a prenatally biased score (>2 prenatal/postnatal) (9,953); (Right) shows the significance of overlap, shown by 1000 iterations of overlap with shuffled coordination of the larger dataset, $p<0.001$.

FIG. 16E (Left) shows the overlap of the strong novel human embryonic heart enhancers (12,395) with the enhancers from the Dickel compendium that came from peaks in human fetal data (5,042); (Right) shows the significance of lack of overlap, shown by 1000 iterations of overlap with shuffled coordination of the larger dataset, $p<0.001$.

FIG. 17 is a heatmap of the signal of putative enhancers differentially marked by H3K27ac and H3K4me2.

FIG. 18A (Top) shows indicates the relative effects of enhancer sequences harboring the alternate allele on luciferase gene expression on HL1 cardiomyocytes across the SCN5A/SCN10A gene locus. Each dot indicates a variant centered amplicon tested in the pGL4.23 luciferase vector ordered by genomic position as tested by Kapoor *et al.* PNAS 2019. Bottom shows individual embryo chromatin state annotations from this study across the

SCN5A/SCN10A gene locus. Samples are ordered from top to bottom based on developmental age, earliest to latest. Chromatin states are indicated by color segments using color convention from **FIG. 1C**. Below chromatin state segmentations is stranded RNA-Seq data from representative CS23 heart sample indicating robust SCN5A expression and virtually no expression from SCN10A. Super enhancers are indicated by labeled orange bars. Variants rs41312411 associated with establishment of resting heart rate and P wave duration, rs3922844 associated with establishment of electrocardiogram traits and measures, and rs11708996 associated with Brugada syndrome are indicated in dashed box roughly encompassing a 13kb LD block (R 2 ≥0.95, D'=1, 1000G CEU). The former variant directly overlaps a strong embryonic heart enhancer segment and an experimentally validated heart enhancer, hs2177, (in dark orange) as shown in panel to right.

FIG 18B shows the UCSC browser shot of Hand2 gene locus, conventions are the same as in **FIG. 18A**.

FIG. 18C shows the UCSC browser shot of MYOCD gene locus, conventions are the same as in **FIG. 18A**.

FIG. 19A shows the scatterplot of the log2 fold enrichment and log10 Bonferroni adjusted significance level of GWAS variants associated with atrial fibrillation in all enhancers segments identified in the strong enhancer states for each embryonic heart sample (bright red), the total reproducible strong enhancers from the whole dataset (dark red) or other tissues in Roadmap Epigenome (blue). All values calculated using only variants with p values < 5x10-8 from GWAS Catalog using GREGOR.

FIG. 19B uses the same conventions as **FIG. 19A**, for QT Interval associations.

FIG. 19C uses the same conventions as **FIG. 19A**, for Chrons associations.

FIG. 19D uses the same conventions as **FIG. 19A**, for Congenital and Conotruncal associations.

FIG. 19E shows a Scatterplot of the odds ratio of Lupus GWAS SNPS using the 1E--8 Threshold by the log10 GARFIELD Bonferroni adjusted p-values. Roadmap blood and immune cell types (triangle symbol) and non-immune cell types (star symbol) are colored by chromatin state as indicated by the color key. Significant enrichments in immune cell types are labeled.

FIG. 19F shows the enrichment of GWAS analysis p-values for atrial fibrillation in strong enhancer state annotations (E13_EnhA1, E14_EnhA2, E15_EnhAc) as determined by

GARFIELD. Lines radiating from the center of the plot show odds ratios at eight indicated GWAS p-value thresholds from full summary statistics. Dots at innermost ring of the outer circle indicate significant enrichment for GWAS variants with p-values < 10-8 while dots at the outermost ring with p-values < 10-5. Samples from this study and Roadmap epigenome are sorted by tissue as indicated by color key. Atrial fibrillation shows greatest enrichment in strong enhancers identified in embryonic heart tissues.

FIG. 19G is the same as in **FIG. 19E** using GWAS summary statistics for systemic lupus erythematosus. Lupus shows greatest enrichment in strong enhancers identified in immune cell types sorted from blood.

FIG. 19H is the same as **FIG. 19E**, for Resting Heart Rate.

FIG. 19I is the same as **FIG. 19E**, for QRS Interval.

FIG. 19J is the same as **FIG. 19E**, for P-wave duration.

FIG. 19K is the same as **FIG. 19F**, for Crohns.

FIG. 20A shows the tSNE plot using all genes annotated by Gencode (v25) quantified by Rail-RNA form this study or GTEx and retrieved from recount2. All samples profiled in this study (red) cluster well with one another relative to other human tissues including adult heart (purples) and brain (greens).

FIG. 20B shows box plots of expression values based on log10 scaled counts for samples from embryonic heart (red), adult heart (purple), and adult brain (green). Genes shown are those identified with a tissue specificity score (GINI) >= 0.9 and highest expression in the embryonic heart.

FIGS. 20C–E show PCA plots for human embryonic heart RNA-Seq Data. Colored and labeled by for C. CS Stage. D. Gender E. RIN score of RNA.

FIG. 21A shows a heatmap of all genes found to be differentially expressed across the time series.

FIG. 21B is a horizontal bar plot of the number of genes differentially expressed across time series. The blue bars indicate downregulated genes and pink for upregulated. Note the early time points have the most amount of differentially expressed genes when compared to the later time points.

FIG. 22A shows the gene ontologies for the CS16 (pink cluster, left panel) time point. The gene ontologies for the CS17 (green cluster, right panel) time periods.

FIG. 22B shows the violin plot of the expression of genes with greater than 15 embryonic heart-specific enhancers in embryonic heart samples, as well as all other GTEx tissues. Significance of differences were calculated using the Mann-Whitney test and are shown at top (p-value ≤ 0.05 = *, ≤ 0.01= **, ≤ 0.001= ***, ≤ 0.0001 = ****).

FIG. 23A inset shows the cyclical pattern of gene expression identified for Group 2 genes from WGCNA; bars indicate normalized enrichment scores for significantly enriched functional categories from gene set enrichment analysis based on comparisons of gene expression between CS16 and CS18 (left) or CS18 and CS20 (right). For all but one category showing enrichment (platelet morphogenesis) inverse trends are shown between CS16 vs CS18 vs CS20.

FIGS. 23B-C show the enrichment plots of genes across heart valve development category based on ranked order from pairwise comparisons between CS16 and CS18 (**FIG. 23B**) or CS18 and CS20 (**FIG. 23C**).

FIG. 24 shows the enrichment of curated gene lists in WGCNA of human time-series brain RNA-seq. The enrichment of heart specific lists (EHE, GINI) do not line up with cardiomyocytes like was observed for the embryonic heart network. Also, there is significant enrichment of embryonic heart specific genes in their null module (unassigned, grey). As expected, a subset of modules get enrichment for constrained genes (LoF d1/d2, pLI).

FIG. 25 shows the full violet module network.

FIG. 26A shows the brown WGCNA module network.

FIG. 26B shows the grey60 WGCNA module network.

FIG. 26C shows the mediumpurple3 WGCNA module network.

FIG. 26D shows the green WGCNA module network.

DEFINITIONS

Although any methods and materials similar or equivalent to those described herein can be used in the practice or testing of embodiments described herein, some preferred methods, compositions, devices, and materials are described herein. However, before the present materials and methods are described, it is to be understood that this invention is not limited to the particular molecules, compositions, methodologies or protocols herein described, as these may vary in accordance with routine experimentation and optimization. It is also to be understood that

the terminology used in the description is for the purpose of describing the particular versions or embodiments only, and is not intended to limit the scope of the embodiments described herein.

Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. However, in case of conflict, the present specification, including definitions, will control. Accordingly, in the context of the embodiments described herein, the following definitions apply.

As used herein and in the appended claims, the singular forms "a", "an" and "the" include plural reference unless the context clearly dictates otherwise. Thus, for example, reference to "a peptide amphiphile" is a reference to one or more peptide amphiphiles and equivalents thereof known to those skilled in the art, and so forth.

As used herein, the term "comprise(s)" and linguistic variations thereof denote the presence of recited feature(s), element(s), method step(s), etc. without the exclusion of the presence of additional feature(s), element(s), method step(s), etc. Conversely, the term "consisting of" and linguistic variations thereof, denotes the presence of recited feature(s), element(s), method step(s), etc. and excludes any unrecited feature(s), element(s), method step(s), etc., except for ordinarily-associated impurities. The phrase "consisting essentially of" denotes the recited feature(s), element(s), method step(s), etc. and any additional feature(s), element(s), method step(s), etc. that do not materially affect the basic nature of the composition, system, or method. Many embodiments herein are described using open "comprising" language. Such embodiments encompass multiple closed "consisting of" and/or "consisting essentially of" embodiments, which may alternatively be claimed or described using such language.

For the recitation of numeric ranges herein, each intervening number there between with the same degree of precision is explicitly contemplated. For example, for the range of 6-9, the numbers 7 and 8 are contemplated in addition to 6 and 9, and for the range 6.0-7.0, the number 6.0, 6.1, 6.2, 6.3, 6.4, 6.5, 6.6, 6.7, 6.8, 6.9, and 7.0 are explicitly contemplated.

The term "congenital heart defect" or "CHD" as used herein refers to an abnormality in the heart that is present in an infant at birth. A congenital heart defect may affect the structure of the heart and/or affect the way the heart operates.

The term "infant" when used herein in reference to a human subject refers to a subject less than 1 year of age. The term "child" as used herein in reference to a human subject refers to a subject between 1-17 years of age. The term "adult" when used herein in reference to a human subject refers to a subject that is 18 years of age or older.

As used herein, the term "sample" is used in the broadest sense and is inclusive of many sample types that may be obtained from the subject. Samples may be obtained from animals (including humans) and encompass fluids (*e.g.,* urine, blood, blood products, sputum, saliva, etc.), solids, tissues, and gases. In some embodiments, the sample is a blood sample, a serum sample, or a plasma sample.

The term "single nucleotide polymorphism" or "SNP" refers to a variation in the sequence of a gene in the genome of a population that arises as the result of a single base change, such as an insertion, deletion or, a change in a single base.

The term "subject" refers to any vertebrate, including, but not limited to, a mammal (*e.g.*, cow, pig, camel, llama, horse, goat, rabbit, sheep, hamsters, guinea pig, cat, dog, rat, and mouse, a non-human primate (for example, a monkey, such as a cynomolgous or rhesus monkey, chimpanzee, etc.) and a human). In some embodiments, the subject may be a human or a non-human. In some embodiments, the subject is an infant. In some embodiments, the subject is a child. In some embodiments, the subject is an adult. The subject may be suspected as having a congenital heart defect.

"Treat," "treating," or "treatment" are each used interchangeably herein to describe reversing, alleviating, or inhibiting the progress of a disease and/or injury, or one or more symptoms of such disease, to which such term applies. Depending on the condition of the subject, the term also refers to preventing a disease, and includes preventing the onset of a disease, or preventing the symptoms associated with a disease. A treatment may be either performed in an acute or chronic way. The term also refers to reducing the severity of a disease or symptoms associated with such disease prior to affliction with the disease. Such prevention or reduction of the severity of a disease prior to affliction refers to administration of a pharmaceutical composition to a subject that is not at the time of administration afflicted with the disease. "Preventing" also refers to preventing the recurrence of a disease or of one or more symptoms associated with such disease.

DETAILED DESCRIPTION

Congenital heart defects (CHDs) are among the most common birth defects, affecting ≈1% of live births worldwide and remain the leading cause of infant mortality in developed nations. Despite evidence suggesting a strong genetic component, ≈60% of CHD cases remain unexplained. Environmental causes, familial forms of CHD, and de novo damaging mutations in over 400 genes each explain <10% of cases, while chromosomal abnormalities including aneuploidies and large structural variations are implicated in over 20% of CHD. These findings suggest that CHD arises through combinations of otherwise benign mutations in a large number of genes, unappreciated genetic-environmental interactions, or disruption of regulatory sequences that control heart development. There are indications that regulatory regions are causative for CHDs. Patients homozygous for rare variation in a heart-specific regulatory sequence controlling the cardiac TF (transcription factor) TBX5 (T-box TF 5) have isolated CHD.

Recent analysis of gene expression from the heart at multiple stages of human development indicates the dynamics of gene expression occur primarily during the embryonic period of human development or the first 8 postconception weeks, aligning well with known structural and functional changes in the developing heart. Due to the limited nature of embryonic and fetal tissues and difficulty with performing functional genomics on small amounts of primary material, several groups have used directed differentiation approaches in human and mouse stem cells to model cardiogenesis, providing insight into global binding of multiple cardiac TFs, conserved cardiac regulatory networks, and cooperativity among these specific TFs. However, these data have not been systematically compared with human cardiomyocytes, and it remains unclear how faithfully these systems reflect normal human heart development at the gene regulatory level. Regulatory annotation of the human genome from the ENCODE Project and Roadmap Epigenome includes primary tissues of the human heart. However, these data are exclusively from fetal and adult stages after these major molecular and morphological changes have occurred. These two large consortia used chromatin immunoprecipitation followed by next-generation sequencing (ChIP-seq) of several histone modifications in combination with machine learning approaches in over 100 tissues and cell types. This approach resulted in the

learning of several hidden Markov models of ChromHMM (chromatin state) using the combined tissue dataset.

The number of states, and, therefore, the complexity of the model, depends upon the histone modifications used. A 15-state model can be learned from the histone modifications H3K4me1, H3K4me3, H3K9me3, H3K27me3, and H3K36me3 (frequently associated with active enhancers, active promoters, stable heterochromatin, facultative heterochromatin and active transcription, respectively). When H3K27ac (associated with active chromatin with functions ranging from active transcription to active enhancers) is included, an increasingly complex model of 18 different chromatin states can be called. The addition of five other histone modifications (H2A.Z [histone 2A variant Z], H3K4me2, H3K9ac, H3K79me2, H4K20me1) and DNA accessibility (DNAse hypersensitivity) provide the ability to call 25 chromatin states. Through the use of such an ensemble, multitissue approach, various types of active and repressed features are mapped onto the genome and cell type–specific or stage-specific enhancers can be readily distinguished.

Thus, provided herein is the systematic characterization of chromatin states from human embryonic hearts at 4 to 8 postconception weeks using hidden Markov models of chromatin state in a manner that allows for direct comparison with over 100 tissues profiled by Roadmap Epigenome. Herein, it was confirmed previously validated *in vivo* heart enhancers and identified thousands of previously unknown embryonic heart–specific regulatory sequences. Integration of chromatin and gene expression through coexpression analysis identified groups of genes that are coordinately expressed during early heart development and likely regulated by these heart-specific enhancers. The present multilayered analysis also predicts new genes to be involved in CHD.

Thus, in some aspects, provided herein are methods of identifying a subject as at risk of having a congenital heart defect. In some embodiments, the method of identifying a subject as at risk of having a congenital heart defect comprise assessing a plurality of genes in a sample obtained from the subject. In some embodiments, the plurality of genes are selected from the genes listed in Table 1. In some embodiments, the plurality of genes comprise at least 5 genes listed in Table 1. For example, the plurality of genes may comprise at least 5, at least 10, at least

15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 60, at least 70, at least 80, at least 90, at least 100, at least 120, at least 130, at least 140, at least 150, at least 160, at least 170, at least 180, at least 190, at least 200, at least 210, at least 200, at least 230, at least 240, or at least 250 genes listed in Table 1. In some embodiments, the plurality of genes comprises all of the genes listed in Table 1.

Table 1. Genes that may be assessed to determine risk of CHD

MED13 NCKAP1 COL5A1 TNPO1 ANKRD17 TJP1 PTCH1 CHD7 GIGYF2 PHF21A DYNC1H1 SMARCAD1 BRWD3 TRIM33 MN1	UBP1 SPTBN1 PREX1 MED14 PLEKHA5 COL1A1 ANK2 IPO7 ZEB2 PHIP ADGRL2 AGPS PRKAR1A ZMYND11 ATP1A1 NCOA2	IPO5 TEAD1 SOBP RASA1 CDC5L UBR3 MLLT3 NPAS2 MAPKAPK2 MEIS1 UBE2O CUL1	HOMER1 BAZ2A SP3 PKN2 ZNF609 ARHGAP42 ATAD2B DENND5B XPR1 HIPK2 SAP130 SMAD5 PPP4R2 OPHN1 CHD1 FAM120A ZNF608	RPL5 ANK1 DYRK2 CHD9 KIF1B TRIM71 TET1 CELF2 CLASP1 EXOC5 PDS5B PTPRT FAT4 CTNND1	GRIN2A ASAP1 YAP1 CEP170 EBF2 ROCK1 XRN1 RAPGEF2 NFIA JPH3 COL1A2 SMARCA2 UBE3A HNRNPR ASTN1 FRY MAP3K1 STRBP RXRB KDM3A
PRKAR2B MGAT1 PROX1 ZBTB38 CLCN3 DNAJC6 CREB1 IREB2 NFIB MIER3 RAD23B PIK3R1 REL DENND1A	SLC2A1 SHH DCHS1 NPTN HDAC9 HGF RYR2 STT3B ZC3H12C AKAP6 NEGR1 ADCY5 EDNRA LRP12	ST7 ERBB4 AMOT PPP1R13B SRPK2 GRHL2 TRIM2 TGFB3 SLC8A1 FBXO28 FRYL RBPMS TAOK1 VWC2	NRIP1 KCND3 COL12A1 ATF2 RIMS1 PHF3 MYO18A KLF12 BAZ2B NRK NF1 ARNT2 AR KIAA1549	NKX2-3 STAU2 ELOVL6 SPRED1 NR3C1 THAP12 BNC1 ABLIM1 MAGT1 NKX2-5 DCLK2 SYT7 SPECC1L ZMYM2	ANGPT1 GFPT1 PLCH1 LDB2 SIK1 HMBS COPZ1 MTUS2 OXR1 DHX36 PER2 THBS2 EHBP1 NR3C2

STK39 GNL3L YIPF5 CDYL AZIN1 EVL TLE1 ACTN2 SH3GLB1	HECW2 NRXN2 SMAD2 HMGCR DAAM1 RPS6 HUNK ARHGAP5	KDM1A PDZD2 SLC39A10 CNP	ENSG00000273217 FOXF1 CHMP4B PCGF5 SEL1L EEA1 MET	HK1 RASAL2 RBFOX1	ATXN1L LRRC4C PTPRK CDC40 NNT ABCE1 G2E3
GRIP1 NXPH2 PPP2R3A SNX13 RAB11FIP2	FAM13B STRN DMTN NECTIN3	HIPK3 SRSF6 BBX TFAP2C MAP4K3 CD47 ANLN	SECISBP2L FHOD3 SNAP23 ST14 CLASRP MAP7 TMEM204 EPAS1	MAST4 LGR4 ZHX1 RPL27A PMP22 NEK7 NDUFS7 APBA1 TMEM135 ONECUT2 RTN4 CLINT1 CAP2	HEY2 PTOV1 MEX3C C1orf21 SLC25A12 NEDD1 ACO2 MYO6 PCDH7

In some embodiments, the method further comprises identifying the subject as at risk of having a congenital heart defect based upon the assessment of the plurality of genes. The term "assessing", or "assessment" is used in the broadest sense and is inclusive of many types of evaluations that may be performed that are indicative of gene expression and/or mutation(s) in a gene. For example, in some embodiments assessing the plurality of genes comprises determining a copy number for at least one of the plurality of genes. The subject may be identified as at risk of having a congenital heart defect based upon the assessment of the copy number for the at least one gene. In some embodiments, the copy number for at least one gene, at least 2 genes, at least 3 genes, at least 4 genes, at least 5 genes, at least 6 genes, at least 7 genes, at least 8 genes, at least 9 genes, at least 10 genes, or more than 10 genes in the panel of genes (*e.g.* at least 10, at least 25, at least 50, etc.) is assessed and used to determine the risk of having a congenital heart defect. In some embodiments, the subject may be identified as at risk of having a congenital heart defect if the copy number for the at least one gene is increased. For

example, the gene may be associated with increased risk of congenital heart defects, and therefore an increased copy number of the gene may identify the subject as at risk of having a congenital heart defect. In some embodiments, the subject may be identified as at risk of having a congenital heart defect if the copy number for the at least one gene is decreased. For example, the gene may be associated with protection from (*e.g.*, resistance to) congenital heart defects, and therefore a decreased copy number or the absence of a copy of the gene may indicate that the subject is at risk of having a congenital heart defect.

In some embodiments, assessing the plurality of genes comprises detecting the presence of one or more single nucleotide polymorphisms (SNPs) for at least one of the plurality of genes. The subject may be identified as at risk of having a congenital heart defect if at least one gene contains a SNP. For example, the presence of one or more SNPS for at least one gene, at least 2 genes, at least 3 genes, at least 4 genes, at least 5 genes, at least 6 genes, at least 7 genes, at least 8 genes, at least 9 genes, at least 10 genes, or more than 10 genes in the panel of genes (*e.g.* at least 10, at least 25, at least 50, etc.) may be assessed to determine the risk of a subject having a CHD.

Any suitable method for assessing the plurality of genes (*e.g.* assessing copy number variation, assessing SNPs, measuring expression of a gene product, etc.) may be used. Suitable methods for assessing a gene (*e.g.* for number variants, SNPs, etc.) include, for example, hybridization methods, polymerase chain reaction (PCR)-based methods, and sequencing methods.

The term "polymerase chain reaction" ("PCR") refers to the method of K.B. Mullis U.S. Patent Nos. 4,683,195, 4,683,202, and 4,965,188, that describe a method for increasing the concentration of a segment of a target sequence in a mixture of genomic or other DNA or RNA, without cloning or purification. This process for amplifying the target sequence consists of introducing a large excess of two oligonucleotide primers to the DNA mixture containing the desired target sequence, followed by a precise sequence of thermal cycling in the presence of a DNA polymerase. The two primers are complementary to their respective strands of the double stranded target sequence. To effect amplification, the mixture is denatured and the primers then annealed to their complementary sequences within the target molecule. Following annealing, the

primers are extended with a polymerase so as to form a new pair of complementary strands. The steps of denaturation, primer annealing, and polymerase extension can be repeated many times (*i.e.*, denaturation, annealing and extension constitute one "cycle"; there can be numerous "cycles") to obtain a high concentration of an amplified segment of the desired target sequence. The length of the amplified segment of the desired target sequence is determined by the relative positions of the primers with respect to each other, and therefore, this length is a controllable parameter. By virtue of the repeating aspect of the process, the method is referred to as the "polymerase chain reaction" ("PCR"). Because the desired amplified segments of the target sequence become the predominant sequences (in terms of concentration) in the mixture, they are said to be "PCR amplified" and are "PCR products" or "amplicons." Those of skill in the art will understand the term "PCR" encompasses many variants of the originally described method using, *e.g.*, real time PCR, nested PCR, digital PCR, droplet-digital PCR, reverse transcription PCR (RT-PCR), single primer and arbitrarily primed PCR, *etc.*

The term "sequencing" as used herein is used in the broadest sense and is inclusive of many types of methods that may be performed to determine a nucleic acid sequence. Sequencing may be performed to determine a DNA sequence. Sequencing may be performed to determine an RNA sequence. Sequencing methods generally involve amplifying the target sequence (*e.g.* by PCR, as described above), purifying the amplicon, sequencing the amplicon, and analyzing the sequence to detect. RNA sequencing methods involve a first step of converting the desired RNA into complementary DNA fragments (*e.g.* cDNA), prior to amplifying and isolating the desired amplicon. Analyzing the sequence may identify, for example, copy number variants or single nucleotide polymorphisms present for a desired sequence. There are many examples of sequencing methods that may be performed, including Sanger sequencing methods, capillary electrophoresis methods, or next generation sequencing (NGS) methods, also referred to as massively-parallel sequencing, sequencing-by-synthesis, real-time (*e.g.*, single-molecule) sequencing, bead emulsion sequencing, pyrosequencing, nanopore sequencing, etc.

The term "hybridization methods" refers to a variety of methods that involve the use of probes (*e.g.*, DNA probes) that are complementary to a given SNP site. For example, hybridization methods may involve the use of a labeled probe, which binds to a given SNP site and thereby gives a signal indicating the presence of a given SNP in a sample. In some

embodiments, the hybridization method is dynamic allele-specific hybridization. In some embodiments, the hybridization is an array, such as a high-density oligonucleotide SNP array. The use of such an array allows for the investigation of multiple SNPs simultaneously.

In some embodiments, assessing the plurality of genes comprises measuring expression of a gene product. For example, assessing the plurality of genes comprises measuring expression of a gene product, such as a protein, and determining the risk of having a congenital heart defect based upon the expression of said gene product. For example, in some embodiments the method comprises assessing the quantity or amount of a given protein in a biological sample, and using the amount of protein to infer the expression of a gene or the presence of one or more hyperactive (*e.g.*, activity increasing) or hypoactive (*e.g.* activity decreasing) mutations in the gene. Such information may thereby be used to assess the risk of CHD in the subject.

In some embodiments, a single assessment type is performed to determine the risk of the subject having a congenital heart defect. In other embodiments, multiple assessment types are performed to determine the risk of the subject having a congenital heart defect. For example, assessment of both copy number variation in one or more genes, and single nucleotide polymorphisms in one or more genes, may be performed to determine the risk of CHD in the subject.

The sample may be any suitable sample obtained from the subject. Suitable sample types include, for example, biological fluids (*e.g.*, urine, blood, serum plasma, sputum, saliva, etc.), solids, tissues, and gases. In some embodiments, the sample is a urine sample. In some embodiments, the sample is a blood sample. In some embodiments, the sample is a serum sample. In some embodiments, the sample is a plasma sample. In some embodiments, the sample is a saliva sample.

In some embodiments, the subject is a human subject. In some embodiments, the subject is an infant. In some embodiments, the subject is a child. In some embodiments, the subject is an adult. For example, the subject may be a human infant or a human child suspected of having a congenital heart defect based upon the presence of one or more symptoms indicative of a potential CHD. The term congenital heart defect is inclusive of many types of defects, including Atrial Septal Defect, Atrioventricular Septal Defect, Coarctation of the Aorta, Double-outlet

Right Ventricle, d-Transposition of the Great Arteries, Ebstein Anomaly, Hypoplastic Left Heart Syndrome, Interrupted Aortic Arch, Pulmonary Atresia, Single Ventricle, Tetralogy of Fallot, Total Anomalous Pulmonary Venous Return, Tricuspid Atresia, Truncus Arteriosus, and Ventricular Septal Defect. Signs and symptoms for CHDs depend on the type and severity of the particular defect. Some defects might have few or no signs or symptoms. Others might cause a baby to have blue-tinted nails or lips, fast or troubled breathing, tiredness when feeding, and/or sleepiness. In some embodiments, the assessment of risk of having a CHD performed using the panels and/or methods described herein may be corroborated by additional investigations of the infant. For example, additional procedures such as monitoring any one of the above signs/symptoms of CHD may be performed. As another example, imaging procedures may be performed to evaluate heart contractions, rhythms, blood flow, etc.

In some embodiments, provided herein is a panel of genes. In some embodiments, provided herein is a panel of genes for assessing risk of congenital heart defects in a subject. In some embodiments, the panel comprises at least 10 genes listed in Table 1. For example, in some embodiments the panel comprises at least 10, at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 60, at least 70, at least 80, at least 90, at least 100, at least 120, at least 130, at least 140, at least 150, at least 160, at least 170, at least 180, at least 190, at least 200, at least 210, at least 200, at least 230, at least 240, or at least 250 genes listed in Table 1. In some embodiments, the panel comprises all of the genes listed in Table 1. The panel may be used to assess the risk of congenital heart defects in a subject. For example, the panel may be used to assess the risk of congenital heart defects in the subject by any of the assessments performed above. For example, the panel may be used to assess the risk of congenital heart defects in a subject by assessing one or more genes in the panel for changes in copy number. As another example, the panel may be used to assess the risk of CHD in the subject by assessing the one or more genes in the panel for mutations, such as SNPs. As described above, multiple assessment types (*e.g.*, copy number variation, mutations such as SNPs, etc.) may be performed using the panel to evaluate risk of CHD.

In some embodiments, the methods for assessing the risk of congenital heart defects may further comprise providing an appropriate treatment to the subject, if the subject is determined as having a high risk of a congenital heart defect. Suitable treatments for congenital heart defect

may depend on the nature and/or severity of the defect, along with the subject's age, size, and general health. In some embodiments, treatment includes surgical procedures, cardiac catheterizations, and heart transplant. In some embodiments, treatment includes therapeutic agents, such as anticoagulants, agents to control blood pressure, agents to control heart rate/rhythm, etc. Selection of the appropriate treatment may be controlled by a physician, such as a physician having knowledge of the assessments performed using any of the methods described herein.

In some embodiments, provided herein are kits. The kits may be used to assess the panel of genes described herein. The kits may be used to perform any of the methods for assessing risk of congenital heart defect in a subject as described herein. The kit comprises at least one component for assessing risk of congenital heart defect. For example, the kit may comprise components for a hybridization based assay, sequencing, a PCR-based assay, etc. to assess a plurality of genes or gene products in a sample (*e.g.* primers, probes, labels, buffers, enzymes, plates, tubes, etc.). The kit may further comprise a means for obtaining and/or storing a sample, or components for processing a sample (*e.g,.* sample collection tubes, sample storage tubes, preserving agents, buffers, etc.). In some embodiments, the kit further comprises instructions for assessing the plurality of genes in sample. For example, the kit may comprise instructions for performing a sequencing based method, a PCR based method, or a protein expression based method for evaluating the plurality of genes. Instructions included in kits can be affixed to packaging material or can be included as a package insert. While the instructions are typically written or printed materials they are not limited to such. Any medium capable of storing such instructions and communicating them to an end user is contemplated by this disclosure. Such media include, but are not limited to, electronic storage media (*e.g.*, magnetic discs, tapes, cartridges, chips), optical media (*e.g.*, CD ROM), and the like. As used herein, the term "instructions" can include the address of an internet site that provides the instructions.

EXAMPLES

Example: Epigenomic and Transcriptomic Dynamics During Human Heart Organogenesis

In this example, the systematic characterization of chromatin states from human embryonic hearts is described at 4 to 8 postconception weeks using hidden Markov models of chromatin state in a manner that allows for direct comparison with over 100 tissues profiled by Roadmap Epigenome. Previously validated in vivo heart enhancers were confirmed and thousands of previously unknown embryonic heart–specific regulatory sequences were identified. Integration of chromatin and gene expression through coexpression analysis identified groups of genes that are coordinately expressed during early heart development and likely regulated by these heart-specific enhancers.

Methods

Data Availability

The anonymized data have been made publicly available at the Gene Expression Omnibus and can be accessed at via accession numbers GSE137731 (ChIP-seq) and GSE138799 (RNA sequencing [RNA-seq]).

Preparation of Human Tissue for ChIP-Seq and RNA-Seq

Human embryonic craniofacial tissue was collected, staged, and provided by the Joint MRC/Wellcome Trust Human Developmental Biology Resource (www.hdbr.org). Tissues were flash frozen upon collection and stored at -80℃. For both ChIP-seq and RNA-seq frozen human embryonic heart tissues frozen were removed from tubes to a petri dish with cold PBS using 1ml of cold PBS and a cut pipette tip. Hearts were photographed from at least two aspects and the tube with embryo ID photographed for records. Hearts were homogenized by mechanical disruption and divided between samples for RNA-seq and ChIP-seq as necessary. For ChIP-seq, tissues were fixed by incubation in 1% formaldehyde for 15 minutes at room temperature with agitation before being quenched by addition of 2.5M glycine to a concentration of 150mM with rotation/agitation for 10 minutes. Tissues were spun down, washed once with PBS, and placed in

a dry ice- ethanol slurry to flash freeze. For RNA-Seq homogenized tissue was added to Qiazol (Qiagen) in a non-stick 1.5ml eppendorf tube, inverted to mix and placed in a dry ice- ethanol slurry to flash freeze.

ChIP-Seq

Fixed cells pellets were processed for ChIP as previously described.105 Briefly, samples were thawed in 1 mL of 1x Cell Lysis buffer and incubated on ice for 20 minutes. Cells were lysed with dounce homogenization and nuclei were collected by centrifugation (5 min, 2500g, 4°C). Nuclei were resuspended in 300 µL of 1x Nuclear Lysis buffer + 0.3% SDS + 2 mM sodium butyrate and incubated on ice for 20 minutes. Chromatin was sheared with a Qsonica Q800R1 sonicator system operating at amplitude 23 and 2°C for 30 minutes (10 seconds duty, 10 seconds rest). Samples were cleared by centrifugation (5 min, 20,000g, 4°C) and soluble chromatin was transferred equally into seven separate tubes with 10% reserved as an input control. SDS concentration was reduced to 0.18% with ChIP-seq Dilution buffer. Protein G Dynabeads (ThermoFisher) separately preloaded with 2.5–5 µg of antibodies were added to each chromatin aliquot. Antibodies used in this study were as follows: anti-H3K27ac (C15410196, Diagenode), anti-H3K4me1 (C15410194, Diagenode), anti-H3K4me2 (ab7766, Abcam), anti-H3K4me3 (C15410003, Diagenode), anti-H3K27me3 (C16410195, Diagenode), anti-H3K9me3 (C15410193, Diagenode) and anti-H3K36me3 (C15410192, Diagenode). All Diagenode antibodies came pre-validated for ChIP, and the antibody from abcam (H3K4me2) was validated using Absurance H3 Histone Peptide Array (16-667, Millipore). ChIP samples were incubated overnight at 4°C on a rotisserie. The chromatin was then immunoprecipitated on a magnet and the supernatant was discarded. Beads were washed 8 times with 1 mL of 500 mM LiCl ChIP-Seq Wash Buffer and once with 1 mL of TE. Chromatin was eluted from the beads twice with ChIP Elution buffer at 65°C for 10 minutes with constant agitation. Combined eluates for each ChIP were subjected to crosslink reversal overnight at 65°C. Samples were then sequentially treated with RNAse A and proteinase K, purified with a PCR Purification Kit (Qiagen), and eluted in 40 uL of EB. ChIP samples were then quantified with picoGreen (ThermoFisher) and ChIP-seq libraries were prepared (SMARTer® ThruPLEX® DNA-seq 48S Kit, R400427, Takara Bio USA), then quantified by qPCR (NEBNext Library Quant Kit for Illumina E7630L), multiplexed, and sequenced for 75 cycles across multiple flow cells on an

Illumina NextSeq 500 instrument using a NextSeq 500/550 High Output v2 kit (75 cycles, Cat No. FC-404-2005).

ChIP-Seq Data Analysis

Quality control was performed on ChIP-seq reads using FastQC (version [v.] 0.11.5) and MultiQC (v.1.1). Trimming for adapters, quality and length was performed using Trimmomatic (v.0.36) for single end data. ChIP-seq reads were aligned to the human genome (hg19) using Bowtie2 (v. 2.2.5). Fragment sizes of each library were estimated using PhantomPeakQualTools (v.1.14). P value-based signal tracks were generated relative to appropriate input controls based on estimated library fragment size using MACS2 (2.1.1.20160309). Bedgraph files for all p-value signals from primary ChIP-Seq data were converted to 25 bp resolution and processed for model training and generation of imputed signals for all samples using ChromImpute (v1.0.3) as previously described.37 Resulting imputed signal tracks were converted to bigWig format for display in UCSC genome browser and converted for use with ChromHMM (v1.12), using ChromImpute's ExportToChromHMM. Signal files for individual chromosomes for each epigenome were binarized and segmentation was performed using the previously published 25-state chromatin models using ChromHMM as previously described. Following segmentation, annotation of states and generation of genome browser files was performed based on annotations provided by Roadmap Epigenome. The accession number for the ChIP-seq signals, imputed signal files and chromatin state segmentations reported in this paper is GEO: GSE137731.

Global Multi-tissue Comparisons of ChIP-Seq Signals

To qualitatively assess the similarity per mark of human embryonic heart ChIP-signals generated and imputed in this paper across all time points, primary and imputed signals were extracted from all samples for all marks in 10kb bins across the genome using the multiBigwigSummary command from DeepTools (v2.0)110 excluding all regions blacklisted by ENCODE, and outputting raw counts. The PCA plot was made using the prcomp() function on the transformed matrix of these counts using the built-in R stats (v3.4.1) package. To qualitatively assess similarity of human embryonic heart samples generated here to other tissues throughout the human body, we assembled a list of all enhancer states (states 13 through 18) from 127 tissues profiled by Roadmap Epigenome, craniofacial samples previously profiled from

our laboratory and those identified from human embryonic heart profiled here. Next, the imputed H3K27ac signals were extracted from all samples at all enhancer regions using the multiBigwigSummary command from DeepTools (v2.0) excluding all regions blacklisted by ENCODE. The resulting signal matrix were filtered to remove regions were signal was low (>10) across all samples (n = 163) and log10 transformed. This transformed matrix was used to calculate the Euclidean distance between each sample. The resulting distance matrix was then processed for t-Distributed Stochastic Neighbor Embedding using the Rtsne package (v0.15, https://github.com/jkrijthe/Rtsne) using options "is_distance = true, perplexity = 10, theta = 0.5, dims = 2, max_iter = 1000". The x and y dimensions were combined with sample and group labels for plotting with ggplot2 in R. Also identified were the super-enhancer regions using H3K27ac ChIP-Seq reads at all heart enhancer segments with default parameters in ROSE.50

Differential Regulatory Site Activation and Motif Enrichment

To identify putative regulatory elements that are differentially utilized between Early (C13), Mid (CS16 and CS17) and Late (CS23) human embryonic heart samples, H3K27ac and H3K4me2 signals were compared at enhancer chromatin state segmentations independently using DiffBind (V2.6.6) in R (V3.4.1).111 For a specific chromatin signal, uniquely aligned reads from two to four replicates of each time period were quantified and normalized for input signal at enhancer segments (states 13 through 18 from 25 state model) using fragment sizes determined by phantompeakqualtools (v.1.14)107, and the DBA_SCORE_TMM_MINUS_FULL_CPM function of DiffBind. Differential signals were determined by DiffBind using DESeq2 and filtered for a false discovery rate less than 0.1. Known motif enrichment in differentially activated regions for each histone modification were determined using HOMER with the options "-size given -len 8,10,12,14 -mask -gc" (v4.9).112 Resulting HOMER output files were loaded into R using homerkit (https://github.com/slowkow/homerkit) and -log10 transformed p-values for each motif were compared between regions more active in each time point vs each other time point, comparisons were as follows: early up versus mid (EVM), early up versus late (EVL), mid up versus early (MVE), mid up versus late (MVL), late up versus mid (LVM), late up versus early (LVE). Z-score were calculated for each motif across the comparisons and plotted as heatmaps. Differentially enriched regions were assigned to the single nearest gene up to 1 Mb away and

resulting gene lists were assessed for gene ontology enrichments using GREAT.113 All results from GREAT were retrieved programmatically using rGREAT (V1.17.1 https://doi.org/doi:10.18129/B9.bioc.rGREAT). Z-score were calculated for each -log10 transformed p-value for each gene ontology enrichment across a single comparison and plotted as a heatmap.

Enrichment of Chromatin States at Cardiomyocyte Chromatin Loops

The non-promoter anchor points from published high-resolution promoter capture Hi-C (PCHi-C) data from iPSC-derived cardiomyocytes (CMs) were overlapped with the chromatin state segmentations from our human embryonic heart samples and chromatin states from Roadmap brain samples (E053, E054, E067, E068, E069, E070, E071, E072, E073, E074, E081, E082, E125). The fold enrichment of overlap was calculated by taking the fraction of overlap of the non-promoter anchors with the desired tissue type state annotations over the fraction of overlap of randomly selected segments to match the segments chromatin state annotations with the same non-promoter anchors. Significance of difference between human embryonic heart and adult brain fold enrichments were calculated using the Mann-Whitney test (p-value ≤ 0.05 = *, ≤ 0.01= **, ≤ 0.001= ***, ≤ 0.0001 = ****).

Enrichment of GWAS signals in enhancer chromatin state segmentations

Two linkage-disequilibrium (LD) aware approaches were used to determine enrichment of cardiovascular and cardiac development related GWAS signals in enhancer chromatin state signals. In the first approach only variants with genome-wide p-values < 5x10-8were selected from the GWAS catalog or published literature related to atrial fibrillation, resting heart rate, QRS interval, and P-wave duration, aortic root size, congenital heart defects, and coronary artery disease (See Key Resources Table for Term Accessions). Enhancer states (ChromHMM states 13 through 18) identified in embryonic heart samples, embryonic craniofacial samples, and all tissues profiled by Roadmap Epigenome were combined into a single annotation per embryo or tissue. Variant enrichment was then determined using GREGOR for each embryo or tissue dataset. Testing was done by using $r2 \geq 0.8$, maximum LD window of 1 Mb, and 500 minimum neighboring SNPs for each variant based on variants found in the 1000 Genome Project samples

of European ancestry. In the second approach we first collected full GWAS summary statistics from cardiac trait related studies of atrial fibrillation, resting heart rate, QRS interval, and P-wave duration along with putative negative controls from immune disease related studies of systemic lupus erythematous and Crohn's disease. GWAS summary statistics were prepared for processing by GARFIELD using garfield-create-input-gwas.sh script for each study. Individual chromatin state annotations for the same embryos and tissues described above were prepared for processing by GARFIELD using garfield_annotate_uk10k.sh. Finally, GARFIELD was used to determine enrichment of SNPs at multiple GWAS significance thresholds ($p<10\text{-}5$ to $p<10\text{-}8$) in each chromatin state across all embryo and tissue samples using settings as described by Iotchkova *et al.* (GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. Nat Genet. 2019; 51:343–53).

ROC Curves

The ROC curves for each of the three types of data (ChromHMM, EMERGE, Dickel) used enhancers verified by the Vista Enhancer browser. The full set of elements that had enhancer activity, n=281 for true heart positives and n=846 were used for true heart negative. For ChromHMM prediction values for enhancers in heart versus non heart, the posterior probabilities were used. Each chromatin state has a posterior probability output file from ChromHMM segmented into 200bp bins. The mean of the sum of the posterior probabilities for states 13,14,15, and 18 were calculated for the ROC curves. For the EMERGE data, a bedgraph prediction track was provided. Peaks were called on this bedgraph using MACS. The command used was mac2 bdgpeakcall with -c = 0.05. The resulting peaks were then merged with a gap distance of 75bp. Overlapping the blacklist from ENCODE and all peaks that intersected with a known TSS were subtracted out. To get the final prediction score for each EMERGE peak, the prediction scores from the bedgraph track that spanned the peaks were summed. The sum of the prediction scores from all peaks that intersected the true positive and true negative elements were used for the ROC curves. From the Dickel *et al.* resource, the score_Prenatal column of the provided putative enhancers with overlapping TSS rows removed was used. The sum of the prenatal score across the true positive and true negative elements were used for the ROC curves. The R package plotROC (v2.2.1) was used to create the ROC curves and calculate the AUC values.

RNA-Seq

RNA was extracted using miRNeasy RNA extraction kit with on-column DNAse treatment according to the manufacturer's protocol (Qiagen). RNA integrity was checked using Agilent Tapestation 2200. RNA-seq libraries were prepared from 100–200ng total RNA using the TruSeq stranded mRNA kit (Illumina). Libraries were quantified using NEBNext Library Quant Kit for Illumina and library quality checked using Agilent Tapestation 220. Libraries were pooled and diluted to 1.8pm and sequenced on the NextSeq500 Illumina platform using 75bp paired end sequencing according to manufacturer's recommendations. The accession number for the RNA-seq bigWigs, and counts matrices reported in this paper is GEO:138799.

RNA-Seq Data Processing

Quality control was performed on RNA-seq reads using FastQC (version [v.] 0.11.5) and MultiQC (v.1.1). Trimming for adapters, quality and length was performed using Trimmomatic (v.0.36). Trimmed fastqs were aligned with Rail-RNA77 using human assembly GRCh38/hg38. The coverage bigWig files output by Rail-RNA were used as input for the generation of counts tables by following the instructions and pipeline from recount https://github.com/leekgroup/recount-contributions), where the comprehensive gencode v.25 annotation was used.76 The "level 3" genes defined by gencode were excluded. The recount rse_gene objects for each sample were combined into one rse_gene object and transformed with scale_counts() from recount (v1.8.2).

RNA-Seq Differential Expression

The scaled rse_gene recount object was made into a DESeq2 (v1.22.2) object. Low gene counts were filtered by removing all genes whose sum of counts across all samples was less than 100. This left a total of 26,122 genes for downstream analysis. Batch effects were mitigated by using the sva package in R. (v3.30.1) Pairwise, differential analysis of each of the eight carnegie stages was performed with DESeq2 including the surrogate variables identified from sva. For the heatmaps and PCA plots the counts table was transformed using the surrogate variables, see provided code (https://github.com/cotneylab) for these calculations. The PCA plots in the supplement were made using the prcomp() function from the built-in R stats (v3.6.3) package.

The genes shown in the heatmaps and used for the PCA plot generation had a Benjamini Hochberg adjusted p-value less than 0.05 and a fold change greater than the absolute value of 1. heatmap.2() from the R library package gplots(v3.0.1.1) was used to generate the heatmaps of the normalized SV corrected counts by setting the scale option to "row". The distance matrix of one minus the transpose of the kendall correlation of the SV corrected counts was clustered using hclust() with method = "complete". The resulting hierarchical clustering was used in the heatmap.2() option Rowv to organize the gene rows. clusterProfiler (v3.12.0) was used to obtain the gene ontology enrichments from the dendrogram groups using the function enrichGO() with standard options. The enrichment map plots were made using enrichplot (v1.6.0) function emaaplot(). Standard options were used with manual code changes to the color and p-value settings.

GTEx tSNE Analysis

The rse_gene R data object, or counts table of all GTEx tissues was retrieved from the Recount2 database (https://jhubiostatistics.shinyapps.io/recount/). The GTEx counts table was generated using the same Rail-RNA & Recount pipeline we used to generate the counts tables for our embryonic heart data which is described above. The GTEx data contained 9,662 samples that were combined with the 24 embryonic heart samples to make one matrix containing 9,686 total samples by 58,037 genes. The meta-data for GTEx is provided in a link under the phenotype column from the Recount2 database, and the tissue assignments located under the column named smts was used, resulting in a total of 31 unique tissues. The counts were transformed using the scale_counts() function from the R library recount (v1.8.2) as is recommended from the workflow in the Recount2 F1000 paper.114 Genes whose mean across all GTEx and embryonic heart tissues was lower than or equal to 1 were removed, resulting in 36,990 genes. The filtered counts matrix was transformed by log10() with a pseudo count of 1 added to all values. The transpose of the log10 transformed matrix was then converted to a distance matrix using the dist() function in R. This distance matrix was used as input for the tsne model generated by using the Rtsne() function from the R package Rtsne (v0.15). The parameters used in Rtsne() were the following: perplexity=10, max_iter=1000, theta=0.5, dims=2.

Tissue Specificity of Gene Expression (GINI)

The combined, filtered GTEx and human embryonic heart counts matrix used for the tsne analysis was also used to calculate the tissue specificity for each gene. The Gini Index for each gene was calculated using the Gini() function from R library Ineq (v0.2-13) on the average counts per tissue. A gene was given a tissue assignment based on the tissue with the maximum count for that gene. To create the heatmap of the various tissues, the distance matrix of one minus the transpose of the pearson correlation of the average expression per tissue for all genes with a Gini score of 0.5 and above was clustered using hclust() with method = "complete". The z-scores of this average expression per tissue matrix was plotted using heatmap.2 with the gene rows organized by the dendrogram calculated from hclust(). The genes for the embryonic heart, spleen and adult heart were determined by cutree() such that it would result in 25 groups to correspond to the 25 tissues. RDAVIDWebService (v1.22.0) was used to obtain the gene ontology enrichments using the original 36,990 genes as the background.

Novel Heart Enhancer Effects on Gene Expression

Assignment of enhancers to genes was made using GREAT v.4.0.4 using human embryonic heart specific enhancers on hg19 with the whole genome as background and using the single nearest gene association rule setting. The line graph of the enhancers versus expression used the average expression per tissue type from GTEx and our heart enhancer data using geom_smooth() from ggplot2 with the method set to "loess". The corresponding violin plot included only average gene expression from genes with more than 25 enhancers targeting. Stat_compare_means() was used to compare all tissues against the embryonic heart, using the Mann-Whitney test (p-value ≤ 0.05 = *, ≤ 0.01= **, ≤ 0.001= ***, ≤ 0.0001 = ****).

Weighted Gene Co-Expression Network Analysis

Co-expression networks were generated using the WGCNA Rpackage based on recommendations put forth by the Horvath group.

Network Construction

For the WGCNA a soft-thresholding power of 8 was chosen assuming an unsigned network and based on recommendations put forth by the Horvath group for samples between 20 and 30. The modules were detected from the network from the cutreeDynamic() function from the WGCNA package with the following parameters, minClusterSize=100, deepSplit=2. Detected modules were merged based on their eigengene correlation. To do this a dendrogram of the module eigengenes was generated and a threshold value of 0.18 was chosen as input for the function mergeCloseModules(). The intra-modular connectivity of each gene was calculated using the intramodularConnectivity() function from the WGCNA library that determined the hub and non-hub designation. The resulting network contained 29 modules.

Plotting of Modules

A multidimensional scaling of the module eigenvectors output from WGCNA was generated to plot the modules in 2-D space using the function cmdscale() from the stats v3.6.1 package. A pearson correlation of the module eigenvectors was calculated for the edges. Positive correlations of 0.5 and greater were included. Modules were plotted that fulfilled the criteria of having significant adjusted p-values (< 0.05) from the GO analysis, significant permutation p-values (< 0.05) of embryonic heart specific enhancers, and/or of embryonic heart specific Gini genes, this resulted in exclusion of 9 modules from the plot. The module eigenvectors of each of the 20 modules were plotted using ggplot2, geom_smooth() function with the loess smoothing method. The confidence intervals were removed for ease of visualization.

Gene Ontology and Functional Enrichments

RDAVIDWebService (v1.22.0) was used to obtain gene ontology enrichment of the genes within each of the 29 WGCNA modules. The gene background list used was all the genes input into the WGCNA. The module enrichments of embryonic heart specific Gini genes, embryonic heart specific enhancers, various disease gene lists, and the NKX2-5 bound gene lists were determined by a permutation test with 1000 iterations. The single-cell RNA-seq differential expression for genes with an adjusted-p value <= 0.05 were used. For the LOEUF enrichment analysis among the hub vs non-hub genes, the non-hub genes were randomly sampled using the

R function sample(). The number of non-hub genes sampled were the same number of total hub genes within the network- 10% of 26,122 genes or 2,612 from the 23,510 non-hub gene list. This process was iterated 1000 times to get a mean (gray bars) and standard deviation (shown as error bars) for each LOEUF decile. The LOEUF score and decile designation for each gene is freely available through gnomAD v2.1.1

Protein-protein interaction analysis

To generate the ppi histogram, 100 randomized versions of the WGCNA network were made. This was done by randomly assigning the 26,122 genes to 29 modules of equal gene sizes to the original network using the R function sample(). The ppi enrichment of up to 500 randomly chosen genes for each module of each of the 100 randomized versions was then determined using the STRINGdb (v1.24.0) package. Up to 500 genes were used due to constraints from the STRINGdb package. The ppi database was loaded by using STRINGdb$new with version ="10" and score_threshold=0.4. For each iteration, the output p-value of the STRINGdb call get_ppi_enrichment() was adjusted using the Bonferroni method. The number of modules that met the adjusted p-value cut-off of 0.05 was counted for each iteration to produce frequency values.

Data display

Throughout this work we make heavy use of the ggplot2 package (v3.2.0) in R(v3.6.1) package.

Results

Chromatin State Profiling of Human Embryonic Heart Development

As demonstrated by Roadmap Epigenome, ensemble approaches using many different biochemical markers are required to more completely characterize how the genome is utilized in a given biological condition. To systematically profile chromatin states of the developing human embryonic heart, ChIP-seq was used with antibodies against 7 histone H3 posttranslational modifications on individual human hearts from 18 embryos spanning the organogenesis phase of heart development (**FIG. 1A**; **FIG. 9**, Table 2). Overall, these raw data were of high quality with

high correlation between experiments performed with the same antibody across separate individual embryos and generally clustered by marks associated with genome activation (H3K4me1, H3K4me2, H3K4me3, H3K27ac, and H3K36me3) versus repression (H3K9me3 and H3K27me3; **FIG. 1B**, **FIG. 10A**). These data were uniformly processed to allow direct comparison of 127 human cell types profiled by Roadmap Epigenome and 21 human developing craniofacial samples. An imputation approach was then used to generate a complete set of 12 global epigenomic signals for each sample using chromatin signal imputation as previously described by Roadmap Epigenome and used for human craniofacial development.19,21 The imputed signals correlated well with primary signals and predominantly clustered by known biological function as was observed in Roadmap Epigenome and human craniofacial epigenomic data (**FIG. 1B**, **FIG. 10B**).

Table 2. Embryo ID and Karyotype. The embryo ID and karyotype of embryo used in this example as reported from HBBR, related to FIG. 1.

Table 2

Sample ID	Stage	Date of processing	Karyotype	Collection Site	Collection Date
14131	CS12	1/11/2019	rsa(13,15,16,18,21,22,X)x2	Newcastle	4/27/2018
12383	CS13	7/26/2018	46,XY	London	1/13/2015
12690	CS13	7/26/2018	46,XY	London	6/23/2015
14401	CS13	1/11/2019	rsa(13,15,16,18,21,22)x2,(X,Y)x1	London	10/16/2018
14479	CS13	1/11/2019	rsa(13,15,16,18,21,22)x2,(X,Y)x1	Newcastle	11/22/2018
12408	CS14	1/11/2019	46, XY	London	1/20/2015
14135	CS14	1/11/2019	rsa(13,15,16,18,21,22,X)x2	Newcastle	5/3/2018
12997	CS16	1/11/2019	46,XY	London	1/14/2016
14213	CS16	1/11/2019	rsa(13,15,16,18,21,22,X)x2	Newcastle	6/21/2018
14315	CS16	1/11/2019	rsa(13,15,16,18,21,22,X)x2	Newcastle	8/16/2018
14209	CS16	1/11/2019	rsa(13,15,16,18,21,22)x2,(X,Y)x1	Newcastle	8/19/2018
12291	CS17	6/21/2018	46,XY	Newcastle	1/1/2017
12331	CS17	6/21/2018	46,XX	Newcastle	1/1/2017
12752	CS17	6/21/2018	46,XX	Newcastle	1/1/2017
12059	CS18	7/23/2018	46,XY	London	5/21/2014
12456	CS18	7/23/2018	46,XY	London	12/6/2016
13474	CS18	7/23/2018	46,XX	London	2/11/2015
11914	CS19	6/18/2018	46,XX	London	1/8/2014
12135	CS19	6/18/2018	46,XY	London	7/23/2014
12248	CS19	6/20/2018	46,XY	London	10/8/2014
12448	CS20	7/24/2018	46,XY	London	2/11/2015
12451	CS20	7/24/2018	46,XX	London	2/11/2015
13068	CS20	7/24/2018	46,XX	London	2/23/2016
11849	CS21	7/25/2018	46,XY	London	10/2/2013
12093	CS21	7/25/2018	46,XX	London	8/18/2014
12210	CS21	7/25/2018	46,XX	London	9/10/2014
12058	CS23	7/27/2018	46,XY	London	5/21/2014
12151	CS23	7/27/2018	46,XX	London	8/6/2014
12193	CS23	7/27/2018	46,XY	London	9/3/2014

Having uniform epigenomic datasets for each heart sample across the developmental series, we then applied the previously generated 15-, 18-, and 25-state models of chromatin

activity developed by Roadmap Epigenome to segment the genome into chromatin states. The individual state classifications and color coding for each model are provided in **FIG. 11** for easy reference. The number of segments identified for each of the 25 chromatin states was similar across all of our 18 samples (**FIG. 1C**). The pattern of chromatin state segments identified in our human embryonic heart samples was similar to that identified in 127 tissues from Roadmap Epigenome, with the one exception of significantly increased numbers of poised promoter segments (state 22 from 25-state model; **FIG. 1D**). These findings suggest our imputation and segmentation approaches generated data globally similar to all other human tissues from Roadmap Epigenome, allowing us to make direct comparisons of chromatin state utilization between tissue types and stages.

Genomic enrichments and biological features of individual chromatin states have been thoroughly explored in adult tissues but have yet to be extensively explored during organogenesis. Two groups have recently published resources aimed at comprehensive identification of active regulatory sequences in heart development. However, they have profiled a limited set of active functional marks from tissue after human heart organogenesis has largely completed or applied a novel computational framework tuned for prediction of heart regulatory sequences that has not been extensively compared across human tissues or more generalized functional annotation tools. While these resources have both shown enrichment for in vivo developmental heart enhancers, they did not characterize other regulatory states such as repression. Therefore, inventors aimed to understand what the various chromatin states that have been annotated mean biologically during heart development. To achieve this, structural and functional enrichments genome wide were characterized for each chromatin state and compared them to regions identified in a previously published compendium of heart enhancers and the EMERGE prediction framework.

The positions of each chromatin state segment were first characterized relative to known functional positions in the genome. It was found as expected that segments of the genome annotated as state 1 from the 25-state model (1_TssA) previously classified as active transcription start sites (TSS) were located primarily within 10 base pairs of a TSS based on GENCODE, version 25, annotation of the human genome (FIG. 4A). Promoter-associated states 2 through 4 (2_PromU, 3_PromD1, and 4_PromD2) and bivalent promoter state 23 (PromBiv)

were generally located within 1000 bp of known TSS. The remaining states from 5 to 21 were progressively more distant from known TSS with heterochromatic segment annotations (21_Het) being the most distant. Transcription-associated states (5_Tx5, 6_Tx, 7_Tx3, and 8_TxWk) and regulatory states located in introns of active genes (9_TxReg, 10_TxEnh5, 11_TxEnh3, and 12_TxEnhW) were closer to TSS than other active regulatory states.

Among the 7 putative enhancer states (13_EnhA1, 14_EnhA2, 15_EnhAF, 16_EnhW1, 17_EnhW2, 18_EnhAc, and 19_Dnase), those segments of the genome annotated to be strong enhancer states (states 13 through 15) were closer to TSS than weak enhancer states (states 16, 17, and 19) with state 13 being systematically the closest to known TSS. State 18 (18_EnhAc), which was defined primarily by H3K27ac signals alone, showed similar distributions of distances relative to TSS as other strong enhancer states (**FIG. 12A**). Regions assembled in a compendium of heart enhancers by Dickel et al showed similar distributions of distance to TSS as strong enhancer states 14, 15, and 18 but were more distant than state 13. These states were driven in the original Roadmap model by H3K27ac and H3K4me1 with the Dickel resource being heavily biased toward H3K27ac and its depositor P300. The strongest enhancer state, state 13, was identified in Roadmap Epigenome by many more features including the presence of H3K27ac, H3K4me1, H3K4me2, and DNase and conversely the absence of H3K9me3, H3K27me3, H3K36me3, and H3K79me2, which are uniquely captured in our study. Interestingly, regulatory regions predicted by EMERGE showed a bimodal distribution with a significant fraction of regions located directly overlapping known TSS (**FIG. 12A**). When we analyzed the size of segments annotated as each chromatin state, we saw similar trends across all states with median values typically around 1 kb. Larger distributions were observed at heterochromatic regions (21_Het) and portions of the genome with no detectable activity or quiescent (25_Quies). When we analyzed the Dickel compendium and regions annotated by EMERGE, we observed significant excursions from the values we observed for our chromatin state segments. The Dickel compendium identified uniformly larger regions than enhancer chromatin states identified from our data, while most regions identified by EMERGE were significantly smaller (**FIG. 12B**).

Sequence conservation has been frequently used to identify functional regions of the genome, particularly those with gene regulatory function in intronic or intergenic spaces of

vertebrate genomes. Regions associated with active TSS or assigned active or bivalent promoter states were most conserved based on multispecies alignments from 100 vertebrate species (**FIG. 12C**). Heterochromatic and polycomb repressed states were the least conserved among the 25 chromatin states. Among putative enhancer chromatin states, strong enhancer state 13, weak enhancer state 16, and DNase accessible regions state 19 were the most conserved. Median conservation scores of regions identified by Dickel et al were comparable to weak enhancer state 16. Notably EMERGE regions showed among the lowest median conservation scores across all categories greater only than enhancer state 15.

While conservation is an indicator of selective pressure across species, a large portion of the human genome does not show significant conservation across all vertebrates and might be uniquely involved in human disease. Several groups have attempted to bridge this knowledge gap and developed machine learning techniques to predict deleteriousness of all 8.6 billion possible single-nucleotide variants in the human genome. These approaches leverage many different functional genomics datasets, TF-binding site predictions, and sequence conservation, among others and are informative for identifying regulatory variants linked to disease. When we interrogated our chromatin states with 2 of these scoring metrics (Combined Annotation Dependent Depletion [CADD] and Linear Model Inference of Natural Selection From Interspersed Genomically Coherent Elements [LINSIGHT]), states 20 (ZNF_Repeats) and 21 (heterochromatin) had the lowest median scores (**FIG. 12D** and **FIG. 12E**). TSS and promoter-associated states 1, 2, 3, and 23 had the highest scores predicted by these methods. Transcribed regions (states 5 through 8) despite being nearer TSS sites had uniformly depressed Combined Annotation Dependent Depletion and Linear Model Inference of Natural Selection From Interspersed Genomically Coherent Elements scores. Strong enhancer state 13 had the highest median Combined Annotation Dependent Depletion and Linear Model Inference of Natural Selection From Interspersed Genomically Coherent Elements scores among distal regulatory chromatin states, significantly higher than all other enhancer states and regions annotated by both Dickel et al and EMERGE (**FIG. 12D** and **FIG. 12E**).

Combined, these results strongly corroborate the functional labels assigned to each of the chromatin states applied by Roadmap Epigenome. These systematic analyses revealed many expected distributions for chromatin states relative to known features of the genome and elevated

sequence conservation among active regulatory regions. Our results also revealed a number of similarities between our annotations and previous databases of heart regulatory sequences. However, our enhancer chromatin states, particularly the strong categories, are significantly more conserved and harbor significantly more positions predicted to be involved in human disease than either of these 2 catalogs of putative heart regulatory sequences. While indicative of function, these metrics are largely generic and not necessarily related to the heart. Thus, we next aimed to understand the direct relevance of these chromatin state annotations for heart development and disease.

Identification of Novel Human Embryonic Heart Enhancers

Comparison of enhancers across many different tissues has shown them to be far more tissue specific in their activation than promoters. To determine whether the enhancer segments from our developing heart chromatin state segmentations are indeed tissue specific, we aimed to examine analogous chromatin state annotations quantitatively across as many tissues as possible. To achieve this, we first assembled all segments identified as an enhancer chromatin state based on the 25-state chromatin model in any tissue profiled by Roadmap Epigenome, this study, previously generated human craniofacial development samples, and chromatin data for nuclei isolated from fetal, infant, and adult cardiomyocytes. In our human embryonic heart samples alone, we identified a total of 177 412 segments that were reproducibly annotated as 1 of 6 putative enhancer states (states 13–18, EnhA1, EnhA2, EnhAF, EnhW1, EnhW2, and EnhAc) in at least 2 samples.

The chromatin state segmentations are powerful tools for genome-wide annotation, but they have limited utility on their own as the quantitative nature of the underlying data is discarded. To overcome this obstacle, we leveraged H3K27ac signals from each of these samples since it has been frequently shown to be highly tissue specific in its distribution and generally associated with enhancer activation. Due to the overall better performance of normalized, imputed signals relative to primary ChIP-seq data, we extracted imputed H3K27ac *P* signals from 174 samples at 444 413 enhancer segments across the genome. We found the strongest global correlations between related tissue types, such as immune cell types, brain region tissues, and the embryonic heart samples. These data separated largely into adult versus embryonic

groups and subsequently by tissue type (**FIG. 2A**; **FIG. 13A** and **FIG. 13B**). The embryonic heart samples clustered well with one another and also showed strong global correlation in H3K27ac signals with the fetal and adult heart tissue samples profiled by Roadmap Epigenome. Interestingly, the isolated cardiomyocyte nuclei formed their own distinct cluster and do not separate into adult versus developing samples. They also showed lower global correlation values with the heart tissues profiled by Roadmap Epigenome suggesting isolation of nuclei had significant effects on chromatin state or organization in those experiments. A tSNE (t-distributed stochastic neighborhood embedding) projection of these data further confirmed these findings (**FIG. 2A**) and showed the distinct nature of the isolated cardiomyocyte H3K27ac data (**FIG. 13B**). Given the potential for strong batch effects in the cardiomyocyte nuclei data due to isolation and sorting of nuclei before ChIP and the decreased correlation with fetal and adult heart samples, we excluded these samples from downstream analysis. These results confirmed the early developmental stage identity and the cardiac origins of the tissues profiled in an unbiased, global fashion and further demonstrated the tissue-specific nature of the putative enhancer sequences and H3K27ac signals.

To further test the validity of our annotations, particularly the putative enhancer states, we assessed the number of experimentally tested and validated developmental enhancers by the Vista Enhancer Browser captured by each of our chromatin states from the 25-state model (**FIG. 2B**). Overall, 92% of all validated heart-positive enhancers (n=281) were identified from the Vista Enhancer Browser, a 7.5-fold enrichment versus active enhancers lacking activity in the heart (heart negative; $P<2.2\times10^{-16}$). When each of the chromatin states ability to specifically identify heart enhancers were compared, it was observed that enhancer chromatin states captured a significantly larger fraction of heart-positive enhancers than other chromatin states (**FIG. 2B**). Strong enhancer state 13 showed the greatest fraction of overlap with validated heart enhancers and the greatest fold difference relative to overlap with heart-negative enhancers. The heart specificity of state 13 was higher than all other chromatin states and significantly greater than regulatory regions annotated by EMERGE or Dickel *et al.*

Interestingly, weak enhancer states 16 and 17, DNase accessibility state 19, and regions harboring repressive chromatin marks including poised and bivalent promoters and polycomb repressed states showed an opposite trend in capturing heart-positive versus heart-negative

enhancers. This indicated that these 2 weak enhancer states (16 and 17) are not informative for heart regulatory sequences and generally should be interpreted with caution in other tissues where large numbers of validated enhancers do not yet exist. When the performance of chromatin states identified by the 15- and 18-state models was compared, it was found decreasing levels of specificity as number of states decreased (**FIG. 14A** and **FIG. 14B**). Strong enhancer segments identified by the 18-state model showed slightly higher levels of specificity versus EMERGE and Dickel et al (**FIG. 14B**). Enhancers identified by the 15-state model performed the poorest by this metric with EMERGE showing significantly higher specificity (**FIG. 14A**).

These results suggest the 25-state chromatin model is best able to identify heart-specific regulatory sequences, particularly state 13 enhancers. However, this analysis does not take into account the ranking of enhancers for heart specificity provided by both Dickel et al and EMERGE. It is thus possible that higher ranking regulatory sequences in either resource may be better able to identify true heart positives. To test whether this was indeed the case, the ability of strong enhancer states from the 25-state model to recover heart-positive versus heart-negative enhancers compared with performance of ranked lists from both EMERGE and Dickel et al was measured. Area under the curve for receiver operating characteristic curves revealed consistently higher area under the curve for enhancers identified by the 25-state model across all time points versus EMERGE with the exception of the earliest Carnegie Stage (CS) 13 samples. Enhancers identified by Dickel et al had the lowest area under the curve values in this analysis (**FIG. 15C**). When the regions identified by the chromatin states were directly compared with those by EMERGE, it was found the most significant overlaps of EMERGE peaks with active TSS and active promoter annotations (**FIG. 15A**). The EMERGE signal was overall highly enriched near TSS, and subsequently, the active TSS and promoter states (states 1–4) had the highest EMERGE scores (**FIG. 15B**). When the overlap of all chromatin state segments was compared with the Dickel compendium, it was found much less overlap with active TSS states and higher percentages of sequences that were annotated to be weak enhancer states or even quiescent in human embryonic development (**FIG. 16A**). Together, these results suggest that the highest scoring EMERGE peaks are concentrated close to known TSS and the 25-state chromatin model

is better able to identify tissue-specific regulatory sequences than both EMERGE and Dickel compendium.

While these results suggest that the chromatin state segmentations herein are capable of identifying sequences with heart regulatory capacity, the Vista Enhancer Database has been largely constructed by testing elements identified with H3K27ac deposition or P300 binding. The Dickel compendium is based almost entirely on these 2 marks, and the EMERGE framework has been tuned to find positives from this database. To determine whether chromatin state segmentations herein are capable of annotating regulatory sequences identified by different means, a large set of sequences characterized by binding of 7 cardiac TFs in fetal and adult mouse hearts were analyzed. This work demonstrated that regions bound by ≥5 TFs showed significantly more activity than regions with H3K27ac signal but lacking TF binding when systematically tested using a massively parallel reporter assay (MPRA).

When regions tested by MPRA that could be identified in the human genome were interrogated, it was found the highest median RNA-to-DNA ratios for enhancer state segments in our 25-state segmentations. Specifically, it was found significantly higher MPRA activity relative to negative controls in strong enhancer states 14, 15, and 18 (**FIG. 12F**). Weak enhancer state MPRA signals were not significantly different from negative controls. Neither the Dickel compendium nor EMERGE regions showed significantly different MPRA signals from negative controls. These findings indicate the chromatin state segmentations herein are capable of identifying regions that are driven by combinatorial TF binding and not necessarily by H3K27ac.

Thus far, these analyses have indicated the strong enhancer states (states 13–15 and 18) are the most specific for developing heart. Thus it was reasoned that segments annotated as strong enhancer states in the developing heart but not so in other tissues might reveal previously unknown heart-relevant gene regulatory information. When the full set of strong embryonic heart enhancer segments were compared with analogous annotations from all tissues in Roadmap Epigenome, 12 359 segments were identified that had not been previously annotated as strong enhancer chromatin states. These putative enhancers are referred to hereafter as embryonic heart–specific enhancer segments (EHEs). When these regions were compared with identified in the Dickel compendium with prenatal biased scoring enhancers (enhancers with a scoring ratio >2 prenatal/postnatal) or enhancers in the Dickel compendium identified in the single human

fetal sample, small but significant sharing of annotations were observed (**FIG. 16B** through **FIG. 16E**). Combined, these results indicated that the vast majority of our putative EHEs (75.6%) have never been previously identified in any other human tissue or stages of heart development.

Only a handful of the putative EHEs have been functionally tested, making it difficult to assess their relevance for heart development from an in vivo perspective. While such putative enhancers might be novel, it was reasoned they could potentially target genes known to be involved in heart development or function. To address this, we assigned EHEs to the single nearest gene within 1 Mb using the Genomic Regions Enrichment of Annotations Tool and found significant enrichment of biological processes related to heart development and morphogenesis (**FIG. 2C**). The associated mouse phenotypes are heavily fortified with those related to abnormal development, morphology, size, and function of the heart. Additionally, the enriched molecular functions contain multiple terms related to microfibril and tubulin binding along with voltage-gated channel activity in the atrioventricular node.

The sequence content of the EHEs for enrichment of TF-binding sites was subsequently analyzed using Hypergeometric Optimization of Motif Enrichment (**FIG. 2D**). We identified significantly enriched motifs that matched binding sites of TFs involved in heart development such as the GATA family, the MEF2 (myocyte enhancer factor-2) family, and TBX20 (T-box TF 20) among others (**FIG. 2D**, top). When de novo motif enrichment was performed using Hypergeometric Optimization of Motif Enrichment, enrichment of the same families of TFs was seen with the addition of factors also known to contribute to heart development, such as NKX2-5 (NK2 homeobox 5) and SMAD (SMA and mothers against decapentaplegic gene family) signaling (**FIG. 2D**, bottom). Through this global analysis, TFs likely involved in human embryonic heart enhancer activation were identified. Functional and motif enrichment within these EHEs identified from this multitissue, unbiased analysis demonstrated the power of these approaches and suggested that many of the novel sequences identified are bona fide enhancers and likely important for normal heart development.

Differential Motif Utilization of Embryonic Heart Enhancers Across Development

Having demonstrated enrichment of heart-related TF-binding sites in putative EHEs, it was contemplated whether this was a general trend across all of heart development or if there

were temporal shifts in enhancer activation modulated by different TFs during the embryonic period of heart development. To simplify our analysis, we consolidated the developmental time course into 3 general periods: early (CS13), mid (CS16–17), and late (CS23) development (**FIG. 3A**). Differentially activated putative enhancer segments across this trajectory by comparing H3K27ac and H3K4me2 signals at all reproducible enhancer segments between each of the 3 stages of embryonic heart development (**FIG. 3B**, **FIG. 3C**, and **FIG. 17**) were identified.

For putative enhancer segments differentially activated based on H3K27ac signals, the greatest differences were observed in motif enrichment between early and later stages of development (**FIG. 3D**). Putative enhancer segments active early were specifically enriched for pluripotency-related TFs like *SOX2* and *OCT4* and multiple members of the KLF (Kruppel-like factor) and Forkhead families of TFs. Motifs enriched in late putative enhancer segments included many zinc finger–containing TFs and multiple members of the T-box, GATA, and PAX (paired box) families of TFs. Notably, enhancer segments more strongly active in the mid period of heart embryonic development based on H3K4me2 signals showed the most pronounced enrichment of TF motifs (**FIG. 3E**). Many of the same TF motifs enriched in early and late putative enhancer segments based on H3K27ac were shifted to enrichment in the mid period, suggesting dynamics in TF utilization related to chromatin state.

Based on these differences in differentially activated putative enhancer segments, it was hypothesized that the location of these segments and subsequently the genes that they control might uncover distinct patterns of biological pathway utilization during this developmental series. Differentially activated putative enhancer segments from each comparison to the nearest gene were assigned and the most significantly enriched gene ontology terms were identified (**FIG. 3F**). In all comparisons, terms associated with cardiac chamber and septum morphogenesis and development were highly enriched. Genes near putative enhancer segments most strongly active during the early period were uniquely enriched for pathways related to bone growth and fibroblast growth factor (FGF) receptor signaling. Terms associated with cardiomyocyte differentiation, muscle cell development, and muscle cell contraction were observed for genes assigned putative enhancers more strongly active in the late period. These results agree with known dynamics in relative ratios of cardiomyocytes versus other cell types during these different developmental periods. These bulk tissue experiments are incapable of

disentangling such cell type–specific regulatory element utilization. These differentially active putative enhancer lists are likely enriched for such sequences but require single-cell chromatin accessibility methods to confirm.

Identification of Embryonic Heart–Specific Super Enhancers and Long-Range Chromatin Interactions

To better understand the role these putative enhancer segments might play in cardiac-related diseases, the EHEs were interrogated for known heart-related loci. Defects in the cardiac homeobox TF *NKX2-5* have been implicated in many different CHDs including conotruncal malformations, septal defects, and atrioventricular conduction block. Most human patients with *NKX2-5*–associated CHD have heterozygous loss-of-function mutations, and haploinsufficient mice have significantly depleted protein levels in the heart. Depletion or complete knockout of *NKX2-5* results in dysregulation of many downstream target genes in mice and human cardiomyocytes, suggesting it is a master regulator of cardiac development. Therefore, identifying regulatory sequences that control *NKX2-5* expression specifically in the developing heart could be valuable information for understanding the unknown genetic causes of CHDs.

When the genomic locus containing this gene was inspected, it was found that it was surrounded by a plethora of strong enhancer state segments including an EHE immediately downstream of the coding exons (**FIG. 4A**). There are a number of strong enhancer segments identified uniformly from CS14 to CS23 ≈50 kb upstream that are largely repressed in fetal and adult hearts. These regions are particularly interesting as they cannot be readily identified through sequence conservation based on comparisons of 100 vertebrate genomes (**FIG. 4A**). Moreover, these coordinately activated strong enhancer segments are predicted as an embryonic heart super enhancer region that encompasses over 200 kbs surrounding *NKX2-5* (n=4215). This is ≈4× larger than any other super enhancer annotation for this region based on fetal or adult human heart samples. Such regions have been associated with tissue-specification loci, further reinforcing the central role *NKX2-5* plays in heart development and the novel information our resource has identified.

The locus harboring *SCN5A* was then inspected, which has been implicated in multiple cardiac diseases and specifically known to cause ≈20% to 30% of cases of Brugada syndrome (**FIG. 18A**). Here, it was observed several clusters of strong enhancer segments upstream, within, and downstream of the gene in heart tissues at multiple stages (adult, fetal, and embryonic tissue). The downstream cluster of strong enhancer segments has been previously annotated as a super enhancer region in mouse cardiomyocytes and deletions had significant effects on *SNC5A* expression in the heart. This analysis confirmed this finding but also annotated a super enhancer that encompassed the strong enhancer segment cluster upstream of the gene. A strong enhancer segment in the 16th intron completely encompassed a validated enhancer with specific activity in the mouse embryonic heart. Additionally, a 13-kb linkage disequilibirium (LD) block ($R2 \geq 0.95$, $D'=1$, 1000G CEU) that contains multiple variants associated with cardiac phenotype–related variants overlaps this tested enhancer region and adjacent loci (**FIG. 18A**, dashed box). These variants included rs41312411 associated with establishment of resting heart rate and P-wave duration, rs3922844 associated with establishment of ECG traits and measures, and rs11708996 associated with Brugada syndrome.

Variants in potential regulatory sequences across this entire locus have been tested for effects on enhancer activity in cultured cardiomyocytes. When an in vitro reporter assay data was overlaid for putative regulatory sequences harboring alternative alleles on the chromatin state segmentations, it was generally observed the strongest effects for sequences overlapping strong putative enhancer segments (**FIG. 18A**). Sequences that overlapped quiescent regions in our segmentations had minimal effects on reporter gene expression indicating they likely have limited regulatory capacity in vivo. In addition to these well studied loci, it was found that other genes important for normal cardiac function and development, such as *HAND2* and *MYOCD*, were within super enhancers (**FIG 18B** and **FIG 18C**).

Having demonstrated that genes important for normal cardiac function reside within super enhancers and EHEs are enriched for heart-relevant biology, it was reasoned heart-specific super enhancers could also identify important, previously unknown regulatory landscapes or important cardiac genes. When super enhancer regions were compared globally between the embryonic heart and analogous annotations from many different tissues using dbSUPER, we identified 1611 human embryonic heart–specific super enhancers. These regions were strongly

enriched for genes related to transcription regulation or roles in cell junctions. One such example is the cardiac T-box TF *TBX20,* which is required for cardiomyocyte proliferation in mice and linked to dilated cardiomyopathy in humans. The large noncoding region adjacent to this gene contains 60 putative enhancers, nearly one-third are EHEs, and multiple heart-positive in vivo enhancers (**FIG. 4B**). The specific nature of our annotations is readily apparent at this location with the distal noncoding region, and the sequences surrounding the *TBX20* TSS are sparsely annotated across 127 tissues and cell types in Roadmap Epigenome.

Another putative embryonic heart–specific super enhancer of note is ≈200 kb in length and resides in the large noncoding region upstream of gap junction protein *GJA1*. Sites throughout this ≈1-Mb region form long-range interactions with *GJA1* in human induced pluripotent stem cell–derived cardiomyocytes. Deletion of heart-specific enhancers of *Gja1* in mice is sufficient to decrease its expression, which has in turn been previously linked to arrhythmias. This set of embryonic heart–specific super enhancers includes many additional loci that are not currently known to play a role in cardiac development making them good candidates for future study.

While the examples above demonstrated cis-regulatory landscapes surrounding a single gene, as indicated for enhancers of *GJA1*, such regulatory sequences can interact with their targets over long distances through chromatin looping. Such loops can be difficult to predict in silico, and given the tissue-specific nature of enhancers, appropriate tissues or cell types have to be utilized to identify biologically relevant interactions. Although the developing heart is made up of many different cell types, recent single-cell RNA-seq analysis of CS16 hearts revealed a significant proportion of cardiomyocytes. It was therefore hypothesized that interaction data from cardiomyocytes might allow us to better understand the physical relationship between our chromatin state segmentations and target genes in a cardiac relevant context. Thus, these annotations were integrated with previously published high-resolution promoter capture Hi-C data from induced pluripotent stem cell–derived cardiomyocytes. The distal anchor points were integrated from cardiomyocytes with the functional annotations from the human embryonic heart samples herein using Roadmap adult brain samples as a control. The fold enrichment of interactions for each chromatin state (**FIG. 4C**) was then calculated. The largest enrichments in distal anchor points were in strong enhancer and transcription regulatory states (specifically

states 9–11 and 13–15) for both embryonic heart and brain. However, the largest degree of specificity between embryonic heart versus brain in these states was identified for strong enhancer states 13 and 14. Little-to-no significant change in fold enrichment of interactions between these 2 tissues was seen in poised, repressed, or quiescent states (states 22–25). Overall, these findings confirmed many groups' observations that strong, tissue-specific enhancers interact with their target gene promoters in a tissue-specific fashion and further demonstrated the relevance of these annotations for human embryonic heart development.

Systematic Enrichment of Heart Phenotype and Defect-Associated Variants in Embryonic Heart Enhancers

Since an overlap of putative enhancer segments with disease-associated variants at known cardiac disease–related genes was observed, it was set out to determine whether this finding was generalizable across all strong putative enhancer segments regardless of their proximity to a known heart gene. Variants significantly associated with variation in normal heart phenotypes and CHDs from the genome-wide association study (GWAS) catalog were compiled. Then enrichment of these variants in strong enhancer state annotations from all tissues profiled by Roadmap Epigenome and the developing human heart were assessed. A significant enrichment of variants associated with systolic blood pressure, electrocardiograph traits and measures, resting heart rate, QRS characteristics, and QT interval in strong enhancer segments identified in most embryonic heart samples (**FIG. 5A-FIG. 5D**; **FIG. 20B**) were observed. The earliest samples we profiled (CS13) did not show enrichment for any of these traits suggesting that gene regulatory programs influencing these phenotypes are not active until after 4.5 postconception weeks. As a negative control, variants associated with Crohn disease that have been previously shown to be enriched in enhancers identified in immune cell types were interrogated. This finding was confirm and minimal enrichment of these variants in strong enhancer segments identified during human heart development (**FIG. 20C**) was seen. Common variants associated with CHD incidence (**FIG. 20D**) were then looked into. Unexpectedly, it was found that once *P* was corrected for multiple testing, no set of tissue putative enhancer segments were significantly enriched.

Recent findings in atrial fibrillation (AF) have suggested that sites accessible during fetal heart development are enriched for variants associated with this disorder. However, the functional nature of these sites is unknown, as significant enrichment in fetal heart was observed only for H3K4me1 peaks—a mark typically associated with poised enhancers. This could suggest that enhancers primed in fetal heart development but not yet fully active are important for AF, but it is unclear whether regulatory elements active even earlier or in different chromatin states may play a role. The availability of full summary statistics for this particular disease phenotype allowed us to leverage more systematic, genome-wide analysis of >8 million positions in the genome in a linkage disequilibrium aware fashion.

To address this issue, full summary statistics for a large, multiethnic GWAS for AF were retrieved and assessed enrichment of associated variants across all 25 chromatin states from all tissues in Roadmap Epigenome, craniofacial development, and embryonic heart development. The findings confirm enrichment of AF-associated variants in fetal and adult heart active regulatory sequences. Surprisingly, it was observed enrichment of AF-associated variants across virtually all embryonic heart strong enhancer states (**FIG. 5E**; **FIG. 19A** and **FIG. 19F**). The greatest fold enrichment was observed in strong enhancer segments (Hidden Markov model of chromatin state 13) active at the earliest time points we profiled (CS13; OR, 6.65; P=8.69×10^{-06}). Using this approach, it was also observed enrichment for variants associated with resting heart rate, QRS interval, and P-wave duration consistently across strong enhancer states identified in the embryonic human heart (**FIG. 19H** through **FIG. 19J**). Using full GWAS summary statistics from 2 immune-related diseases (systematic lupus erythematous and Crohn disease, as negative controls), enrichment in any embryonic heart enhancer states (**FIG. 5F**; **FIG. 19G** through **FIG. 19K**) was not observed. Conversely, the enrichment of lupus-associated variants in active TSS (state 1) and strong enhancer state segments from immune-related cell types (**FIG. 19E**) was observed. Surprisingly, when all chromatin states were assessed and identified in embryonic heart samples, it was found significant enrichment of lupus-associated variants uniformly across polycomb repressed segments (state 24) and bivalent promoter segments (state 23) identified in the embryonic and fetal heart but not adult heart samples (**FIG. 5F**).

Gene Expression Profiling of Embryonic Heart Development

The transcriptomes of 3 biological replicates at each of 8 distinct embryonic stages (CS13, 16–21, and 23), largely overlapping the time points profiled for chromatin state were profiled. This window of time has the greatest dynamics of gene expression based on comparisons of transcriptomes of developing hearts from multiple species. To leverage a large number of previously published data sets, we processed our data using a uniform analysis pipeline used by the recount2 database.

Processing data in this way allowed us to directly compare our data to all tissues profiled by Genotype-Tissue Expression Project (GTEx), including 489 samples from adult heart tissue. To assess the validity of this approach, all samples from GTEx (n=9662) were clustered and the embryonic heart samples (n=24) based on expression of 36 990 genes were assessed using t-distributed stochastic neighborhood embedding. Generally good clustering of adult samples by tissue type was observed, including subregions of the brain and the ventricle and atrium of the adult heart. Embryonic heart samples were tightly clustered and distinct from adult heart samples and other GTEx tissues (**FIG. 20A**). Principal component analysis of only the embryonic heart samples showed good clustering by stage and minimal effects from sex and RNA quality in the first 2 components (**FIGS. 20C–E**). Overall, these results suggest that our expression data are of high quality and likely informative for understanding early human heart development.

Genes that are expressed in a limited number of tissues are more likely to be disease-related genes than those with broader expression patterns. Based on these general trends, it was hypothesized that genes expressed specifically during embryonic heart development are likely involved in cardiac defects. To evaluate this, a measure of specificity based on the Gini coefficient—a metric originally used to measure income inequality, which accurately identified genes with tissue- and cell-type specific expression—was used. 347 genes with elevated Gini coefficients (>0.5) and the highest expression in embryonic heart (**FIG. 6A**) were identified. These genes were strongly enriched for genes involved in heart development including *TBX5*, *IRX4*, *HAND1*, *HAND2*, and *FGF12* (**FIG. 6B**, bottom; **FIG. 20B**).

The same tissue-specific functional trends were observed when genes with elevated Gini coefficients and the highest expression in either the brain or the spleen, demonstrating the unbiased nature of this analysis and the specificity of our findings (**FIG. 6B**, top) were examined. Genes with the highest degree of embryonic heart specificity (Gini, ≥0.9) included known heart developmental TFs (*NKX2-5*, *NKX2-6*, and *TBX20*), myosin light chain genes (*MYL3*, *MYL4*, and *MYL7*), the long noncoding RNA *BANCR*, and the sinoatrial node–associated channel gene, *HCN4* (**FIG. 20B**). The single highest Gini coefficient gene assigned to embryonic heart was *LRRC10*—a leucine-rich repeat–containing protein previously identified as having cardiomyocyte-specific expression and linked to human dilated cardiomyopathy.

When this list of genes for potential regulatory effects were analyzed, it was found that more than half of the most significant predicted upstream regulators were genes from this list including *MYOCD*, *TBX5*, *TBX20*, *HAND2*, *GATA4*, *NKX2-5*, and *NKX2-6*. Consistent with these findings, it was found that promoters of embryonic heart elevated Gini coefficient genes were significantly enriched for conserved *NKX2-5* binding site motifs. These findings suggest highly connected direct regulatory effects among these embryonic heart–specific genes. Additionally, it was observed EHEs were enriched for motifs of many embryonic heart elevated Gini coefficient genes (*GATA4*, *GATA6*, *TBX20*, *HAND2*, and *NKX* [NK homeobox]; **FIG. 2D**). Combined, these findings show that the Gini coefficient effectively identifies many heart-related disease genes, enables novel inference of regulatory relationships between these genes, and implicates new genes in heart-related disease phenotypes.

Thus far, these gene expression analyses combined all of the samples herein into a single embryonic heart category, but dynamics of heart gene expression have been reported to be the greatest during the time points we have profiled. To characterize these dynamics, it was first set out to identify genes that are differentially expressed in a pairwise fashion between each of the stages of embryonic heart development. 7167 differentially expressed between all sets of time points (**FIGS. 21A–B**) were identified. The majority of differentially expressed genes were identified relative to the CS13 time point. Few genes were differentially expressed between time points CS17 to CS23, supporting the finding that gene expression dynamics are restricted to early heart development in humans. Hierarchical clustering of these differentially expressed genes identified 4 major groups of genes that show a wave of expression across this developmental series (**FIG. 6C**; **FIGS. 21A–B**). Genes expressed most strongly in early samples

were enriched for functions related to general embryo development/morphogenesis and cholesterol biosynthesis (**FIG. 6D**, left). These genes were also enriched for functions related to neuron generation and differentiation suggesting similar early developmental architecture between the heart and brain. Genes more strongly expressed at the end of the embryonic period were enriched for functions related to ion channel activity, growth factor binding, extracellular matrix organization, and intracellular signaling (**FIG. 6D**, right). The most significant enrichment of genes related to heart contraction and cell-cell adhesion were found in genes most strongly expressed at the CS17 time point (**FIG. 22A**, right). These results indicate much of the dynamics of gene expression during early heart development have been captured but that many important events likely happen even earlier than can be currently examined.

Regulatory Effects of Embryonic Heart Enhancers on Gene Expression

Since it was found that EHEs are strongly enriched for binding sites of heart-related TFs and systematically located near genes with heart-related functions, the impact of such regulatory elements might have on embryonic heart–specific gene expression was set to be understood. To address this question, embryonic heart–specific enhancers were assigned to the single nearest gene. For genes that were assigned an EHE, most had a modest number assigned (mean, 4), but some genes had >30. When gene expression was interrogated, it was found that an increasing number of EHE assignments was associated with greater differences in gene expression in embryonic heart tissues relative to all other tissues from GTEx (**FIG. 7A**). Specifically, genes with >15 EHEs had significantly higher gene expression within embryonic heart samples than all other GTEx tissues (**FIG. 7A**; **FIG. 22B**). It was then asked if these EHEs could be associated with embryonic heart–specific expression by measuring the distance of each enhancer to high Gini coefficient genes identified in the embryonic heart from above. It was found that significant enrichments of EHEs near embryonic heart–specific genes relative to random permutations of the same number of enhancers across distances ranging from 5 to 100 kb (**FIG. 7B**).

These results coupled with the enrichment of physical interactions of putative embryonic heart strong enhancers with promoters active in cardiomyocytes suggested a direct role for EHEs in driving embryonic heart–specific gene expression. Additionally, the sequence content of these

enhancers indicated they are in turn regulated by many of these embryonic heart–specific genes pointing to the existence of coregulated networks of genes activated in early heart development.

Gene Coexpression Networks Reveal Trajectories of Coordinated Expression During Early Heart Development

Genes that are coexpressed across development have been proposed to share similar regulatory mechanisms and form networks that are important for normal development. Genes coexpressed during early brain development, particularly those that have correlations with many different genes or hub genes, are involved in Autism Spectrum Disorder (ASD) risk. It was, therefore, hypothesized that identifying coexpression networks and resulting hub genes could reveal novel candidate CHD genes. To build a coexpression network during embryonic heart development in an unsupervised and unbiased fashion, we used weighted gene coexpression network analysis (WGCNA) using all 24 embryonic heart samples we have profiled. Using this approach, 26 122 genes were distributed across 29 modules, 20 of which showed enrichment for at least one gene ontology category. Modules with gene ontology enrichments expected for early heart development such as embryonic patterning (green, 2573 genes), muscle cell differentiation (brown, 4945 genes), and sarcomere assembly (violet, 1267 genes; **FIG. 7C**) were identified.

Multidimensional scaling of the module eigengenes revealed intermodule coexpression, suggesting some modules were more closely related in their expression than others (**FIG. 7C**). Comparison of trajectories of expression of eigengenes from each module revealed 4 groups of gene expression patterns that reflect positioning of each module in multidimensional scaling space (FIG. 7D). Groups 1 and 3, which are diametrically opposed to one another in multidimensional scaling space, showed downward and upward trends of expression throughout the embryonic period, respectively. Groups 2 and 4, which are also opposite one another in multidimensional scaling space, showed multiphasic but offset patterns of expression. Group 2 showed a particularly strong wave-like pattern between CS16 and CS20. When gene set enrichment analyses were performed across gene expression from CS16, CS18, and CS20, we readily observed cyclical enrichment of a number of pathways (**FIGS. 23A –C**). These included heart valve development, tissue migration, mitochondrial gene expression, and several metabolic processes.

Significance Tests Give Context to WGCNA of Early Developing Heart

To further characterize and validate the WGCNA network, the enrichment of several curated gene lists were evaluated. As we demonstrated above, binding sites for TFs expressed specifically in the embryonic heart were significantly enriched in embryonic heart–specific enhancers. This suggested that coordinated regulation between enhancer sites, the proteins that bind them, and the genes they target might be in action during early heart development. Five modules were significantly enriched for elevated Gini coefficient genes, while 2 modules were significantly enriched for known CHD genes (Benjamini-Hochberg adjusted $P<0.05$; **FIG. 8A**). All of these modules were significantly enriched for genes assigned at least 1 EHE. Analysis of our modules with single-cell RNA-seq on a 6.5 postconception week human heart revealed that despite performing bulk RNA experiments, our modules are organized in cell type–relevant patterns. Analysis of genes that identify the major cell types from CS16 heart revealed that many of the modules in groups 1 through 3 of our network are significantly enriched for embryonic heart cell type–specific genes (**FIG. 8A**). For example, the brown4 and violet modules that are enriched for gene ontologies related to the sarcomere and muscle cell development (**FIG. 7C**) concordantly have significant enrichment for the cardiomyocyte cell types (**FIG. 8A**). The combined enrichment of both specific gene expression and specific enhancer activation in these modules suggest the network that was constructed is particularly meaningful during embryonic heart development (**FIG. 8A**).

To confirm that the WGCNA network is indeed uniquely able to identify heart-relevant biology, a coexpression network constructed for the developing human brain was leveraged. Analysis of embryonic heart–specific genes on this network supported this hypothesis as only 2 modules show enrichment for heart high Gini genes (**FIG. 24**). No modules were significantly enriched for known CHD genes. Analysis of the CS16 heart cell-type associated genes revealed modules primarily in group 5 with enrichment, but overall, the network lacked the cell-type specificity patterns observed on the heart WGCNA network (**FIG. 24**).

Disease Relevance of WGCNA

Functional enrichments and significance tests revealed brown4 and violet modules to be important for normal heart development; however, they lacked enrichment for known CHD-causing genes. Together, these findings indicated these modules are important for normal heart development and thus it was hypothesized might harbor novel CHD genes. To address this question, it was leveraged measures of selection against loss-of-function variants from whole exome and whole genome sequencing of over 140 thousand healthy controls by the Genome Aggregation Database. Genes scoring in the lowest 2 deciles of the loss-of-function observed/expected upper bound fraction (LOEUF) measure were shown to be enriched for known haploinsufficient genes and genes essential for survival in cell culture models. When these gene modules were analyzed, it was found significant enrichment of LOEUF decile 1 and 2 genes in the violet and brown4 modules, as well as many modules within expression groups 1 through 4 but not those outside these groups.

Notably, the violet module is strongly enriched for sarcomere-related genes, and the known CHD gene *NKX2-5* was a top scoring hub node (**FIG. 8B**). This TF has been previously implicated in cardiac cell structure and function, particularly in sarcomere organization, and falls into the second decile of LOEUF scores. Given the purported importance of hub genes from WGCNA analyses in disease and a critical cardiac TF being identified as both a hub gene and intolerant to disruption, it was contemplated whether hub genes in our network might be generally intolerant to loss-of-function mutations in otherwise healthy individuals. When hub genes across the entire network were interrogated, significant enrichment of low LOEUF score genes and significant depletion of genes from the tenth decile was found (**FIG. 8C**). Genes that have not been implicated in CHD but are characterized by high connectivity in our network, heart-specific expression, and low LOEUF scores thus represent novel candidate CHD genes.

WGCNA reveals NKX2-5 regulatory program.

Herein, many of these results have confirmed a central role of *NKX2-5* in human heart organogenesis. Its binding sites are enriched in both EHEs and the promoters of genes with heart-specific expression. In hESC (human embryonic stem cell)-derived cardiomyocytes,

NKX2-5 bound the promoters and directly regulated the expression of many genes involved in heart development, specifically those involved in voltage-gated ion channel activity. Two prominent genes in the violet module that are directly regulated by NKX2-5 in human cardiomyocytes are the TFs *HEY2* and *IRX4* (**FIG. 8B**; **FIG. 25**). Both of these have been shown to play a role in ventricular myogenesis in mice and are linked to heart abnormalities in humans.

The existence of these direct targets in the same module built in an unsupervised fashion from only gene expression data suggests that these modules may reflect direct, physical connections of TFs and their target genes. To determine whether these observations were representative of larger patterns of regulation, we analyzed all WGCNA modules for enrichment of the 228 genes directly regulated by *NKX2-5* in human cardiomyocytes. It was found that indeed the violet module is enriched for direct regulatory targets, which were all well-connected hub genes, assigned embryonic heart-specific enhancers, and had elevated heart-specific expression values (Gini, ≥0.5; **FIG. 8B**).

It was also found enrichment of *NKX2-5* target genes in 4 other modules (green, lightgreen, skyblue3, and mediumpurple3) that are enriched for embryonic patterning, ion channel function, and mitosis (**FIG. 7C**; **FIGS. 26A–D**). Genes that have the same characteristics as *NKX2-5* in our networks, specifically high connectivity, tissue-specific expression, and potential regulation by heart-specific enhancers, are likely integral to normal heart development. Indeed, several genes identified in this data-driven fashion have been clearly linked to CHD as reported in Online Mendelian Inheritance in Man (OMIM) including *NKX2-6* and *TNNI3K*, while others have been implicated in mouse or are pending further findings in human to be considered a CHD gene, such as *IRX4*, *HEY2*, and *HCN4*.

Finally, to further understand the functional relevance of the gene coexpression network, it was reasoned that genes might be coexpressed with one another to physically interact. These modules were assessed for protein-protein interactions, and it was found that 15 of our modules contained significant protein-protein interactions. This was a highly significant result relative to randomly constructed modules (FIG. 8D), indicating this network has identified coherent connections for both gene regulation and physical interactions of proteins.

Discussion

Herein, it was jointly profiled chromatin state and gene expression during the organogenesis phase of human heart development. These efforts have dramatically increased the number of regulatory sequences that are predicted to be active in the developing human heart. The function and specificity of these regulatory elements by leveraging a variety of publicly available data sources including TF binding, in vivo enhancer assays, genome-wide association studies, and chromosome conformation were validated.

Tens of thousands of novel regulatory sequences likely active specifically in the developing human heart were identified. Even though these are newly identified sequences, they display characteristics familiar to the heart development field. For example, they are enriched with binding sites for families of TFs long appreciated to play a role in heart development such as GATA, MEF2, and NKX and are located in close proximity to genes with documented roles in mammalian heart development. These systematic comparisons of gene expression between the developing heart and 25 other human tissues have allowed us to identify a set of genes with high specificity for expression during heart organogenesis. The embryonic heart–specific enhancers we have identified are enriched near these developing heart-specific genes. Furthermore, the number of heart-specific enhancers that can be assigned to a gene is correlated with relative expression level compared with all available human tissues. These findings strongly support a direct role for these regulatory sequences in driving embryonic heart–specific gene expression patterns.

In addition to individual regulatory sequences, the localized landscapes of coactivated enhancers commonly referred to as super enhancers were characterized. Large regulatory landscapes have been implicated in tissue specification and tumorigenesis and encompass genes important for these processes. In this data, thousands of super enhancers across all time points of embryonic heart development were identified, 1611 of which had not been previously annotated. These include embryonic heart–specific super enhancers near important heart genes, *TBX20*, *GATA4*, *HAND1*, *HCN4*, *IRX3*, *IRX4*, and *IRX6*, among others. Additionally, a significant majority of genes we identified with strong heart-specific expression, including *UNC45B*, *METTL11B*, *MYL3*, *MYL4*, *MYL7*, and *AFP*, are directly adjacent or completely encompassed by super enhancers active in the embryonic heart. These findings

demonstrate the large number of regulatory inputs that control tissue-specification genes during heart development and suggest a high level of redundancy to maintain proper gene control.

Numerous studies have shown tissue-specific regulatory regions are enriched for variants associated with tissue-specific phenotypes and diseases. This analysis strongly confirmed these previous results. Specifically, it was shown that variants associated with quantitative differences in a variety of heart phenotypes such as resting heart rate and pattern of beating were enriched most strongly in embryonic heart enhancers. Most strikingly, consistent enrichment of variants associated with atrial fibrillation in strong enhancer segments across all of embryonic heart development were observed. Recent studies found significant enrichment for AF-associated index variants in accessible chromatin from fetal heart and speculated on a fetal origin of AF. In the data herein, variants that confer risk of AF are in established enhancers active even earlier, specifically those active at 4.5 weeks post-conception during the embryonic period of development. This suggests either AF causes a reversion to an embryonic pattern of gene regulation or predisposition to adult onset AF is established extremely early in development. Further study of this phenomenon is clearly warranted and could indicate AF should be considered a CHD.

One of the most interesting findings of the GWAS analysis came from analysis of the autoimmune disorder lupus that had been initially intended as a control. Heart disease is a frequent complication in lupus patients and is typically attributed to inflammation of the heart due to lupus-specific antibodies that direct immune cells to attack this organ. These results could suggest that inflammation activates portions of the genome that are normally repressed in heart development or that inappropriate activation of regions during heart development predisposes one to these complications. Importantly, these results revealed that repressed regions can also be systematically enriched for specific disease-relevant loci, which has not been previously described for chromatin state segmentation data.

It has been confirmed for the first time jointly analyzed gene expression and chromatin state for embryonic human heart development. This WGCNA identified modules of genes in an unbiased way, yet when it was analyzed, these groups of genes, it was uncovered coherent biological functions and expression characteristics across this developmental trajectory. It was found that a subset of modules is significantly enriched for both heart-specific gene expression

and heart-specific enhancer activation. Many of these same modules are also enriched for known CHD genes, and well-connected or hub genes in these modules are generally intolerant to loss-of-function mutations in otherwise healthy individuals. When these networks with binding and functional data for the cardiac TF *NKX2-5* were systematically interrogated, a clear physical regulatory connections both within the *NKX2-5*–containing module and other modules was uncovered. These *NKX2-5* connected modules were enriched for functions this gene has been suggested to regulate, including activation of sarcomere and ion channel genes and repression of neurogenesis. These findings demonstrate that these networks represent real biological connections relevant to heart development. Genes with characteristics similar to *NKX2-5*, such as specificity of expression, highly connected to other genes in WGNCA modules, regulated by heart-specific enhancers, and low tolerance to gene disruption, are, therefore, prime candidates for CHD genes.

All these datasets openly in commonly used formats are directly comparable to other large consortia. These can be downloaded from Gene Expression Omnibus, retrieved via public track hub functionality on the UCSC Genome Browser, or directly from the Cotney Lab website. The Gini index scoring for over 36 000 genes from 31 tissues is broadly useful for studies of all of these tissues and identifies both housekeeping and highly tissue-specific genes. It is anticipated that the regulatory regions and genes identified herein will be useful in research on the regulation of heart development, can be used as tissue-specific drivers of reporters, and provide much needed functional context for noncoding variation in clinical whole genome sequencing.

It is understood that the foregoing detailed description and accompanying example is merely illustrative and are not to be taken as limitations upon the scope of the invention, which is defined solely by the appended claims and their equivalents.

Various changes and modifications to the disclosed embodiments will be apparent to those skilled in the art. Such changes and modifications, including without limitation those relating to the chemical structures, substituents, derivatives, intermediates, syntheses, compositions, formulations, or methods of use of the invention, may be made without departing from the spirit and scope thereof.

We claim:

1. A method of identifying a subject as at risk of having a congenital heart defect, the method comprising:
   a. assessing a plurality of genes in a sample obtained from the subject, wherein the plurality of genes are selected from the genes listed in Table 1; and
   b. identifying the subject as at risk of having a congenital heart defect based upon the assessment of the plurality of genes.

2. The method of claim 1, wherein the plurality of genes comprises at least 5 genes listed in Table 1.

3. The method of claim 1, wherein the plurality of genes comprises at least 10 genes listed in Table 1.

4. The method of claim 1, wherein the plurality of genes comprises at least 25 genes listed in Table 1.

5. The method of claim 1, wherein the plurality of genes comprises at least 50 genes listed in Table 1.

6. The method of claim 1, wherein the plurality of genes comprises at least 100 genes listed in Table 1.

7. The method of claim 1, wherein the plurality of genes comprises at least 150 genes listed in Table 1.

8. The method of claim 1, wherein the plurality of genes comprises at least 200 genes listed in table 1.

9. The method of claim 1, wherein the plurality of genes comprises at least 250 genes listed in Table 1.

10. The method of claim 1, wherein the plurality of genes comprises all of the genes listed in Table 1.

11. The method of any one of claims 1-10, wherein assessing the plurality of genes comprises determining a copy number for at least one of the plurality of genes.

12. The method of any one of the preceding claims, wherein assessing the plurality of genes comprises detecting the presence of one or more single nucleotide polymorphisms (SNPs) for at least one of the plurality of genes.

13. The method of any one of the preceding claims, wherein assessing the plurality of genes comprises measuring expression of a product of at least one of the plurality of genes.

14. A panel of genes for assessing risk of congenital heart defects, the panel comprising at least 10 genes listed in Table 1.

15. The panel of claim 14, comprising at least 25 genes listed in Table 1.

16. The panel of claim 14, comprising at least 50 genes listed in Table 1.

17. The panel of claim 14, comprising at least 100 genes listed in Table 1.

18. The panel of claim 14, comprising at least 150 genes listed in Table 1.

19. The panel of claim 14, comprising at least 200 genes listed in Table 1.

20. The panel of claim 14, comprising at least 250 genes listed in Table 1.

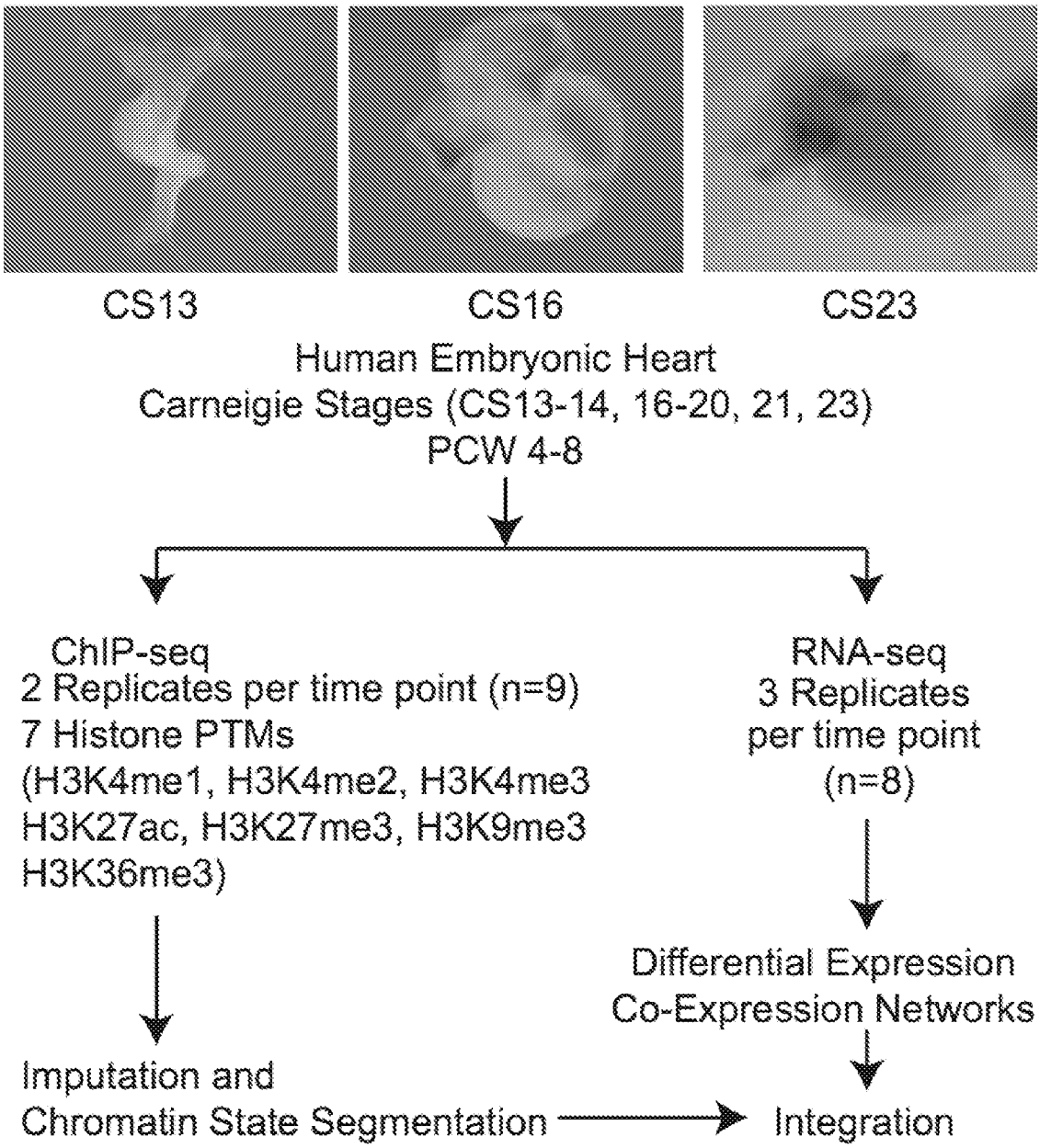

CS13
CS16
CS23
Human Embryonic Heart
Carneigie Stages (CS13-14, 16-20, 21, 23)
PCW 4-8
ChIP-seq
2 Replicates per time point (n=9)
7 Histone PTMs
(H3K4me1, H3K4me2, H3K4me3
H3K27ac, H3K27me3, H3K9me3
H3K36me3)
RNA-seq
3 Replicates
per time point
(n=8)
Differential Expression
Co-Expression Networks
Imputation and
Chromatin State Segmentation
Integration

FIG. 1A

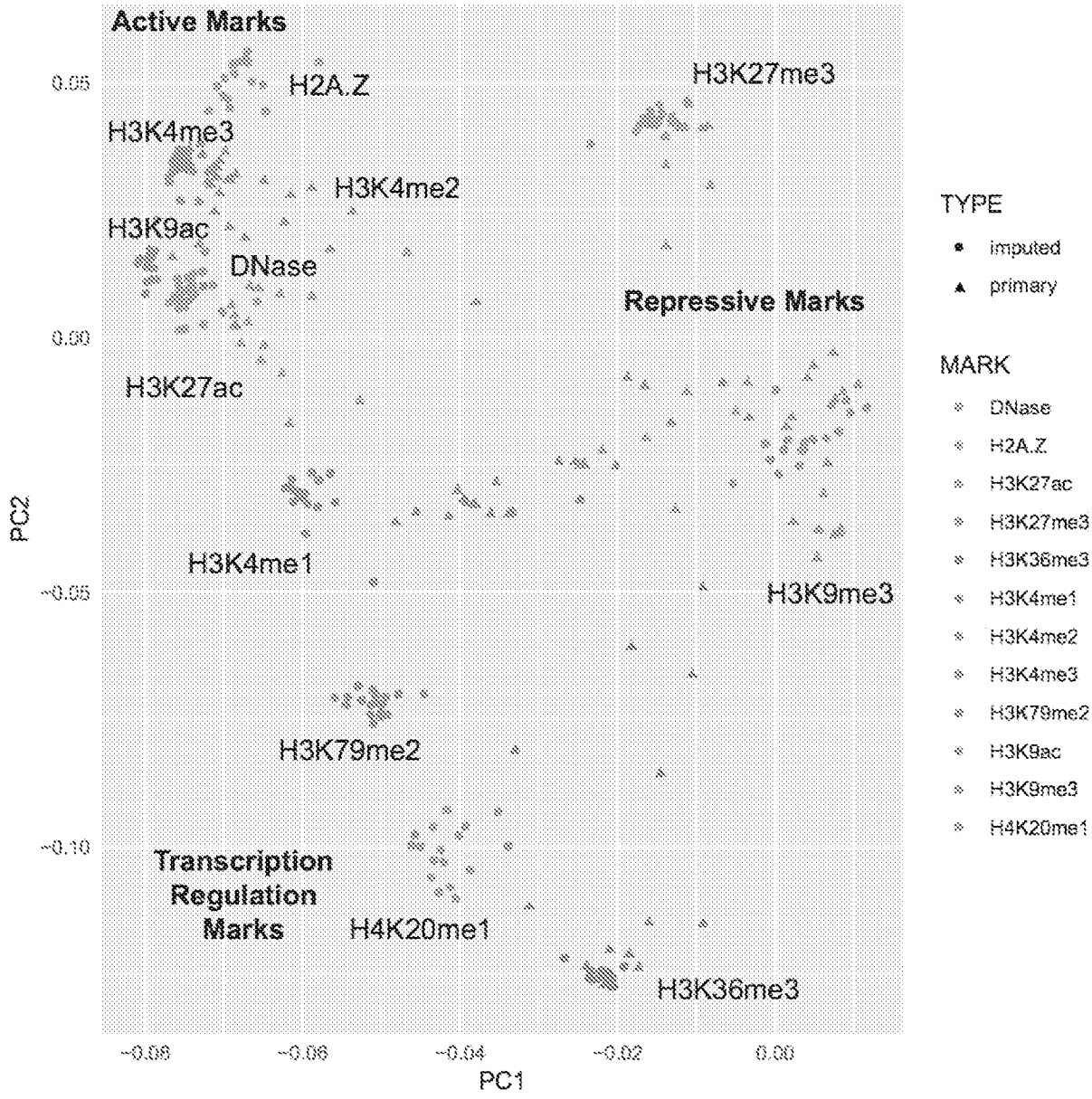

Active Marks
H2A.Z
H3K4me3
H3K4me2
H3K9ac
DNase
H3K27ac
H3K4me1
H3K79me2
Transcription Regulation Marks
H4K20me1
H3K36me3
H3K27me3
Repressive Marks
H3K9me3
0.05
0.00
−0.05
−0.10
PC2
−0.08
−0.06
−0.04
−0.02
0.00
PC1
TYPE
imputed
primary
MARK
DNase
H2A.Z
H3K27ac
H3K27me3
H3K36me3
H3K4me1
H3K4me2
H3K4me3
H3K79me2
H3K9ac
H3K9me3
H4K20me1

FIG. 1B

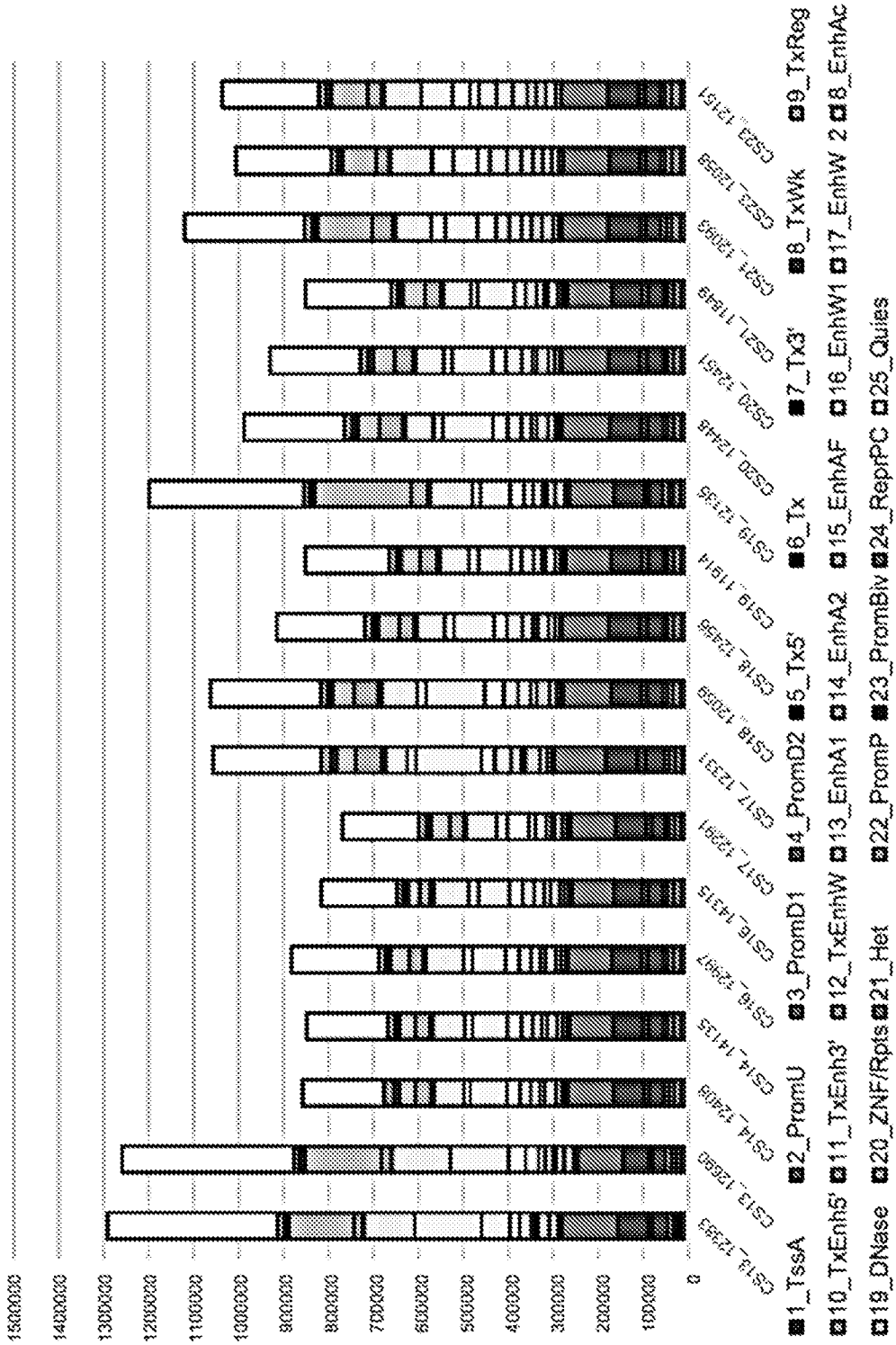

Cumulative Number of Chromatin States
1_TssA
2_PromU
3_PromD1
4_PromD2
5_Tx5'
6_Tx
7_Tx3'
8_TxWk
9_TxReg
10_TxEnh5'
11_TxEnh3'
12_TxEnhW
13_EnhA1
14_EnhA2
15_EnhAF
16_EnhW1
17_EnhW 2
8_EnhAc
19_DNase
20_ZNF/Rpts
21_Het
22_PromP
23_PromBiv
24_ReprPC
25_Quies

FIG. 1C

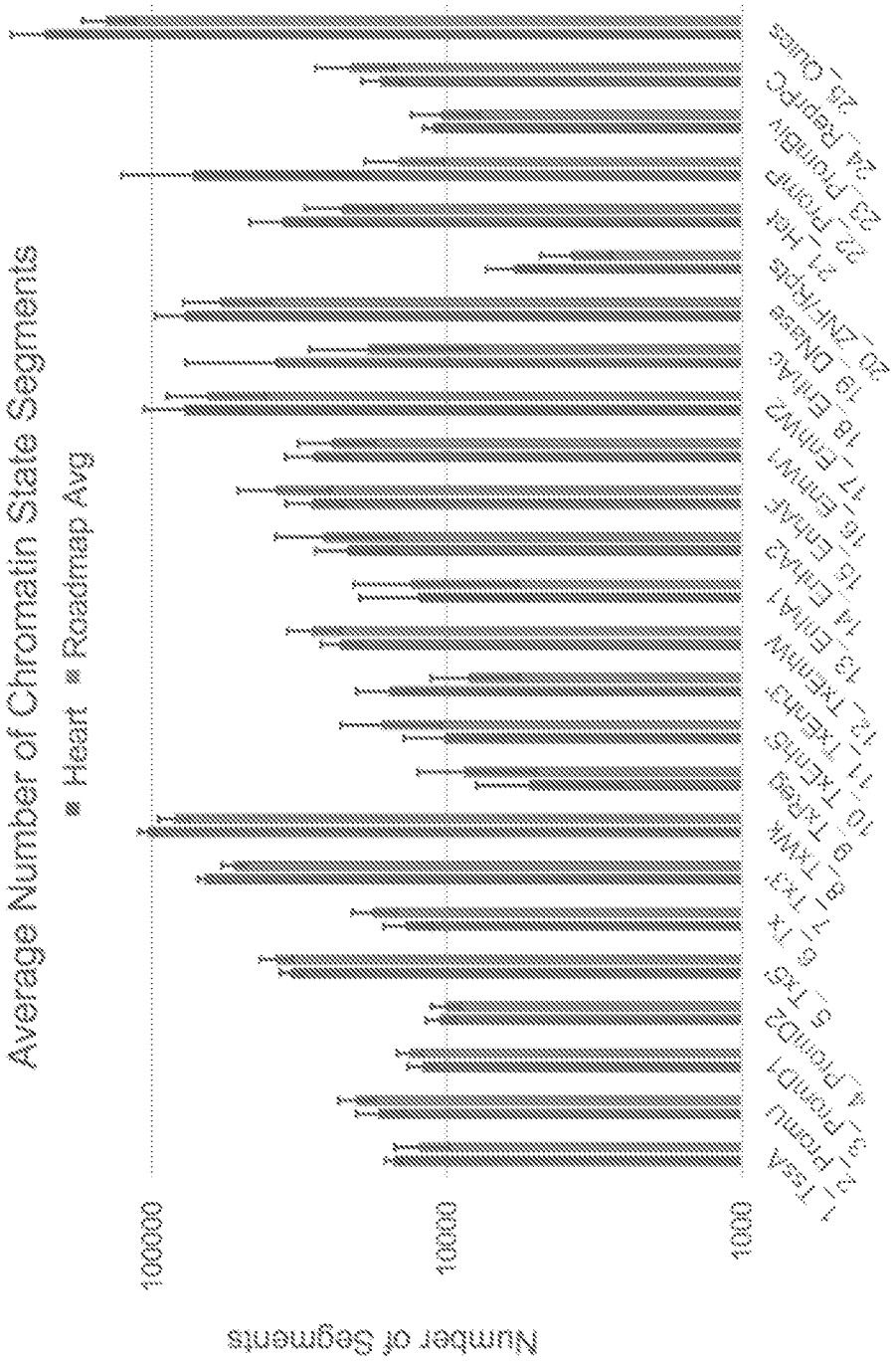

Average Number of Chromatin State Segments
Heart
Roadmap Avg
Number of Segments
100000
10000
1000
1_TssA
2_PromU
3_PromD2
4_PromD1
5_TssF
6_Tx
7_Tx3
8_TxWk
9_TxReg
10_TxEnh5
11_TxEnh3
12_TxEnhW
13_EnhA1
14_EnhA2
15_EnhAF
16_EnhW1
17_EnhW2
18_EnhAc
19_DNase
20_ZNF/Rpts
21_Het
22_PromP
23_PromBiv
24_ReprPC
25_Quies

FIG. 1D

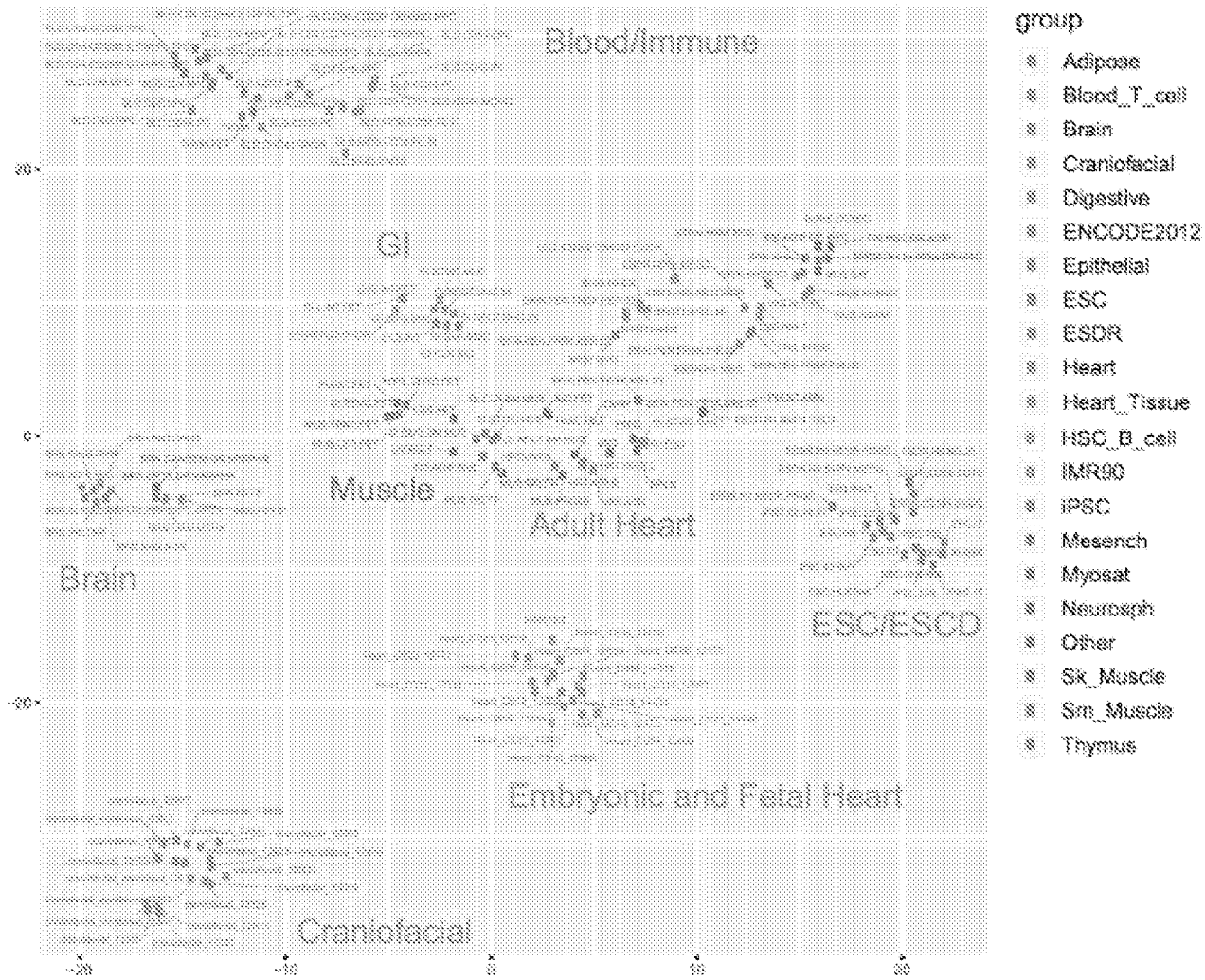

Blood/Immune
GI
Muscle
Adult Heart
Brain
ESC/ESCD
Embryonic and Fetal Heart
Craniofacial
group
Adipose
Blood_T_cell
Brain
Craniofacial
Digestive
ENCODE2012
Epithelial
ESC
ESDR
Heart
Heart_Tissue
HSC_B_cell
IMR90
iPSC
Mesench
Myosat
Neurosph
Other
Sk_Muscle
Sm_Muscle
Thymus

FIG. 2A

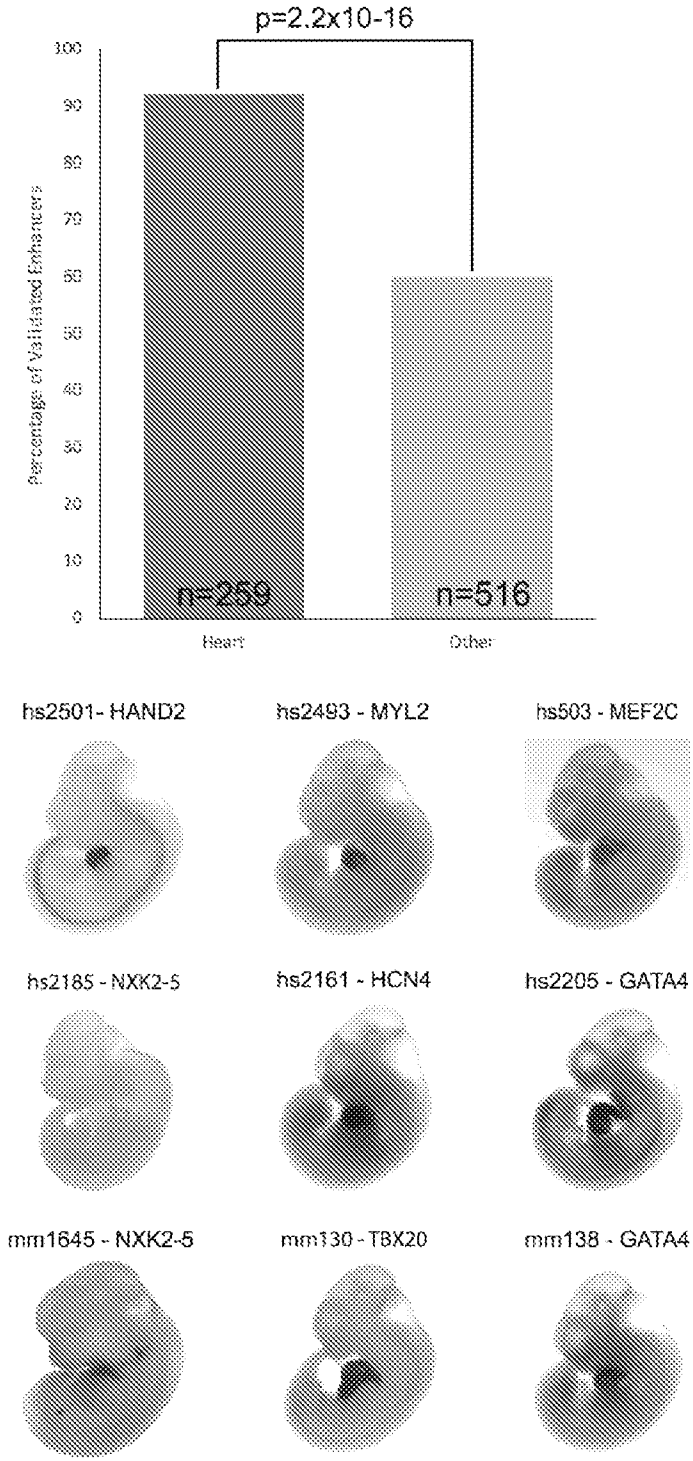

p=2.2x10-16
Percentage of Validated Enhancers
100
90
80
70
60
50
40
30
20
10
0
n=259
n=516
Heart
Other
hs2501 - HAND2
hs2493 - MYL2
hs503 - MEF2C
hs2185 - NXK2-5
hs2161 - HCN4
hs2205 - GATA4
mm1645 - NXK2-5
mm130 - TBX20
mm138 - GATA4

FIG. 2B

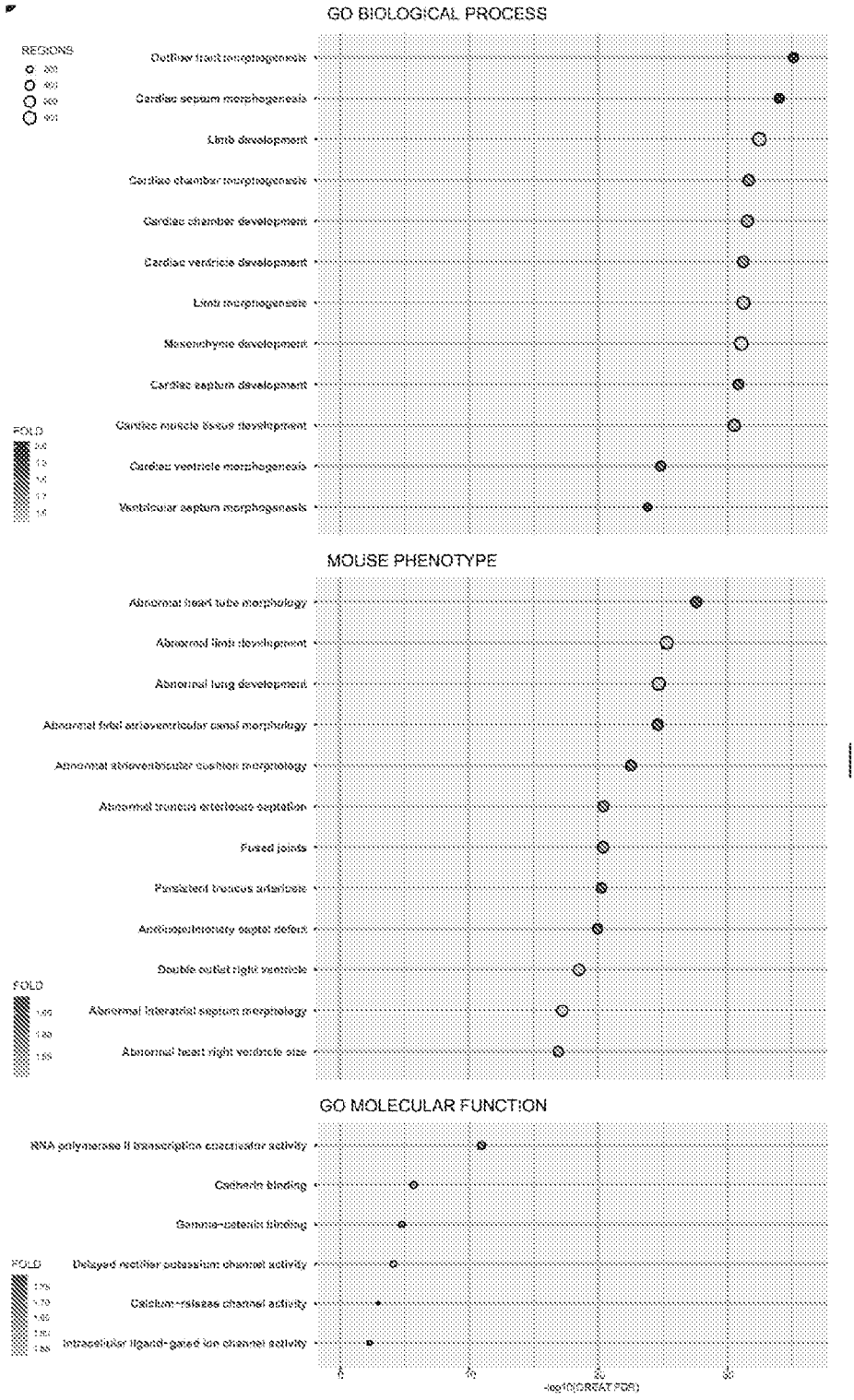

GO BIOLOGICAL PROCESS
REGIONS
Outflow tract morphogenesis
Cardiac septum morphogenesis
Limb development
Cardiac chamber morphogenesis
Cardiac chamber development
Cardiac ventricle development
Limb morphogenesis
Mesenchyme development
Cardiac septum development
Cardiac muscle tissue development
Cardiac ventricle morphogenesis
Ventricular septum morphogenesis
FOLD
MOUSE PHENOTYPE
Abnormal heart tube morphology
Abnormal limb development
Abnormal lung development
Abnormal fetal atrioventricular canal morphology
Abnormal atrioventricular cushion morphology
Abnormal truncus arteriosus septation
Fused joints
Persistent truncus arteriosus
Anomalous coronary ostial defect
Double outlet right ventricle
Abnormal interatrial septum morphology
Abnormal heart right ventricle size
FOLD
GO MOLECULAR FUNCTION
RNA polymerase II transcription coactivator activity
Cadherin binding
Gamma-catenin binding
Delayed rectifier potassium channel activity
Calcium-release channel activity
Intracellular ligand-gated ion channel activity
-log10(GREAT FDR)

FIG. 2C

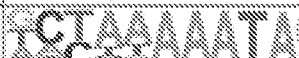
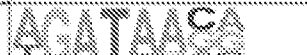
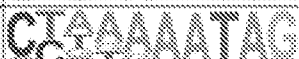
	MOTIF	NAME	P-value
1		Gata2	$1x10^{-137}$
2		Gata1	$1x10^{-129}$
3		Mef2d	$1x10^{-120}$
4		Gata6	$1x10^{-119}$
5		Gata4	$1x10^{-110}$
6		Mef2c	$1x10^{-91}$
7		Tbx20	$1x10^{-75}$
8		Gata3	$1x10^{-64}$
9		Mef2b	$1x10^{-55}$
10		Mef2a	$1x10^{-50}$
11		NF1	$1x10^{-45}$
12		Meis1	$1x10^{-35}$
13		Hand2	$1x10^{-26}$
14		Esrrb	$1x10^{-21}$
15		Pitx1	$1x10^{-21}$

FIG. 2D

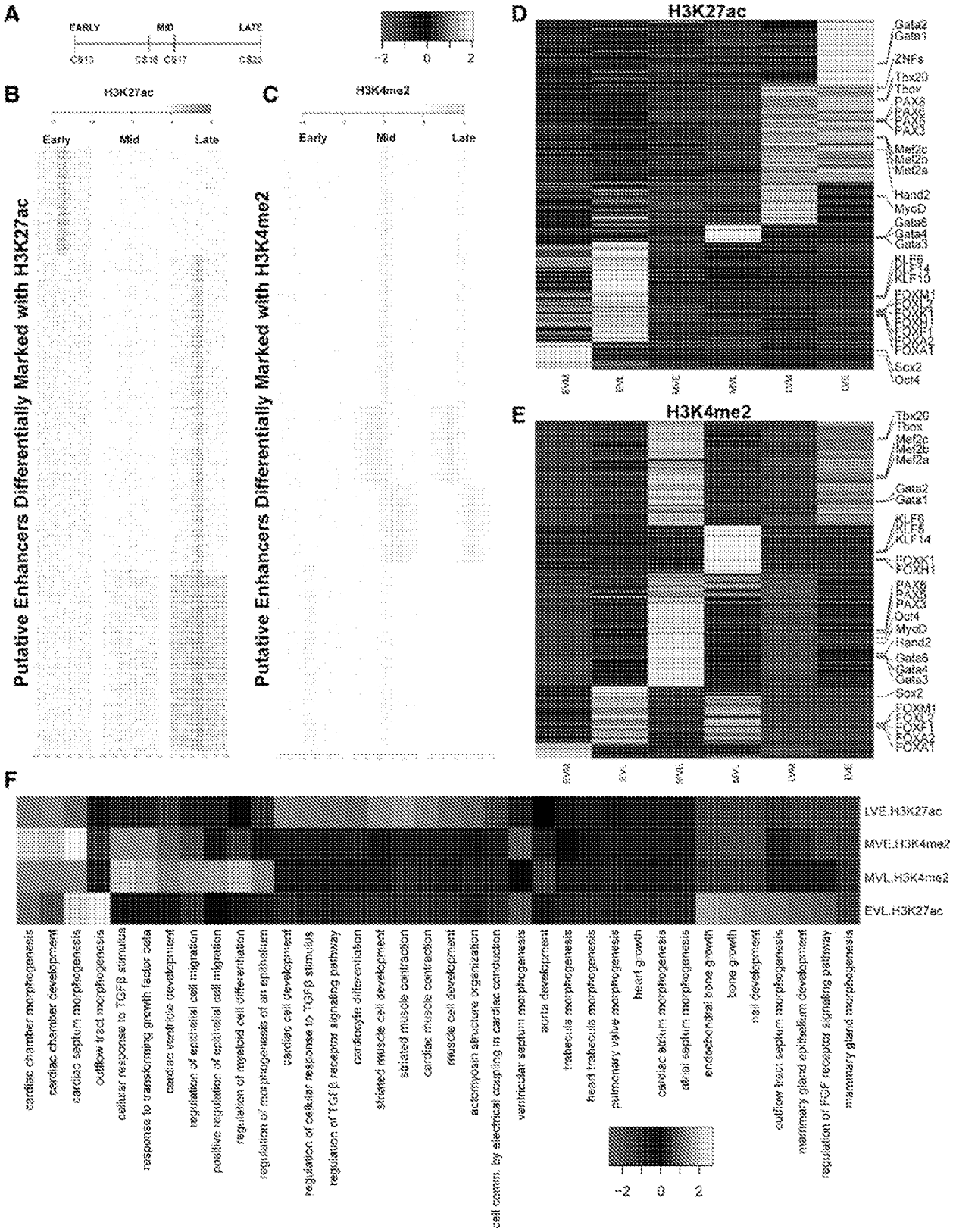

A
EARLY
MID
LATE
CS13
CS16
CS17
CS23
B
H3K27ac
Early
Mid
Late
Putative Enhancers Differentially Marked with H3K27ac
C
H3K4me2
Early
Mid
Late
Putative Enhancers Differentially Marked with H3K4me2
D
H3K27ac
Gata2
Gata1
ZNFs
Tbx20
Tbox
PAX6
PAX5
PAX3
Mef2c
Mef2b
Mef2a
Hand2
MyoD
Gata6
Gata4
Gata3
KLF9
KLF14
KLF10
FOXM1
FOXK1
FOXH1
FOXA2
FOXA1
Sox2
Oct4
EVE
EVL
MVE
MVL
LVE
LVL
E
H3K4me2
Tbx20
Tbox
Mef2c
Mef2b
Mef2a
Gata2
Gata1
KLF9
KLF5
KLF14
FOXK1
FOXH1
PAX6
PAX5
PAX3
Oct4
MyoD
Hand2
Gata6
Gata4
Gata3
Sox2
FOXM1
FOXK1
FOXH1
FOXA2
FOXA1
EVE
EVL
MVE
MVL
LVE
LVL
F
LVE.H3K27ac
MVE.H3K4me2
MVL.H3K4me2
EVL.H3K27ac
cardiac chamber morphogenesis
cardiac chamber development
cardiac septum morphogenesis
outflow tract morphogenesis
cellular response to TGFβ stimulus
response to transforming growth factor beta
cardiac ventricle development
regulation of epithelial cell migration
positive regulation of epithelial cell migration
regulation of myeloid cell differentiation
regulation of morphogenesis of an epithelium
cardiac cell development
regulation of cellular response to TGFβ stimulus
regulation of TGFβ receptor signaling pathway
cardiocyte differentiation
striated muscle cell development
striated muscle contraction
cardiac muscle contraction
muscle cell development
actomyosin structure organization
cell comm. by electrical coupling in cardiac conduction
ventricular septum morphogenesis
aorta development
trabecula morphogenesis
heart trabecula morphogenesis
pulmonary valve morphogenesis
heart growth
cardiac atrium morphogenesis
atrial septum morphogenesis
endochondral bone growth
bone growth
nail development
outflow tract septum morphogenesis
mammary gland epithelium development
regulation of FGF receptor signaling pathway
mammary gland morphogenesis
-2
0
2

FIG. 3

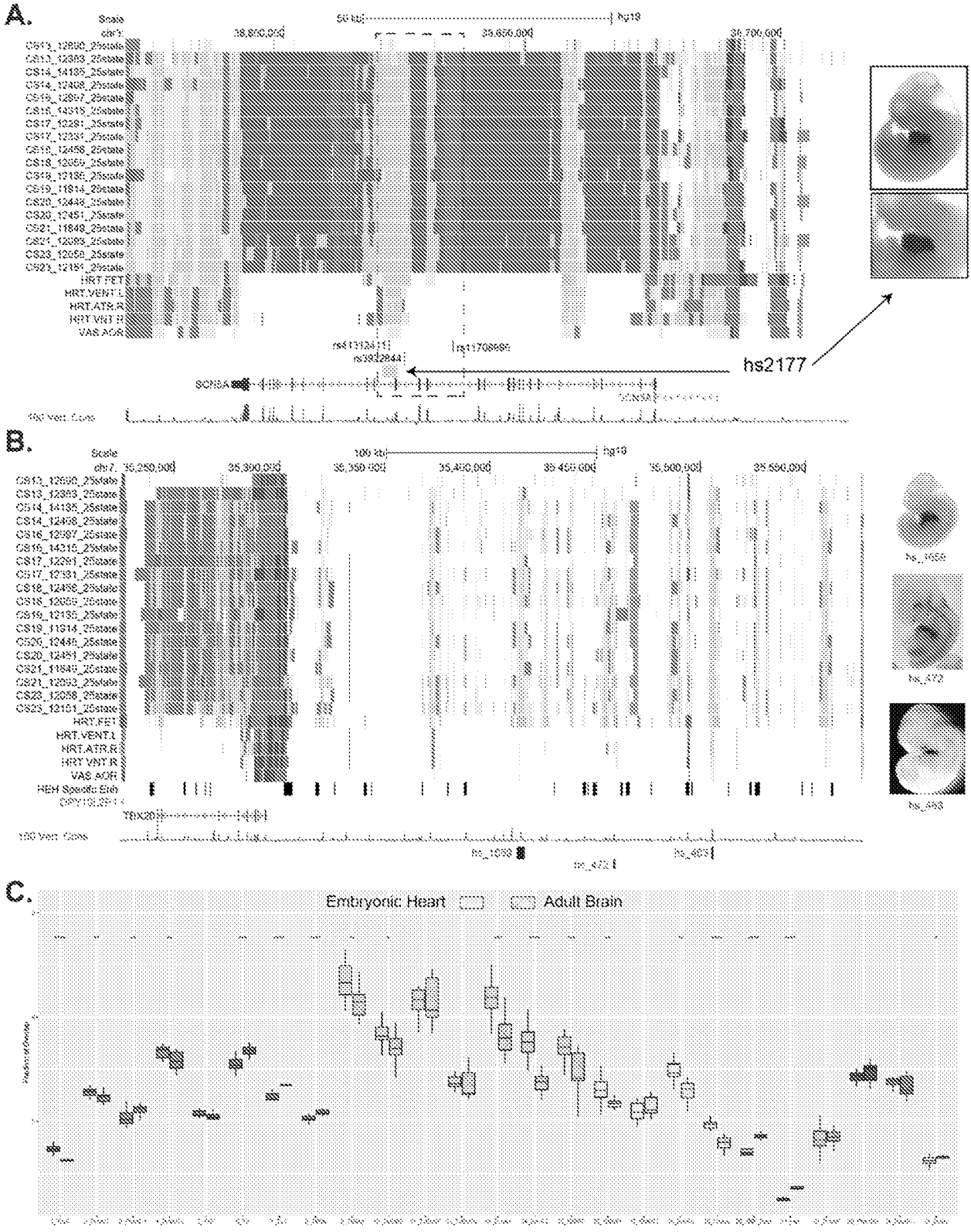


FIG. 4

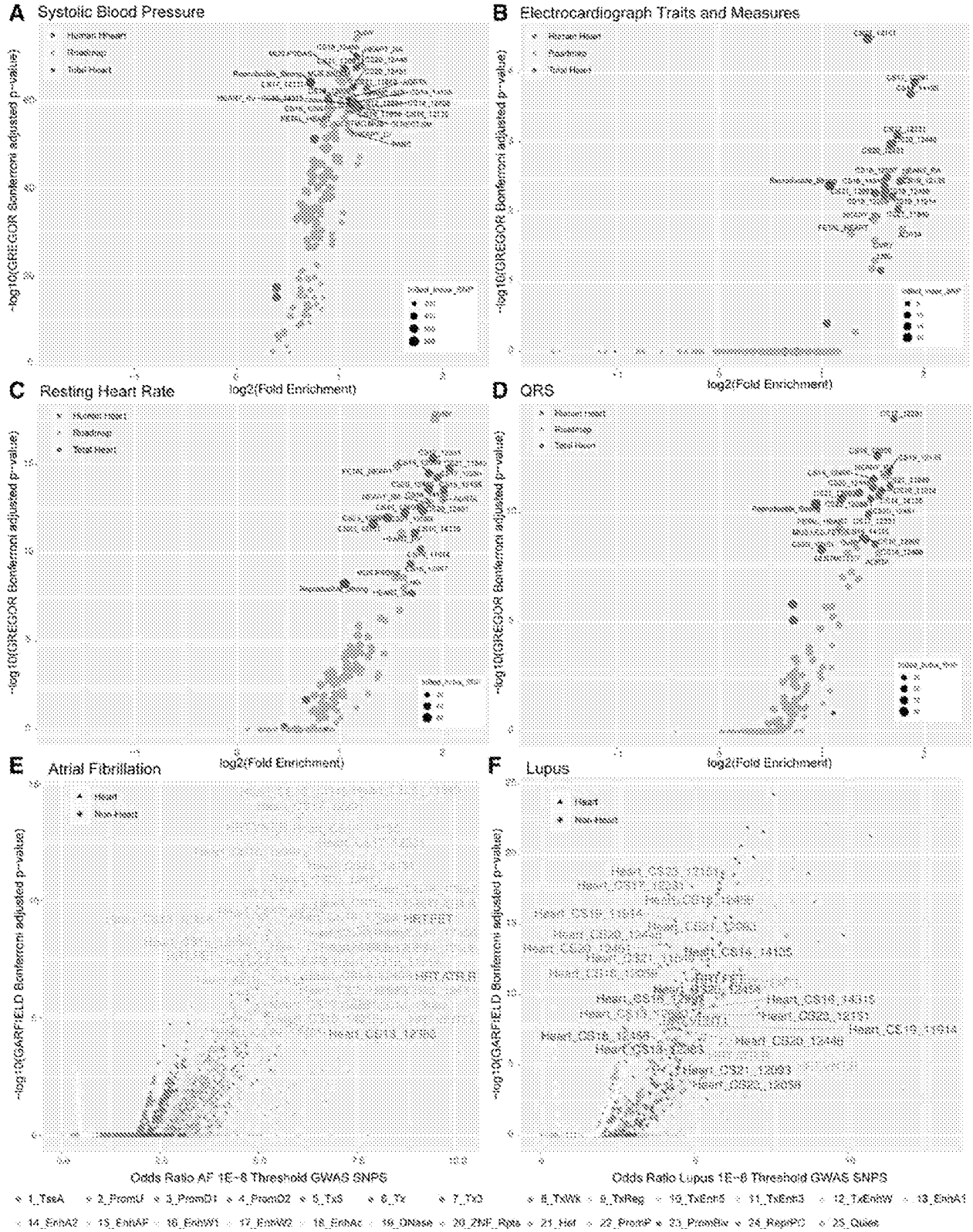

A Systolic Blood Pressure
B Electrocardiograph Traits and Measures
C Resting Heart Rate
D QRS
E Atrial Fibrillation
F Lupus
-log10(GREGOR Bonferroni adjusted p-value)
log2(Fold Enrichment)
-log10(GARFIELD Bonferroni adjusted p-value)
Odds Ratio AF 1E-6 Threshold GWAS SNPS
Odds Ratio Lupus 1E-6 Threshold GWAS SNPS
1_TssA
2_PromU
3_PromD1
4_PromD2
5_TxS
6_Tx
7_Tx2
8_TxWk
9_TxReg
10_TxEnh5
11_TxEnh3
12_TxEnhW
13_EnhA1
14_EnhA2
15_EnhAF
16_EnhW1
17_EnhW2
18_EnhAc
19_DNase
20_ZNF_Rpts
21_Het
22_PromP
23_PromBiv
24_ReprPC
25_Quies

FIG. 5

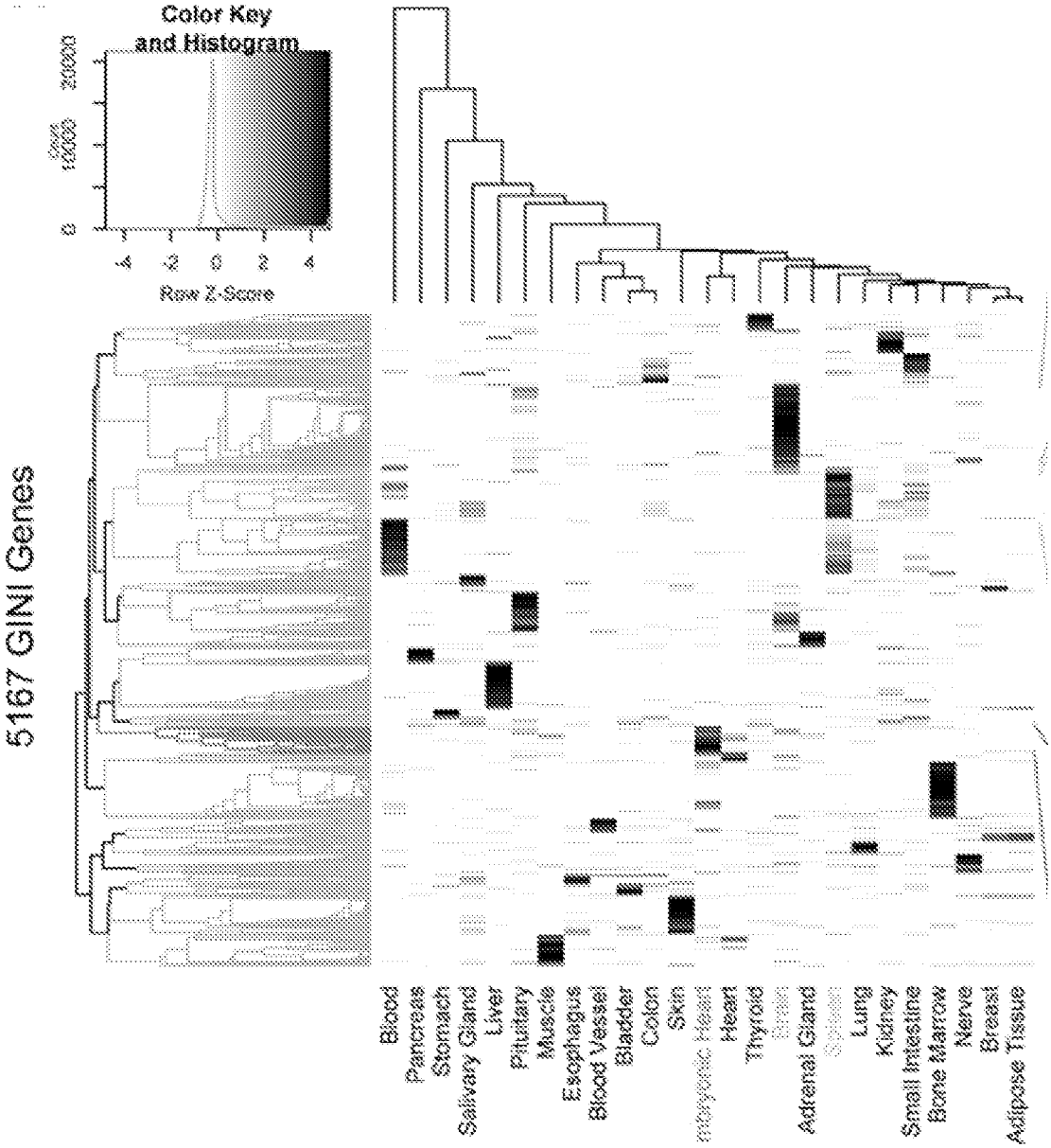

Color Key
and Histogram
Count
0
100000
200000
-4
-2
0
2
4
Row Z-Score
5167 GINI Genes
Blood
Pancreas
Stomach
Salivary Gland
Liver
Pituitary
Muscle
Esophagus
Blood Vessel
Bladder
Colon
Skin
Heart
Thyroid
Adrenal Gland
Lung
Kidney
Small Intestine
Bone Marrow
Nerve
Breast
Adipose Tissue

FIG. 6A

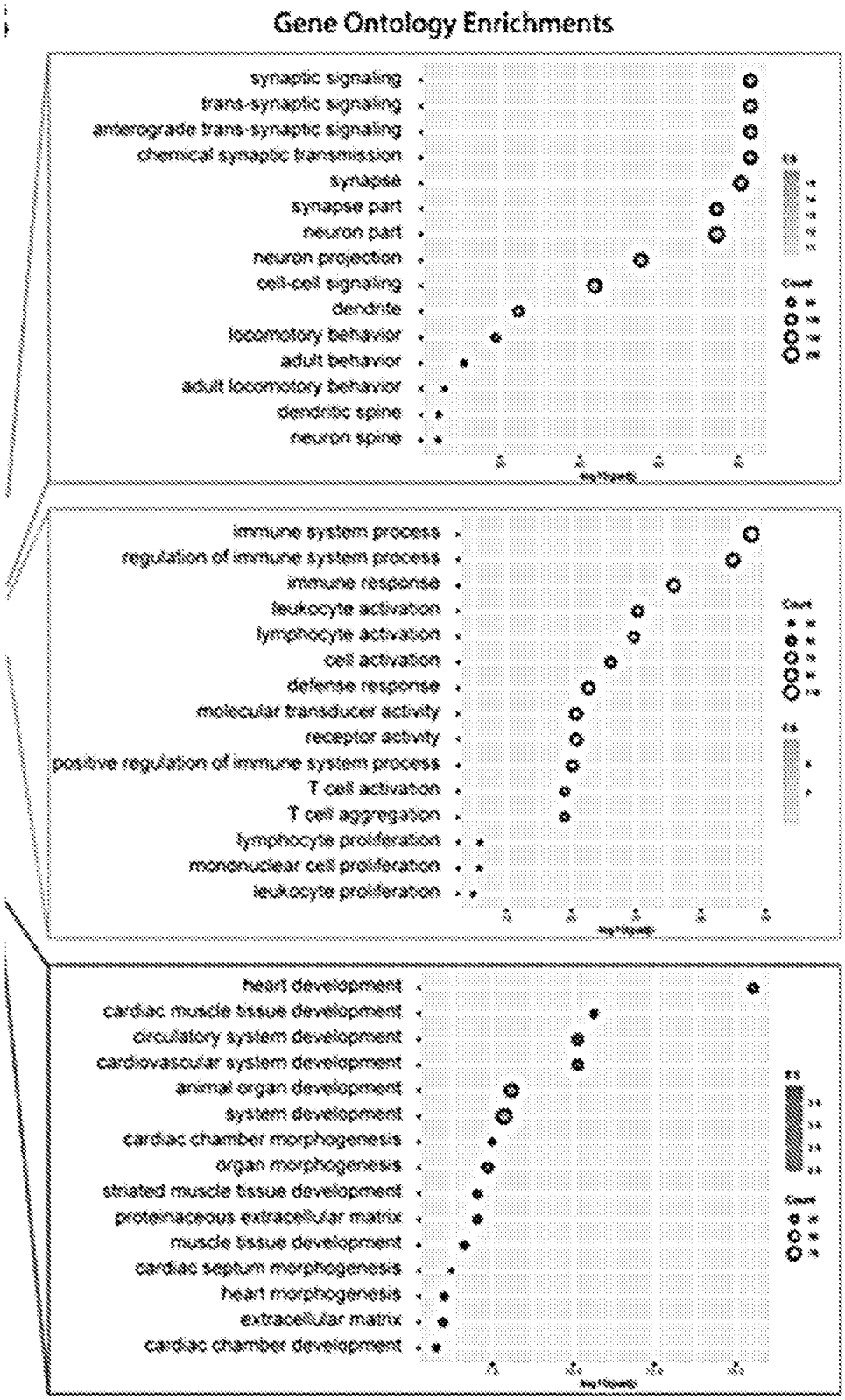

Gene Ontology Enrichments
synaptic signaling
trans-synaptic signaling
anterograde trans-synaptic signaling
chemical synaptic transmission
synapse
synapse part
neuron part
neuron projection
cell-cell signaling
dendrite
locomotory behavior
adult behavior
adult locomotory behavior
dendritic spine
neuron spine
immune system process
regulation of immune system process
immune response
leukocyte activation
lymphocyte activation
cell activation
defense response
molecular transducer activity
receptor activity
positive regulation of immune system process
T cell activation
T cell aggregation
lymphocyte proliferation
mononuclear cell proliferation
leukocyte proliferation
heart development
cardiac muscle tissue development
circulatory system development
cardiovascular system development
animal organ development
system development
cardiac chamber morphogenesis
organ morphogenesis
striated muscle tissue development
proteinaceous extracellular matrix
muscle tissue development
cardiac septum morphogenesis
heart morphogenesis
extracellular matrix
cardiac chamber development

FIG. 6B

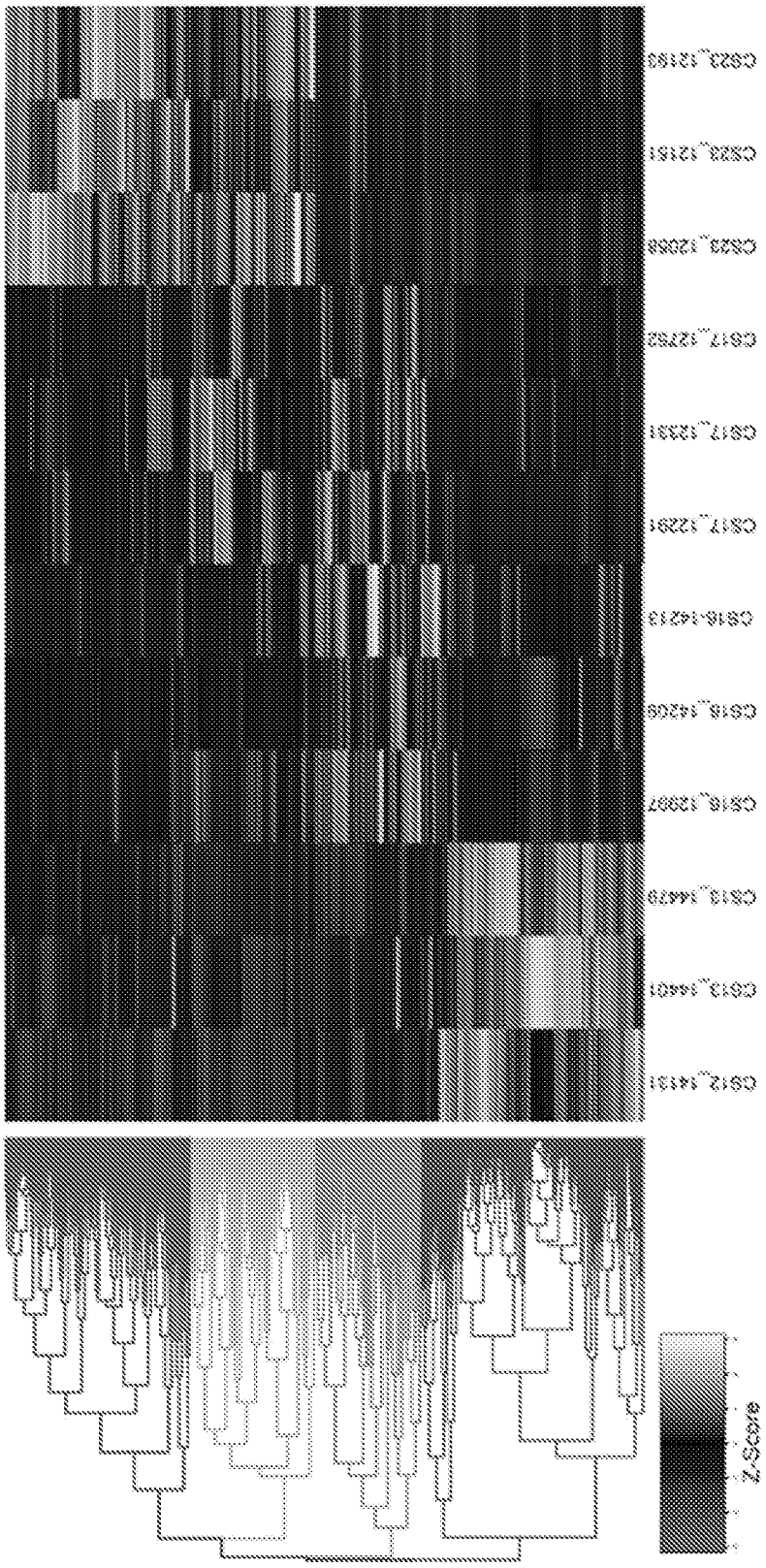

CS23_12183
CS23_12181
CS23_12208
CS17_12282
CS17_12231
CS17_12296
CS16-16213
CS16_14208
CS16_12307
CS13_14479
CS13_14401
CS12_14131
Z-Score

FIG. 6C

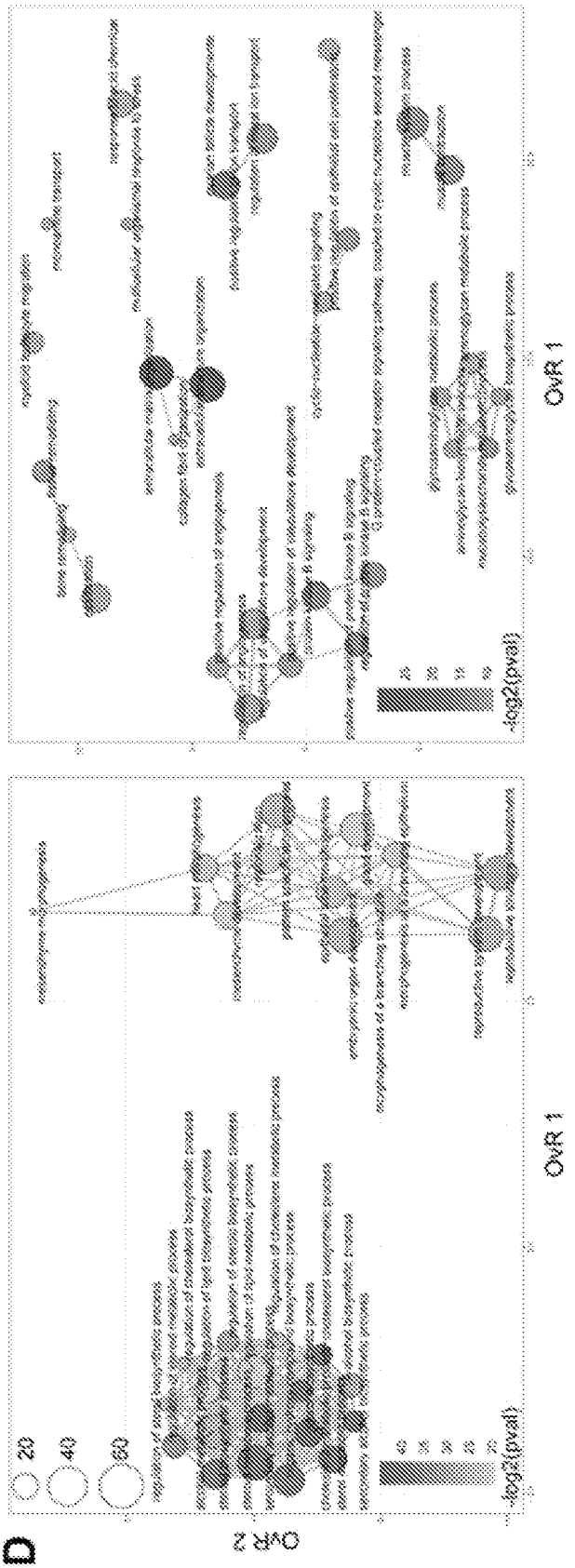


FIG. 6D

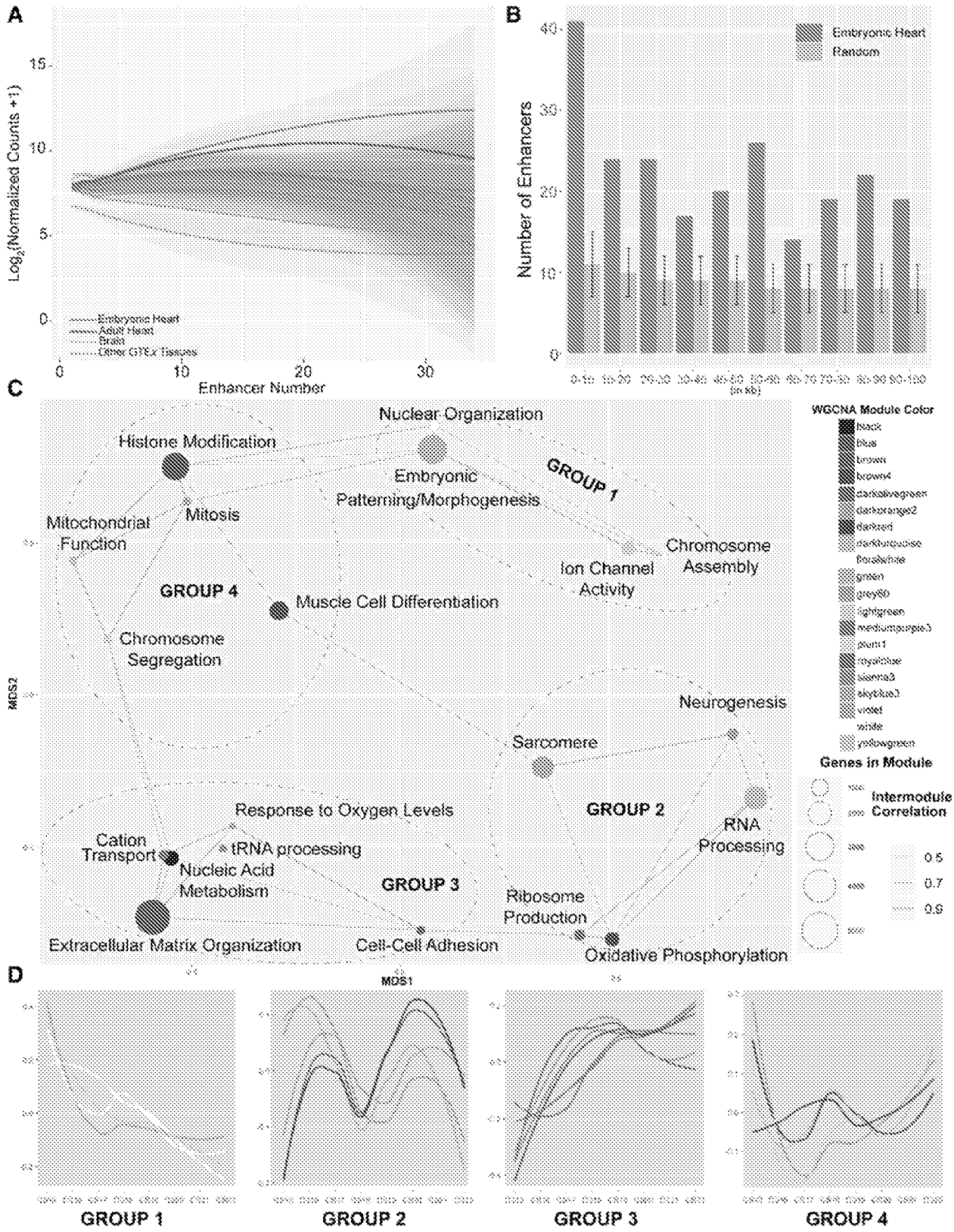

A
Log2(Normalized Counts +1)
Enhancer Number
Embryonic Heart
Adult Heart
Brain
Other GTEx Tissues
B
Number of Enhancers
Embryonic Heart
Random
C
Nuclear Organization
Histone Modification
Embryonic Patterning/Morphogenesis
GROUP 1
Mitochondrial Function
Mitosis
Chromosome Assembly
Ion Channel Activity
GROUP 4
Muscle Cell Differentiation
Chromosome Segregation
Neurogenesis
Sarcomere
Response to Oxygen Levels
GROUP 2
RNA Processing
Cation Transport
tRNA processing
Nucleic Acid Metabolism
GROUP 3
Ribosome Production
Extracellular Matrix Organization
Cell-Cell Adhesion
Oxidative Phosphorylation
MDS1
MDS2
WGCNA Module Color
Genes in Module
Intermodule Correlation
D
GROUP 1
GROUP 2
GROUP 3
GROUP 4

FIG. 7

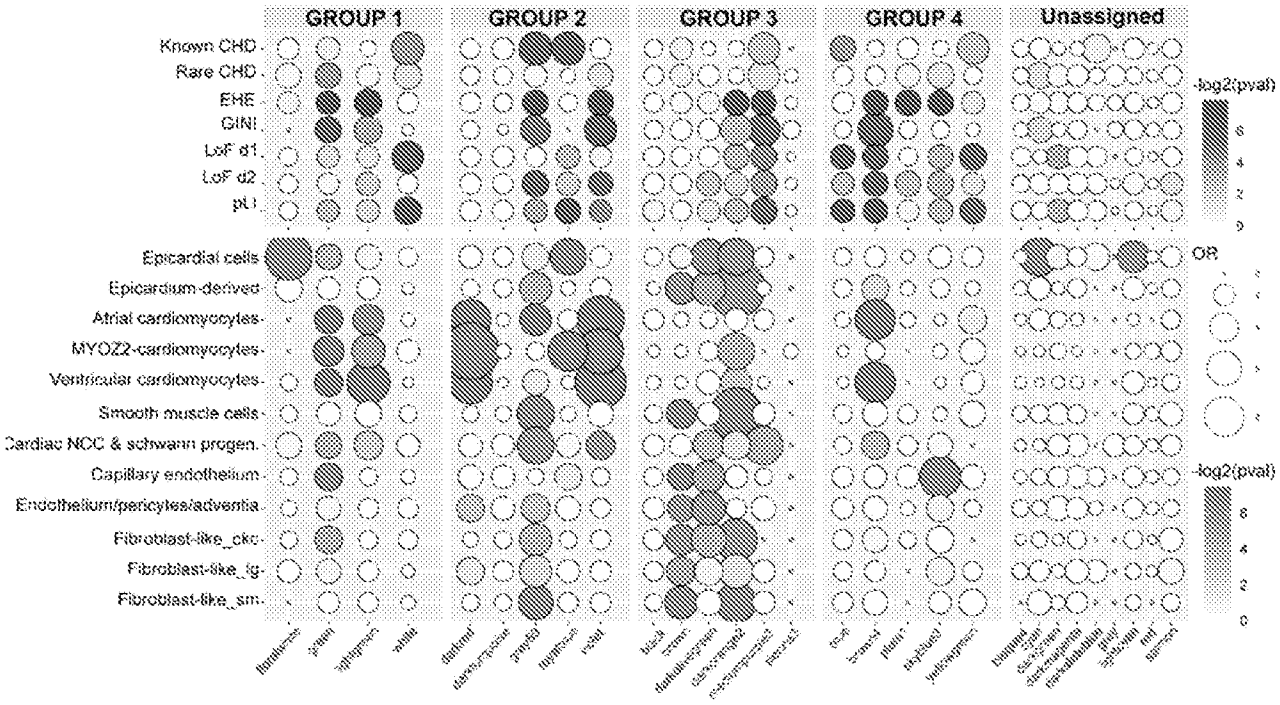

GROUP 1
GROUP 2
GROUP 3
GROUP 4
Unassigned
Known CHD
Rare CHD
EHE
GINI
LoF d1
LoF d2
pLI
-log2(pval)
Epicardial cells
Epicardium-derived
Atrial cardiomyocytes
MYOZ2-cardiomyocytes
Ventricular cardiomyocytes
Smooth muscle cells
Cardiac NCC & schwann progen.
Capillary endothelium
Endothelium/pericytes/adventia
Fibroblast-like_ckc
Fibroblast-like_lg
Fibroblast-like_sm
OR
-log2(pval)

FIG. 8A

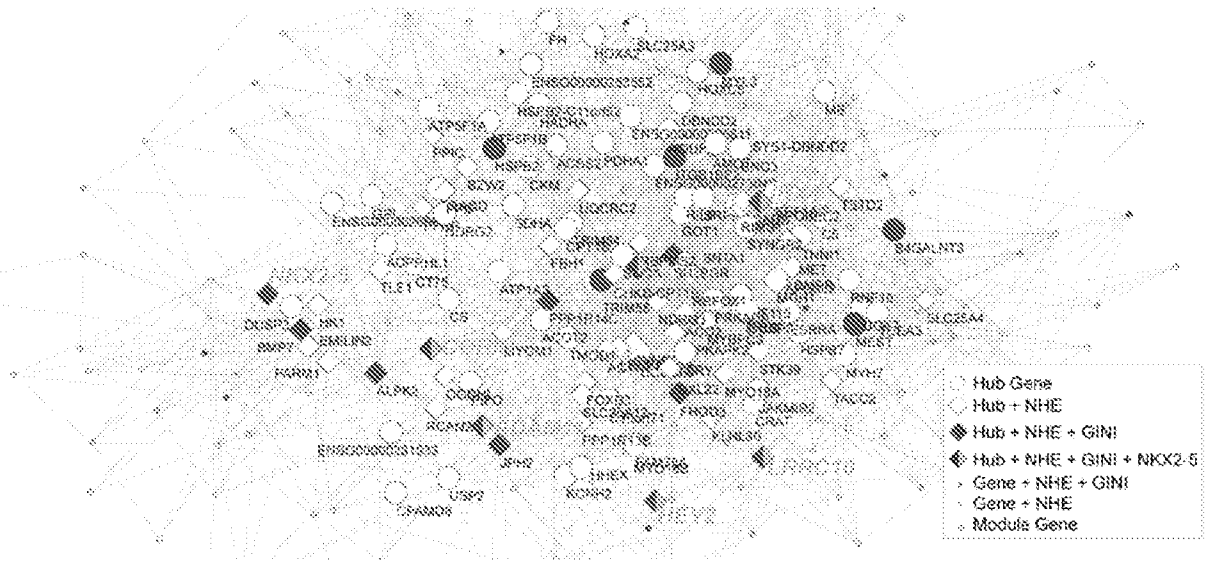

Hub Gene
Hub + NHE
Hub + NHE + GINI
Hub + NHE + GINI + NKX2-5
Gene + NHE + GINI
Gene + NHE
Module Gene

FIG. 8B

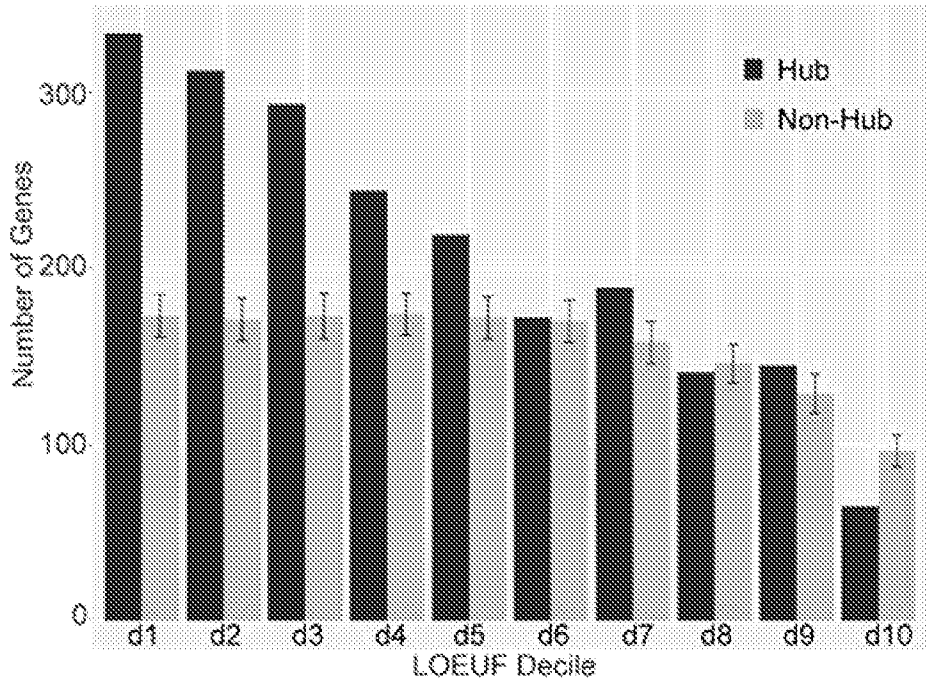

Hub
Non-Hub
Number of Genes
300
200
100
0
d1
d2
d3
d4
d5
d6
d7
d8
d9
d10
LOEUF Decile

FIG. 8C

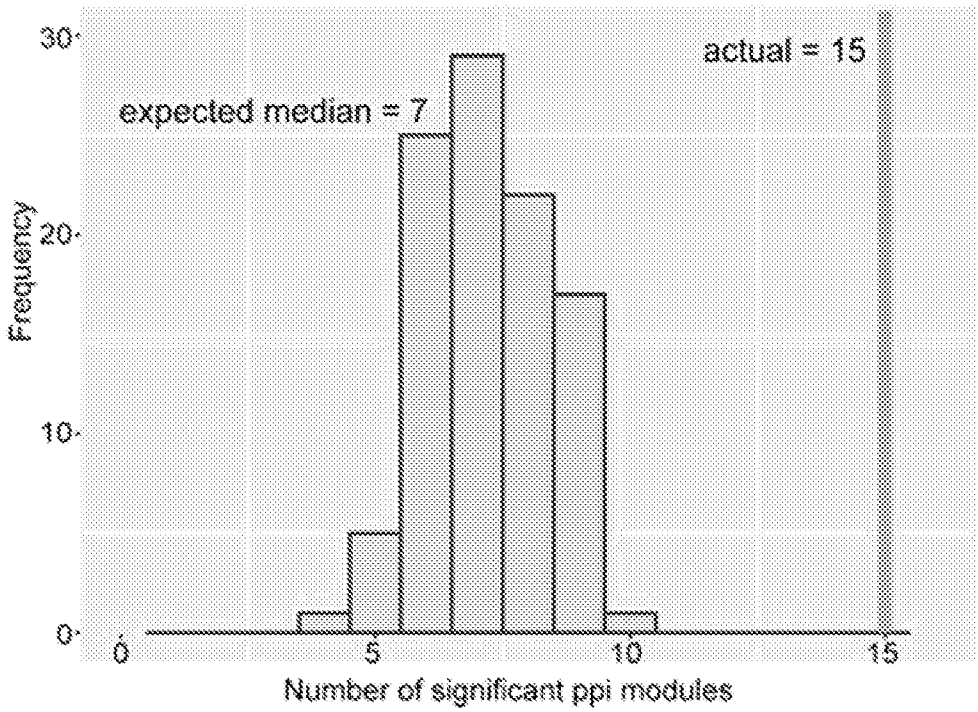

actual = 15
expected median = 7
Frequency
30
20
10
0
0
5
10
15
Number of significant ppi modules

FIG. 8D

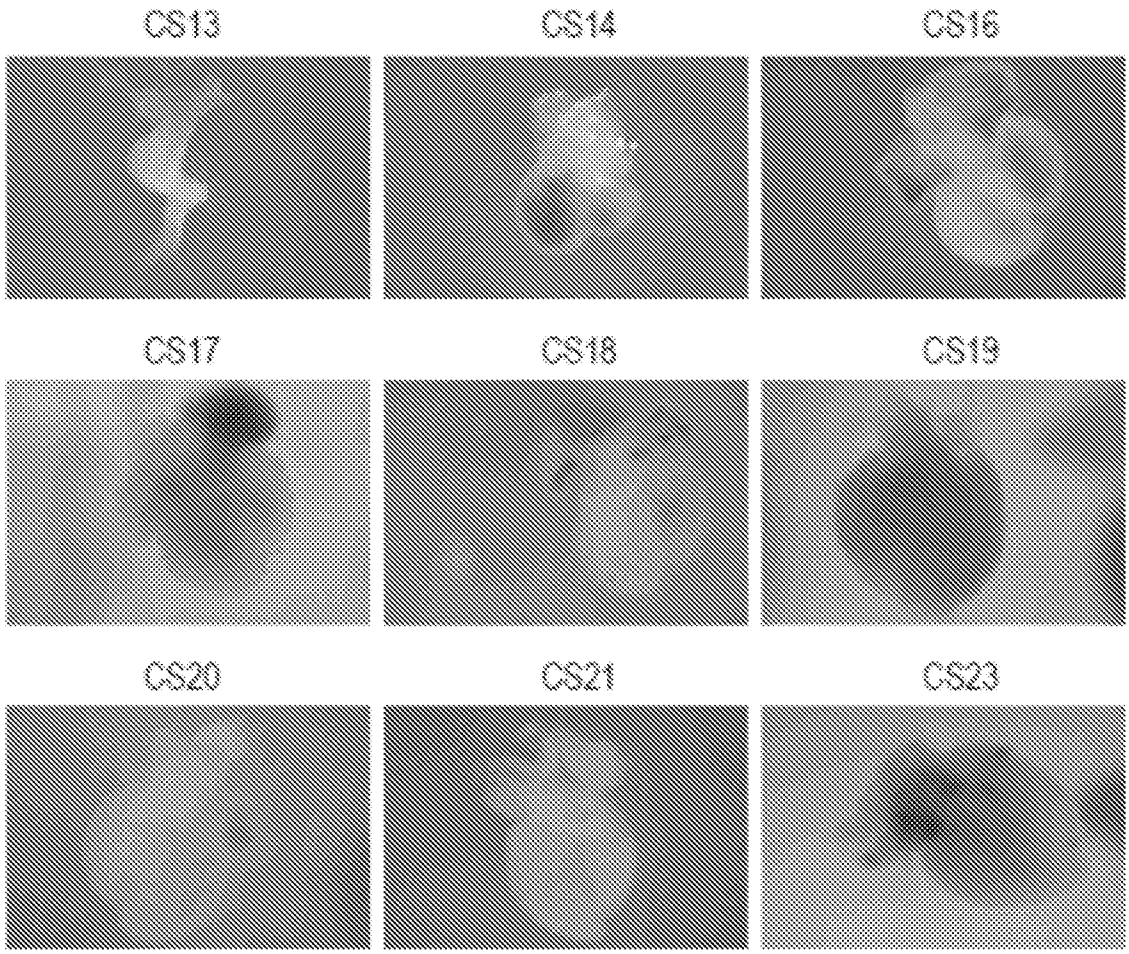


FIG. 9

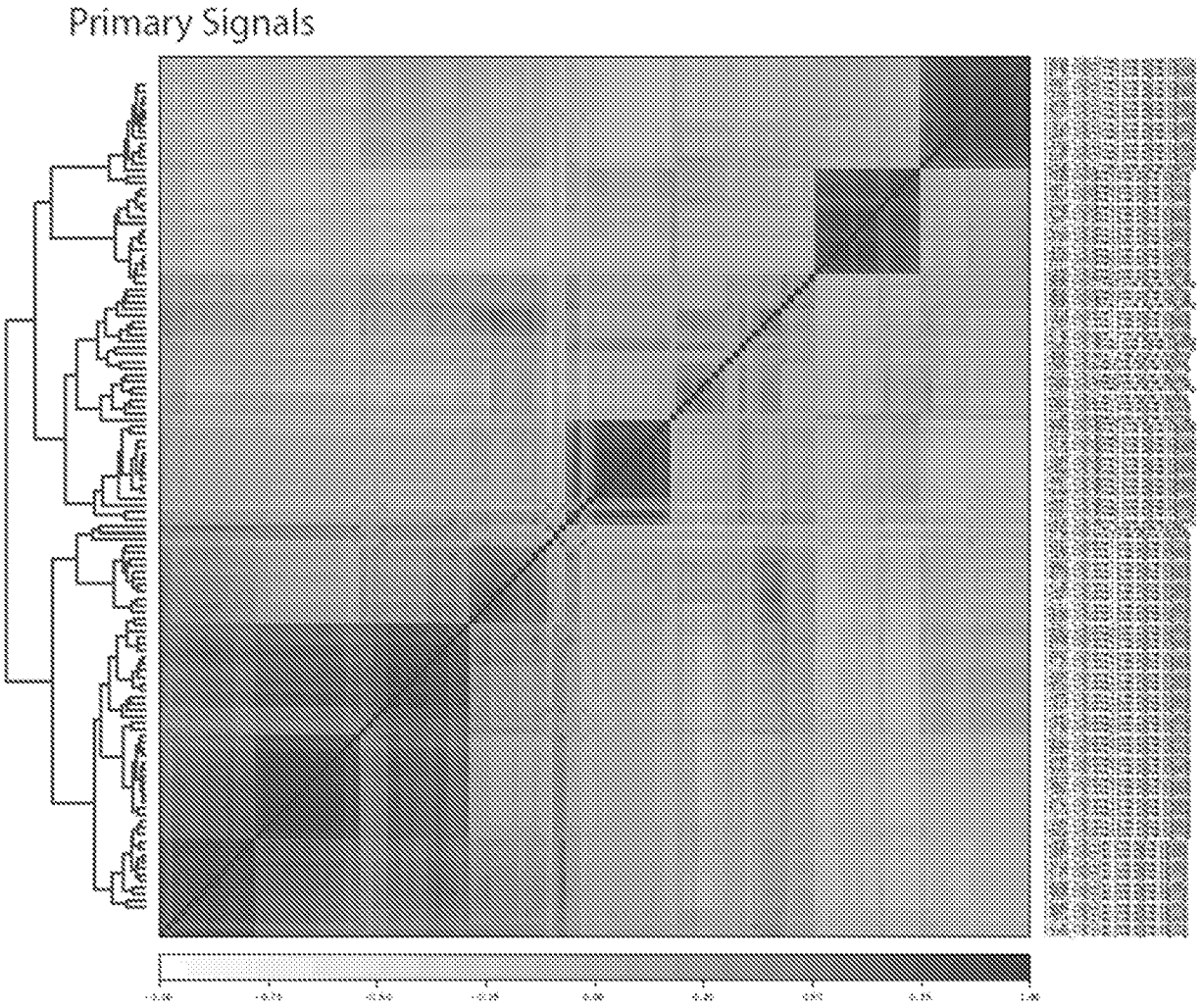


FIG. 10A

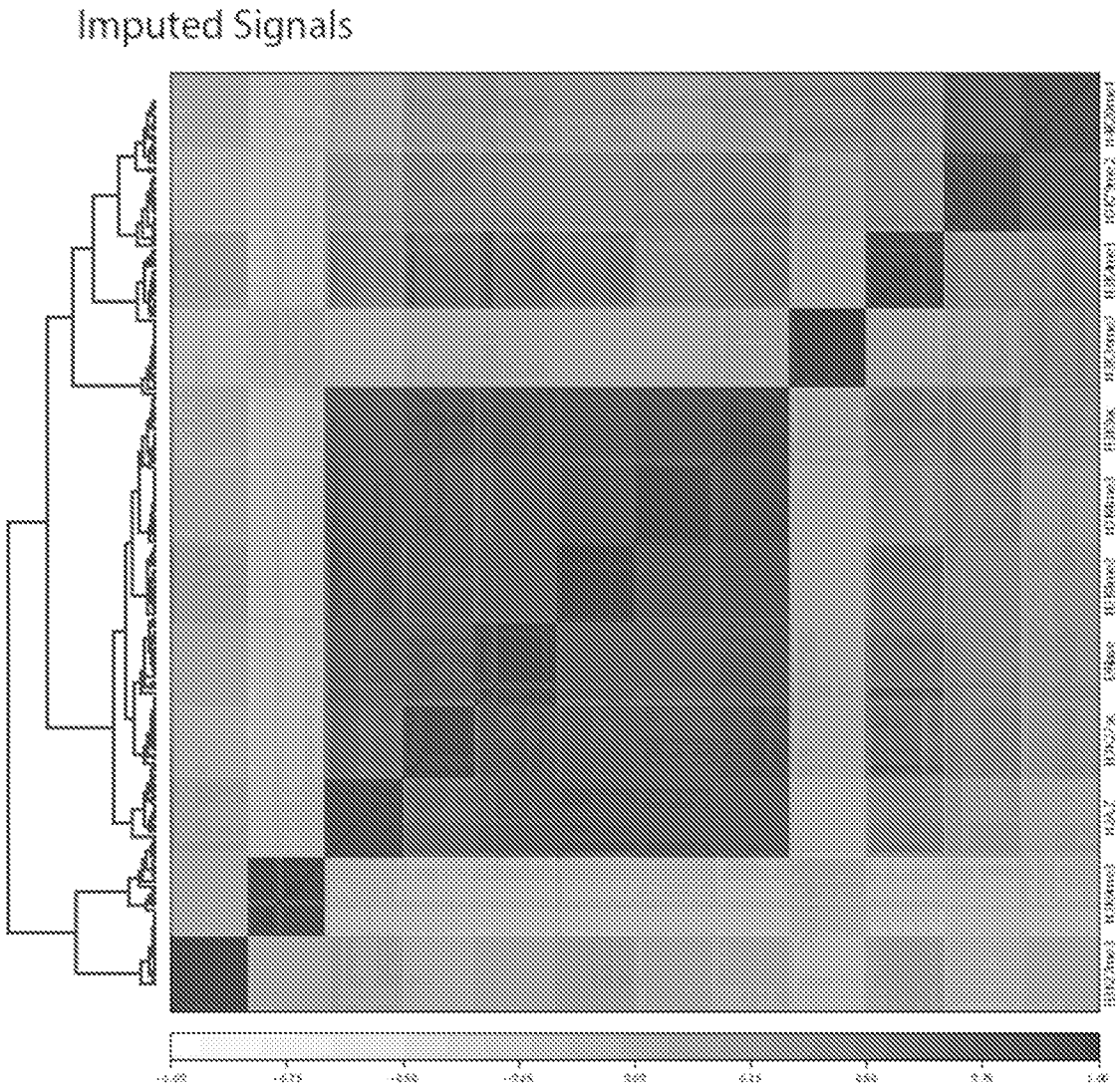


FIG. 10B

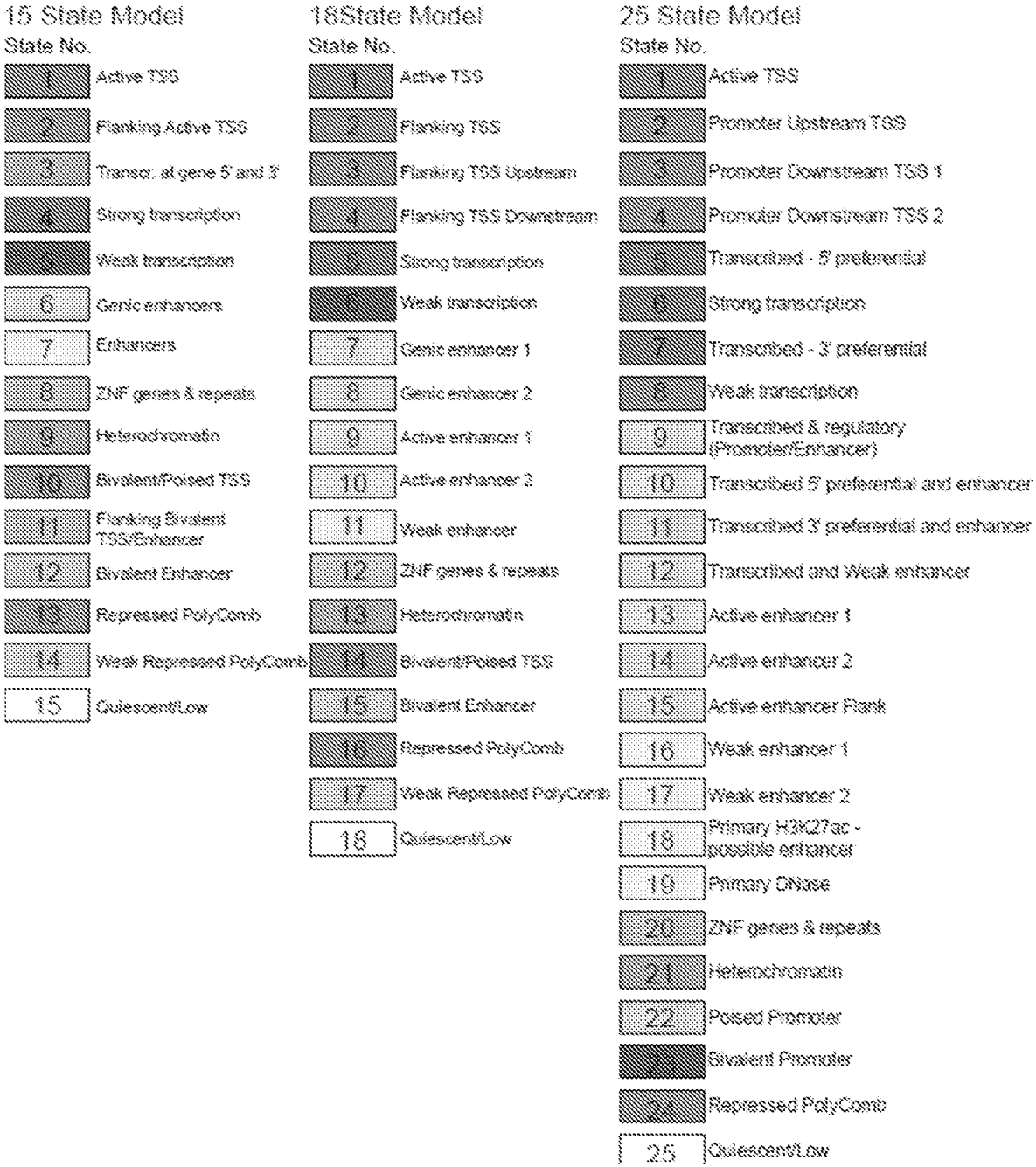

15 State Model
State No.
1 Active TSS
2 Flanking Active TSS
3 Transcr. at gene 5' and 3'
4 Strong transcription
5 Weak transcription
6 Genic enhancers
7 Enhancers
8 ZNF genes & repeats
9 Heterochromatin
10 Bivalent/Poised TSS
11 Flanking Bivalent TSS/Enhancer
12 Bivalent Enhancer
13 Repressed PolyComb
14 Weak Repressed PolyComb
15 Quiescent/Low
18State Model
State No.
1 Active TSS
2 Flanking TSS
3 Flanking TSS Upstream
4 Flanking TSS Downstream
5 Strong transcription
6 Weak transcription
7 Genic enhancer 1
8 Genic enhancer 2
9 Active enhancer 1
10 Active enhancer 2
11 Weak enhancer
12 ZNF genes & repeats
13 Heterochromatin
14 Bivalent/Poised TSS
15 Bivalent Enhancer
16 Repressed PolyComb
17 Weak Repressed PolyComb
18 Quiescent/Low
25 State Model
State No.
1 Active TSS
2 Promoter Upstream TSS
3 Promoter Downstream TSS 1
4 Promoter Downstream TSS 2
5 Transcribed - 5' preferential
6 Strong transcription
7 Transcribed - 3' preferential
8 Weak transcription
9 Transcribed & regulatory (Promoter/Enhancer)
10 Transcribed 5' preferential and enhancer
11 Transcribed 3' preferential and enhancer
12 Transcribed and Weak enhancer
13 Active enhancer 1
14 Active enhancer 2
15 Active enhancer Flank
16 Weak enhancer 1
17 Weak enhancer 2
18 Primary H3K27ac - possible enhancer
19 Primary DNase
20 ZNF genes & repeats
21 Heterochromatin
22 Poised Promoter
23 Bivalent Promoter
24 Repressed PolyComb
25 Quiescent/Low

FIG. 11

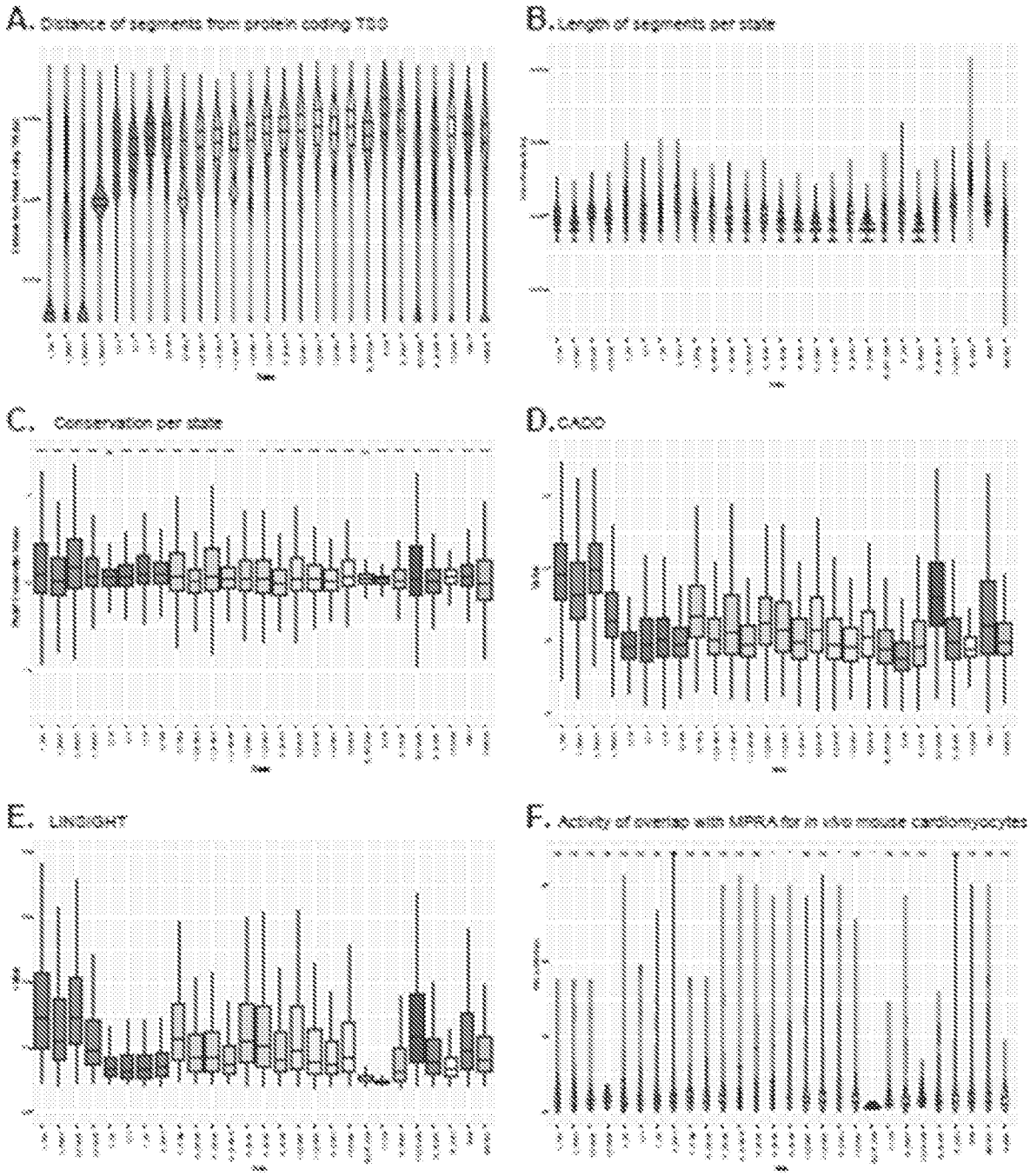


FIG. 12

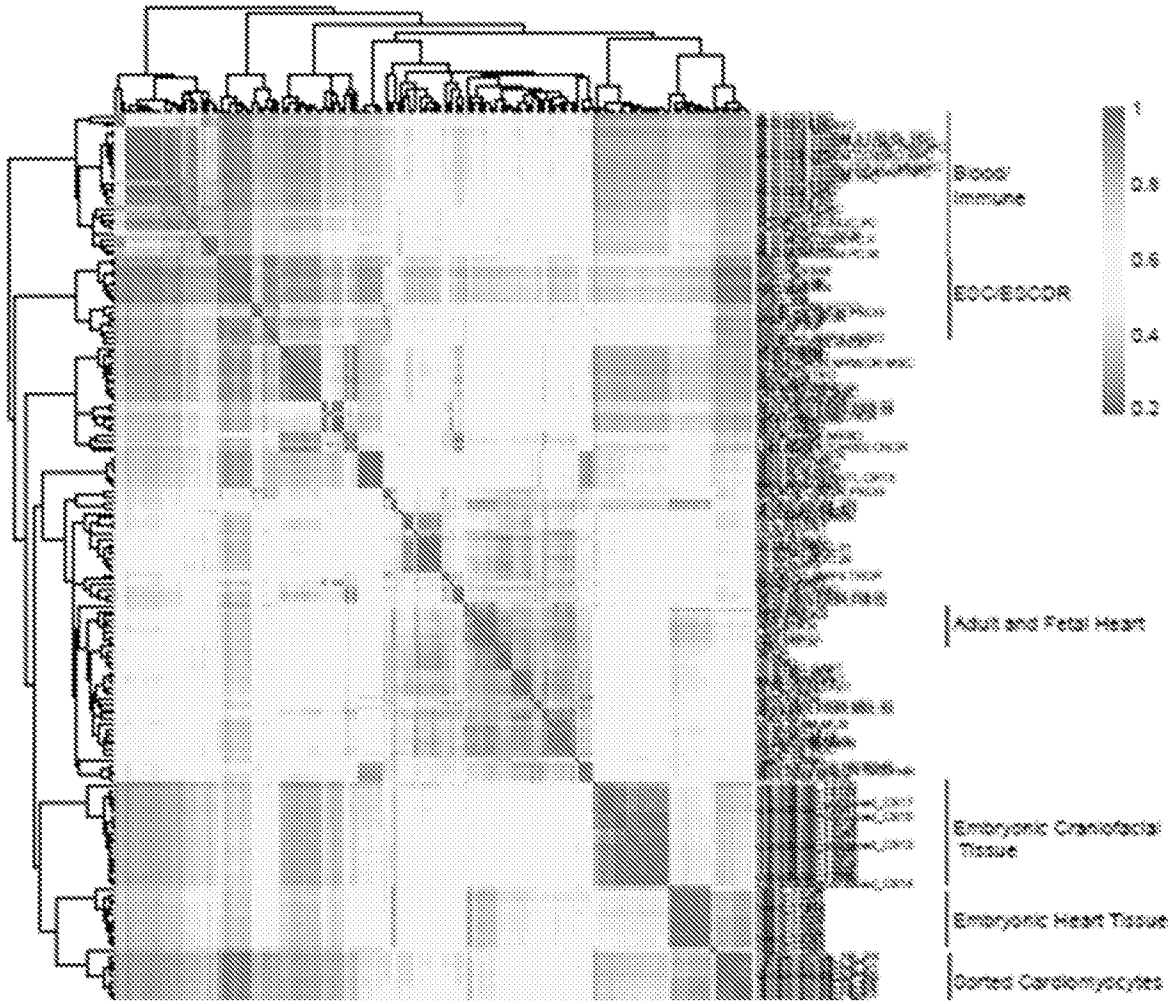

Blood/
Immune
ESC/iPSCDR
Adult and Fetal Heart
Embryonic Cardiotactic
Tissue
Embryonic Heart Tissue
Sorted Cardiomyocytes
1
0.8
0.6
0.4
0.2

FIG. 13A

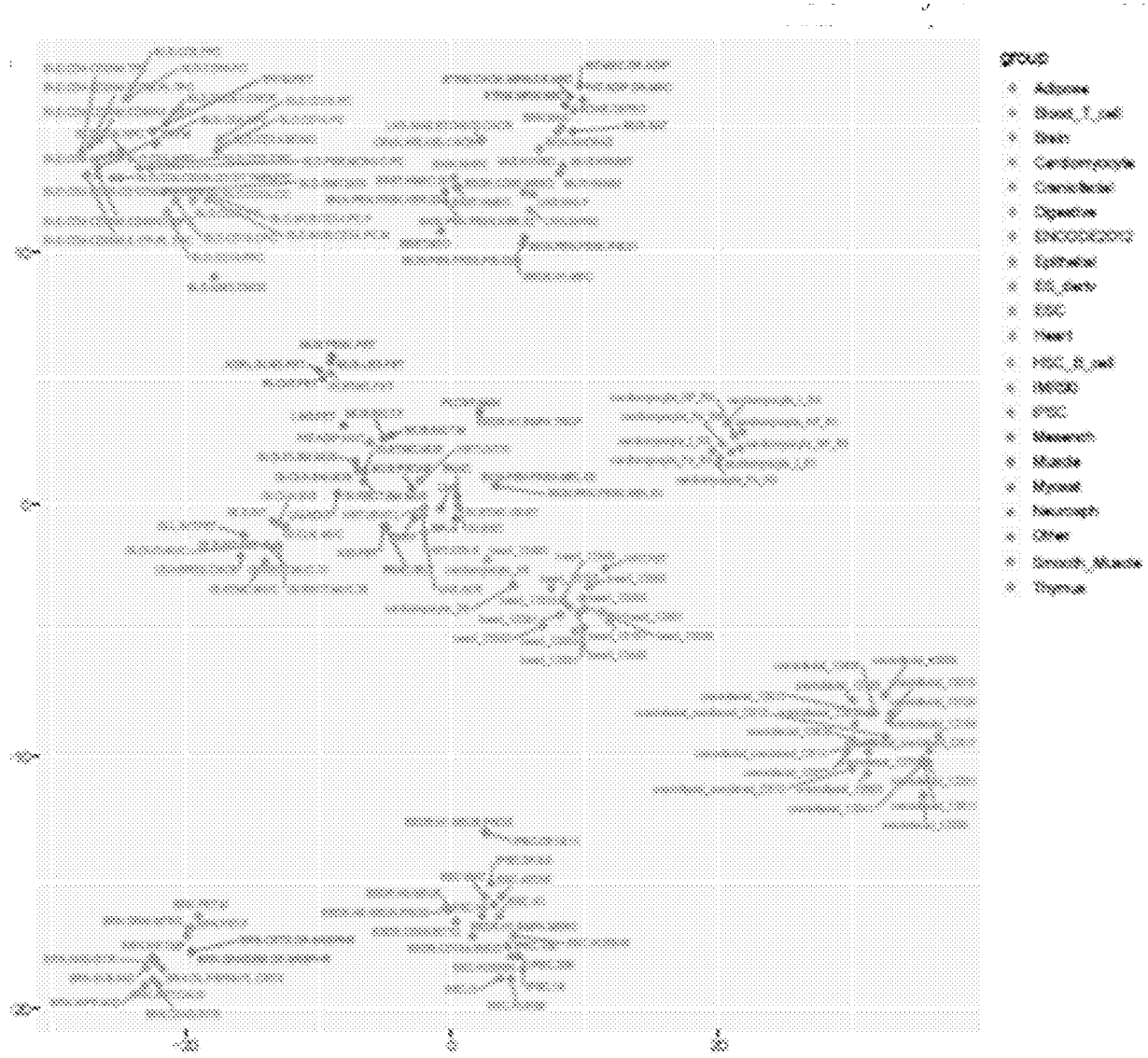


FIG. 13B

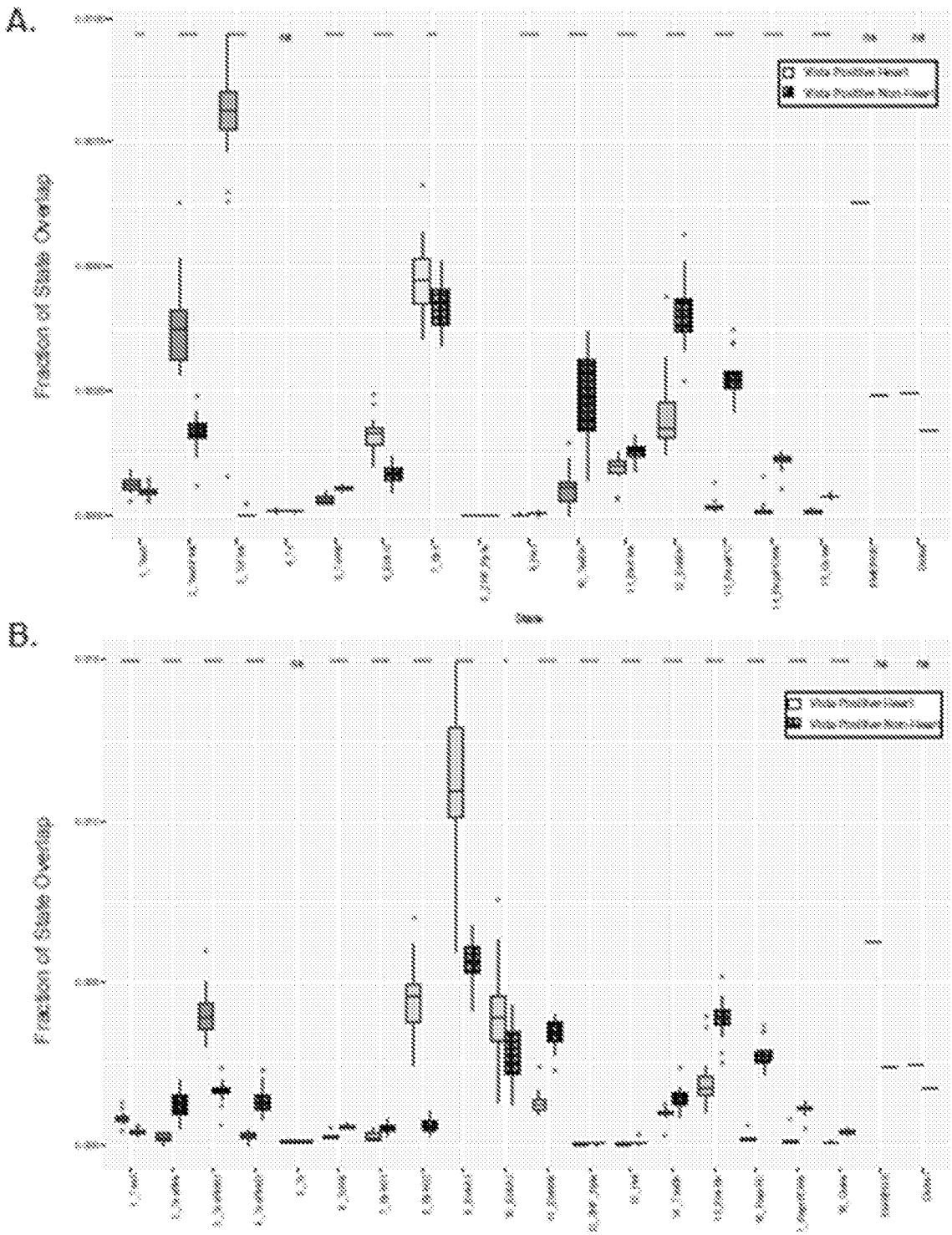


FIG. 14

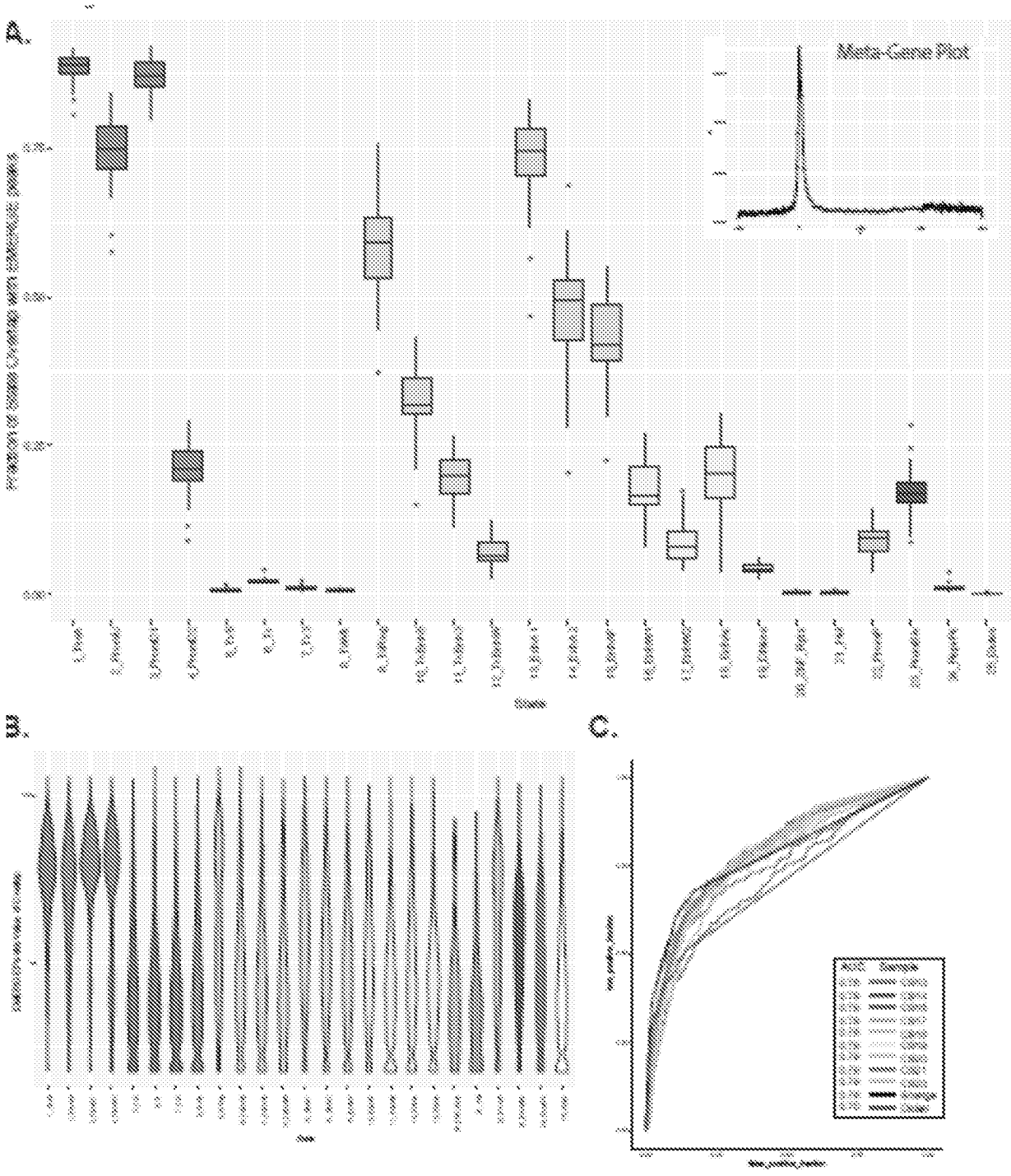


FIG. 15

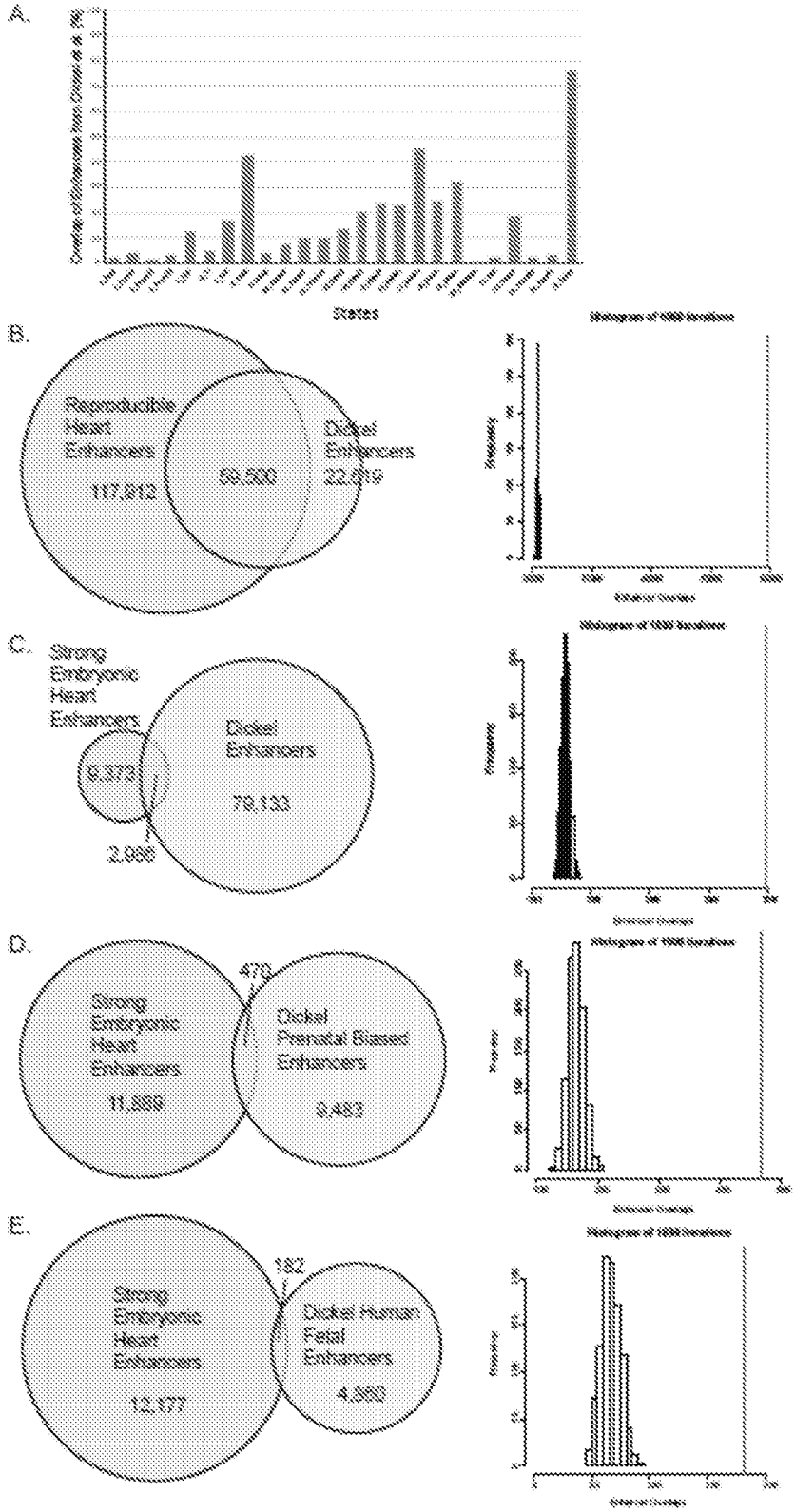

A.
States
B.
Reproducible Heart Enhancers
117,812
59,500
Dickel Enhancers
22,919
C.
Strong Embryonic Heart Enhancers
9,372
2,986
Dickel Enhancers
79,133
D.
470
Strong Embryonic Heart Enhancers
11,889
Dickel Prenatal Biased Enhancers
9,483
E.
182
Strong Embryonic Heart Enhancers
12,177
Dickel Human Fetal Enhancers
4,880

FIG. 16

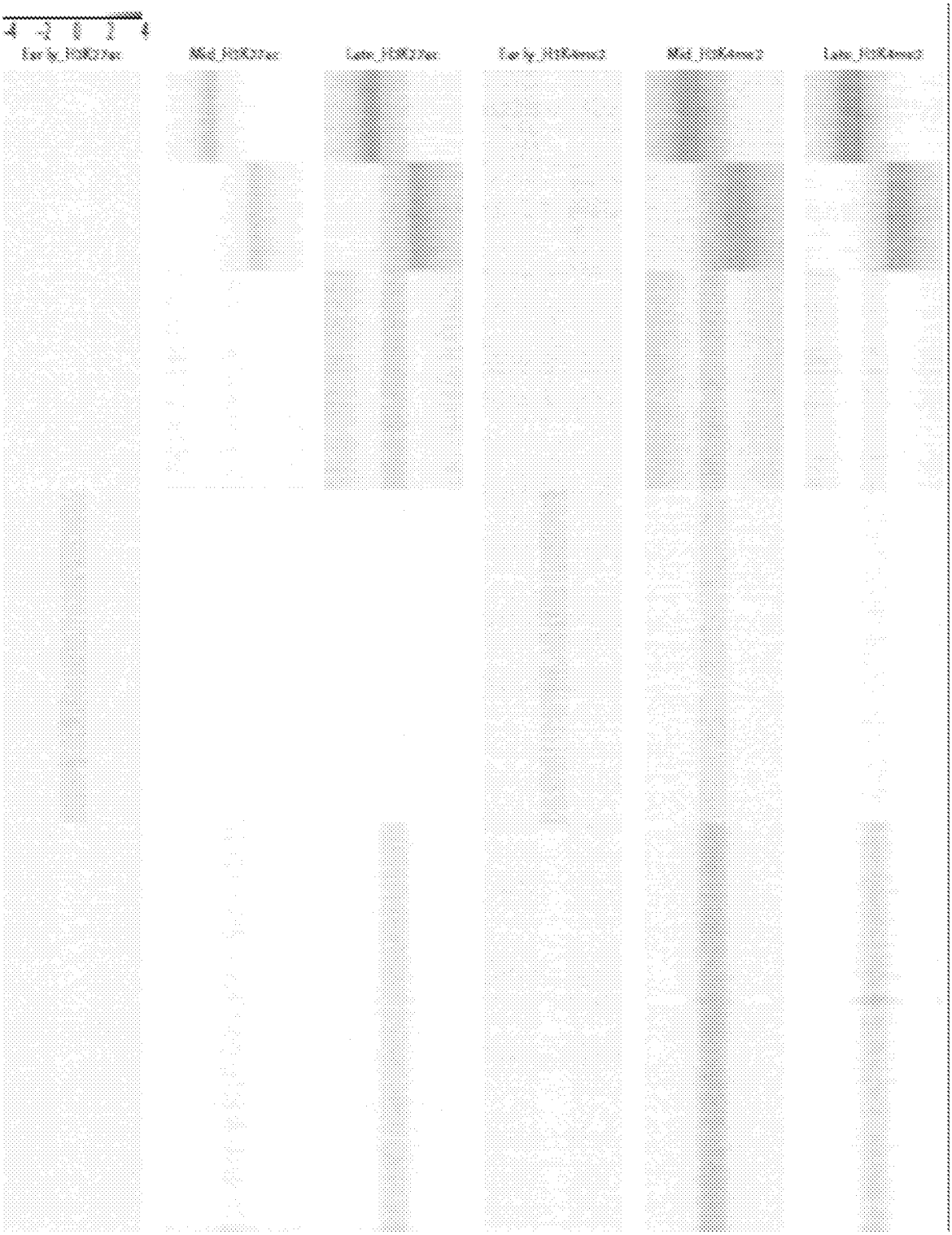


FIG. 17

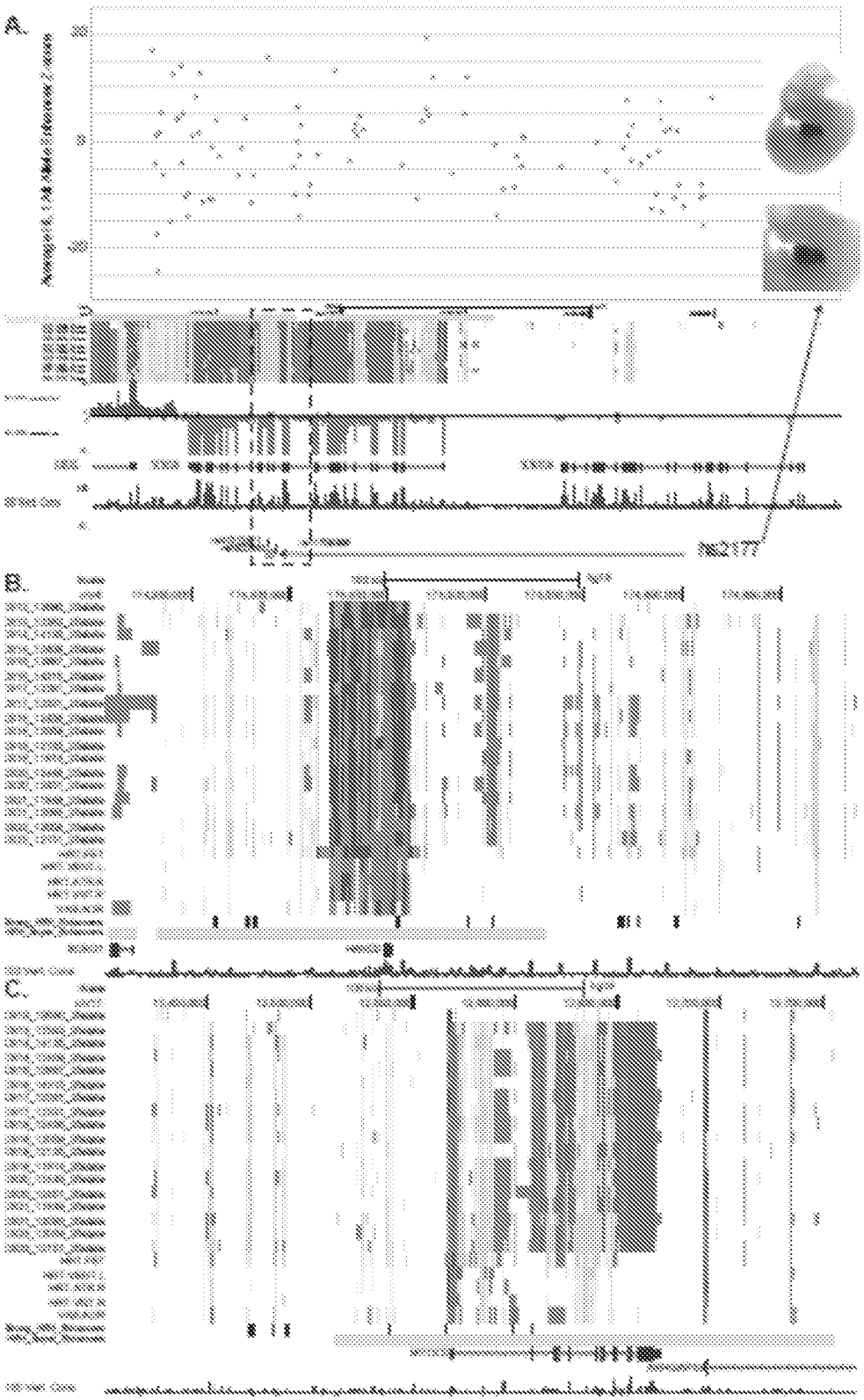


FIG. 18

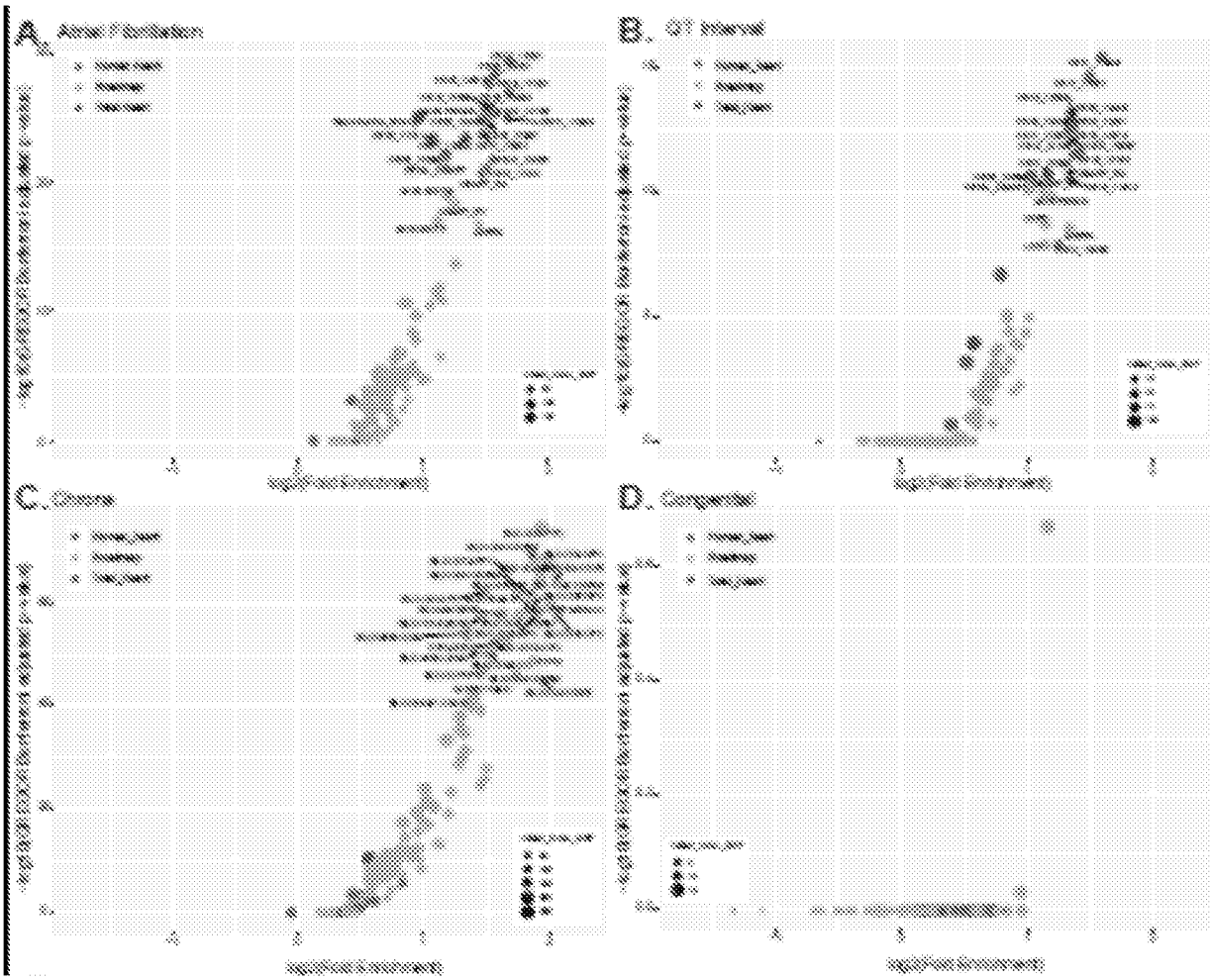


FIG. 19

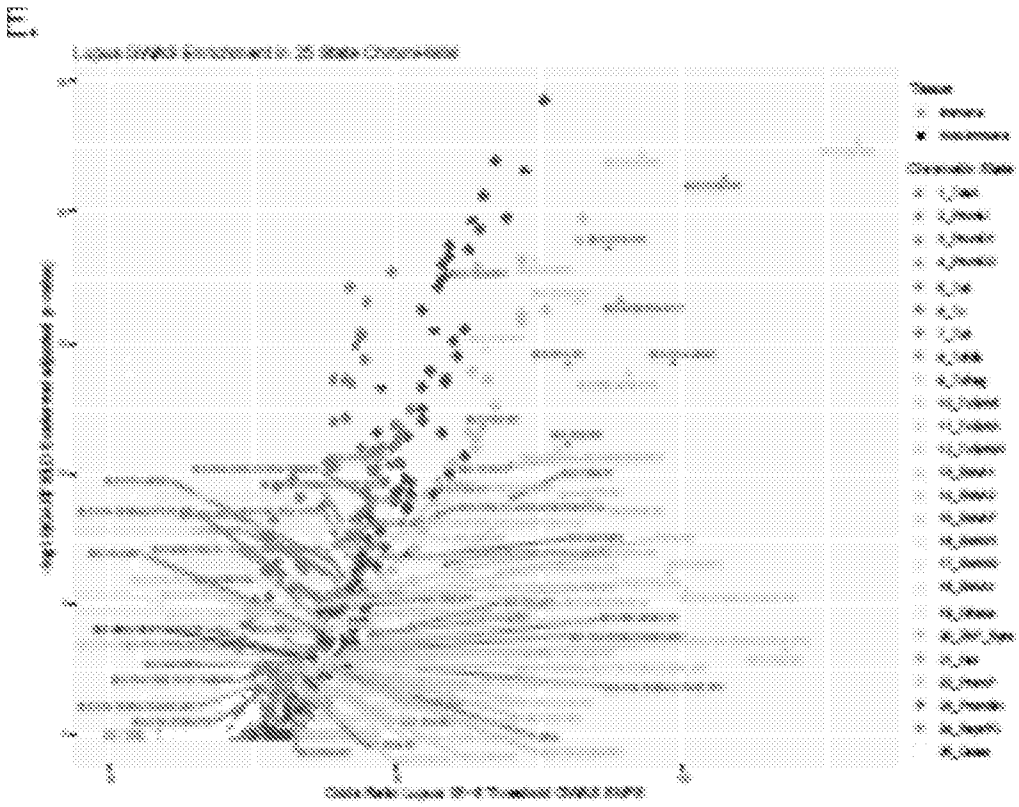


FIG. 19 (cont.)

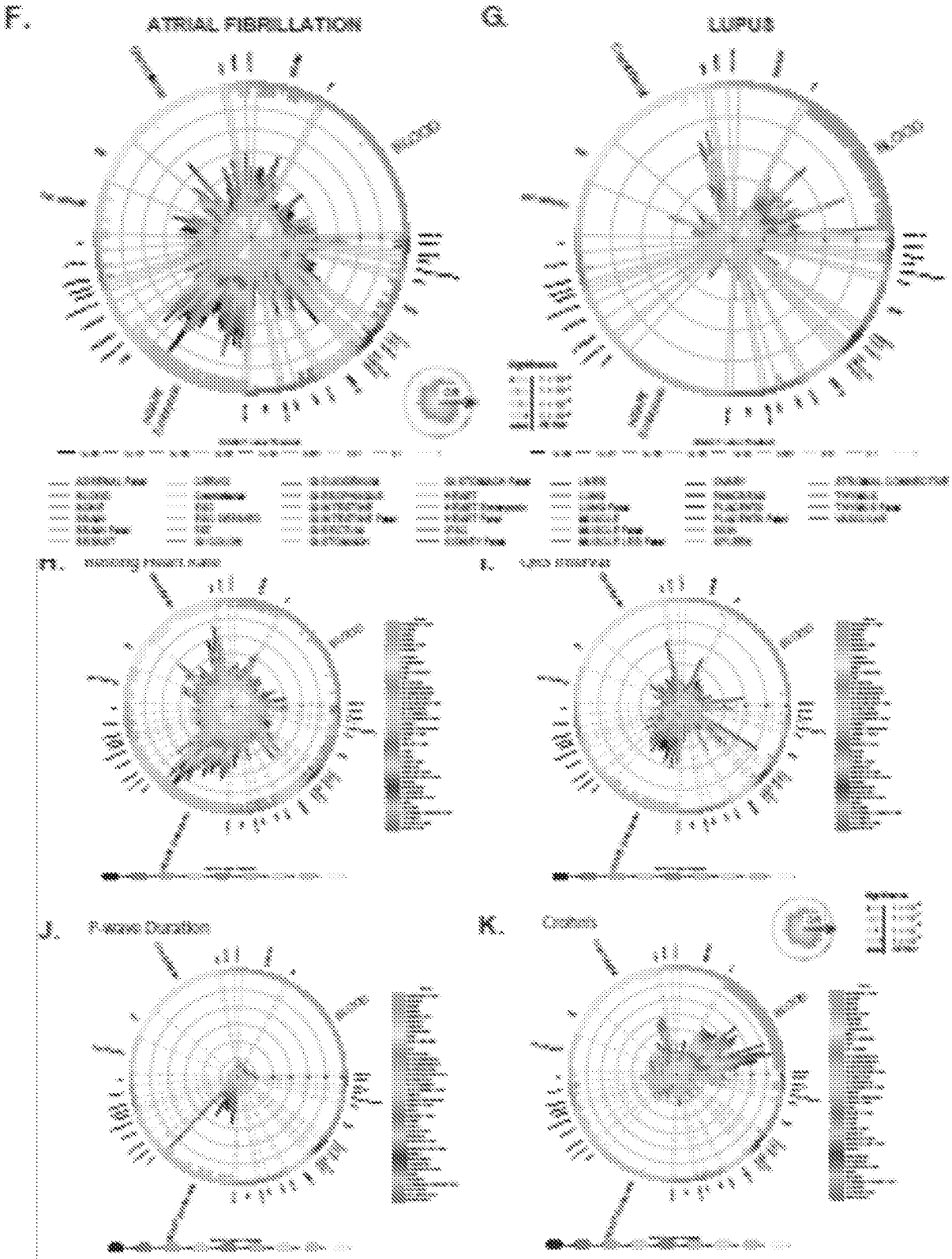


FIG. 19 (cont.)

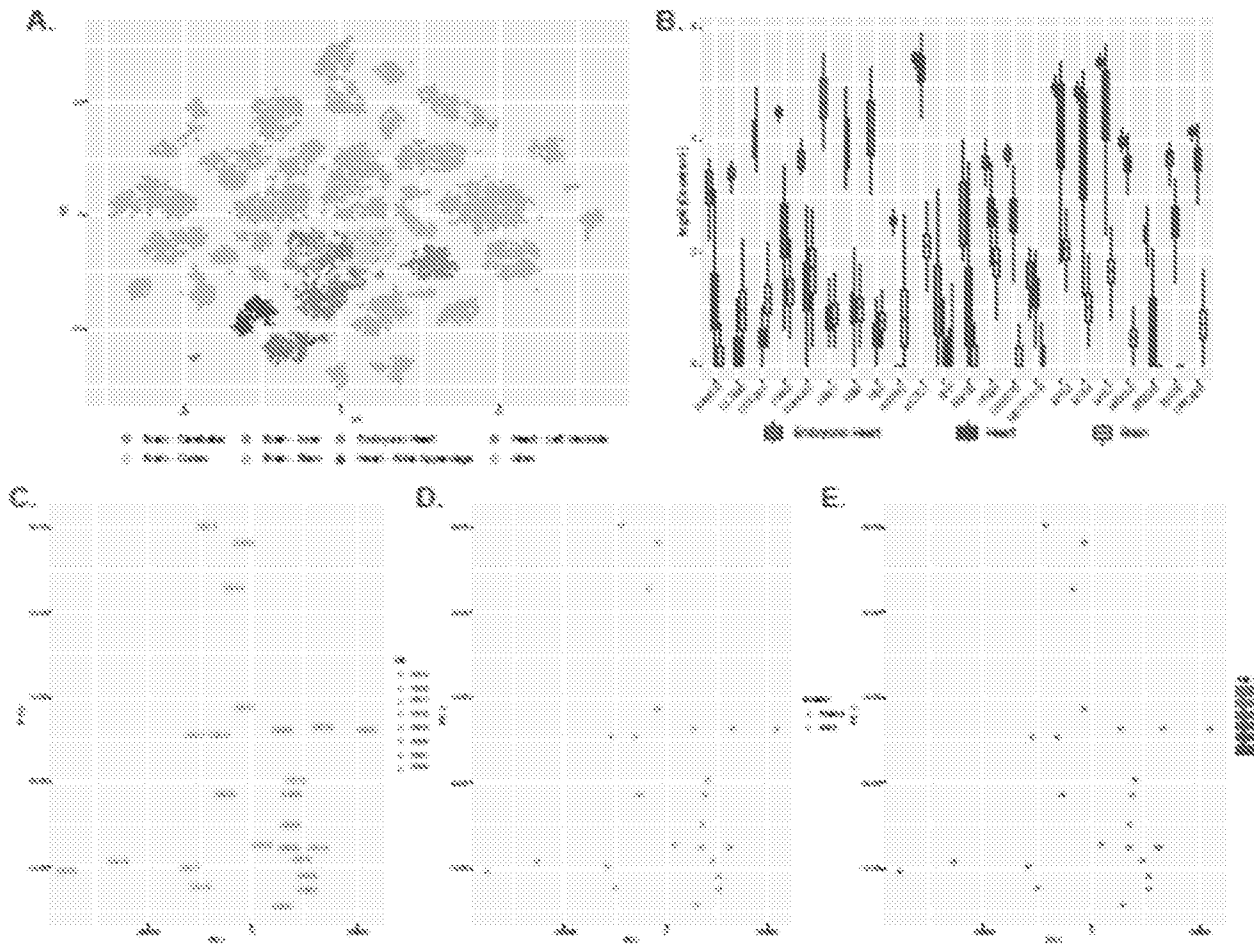


FIG. 20

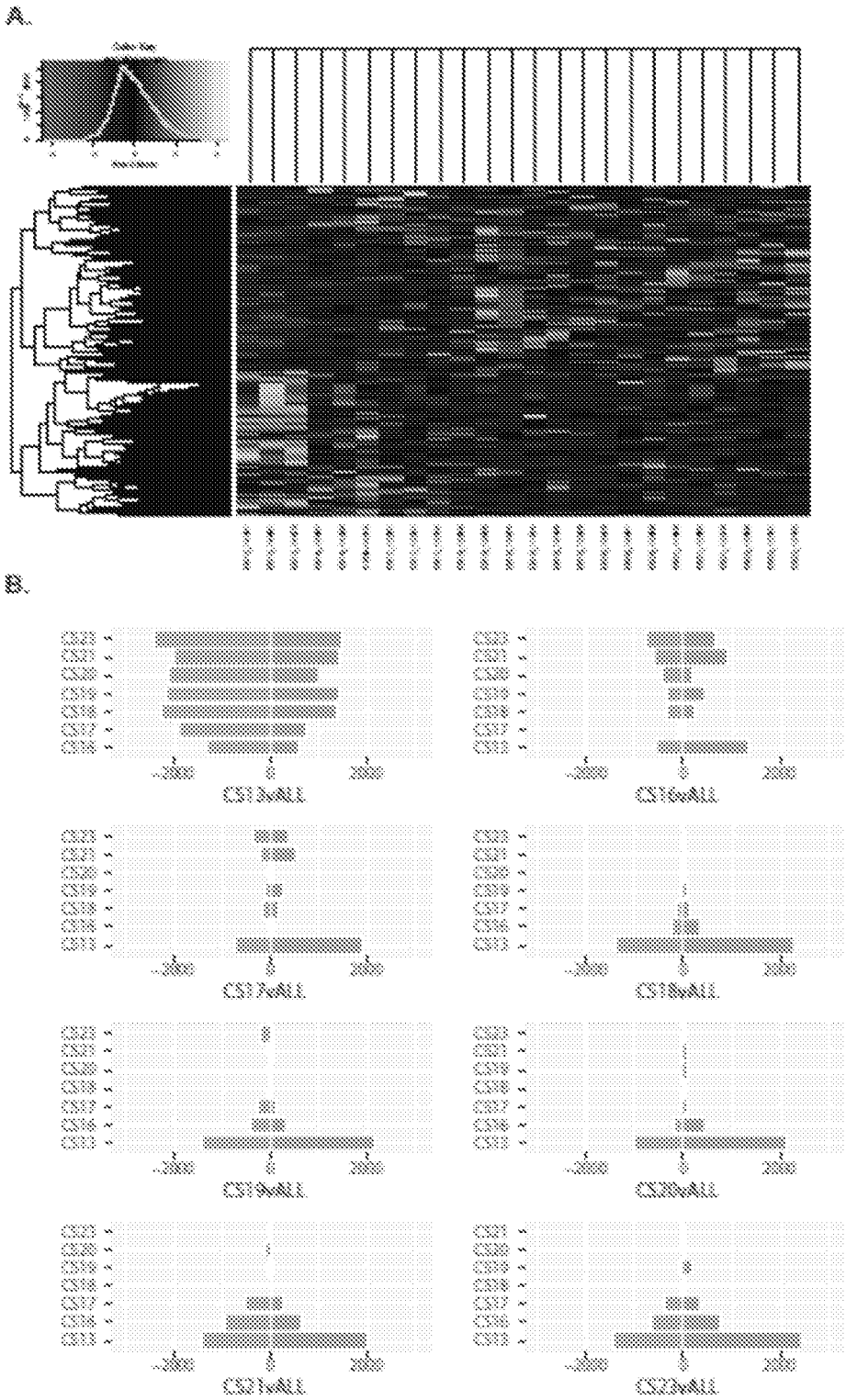

A.
B.
CS15vALL
CS16vALL
CS17vALL
CS18vALL
CS19vALL
CS20vALL
CS21vALL
CS22vALL

FIG. 21

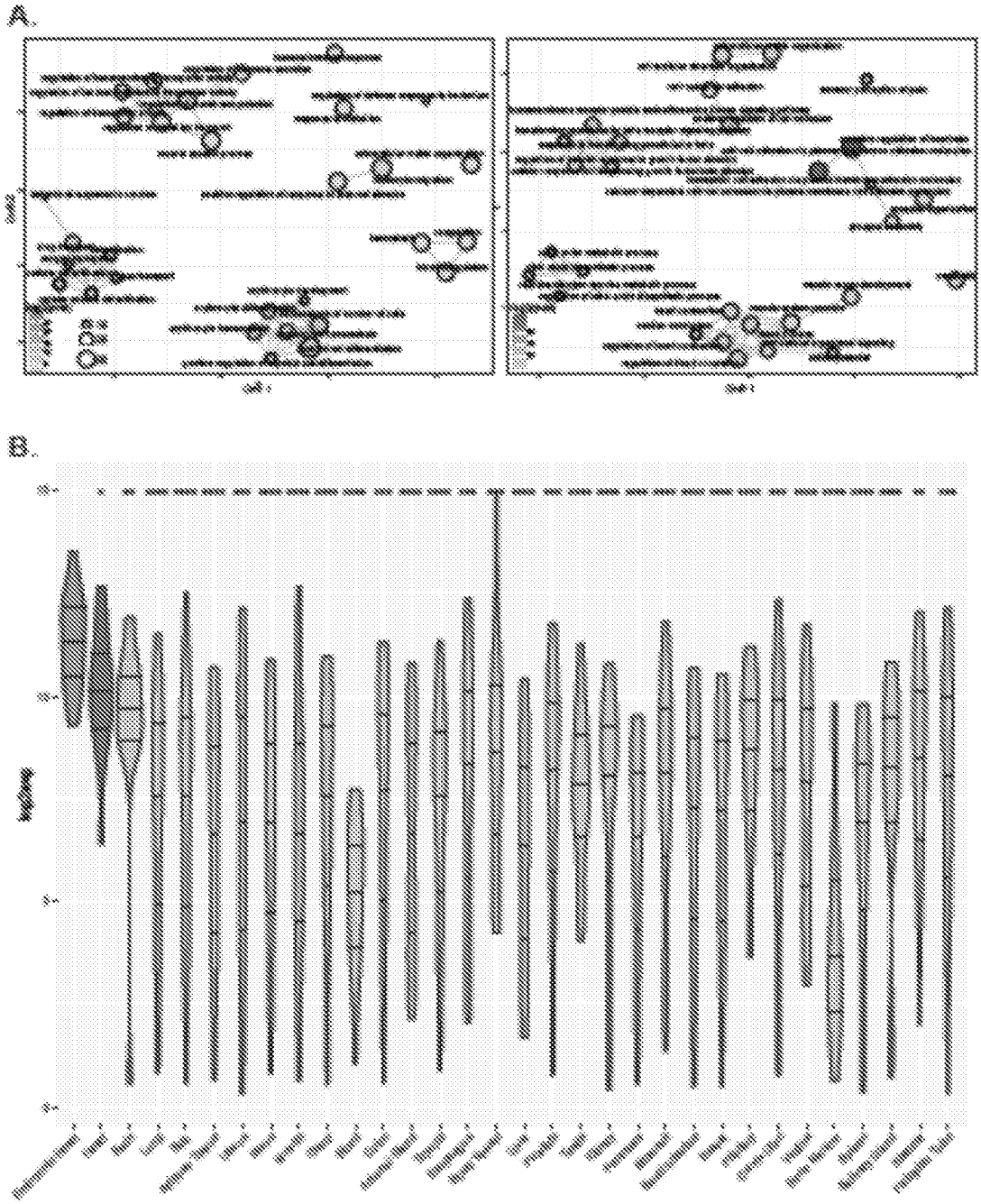


FIG. 22

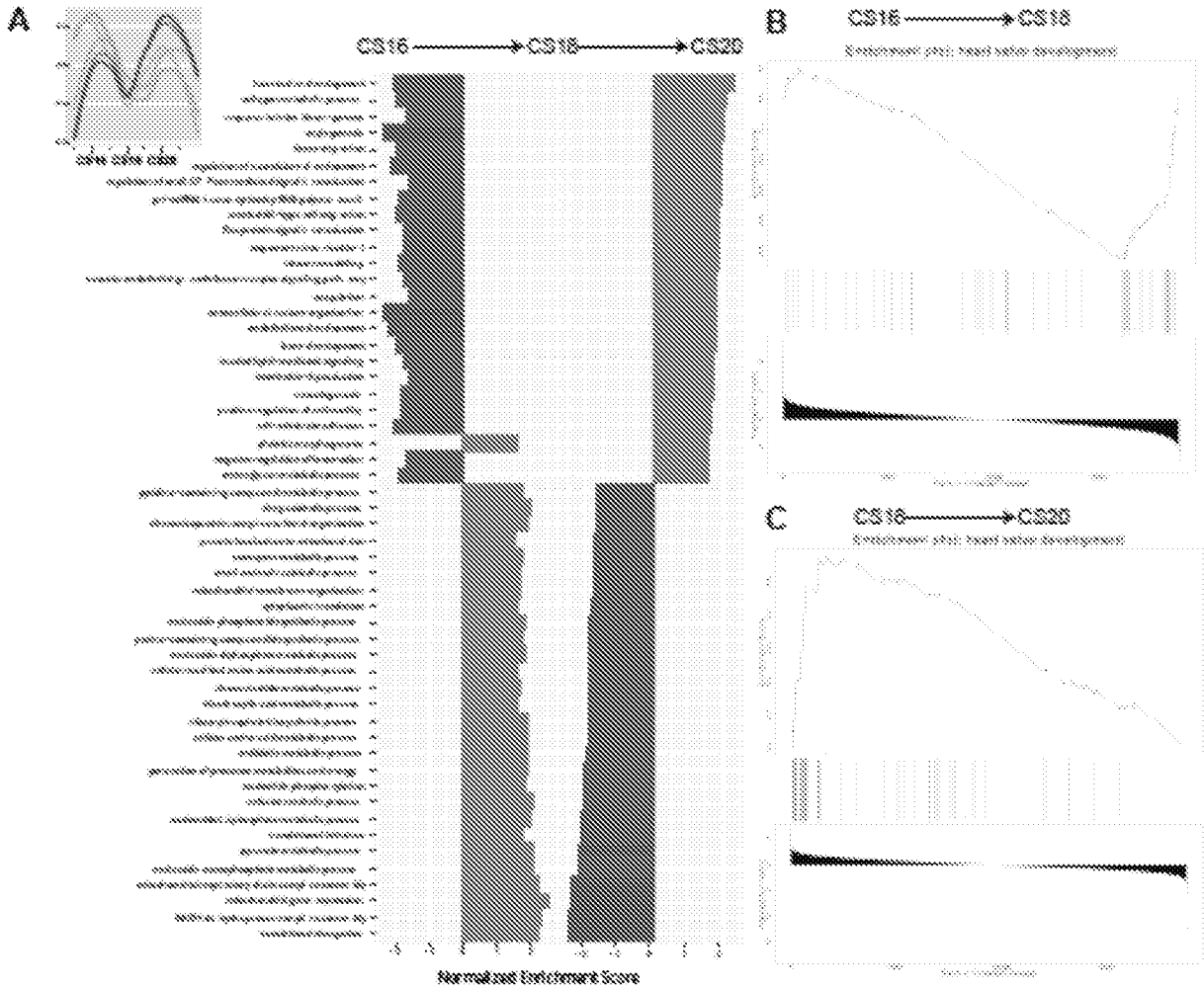


FIG. 23

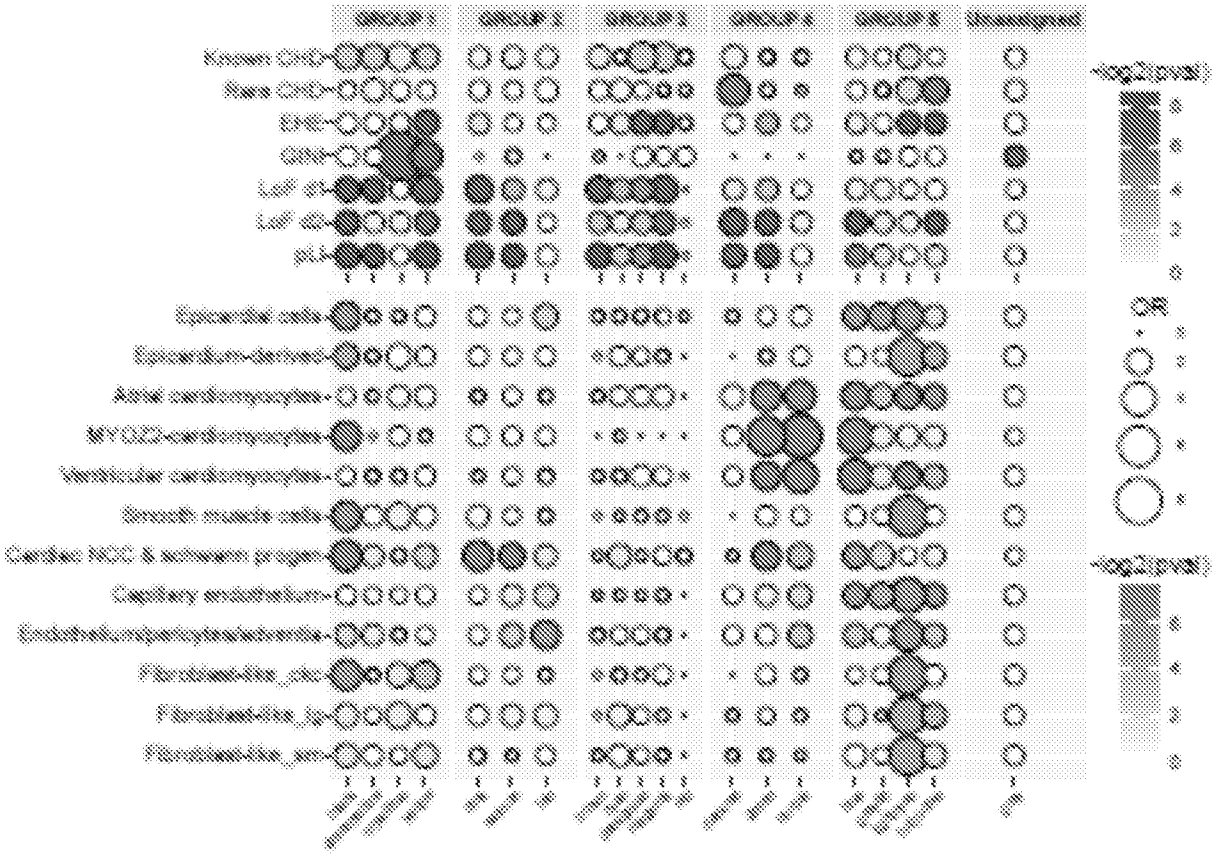


FIG. 24

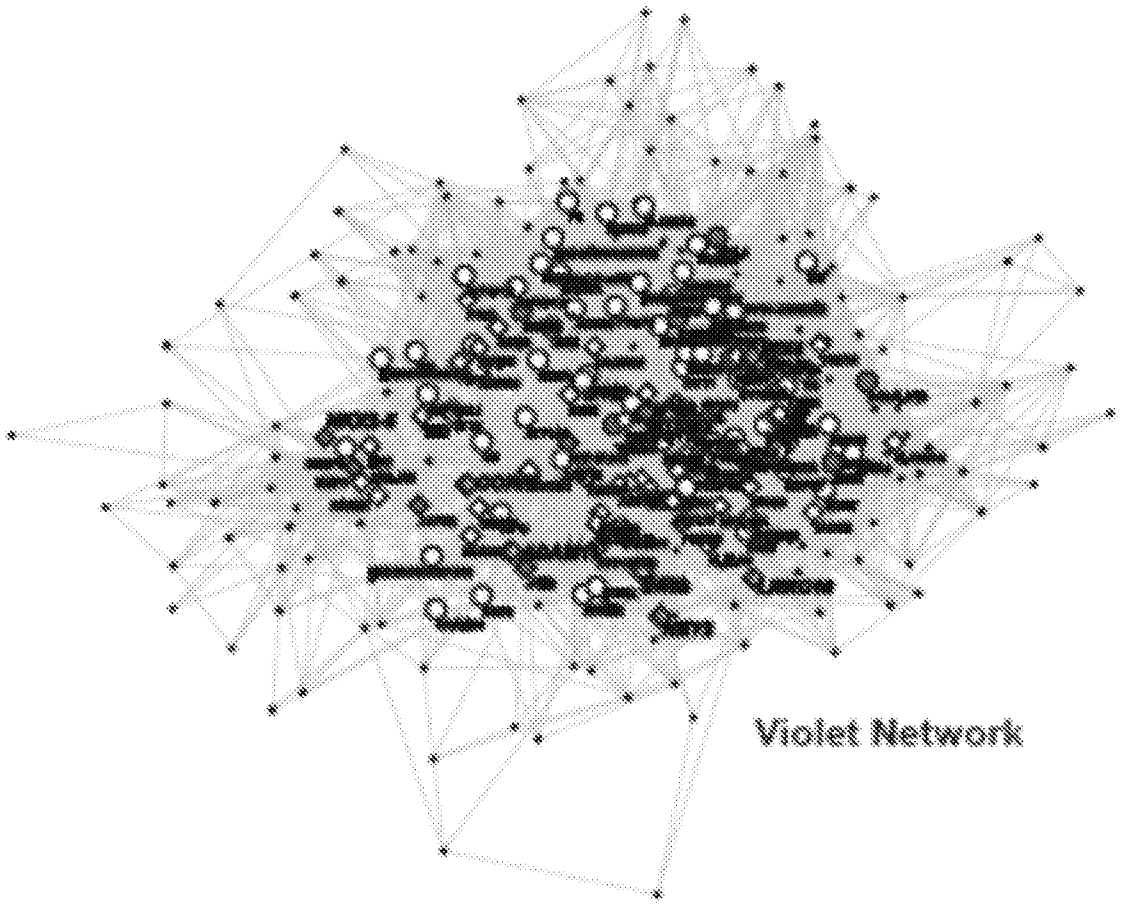


FIG. 25

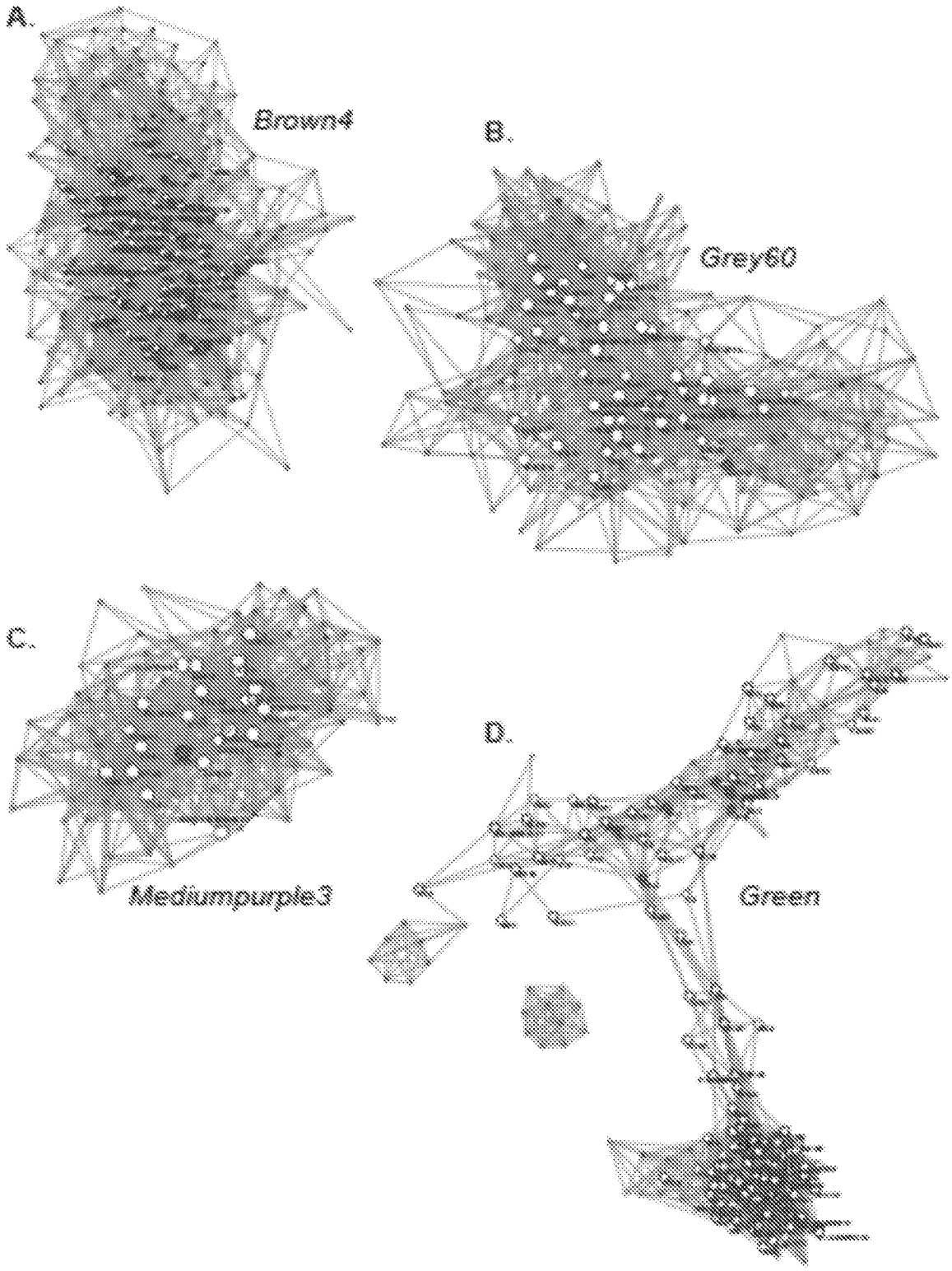


FIG. 26

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2021/041853

A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) - C12Q 1/68; C12Q 1/6879; G06F 19/18; G06F 19/20; G06F 19/24; G16B 25/00 (2021.01)
CPC - C12Q 1/6879; C12Q 1/6886; G16B 25/00; G16B 25/10; G16B 40/00; G16B 40/20; G16Z 99/00 (2021.08)

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
see Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
see Search History document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
see Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2020/0176078 A1 (VERACYTE INC.) 04 June 2020 (04.06.2020) entire document	1
A	WO 2018/204764 A1 (CAMP4 THERAPEUTICS CORPORATION) 08 November 2018 (08.11.2018) entire document	1
A	US 2018/0305689 A1 (MINA THERAPEUTICS LIMITED) 25 October 2018 (25.10.2018) entire document	1
A	US 2013/0231258 A1 (VERACYTE INC. et al) 05 September 2013 (05.09.2013) entire document	1
A	US 2017/0166965 A1 (WILLIAM BEAUMONT HOSPITAL et al) 15 June 2017 (15.06.2017) entire document	1

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
23 November 2021	DEC 27 2021
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, VA 22313-1450 Facsimile No. 571-273-8300	Authorized officer Harry Kim Telephone No. PCT Helpdesk: 571-272-4300

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2021/041853

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☒ Claims Nos.: 12, 13
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

See extra sheet(s).

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
1

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

Continued from Box No. III Observations where unity of invention is lacking

This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1. In order for all inventions to be examined, the appropriate additional examination fees need to be paid.

Group I+: claims 1-11 and 14-20 are drawn to gene panels and methods of use thereof for the assessment of risk of congenital heart defect or disease in a subject.

The first invention of Group I+ is restricted to a method of identifying a subject as at risk of having a congenital heart defect, the method comprising: a. assessing a plurality of genes in a sample obtained from the subject; and b. identifying the subject as at risk of having a congenital heart defect based upon the assessment of the plurality of genes, wherein the plurality of genes are selected to be MED13 and NCKAP1, and panels of genes comprising the same. It is believed that claim 1 reads on this first named invention and thus this claim will be searched without fee to the extent that it reads on MED13 and NCKAP1.

Applicant is invited to elect additional gene combinations to be searched in a specific combination by paying additional fee for each set of election. An exemplary election would be a method of identifying a subject as at risk of having a congenital heart defect, the method comprising: a. assessing a plurality of genes in a sample obtained from the subject; and b. identifying the subject as at risk of having a congenital heart defect based upon the assessment of the plurality of genes, wherein the plurality of genes are selected to be COL5A1 and TNPO1, and panels of genes comprising the same. Additional gene combinations will be searched upon the payment of additional fees. Applicants must specify the claims that read on any additional elected inventions. Applicants must further indicate, if applicable, the claims which read on the first named invention if different than what was indicated above for this group. Failure to clearly identify how any paid additional invention fees are to be applied to the "+" group(s) will result in only the first claimed invention to be searched/examined.

The inventions listed in Groups I+ do not relate to a single general inventive concept under PCT Rule 13.1, because under PCT Rule 13.2 they lack the same or corresponding special technical features for the following reasons:

The Groups I+ formulas do not share a significant structural element responsible for identifying a subject as at risk of having a congenital heart defect, requiring the selection of alternative gene combinations where "method of identifying a subject as at risk of having a congenital heart defect, the method comprising: a. assessing a plurality of genes in a sample obtained from the subject, wherein the plurality of genes are selected from the genes listed in Table 1; and b. identifying the subject as at risk of having a congenital heart defect based upon the assessment of the plurality of genes."

Additionally, even if Groups I+ were considered to share the technical features of a method of identifying a subject as at risk of having a congenital heart defect, the method comprising: a. assessing a plurality of genes in a sample obtained from the subject; and b. identifying the subject as at risk of having a congenital heart defect based upon the assessment of the plurality of genes; and panel of genes for assessing risk of congenital heart defects. However, these shared technical features do not represent a contribution over the prior art.

Specifically, US 2017/0166965 A1 to William Beaumont Hospital et al. discloses a method of identifying a subject as at risk of having a congenital heart defect (predicting Congenital Heart Defect based on measurement of the frequency or percentage methylation of cytosine nucleotides in various identified loci in the DNA of individuals, Para. [0018]), the method comprising: a. assessing a plurality of genes in a sample obtained from the subject ([o]btaining a sample from a patient; B) Extracting DNA from blood specimens; C) Assaying to determine the percentage methylation of cytosine at loci throughout the genome; D) Comparing the cytosine methylation level of the patient to a well characterized population of normal and Congenital Heart Defect groups; and E) Calculating the individual risk of Congenital Heart Defect based on the cytosine methylation level at different sites throughout the genome, Para. [0018]); and b. identifying the subject as at risk of having a congenital heart defect based upon the assessment of the plurality of genes([o]btaining a sample from a patient; B) Extracting DNA from blood specimens; C) Assaying to determine the percentage methylation of cytosine at loci throughout the genome; D) Comparing the cytosine methylation level of the patient to a well characterized population of normal and Congenital Heart Defect groups; and E) Calculating the individual risk of Congenital Heart Defect based on the cytosine methylation level at different sites throughout the genome, Para. [0018]) and panel of genes for assessing risk of congenital heart defects (a panel of cytosine markers for distinguishing individual categories of common CHD from normal cases and also for distinguishing CHD as a group from normal cases without CHD, Para. [0015]).

The inventions listed in Groups I+ therefore lack unity under Rule 13 because they do not share a same or corresponding special technical features.