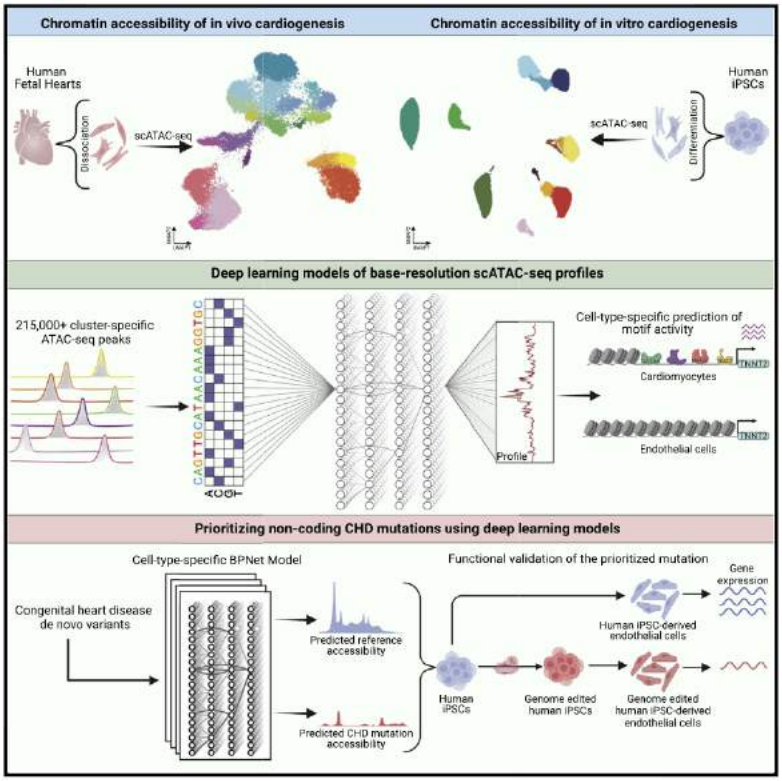
Cell

Integrative single-cell analysis of cardiogenesis identifies developmental trajectories and non-coding mutations in congenital heart disease

Graphical abstract



Highlights

- Single-cell chromatin data dissect distinct TF cardiogenesis regulatory programs
- Dynamic transcription factor activity defines major differentiation trajectories
- Molecular benchmarking with *in vivo* cells enables the optimization of *in vitro* protocols
- Neural networks prioritize non-coding *de novo* mutations in congenital heart disorders

Authors

Mohamed Ameen, Laksshman Sundaram, Mengcheng Shen, ..., Thomas Quertermous, William J. Greenleaf, Anshul Kundaje

Correspondence

ioannis1@stanford.edu (I.K.), kevwang@stanford.edu (K.C.W.), tomq1@stanford.edu (T.Q.), wjg@stanford.edu (W.J.G.), akundaje@stanford.edu (A.K.)

In brief

Cell-type-resolved regulatory atlas of the developing human heart reveals cellular differentiation trajectories in cardiogenesis and an involvement of non-coding genetic variants in congenital heart diseases.

Ameen et al., 2022, Cell *185*, 4937–4953
December 22, 2022
https://doi.org/10.1016/j.cell.2022.11.028

CellPress

Cell

CellPress

Article

Integrative single-cell analysis of cardiogenesis identifies developmental trajectories and non-coding mutations in congenital heart disease

Mohamed Ameen,[1,3,15] Laksshman Sundaram,[2,3,15] Mengcheng Shen,[4] Abhimanyu Banerjee,[3,5] Soumya Kundu,[2] Surag Nair,[2] Anna Shcherbina,[6] Mingxia Gu,[7] Kitchener D. Wilson,[4] Avyay Varadarajan,[8] Nirmal Vadgama,[9] Akshay Balsubramani,[10] Joseph C. Wu,[4] Jesse M. Engreitz,[10] Kyle Farh,[3] Ioannis Karakikes,[4,9,*] Kevin C. Wang,[1,11,12,*] Thomas Quertermous,[13,16,*] William J. Greenleaf,[10,14,*] and Anshul Kundaje[2,10,*]

[1]Department of Cancer Biology, Stanford University, Stanford, CA, USA
[2]Department of Computer Science, Stanford University, Stanford, CA, USA
[3]Illumina Artificial Intelligence Laboratory, Illumina Inc, Foster City, CA, USA
[4]Cardiovascular Institute, Stanford University, Stanford, CA, USA
[5]Department of Physics, Stanford University, Stanford, CA, USA
[6]Department of Biomedical Informatics, Stanford University, Stanford, CA, USA
[7]Center for Stem Cell and Organoid Medicine, CuSTOM, Division of Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA
[8]Department of Computer Science, California Institute of Technology, Pasadena, CA, USA
[9]Department of Cardiothoracic Surgery, Stanford University, Stanford, CA, USA
[10]Department of Genetics, Stanford University, Stanford, CA, USA
[11]Department of Dermatology, Stanford University School of Medicine, Stanford, CA, USA
[12]Veterans Affairs Palo Alto Healthcare System, Palo Alto, CA, USA
[13]Division of Cardiovascular Medicine, Stanford University School of Medicine, Stanford, CA, USA
[14]Department of Applied Physics, Stanford University, Stanford, CA, USA
[15]These authors contributed equally
[16]Lead contact
*Correspondence: ioannis1@stanford.edu (I.K.), kevwang@stanford.edu (K.C.W.), tomq1@stanford.edu (T.Q.), wjg@stanford.edu (W.J.G.), akundaje@stanford.edu (A.K.)
https://doi.org/10.1016/j.cell.2022.11.028

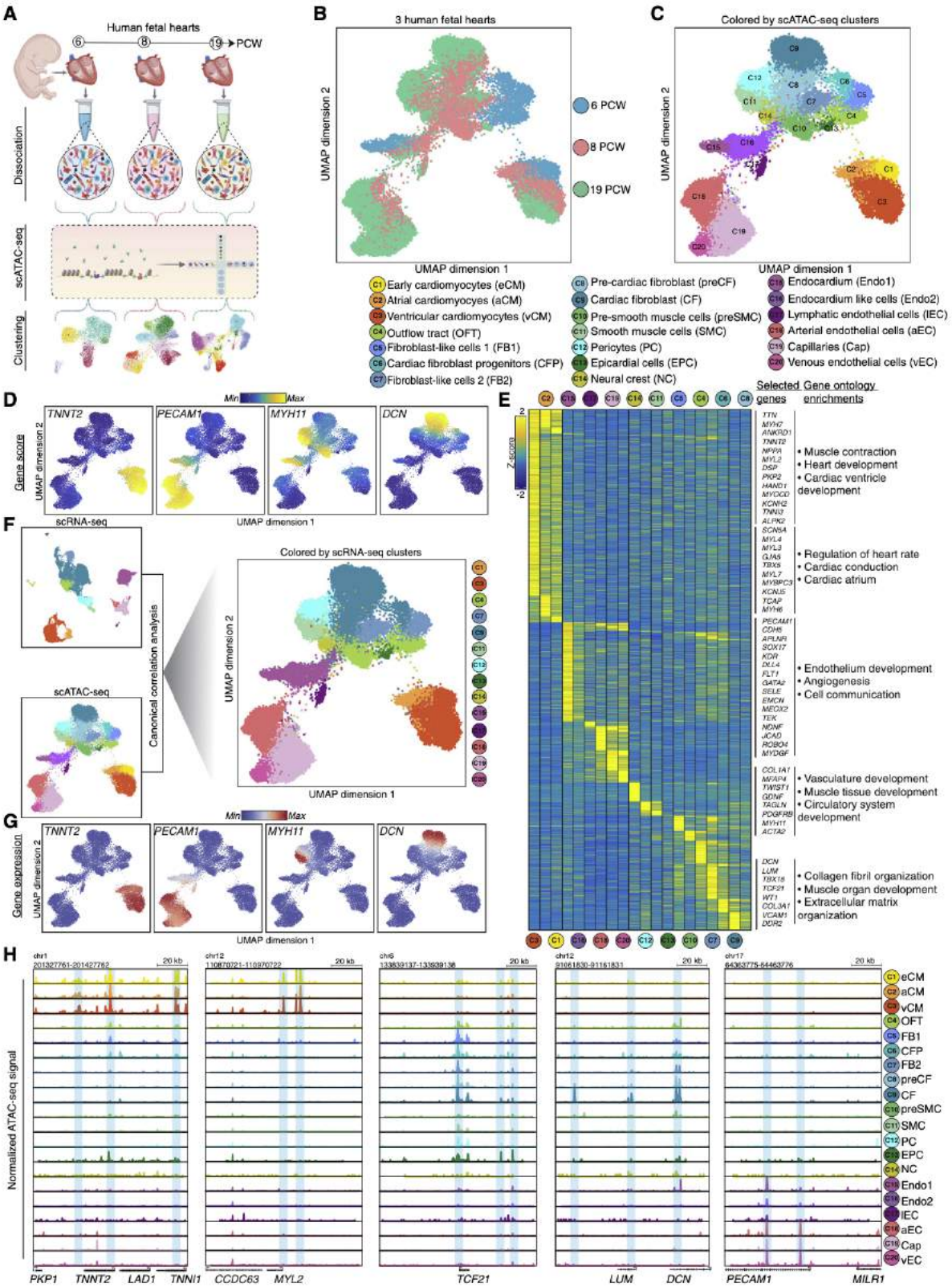
SUMMARY

To define the multi-cellular epigenomic and transcriptional landscape of cardiac cellular development, we generated single-cell chromatin accessibility maps of human fetal heart tissues. We identified eight major differentiation trajectories involving primary cardiac cell types, each associated with dynamic transcription factor (TF) activity signatures. We contrasted regulatory landscapes of iPSC-derived cardiac cell types and their *in vivo* counterparts, which enabled optimization of *in vitro* differentiation of epicardial cells. Further, we interpreted sequence based deep learning models of cell-type-resolved chromatin accessibility profiles to decipher underlying TF motif lexicons. *De novo* mutations predicted to affect chromatin accessibility in arterial endothelium were enriched in congenital heart disease (CHD) cases vs. controls. *In vitro* studies in iPSCs validated the functional impact of identified variation on the predicted developmental cell types. This work thus defines the cell-type-resolved *cis*-regulatory sequence determinants of heart development and identifies disruption of cell type-specific regulatory elements in CHD.

INTRODUCTION

Organogenesis of the heart begins from two distinct mesodermal cellular progenitors that originate from the primary heart field (PHF) and secondary heart field (SHF). These two mesodermal lineages give rise to three major subtypes of heart cells: myocardial, epicardial, and endocardial cells that later integrate with cells from the neural crest to form a functional human heart.[1,2,3] Prior studies that have profiled the single-cell transcriptome of the developing human heart have greatly enhanced our understanding of cell types and genes important for cardiogenesis.[4,5,6] However, a comprehensive resource of cell-type-resolved *cis* and *trans* regulators of gene expression programs across differentiation trajectories in human cardiac development is lacking.

Congenital heart disease (CHD) is the most common form of developmental birth defect, affecting 1% of live childbirths every year.[7] Approximately one-third of children with CHD have a linked genetic etiology accounting for the disorder. Only



(legend on next page)

8% of such cases are attributed to mutations in protein-coding gene regions,[8,9,10,11] suggesting that other causes, including disruption of gene regulation, substantially contribute to the etiology of CHD. The gaps in our understanding of transcriptional regulation of cardiogenesis and its dysregulation by non-coding CHD mutations raise several unresolved questions: (1) What are the dynamic *cis*-regulatory elements (cREs) and target genes that define cell types and cell state transitions in cardiogenesis? (2) What is the combinatorial lexicon of transcription factor (TF) motifs encoded in these dynamic cREs? (3) Are *de novo* non-coding CHD mutations enriched in cRE landscapes of specific fetal heart cell types? (4) What are the TF binding sites, cREs, and target genes impacted by putative causal non-coding CHD mutations? (5) Which *in vitro* differentiated cellular model systems demonstrably reproduce the chromatin landscape of the *in vivo* developing human heart, thereby enabling functional validation of the regulatory impact of mutations?

To address these questions, we derived a joint atlas of integrated single-cell data by generating and combining single-cell assay of transposase accessible chromatin sequencing (scATAC-seq) experiments. These studies profiled the chromatin landscape of three primary human fetal heart samples spanning post-conception weeks (PCW) 6, 8, and 19 and deconvolved 20 distinct cell types spanning three progenitor lineages and neural crest cells. We trained convolutional neural networks (CNN) that predict cell-type-resolved chromatin accessibility profiles from DNA sequence to decipher the dynamic motif lexicon of combinatorial TF binding at all cREs in each cell context.[12,13] We used the optimal transport algorithm to identify 8 major differentiation trajectories, defining the continuous progression of TF activities that promote the formation of primary cell types of the heart.[14] Using this atlas of cell states representing *in vivo* cardiac development, we compared accessible chromatin landscapes of common *in vitro* cellular model systems comprising major cardiac cell types derived from iPSCs. Based on insights from the comparison of *in vitro* and *in vivo* epicardial cells, we optimized the differentiation protocol for iPSC-derived epicardial cells which produced *in vitro* differentiated epicardial cells with substantially greater epigenomic similarity to *in vivo* counterparts. Finally, we used our deep learning models to prioritize, non-coding mutations in CHD trios from the Pediatric Cardiac Genomics Consortium (PCGC)[15] based on their predicted impact on cell-type-specific chromatin accessibility of putative cREs via disruption of TF binding sites. We used CRISPR-based enhancer knockout experiments with *in vitro* differentiated endothelial cells to validate the regulatory impact of a putative cell-type-specific enhancer predicted to harbor a deleterious CHD mutation altering expression of *JARID2*, an important CHD gene. Together, these datasets and predictive models define the *cis*- and *trans*-regulatory landscape of the developing human heart across mid-gestation developmental trajectories, elucidate the fidelity of diverse iPSC-to-lineage *in vitro* differentiations, and provide a deep learning framework capable of specifically nominating non-coding *de novo* mutations in candidate cREs predicted to disrupt TF binding, chromatin state in CHD.

RESULTS

Integrating single-cell ATAC and RNA sequencing data into a unified cell-type-resolved regulatory atlas of the developing human heart

To capture chromatin dynamics in different cell populations throughout fetal heart development, we used the Chromium 10X platform to generate scATAC-seq data[16] from three primary human fetal heart samples at 6-, 8-, and 19-weeks post-conception (PCW) (Figure 1A). We obtained 30,426 high quality scATAC-seq cell barcodes post filtering and quality control (Figure S1, Table S1, STAR Methods). We applied iterative latent semantic indexing (LSI) on accessible chromatin regions to map the cells from all three time points into a multidimensional principal component (PC) space[17,18,19] and used the Leiden clustering algorithm to discover and optimize clusters of cells that potentially correspond to distinct cell types[20] (Figures 1B, 1C, and S1, Table S1, STAR Methods). We deciphered each cluster's likely cell type identity based on chromatin-derived gene accessibility scores (GA-scores) of reference marker genes known to exhibit cell-type-specific gene expression and identified 215,163 putative cREs as scATAC-seq peak regions over all cell types (Figures 1D, 1E, and S1, Table S1, STAR Methods).[17]

To understand the correspondence between the chromatin and gene expression landscapes of these cell types, we analyzed previously published scRNA-seq data from developmental time

Figure 1. A single-cell epigenomic atlas of the developing human heart

(A) Schematic of gestational sample time (post-conception week, PCW) and genome-wide profiling methods represented in this study.

(B) Uniform Manifold Approximation and Projection (UMAP) of cells based on accessible chromatin regions (scATAC-seq). Cells are colored according to sample gestational time.

(C) UMAP of cells based on accessible chromatin regions (scATAC-seq). Cells are colored according to cell types identified.

(D) Single-cell gene accessibility scores (based on scATAC-seq) of *TNNT2*, *PECAM1*, *MYH11*, and *DCN*.

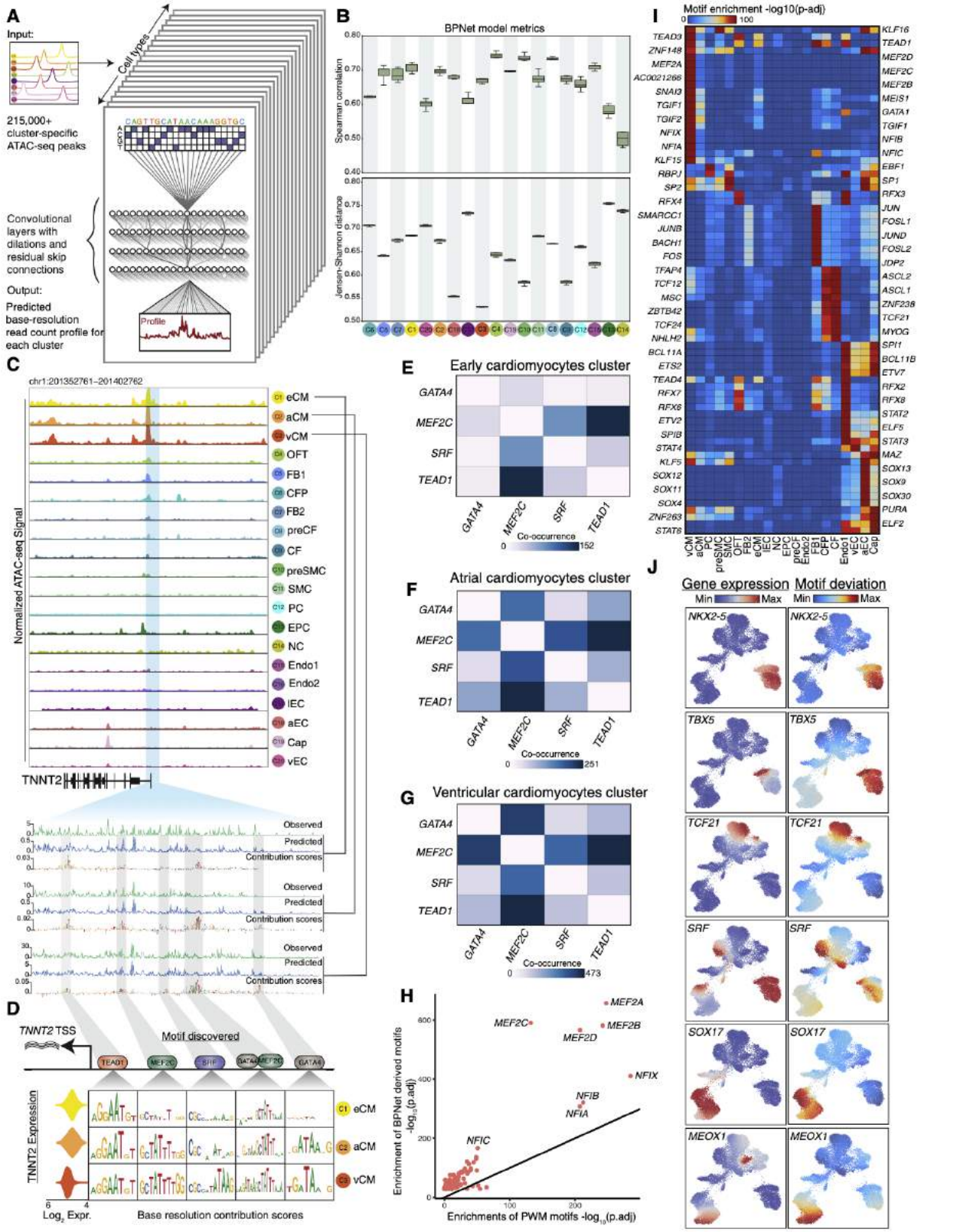
(E) Heatmap of *Z* scores of $\log_2$(scATAC-seq read counts) in 215,163 *cis*-regulatory elements (cREs) across scATAC-seq cell type clusters derived from (B). Representative genes with cluster-specific differential gene accessibility scores are shown to the right. Gene ontology enrichments indicate the statistically significant (adjusted p value < 0.005, Gprofiler Fisher exact test) cellular processes for genes with differential gene accessibility scores associated with the clusters of cell-type-specific cREs.

(F) UMAPs of scRNA-seq and scATAC-seq cells colored by cluster assignment in their respective data modality, and UMAP of scATAC-seq cells highlighted by complementary scRNA-seq clusters.

(G) Single-cell gene expression (scRNA-seq) of *TNNT2*, *PECAM1*, *MYH11*, and *DCN*.

(H) Genome tracks of cell-type-resolved aggregate scATAC-seq data around the *TNNT2*, *MYL2*, *TCF21*, *DCN/LUM*, and *PECAM1* gene loci (left to right). The scale of the tracks (from left to right) range from 0–0.28, 0–0.31, 0–0.18, 0–0.14, and 0–0.2, respectively, in units of fold-enrichment relative to the total number of reads in TSSs per 10K. Highlights indicate the relevant cell-type-specific putative enhancers in each gene locus.

See also Figures S1 and S2.



(legend on next page)

points that closely match those sampled in our scATAC-seq atlas[4,5,6,21] (Figures 1F and S2, Table S1). Cells from our annotated scATAC-seq atlas were then matched with their nearest neighbor cells in the scRNA-seq atlas using canonical correlation analysis (CCA)[22], and we observed highly concordant imputed gene expression of marker genes (Figures 1F, 1G, S2D, and S2E).

Next, we used our integrated atlas to examine the relationship between the expression of well-known lineage-specific marker genes and the chromatin dynamics of their putative cREs. For example, *TNNT2,* a well-known cardiomyocyte marker, exhibited the strongest accessibility at its promoter and putative distal enhancers, specifically in the three cardiomyocyte clusters (Figure 1H). The patterns of accessibility matched the specificity and relative levels of expression of *TNNT2* in the same clusters (Figure S2C). In contrast, *MYL2*, a specific marker of vCMs, exhibited similar distal chromatin accessibility in the three myocardial lineage clusters, while the promoter was not accessible, and the gene was not expressed, in aCMs (Figures 1H and S2C), indicating that accessibility of these distal elements may not be sufficient to drive its expression. In the epicardial cell lineage, we observed increasing chromatin accessibility around the *DCN* marker gene through the cardiac fibroblast cell lineage specification (Figure 1H) concordant with its gene expression dynamics (Figure 1G). We observed analogous dynamics for *PECAM1* and *TCF21* in the endocardial and epicardial lineages, respectively.

Deciphering cell-type-resolved *cis*-regulatory sequence lexicons with deep learning models of base-resolution chromatin accessibility profiles

To decipher the *cis*-regulatory sequence lexicon of TF binding sites in accessible cREs in each cell type, we trained BPNet convolutional neural networks to learn a mapping from 1 kb DNA sequence windows around scATAC-seq peaks and background regions to the corresponding base-resolution, pseudobulk chromatin accessibility profiles[12,13] (Figure 2A). We obtained high, stable Spearman correlation between total observed and predicted Tn5 insertion coverage as well as high concordance between observed and predicted profile shapes at base-resolution in held-out test chromosomes over 5-folds of a chromosome hold-out cross-validation scheme in all cell types (Figure 2B, Table S2).[13]

Next, we interrogated each cell-type-specific BPNet model with the DeepLIFT algorithm to derive the quantitative contribution of every base-pair in each accessible cRE sequence to its predicted accessibility.[23,24] DeepLIFT scores from the eCM BPNet model highlighted short, contiguous stretches of bases with high contribution scores, reminiscent of TF binding motifs, in the accessible promoter of *TNNT2*, a gene critical for sarcomere contractile function of the heart[25] (Figure 2C). Hence, to annotate predictive, active motif instances in all accessible cREs of each cell type, we scanned their sequences for matches to a non-redundant compendium of known TF sequence motifs[26] and restricting to matched instances with high DeepLIFT contribution scores or motif mutagenesis scores derived from each cell-type-specific BPNet model. Although the sequence of a cRE is the same in all cell types, its DeepLIFT contribution score profile can vary across cell types, reflecting cell-type-specific prediction of motif activity by BPNet models of different cell types. For example, the *TNNT2* promoter is highly and equally accessible in all 3 types of cardiomyocytes and drives expression of *TNNT2* in all 3 cell types (Figure 2C). However, the DeepLIFT profiles derived from the eCM, aCM and vCM models for the same promoter sequence highlight distinct combinations of active TF motif instances predicted to regulate accessibility in the three cell types (Figures 2C and 2D). A *TEAD1* motif is predicted to regulate promoter accessibility in all three cell types. A nearby *MEF2C* motif is predicted to be uniquely active in aCM and vCM, while another upstream *MEF2C* motif active in eCM is predicted to be part of a *GATA-MEF* composite motif that is specifically active in aCM and vCM. A *GATA* motif, further upstream, is predicted to be active specifically in aCM and vCM. An *SRF* motif

Figure 2. Cell-type-resolved predictive transcription factor motif syntax derived from deep learning models of base-resolution scATAC-seq profiles

(A) Schematic of the convolutional neural network (BPNet) trained to simultaneously predict base-resolution probability distribution of reads and total read counts of cell-type-resolved pseudobulk scATAC-seq profiles over each 1-kb accessible peak region from 2-kb underlying DNA sequences.

(B) Performance evaluation of BPNet cluster-specific models, computed as the Spearman correlation between observed and predicted total counts (higher is better) across all peaks in each cluster (top) and mean Jenson-Shannon distance (lower is better) between the base-resolution observed and predicted profiles across all peaks in each cluster (bottom). Results are reported on test sets from a 5-fold cross-validation setup.

(C) Top shows the genome tracks of aggregate pseudobulk scATAC-seq around the *TNNT2* locus for each of the cell type clusters. The scale ranges from 0–0.34 in units of fold-enrichment relative to the total number of reads in TSSs per 10K. Bottom zooms into an accessible peak around the *TNNT2* transcription start site and shows the observed base-resolution scATAC-seq read count profiles from the early (eCM), atrial (aCM), and ventricular cardiomyocytes (vCM) clusters, the predicted profiles from the BPNet models of each of the three cell types, and the corresponding DeepLIFT contribution score profiles (height of each base in the sequence is proportional to its contribution score).

(D) Per-base DeepLIFT contribution scores of *TEAD1*, *MEF2C*, *SRF*, and *GATA4* motif locations in the *TNNT2* promoter from eCM, aCM, and vCM (rows from top to bottom). Leftmost column shows the distribution of scRNA-seq expression (in units of $\log_2$[transcripts per 10K]) of *TNNT2* across cells from each of the three clusters.

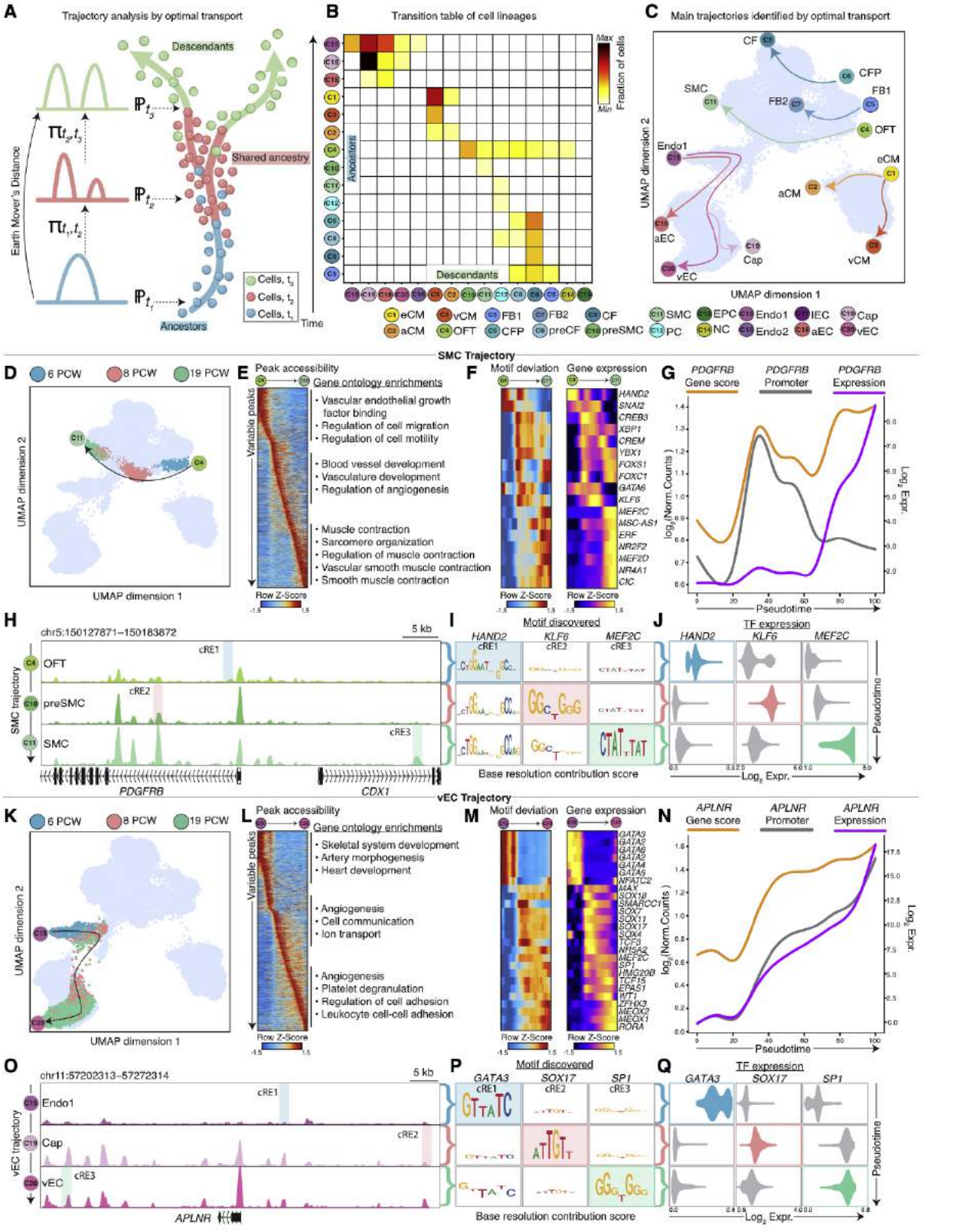
(E–G) Pairwise motif co-occurrence counts for *TEAD1*, *MEF2C*, *SRF*, and *GATA4* motifs based on predicted active motifs across all accessible cREs in eCM, aCM, and vCM, respectively.

(H) Comparison of statistical significance of overlap enrichment ($-\log_{10}$ p value, Wilcoxon rank-sum test) of BPNet model-derived predictive motif instances (y axis) vs. position weight matrix (PWM) based motif instances (x axis) in vCM accessible peaks regions. Predictive motif instances show higher significance of enrichments.

(I) Differential enrichments of BPNET model derived predictive motif instances of transcription factors (rows) in accessible peaks of different cell types (columns).

(J) (left column) scRNA-seq gene expression (in units of $\log_2$[transcripts per 10K]) and (right column) scATAC-seq based ChromVAR motif deviation scores (in units of *Z* scores) for *NKX2-5*, *TBX5*, *TCF21*, *SRF*, *SOX17*, and *MEOX1* shown in the scATAC-seq UMAP representations of all cells.

See also Figure S3.



(legend on next page)

is predicted to be active only in vCM. The higher density of predicted active motifs in the *TNNT2* promoter in aCM and vCM compared to eCM is concordant with the higher expression of *TNNT2* in the former two cell types (Figure 2D). This combinatorial, cell-type-specific motif syntax of these 4 TF at the *TNNT2* promoter is consistent with the genome-wide co-occurrence statistics of their active motifs across all cREs in eCM, aCM and vCM (Figures 2E, 2F, and 2G, Table S2).

We also found that most TFs that are expected to be active in vCMs, including those belonging to the *MEF2* family and *NFI* family, showed significantly stronger enrichment (Benjamini-Hochberg (BH) adjusted hypergeometric test p value < 1e-500 for *MEF2* and < 1e-150 for *NFI*) of active motif instances relative to PWM motif instances in differential, cell-type-specific vCM peaks (Figure 2H). Next, we estimated the enrichment of active motif instances of TFs in accessible cREs of each cell type to identify the TF regulators of cell-type-resolved chromatin accessibility landscapes (Figures 2I and S3A). The cell type specificity of globally predictive TFs identified by the BPNet models was further corroborated by high concordance (Table S2) between TF activity scores (chromVAR[27]) and the expression of the TFs in the scRNA-seq data across developmental timepoints (Figure 2J). Our analyses thus provide a comprehensive resource of cell-type-resolved TF lexicons and annotations of predictive TF sequence motifs in cRE landscapes of human fetal heart development.

Inferring dynamic regulatory control across major cellular differentiation trajectories in human cardiogenesis

Next, we sought to identify major developmental trajectories involving cell state transitions across fetal heart development based on single-cell chromatin dynamics. We used the optimal transport algorithm,[14] previously developed to derive trajectories from scRNA-seq data, to identify the most parsimonious transitions in global chromatin accessibility between cells from PCW6 to PCW19 of fetal heart development (Figures 3A, 3B, 3C, S3B, S3C, and S3D, Table S3, STAR Methods). Overall, we characterized 8 dominant trajectories for all the major cell types at PCW19 (Figures 3B, 3C, and S4). We then characterized genome-wide and locus-specific regulatory dynamics associated with cell state transitions across these trajectories. Below, we present representative case studies contrasting regulation of the development trajectories leading to SMC cell fate.

The SMC trajectory begins with the OFT cells at PCW6 that transition through an intermediate preSMC population in PCW8 to the SMCs at PCW19[28] (Figure 3D). A continuous cascade of dynamically accessible cREs defines cell state transitions across the trajectory (Figure 3E). These dynamic cREs are proximal to genes enriched for temporally relevant vascular developmental processes including cell migration, angiogenesis, and muscle contraction at early, intermediate, and late time points, respectively (Figure 3E). Expression dynamics of several key lineage specifying TFs including *HAND2*, *SNAI2*, *KLF6*, and *MEF2C* were strongly correlated with their chromatin-based motif activity (chromVAR deviation scores) across this trajectory (Figure 3F). Tracking the chromatin accessibility and gene expression of *PDGFRB*, one of the primary marker genes for the SMC population, we observed that initially, the promoter of *PDGFRB* accounts for the majority of accessibility at this locus while gene expression is low (Figure 3G).[29,30] The increase in expression of *PDGFRB* at later time points is associated with increased accessibility of putative intronic

Figure 3. Identifying developmental trajectories in human fetal heart development

(A) Schematic of the optimal transport method used to determine trajectories of cell state transitions using scATAC-seq gene scores of all the cell types identified in Figure 1C.

(B) Cell state transition table of cell lineages identified in the major trajectories obtained through optimal transport. Rows correspond to the parent cell types and columns correspond to the derivative cell types. The heatmap is colored by the fraction of parent cells identified to be ancestors of the derivative cells. (Scale for transition table.: 0.01 to 0.30).

(C) UMAP of scATAC-seq cells highlighting the dominant trajectories identified using optimal transport. The cell types correspond to those in Figure 1C.

(D) UMAPs of scATAC-seq cells in the smooth muscle cell (SMC) trajectory colored by the gestational sample time.

(E) Heatmap of scATAC-seq signal (*Z* score of $\log_2$[reads per 10K]) of variable peaks identified in the SMC pseudotime trajectory. The gene ontology enrichments are calculated using the variable gene scores in the trajectory.

(F) Heatmaps showing *Z* score of ChromVAR motif deviation scores (left) and gene expression $\log_2$(transcripts per 10K), also applicable for all gene expression values plotted in this figure) (right) of TFs with correlated variable activity in cells identified to be in the SMC trajectory, as ordered by pseudotime.

(G) Gene expression, promoter chromatin accessibility $\log_2$(reads per 10K) +/− 1,000 bp TSS and chromatin-derived gene accessibility score ($\log_2$[reads per 10K), applicable for all gene activity values in this figure) dynamics of the *PDGFRB* gene across pseudotime.

(H) Genome tracks of aggregate scATAC-seq data around the *PDGFRB* locus in OFT, preSMC, and SMC clusters. cRE1, cRE2, and cRE3 are three representative cREs with dynamic motif activity further explored in (I) and (J). The ATAC signal range is 0–0.64 in units of fold-enrichment relative to the total number of reads in TSSs per 10K.

(I) Per-base contribution scores of motifs of *HAND2*, *KLF6*, and *MEF2C* in the 3 highlighted cREs in (H). Rows (top to bottom) are per-base contribution scores computed using BPNet models of OFT, preSMC, and SMC, respectively. The columns (left to right) are the highlighted cREs from (H) that are active in OFT, preSMC, and SMC, respectively.

(J) Distribution of scRNA-seq gene expression of *HAND2*, *KLF6*, and *MEF2C* TFs (columns) across cells from OFT, preSMC, and SMC clusters (rows).

(K) UMAPs of scATAC-seq cells in the venous endothelial cell (vEC) trajectory colored by the gestational sample time.

(L) Heatmap of *Z* scores of variable peaks identified in the vEC pseudotime trajectory, similar to (E).

(M) Heatmaps showing *Z* score motif activity (left) and expression (right) of TFs in the vEC trajectory, similar to (F).

(N) Gene expression, promoter chromatin accessibility and chromatin-derived gene accessibility score dynamics of the APLNR gene across pseudotime.

(O) Genome tracks of aggregate scATAC-seq data around the *APLNR* locus in Endo1, Cap and vEC clusters. cRE1, cRE2, and cRE3 similar to (H).

(P) Per-base contribution scores of motifs of *GATA3*, *SOX17*, and *SP1* in the 3 highlighted cREs in (O), similar to (I).

(Q) Distribution of scRNA-seq gene expression of *GATA3*, *SOX17*, and *SP1* TFs (columns) across cells from Endo1, Cap, and vEC clusters (rows).

See also Figures S3 and S4.

enhancers. We then used predictive motif instances derived from cell-type-specific BPNet models to associate inferred TF binding dynamics at specific cREs in the *PDGFRB* locus with TF expression changes across the three timepoints (Figures 3H and 3I).

BPNet models of OFT cells at the PCW6 time point revealed a predictive *HAND2* binding motif (Figure 3I) in a downstream putative enhancer (cRE1 in Figure 3H) that is highly accessible at this early time point. The predicted TF motif dynamics of *HAND2* at this enhancer was correlated with the expression dynamics of *HAND2,* which also peaks in PCW6 and decreases thereafter (Figure 3J). Another cRE (cRE2 in Figure 3H) proximal to the promoter of *PDGFRB*, which showed the highest accessibility in preSMC at the intermediate PCW8 time point, was predicted to be regulated by *KLF6* whose motif showed high contribution scores specifically in the preSMC model (Figure 3I) and whose expression also peaked in preSMCs (Figure 3J). A distal cRE upstream of *PDGFRB* (cRE3 in Figure 3H) with highest accessibility in SMC in PCW19 was predicted to be regulated by *MEF2C* whose motif was specifically predictive in SMC BPNet model (Figure 3I) and whose expression peaked in SMC (Figure 3J). We observed similar dynamics for the vEC and other differentiation trajectories as well (Figures 3K–3Q and S4). Our analysis framework thus provides a lens into the dynamic *cis*-regulatory code of developmental cellular trajectories in human cardiogenesis at basepair resolution.

A systematic comparison of regulatory landscapes of *in vitro* differentiated cardiac cell types and their *in vivo* counterparts in human fetal heart development

Several human induced pluripotent stem cell (iPSC) based *in vitro* cellular models have been developed, including cardiomyocyte (i-CM), endothelial (i-EC), epicardial (i-EPC), cardiac fibroblast (i-CF), and smooth muscle (i-SMC) cells.[31,32,33,34] Our comprehensive, integrated single-cell atlas of *in vivo* cardiac cell types from developing fetal hearts provides an opportunity to investigate the authenticity of these *in vitro* cellular models.

To address this question, we generated iPSC-derived i-CM, i-EC, i-EPC, i-CF, and i-SMC cells through directed differentiation employing established protocols[31,32,33,34] (Figure 4A). We generated scATAC-seq data from all these *in vitro* differentiated iPSC lines at multiple time points using the Chromium (10X Genomics) platform (Figures S5A and 4B, Table S4). Integration and clustering of cells from these scATAC-seq datasets broadly identified nine different cell types, including day 0 iPSC, day 2 mesodermal cells (i-Mes), day 5 i-CP, day 15 i-pCM, and day 30 i-CM, i-EPC, i-SMC, i-CF, and i-EC. Once again, the scATAC-seq derived GA-scores of marker genes were found to be highly specific for the relevant cell types, confirming our cell type annotations[35–37] (Figures 4C and S5B, Table S4).

To evaluate the similarity between chromatin landscapes of the *in vitro* differentiated cell types and their *in vivo* counterparts, we first used the LSI method to project *in vitro* differentiated cells onto the scATAC-seq LSI subspace of all cells from the fetal heart samples[38] (Figure 4D). Majority of Day-15 i-pCMs projected into the PCW6 *in vivo* myocardium-derived eCMs. At day-30, i-CMs projected primarily into the PCW8 *in vivo* vCMs and *in vivo* eCMs, while i-ECs projected across the *in vivo* Endo1, Endo2 and the PCW8 Cap cells. In contrast, *in vitro* epicardium-derived cells, including i-EPC, i-SMC, and i-CF, were distributed across epicardial cell types of the fetal heart without a strong correspondence to their specific *in vivo* counterparts (EPC, SMC, and CF). The day-5 *in vitro* i-CP cells were found to consist of five subclusters that projected across all three distinct lineages of the fetal heart, the myocardium, epicardium, and endocardium, supporting the likely origin of all major differentiated *in vivo* cardiac cell types from a precursor state similar to i-CPs (Figure 4D).

Looking at the differential accessible sites between the *in vitro* cells and their nearest neighbors, we observed that i-pCMs, i-CMs, and i-ECs had the least number of differential peaks relative to their matched *in vivo* cell types (Figure 5A). Consistent with the co-projection analysis, comparison of matched *in vitro* epicardial cell types (i-EPC, i-SMC, and i-CF) and their *in vivo* counterparts revealed more differential peaks relative to corresponding comparisons of cardiomyocytes and endothelial cells. We next identified TF motifs enriched in the differentially accessible scATAC-seq peaks. *AP-1* (*JUN-FOS*, *JDP2*) motifs were strongly enriched in peaks upregulated in most *in vitro* cell types, except cardiomyocytes (Figure 5B). In contrast, downregulated peaks in *in vitro* cell types were most enriched for *SP*, *KLF*, and *WT1* motifs (Figure 5C). Differentially upregulated peaks in i-pCMs and i-CMs were enriched for motifs of classical cardiac TFs including *MEF2* and *NKX*, consistent with their role in cardiomyocyte differentiation.[39] Motifs of *FOX* and *CEBP* TF families, which are involved in epithelial-to-mesenchymal transition (EMT), were enriched in peaks upregulated in *in vitro* epicardium-derived cell types compared to their post-EMT *in vivo* counterparts,[40,41,42,43] suggesting that the *in vitro* epicardial cells may not represent a terminally differentiated state.

Based on these observations, we sought to modulate the EMT pathways active in the iPSC-derived epicardial cells, to improve the development of epicardial-derived cellular lineages. We generated a new differentiation protocol for iPSC-derived epicardial cell lineages to inhibit the EMT activity and promote more faithful recapitulation of *in vivo* differentiation processes. Primarily, this was accomplished by developing a new chemically defined medium that removed the unnecessary components in the commercial medium that might be responsible for the EMT signal (STAR Methods). We observed robust immunofluorescence-based staining for the WT1 marker gene for the new i-EPC cells, confirming the cellular phenotype of these cells and validating our new medium (Figure 5D). We profiled single-cell chromatin accessibility of the new i-EPCs using the 10X Chromium platform. The new i-EPC cells projected more specifically into the *in vivo* epicardial cells of the fetal atlas compared to the original i-EPC cells (Figures 5E and 5F). The *in vivo* EPCs constituted 45% of the nearest *in vivo* cell neighbors of the new i-EPCs compared to only 13% for the original i-EPCs (Figure 5G). Spurious differentially accessible peaks upregulated in the new i-EPCs relative to the *in vivo* EPCs were 35% lower than those between the original i-EPCs and *in vivo* EPCs (Figure 5H). Downregulated differential peaks also showed a 45% reduction (Figure 5I). These observations suggest that the new differentiation protocol produced i-EPCs whose chromatin

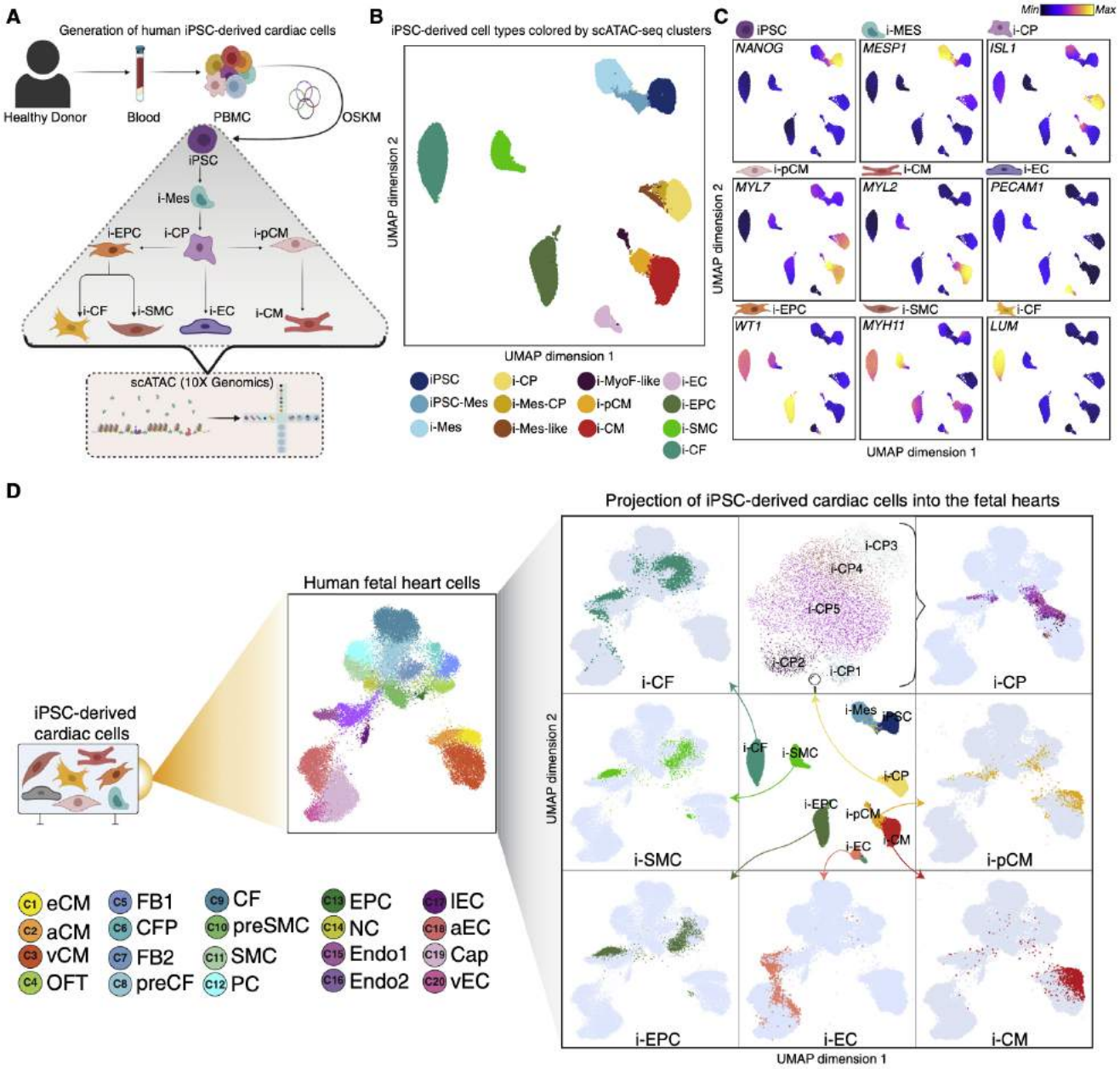


Figure 4. Characterization of *in vitro* iPSC-derived cardiac cell types

(A) Schematic for derivation of human iPS cells, followed by their differentiation to major cardiac cell types and genome-wide scATAC-seq profiling.

(B) scATAC-seq UMAP of all *in vitro* iPSC-derived cells colored according to cell types identified during differentiation (iPSC, induced pluripotent stem cells; iPSC-Mes, partially differentiated mesoderm-like cells; i-Mes, cardiac mesoderm cells; i-CP, cardiac progenitors; i-Mes-CP, partially differentiated cardiac progenitor-like cells; i-Mes-like, partially differentiated mesoderm-like cells; i-MyoF-like, Myofibroblast-like cells; i-pCM, Day 15 iPSC-derived primitive cardiomyocytes; i-CM, Day 30 iPSC-derived mature cardiomyocytes; i-EC, iPSC-derived endothelial cells; i-EPC, iPSC-derived epicardial cells; i-SMC, iPSC-derived smooth muscle cell; and i-CF, iPSC-derived cardiac fibroblast cells).

(C) Gene accessibility scores of marker genes *NANOG*, *MESP1*, *ISL1*, *MYL2*, *MYL7*, *PECAM1*, *WT1*, *MYH11*, and *LUM* projected on the scATAC-seq fetal heart UMAP.

(D) Projection of cells from scATAC-seq experiments profiling *in vitro* iPSC-derived cardiac cell types into the scATAC-seq fetal heart UMAP. Central panel in the 3x3 grid shows the scATAC-seq UMAP of all *in vitro* cardiac cell types. The other panels in the grid are projections of the i-CF (row 1, col 1), i-SMC (row 2, col 1), i-EPC (row 3, col 1), i-EC (row 3, col 2), i-CM (row 3, col 3), and i-pCM (row 2, col 3) cells into the scATAC-seq fetal heart UMAP. Panel in row 1, col 2 shows an scATAC-seq UMAP of 5 subclusters of cells from *in vitro* cardiac progenitors (i-CP1, i-CP2, i-CP3, i-CP4, and i-CP5), which are projected into the scATAC-seq fetal heart UMAP (row 1, col 3).

See also Figure S5.

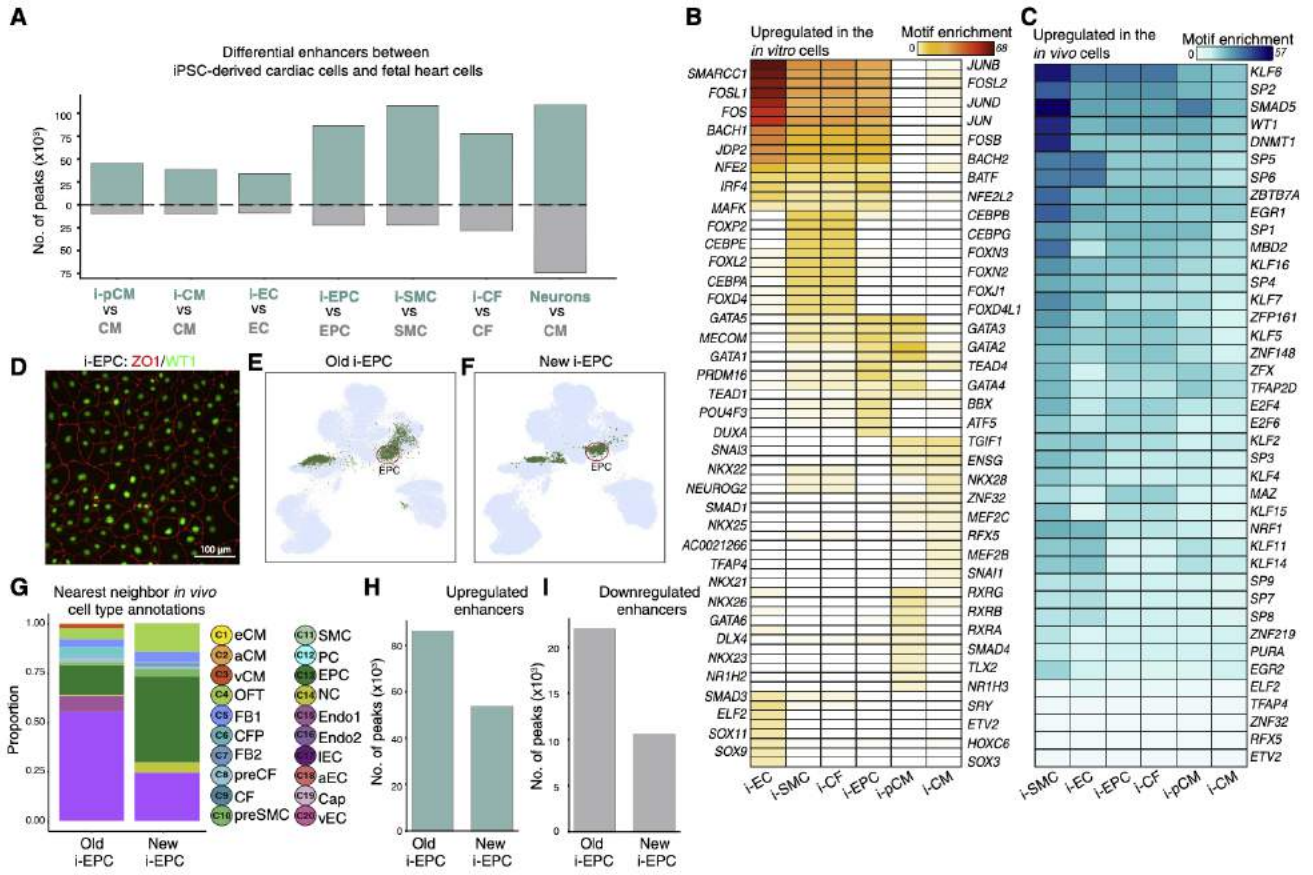


Figure 5. Characterization of *in vitro* iPSC-derived cardiac cell types

(A) Comparison of number of significantly ($\log_2$ fold-change > 1, FDR < 0.05 using two-sided t test) upregulated (in blue) and downregulated (in gray) scATAC-seq peaks in *in vitro* cardiac cell types relative to nearest *in vivo* fetal heart cell types. An analogous differential comparison between *in vivo* ventricular cardiomyocytes from fetal heart and *in vivo* glutamatergic neurons from fetal brain is shown as a reference (rightmost bar).

(B and C) Statistical significance ($-\log_{10}$[adjusted p value], BH-adjusted hypergeometric test) of overlap enrichment of TF motifs in upregulated (B) and downregulated (C) scATAC-seq peaks in *in vitro* cardiac types relative to nearest *in vivo* fetal heart cell types from (A).

(D) Immunofluorescence staining of WT1 in iPSC-derived epicardial cells (i-EPC) from the new epicardial differentiation protocol.

(E and F) Projection of i-EPC cells from scATAC-seq experiments profiling *in vitro* iPSC-derived cardiac cell types onto the scATAC-seq fetal heart UMAP. i-EPC cells from old differentiation protocol (E) and i-EPC cells from the new differentiation protocol (F).

(G) Comparison of cell type annotations of nearest neighbor *in vivo* cells to the old and new i-EPC differentiated cells.

(H and I) Comparison of the number of upregulated differential enhancers (H) and down regulated differential enhancers (I) in old and new i-EPC cells compared to the nearest neighbor *in vivo* cells. Reduction of differential enhancer number is consistent with greater fidelity of representation of epigenetic state, for *in vitro* versus *in vivo* cells. Decreased differential enhancer number suggests more faithful recapitulation of *in vivo* cellular phenotype.

See also Figure S5.

landscapes are substantially more similar to those of *in vivo* epicardial cells in the fetal heart than those derived from the original protocol.

Prioritizing putative causal non-coding *de novo* mutations, TFs, target genes, and cell types in congenital heart diseases

Next, we investigated the utility of our regulatory atlas to decipher single nucleotide, *de novo*, non-coding mutations in congenital heart disease (CHD) patients. We compiled a set of 54,126 *de novo*, non-coding mutations from 763 CHD patients from the Pediatric Cardiac Genomics Consortium[15] (PCGC) (Table S5) and a control set of 110,055 *de novo*, non-coding mutations from healthy controls from the Simons Simplex Collection (n = 1,902 trios) (Table S5). We tested the accessible cRE landscapes of each of the *in vivo* fetal heart cell types for the enrichment of case versus control mutations. Surprisingly, all cell types lacked enrichment (Figure S6A), suggesting that overlapping mutations with cell-type-resolved cREs is insufficient to prioritize potentially causal CHD mutations.

We next used the corresponding cell-type-specific BPNet models to estimate mutation impact scores of candidate case and control point mutations in accessible cREs as the $\log_2$ fold-change in the cumulative predicted scATAC-seq profile probabilities for both alleles over a 100 bp window centered at each mutation (Figure 6A). We observed striking variation of

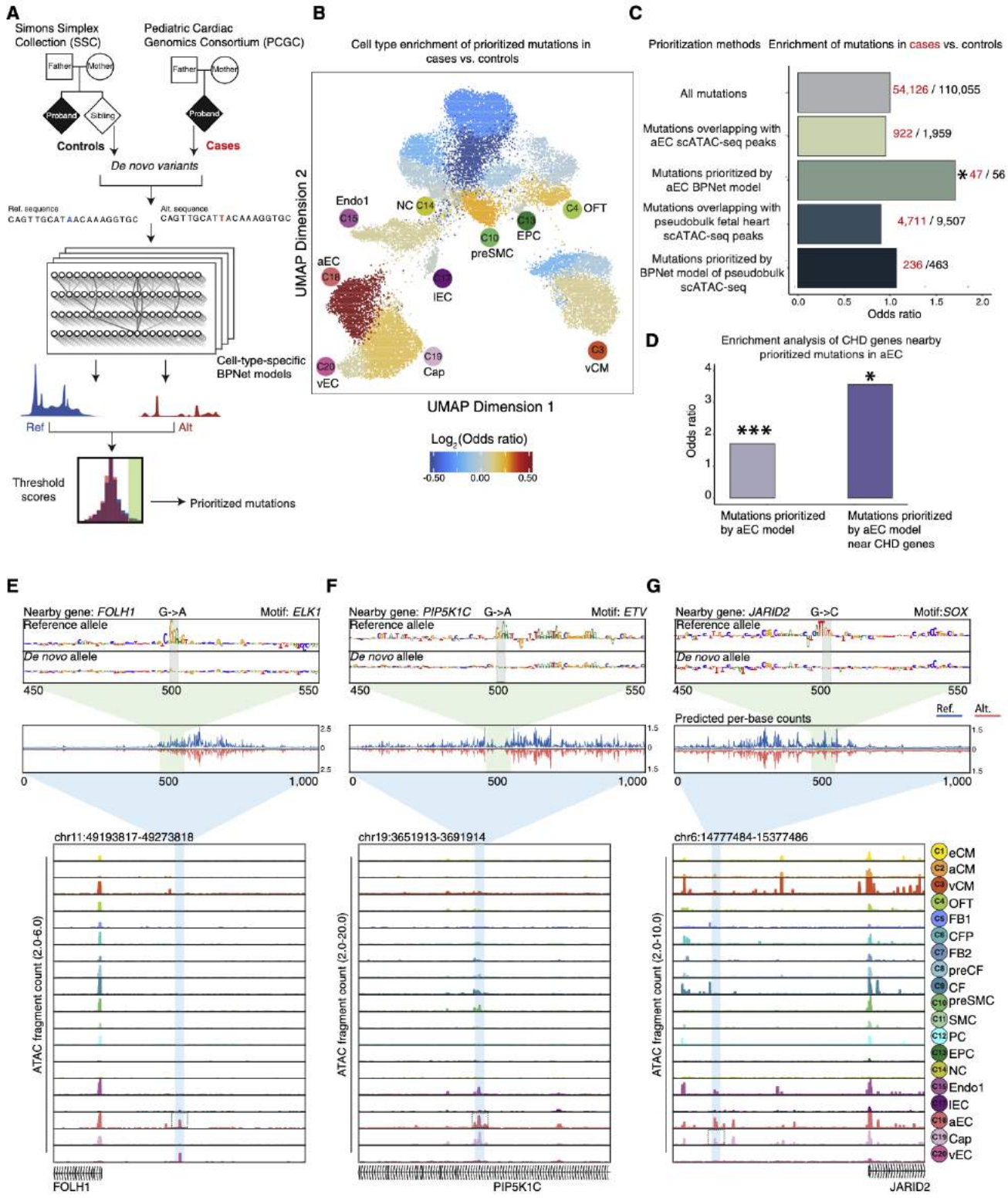


Figure 6. Prioritizing non-coding CHD mutations using deep learning models of scATAC-seq profiles from fetal heart cell types

(A) Schematic of mutation prioritization pipeline that uses cell-type-specific BPNet models to predict scATAC-seq profiles of CHD mutation analysis.

(B) Enrichment (log_2[OR], Fisher Exact Test) of prioritized mutations from each cell-type-specific BPNet model in CHD cases vs. controls plotted on the scATAC-seq UMAP of all fetal heart cells.

(legend continued on next page)

the enrichment of mutations with high predicted mutation impact scores (>95th percentile of the distribution of cell-type-specific impact scores for CHD mutations in peaks) in cases versus controls across cell types (Figure 6B, STAR Methods). Mutations prioritized in several cell types showed weak to moderate enrichments, including NC (OR = 1.016), lEC (OR = 1.033), EPC (OR = 1.042), Endo1/2 (OR = 1.106), vEC (OR = 1.092), vCM (OR = 1.119), Cap (OR = 1.205), OFT (OR = 1.22), and preSMC (OR = 1.307) (Figure 6B, Table S5). The strongest enrichment (Cases n = 47; Control n = 56; OR = 1.707; p value = 0.008, Fisher Exact test) was obtained for mutations prioritized in arterial endothelial cells (aECs) (Figures 6B and 6C, Table S5), which is consistent with the contribution of the endothelial cellular lineage to multiple cardiac structures. These patterns of cell-type-specific enrichment were robust to different measures of mutation impact scores and thresholds for defining high-impact mutations (Figures S6B–S6H).

In contrast, mutation impact scores derived from BPNet models trained on pseudobulk scATAC-seq profiles agglomerated over all fetal heart cell types (OR = 1.01) and HeartENN[15] trained on a large compendium of bulk chromatin data, did not enrich for CHD mutations, indicating that cell type specificity of mutation impact scores is critical for prioritizing *de novo* CHD mutations (Figures 6C and S6I). We further examined whether high-impact mutations prioritized by BPNet in aECs occurred near genes previously associated with CHD based on genetic studies in human cohorts or mouse models obtained from Richter, et al.[15] (744 total CHD-associated genes). We observed a 3-fold enrichment (p value = 0.0486, Fisher Exact test) of predicted high-impact aEC mutations proximal to previously implicated CHD genes in cases (n = 7) compared to controls (n = 4) (Figure 6D).

Next, we performed deeper investigations of the causal chain of TF binding sites, cREs and target genes potentially affected by a subset of high-impact CHD mutations prioritized in aECs that are in close proximity (<200 bp) to summits of high coverage aEC scATAC-seq peaks (Table S5). We used the active motif annotations derived from the cell-type specific BPNet models and the corresponding allele-specific base-resolution contribution scores of cRE sequences harboring these mutations to infer potentially disrupted TF binding sites (Figures 6E, 6F, and 6G). A prioritized G-to-A *de novo* mutation was predicted to ablate an *ELK/ETV* TF motif in a cRE that is exclusively accessible in endothelial cells (aEC, Cap, vEC and lEC) and ~25 kb upstream of a folate hydrolase gene *FOLH1*. *FOLH1* is expressed in endothelial cells (Figure S6J) and has been associated with loss of normal structural endothelial cell integrity[44,45] (Figure 6E). Another G-to-A mutation was predicted to disrupt an *ELK/ETV* TF motif in an endothelial cRE in the intron of the *PIP5K1C* gene, an important developmental TF strongly expressed in endothelial cells (Figure S6K) and implicated in cardinal vein and right ventricular development and CHD[15,44,46] (Figure 6F). Interestingly, several other prioritized mutations were also predicted to disrupt *ELK/ETV* binding sites in accessible aEC cREs proximal to the *MGAT1*, *TIMP3*, *TBX3*, and *NEK3* genes (Table S5), all of which have been previously associated with CHD or cardiovascular defects in human genetic studies or mouse models.[15,44,47,48,49,50] We also found a G-to-C mutation in an accessible cRE distal to the *JARID2* gene predicted to disrupt a *SOX* TF motif in aEC and Cap cells (Figure 6G). JARID2 is an important endothelial TF (Figure S6L) during early heart development, and coding mutations in *JARID2* have been implicated in CHD by previous studies, especially for tetralogy of Fallot.[51,52,53,54]

We used CRISPR/Cas9 to delete 352 bp around the *JARID2* mutation in iPSCs, selected single clones with biallelic deletions of the targeted locus, differentiated these clones into endothelial cells and measured expression of *JARID2* (Figures 7A and S6M). We observed a significant decrease (1.3-fold, p value <0.001, two-sided t test) of *JARID2* expression (Figure 7B) in edited iPSC-derived ECs compared to isogenic controls, thereby verifying transcriptional regulation of *JARID2* by this cRE in the nominated cell type. To further characterize the phenotypic impact of knocking out of this *JARID2* cRE, we compared wild-type (Figure 7C) and *JARID2* cRE KO (Figure 7D) iPSC-derived endothelial cells (i-ECs) for their ability to undergo angiogenesis (tube formation) in an *in vitro* assay. We observed a significant depletion of tubes in the cRE KO cells compared to the wild-type isogenic cells (Figure 7E). We also assayed the allelic impact of the prioritized G-to-C point mutation on transcriptional activity in a luciferase reporter plasmid in i-ECs by cloning the mutated and wild-type 500-bp sequence of the *JARID2* cRE and measuring their luciferase signal. We observed a significant (1.3-fold, p value <0.0004, two-sided t test) decrease in the mutant transcriptional reporter activity compared to the isogenic control promoter, further confirming transcriptional disruption by the point mutation (Figure 7F). In principle, the impact of such non-coding mutations on the expression of critical transcription factors could cause significant downstream cascades of transcriptional dysregulation that in turn affect cellular phenotypes leading to CHD. Our analysis framework thus provides a promising avenue to prioritize putative causal, *de novo* non-coding CHD mutations, their putative target TF binding sites, and genes

(C) Enrichment of mutations in CHD cases vs. controls prioritized using different methods. Mutations prioritized by BPNet models trained on arterial endothelial (aEC) scATAC-seq profiles are enriched in cases vs. controls (OR = 1.707, p value = 0.008, Fisher exact test).

(D) Enrichment of mutations prioritized by aEC BPNet model in cases vs. controls (gray bar) compared to enrichment of mutations prioritized by aEC BPNet model proximal to CHD associated genes (blue bar) in cases vs. controls. (***p value = 0.008, Fisher Exact test, *p value = 0.0486, Fisher Exact test)

(E–G) Case studies of three prioritized *de novo* CHD mutations in endothelial cREs in the (E) *FOLH1*, (F) *PIP5K1C*, and (G) *JARID2* gene loci, respectively. Topmost panel shows contribution scores derived from cell-type-specific BPNet models (aEC for (E and F) and Cap for (G)) of each nucleotide in a 100 bp sequence window containing each allele of the mutation. The changes in contribution scores highlight disruption of active TF motifs (ELK/ETV motifs for (E and F) and SOX motif for (G)). The panel below shows corresponding predicted base-resolution scATAC-seq count profiles in a 1 kb window containing reference (blue) and alternate (red) allele of the mutation (the red tracks for the alternate alleles are inverted along the x axis). These tracks highlight local disruption of predicted scATAC-seq profiles by the mutations. The last panel shows observed cell-type resolved pseudobulk scATAC-seq coverage profiles for all cell types at each locus. Scale of tracks is 2.0–6.0 (*FOLH1*), 2.0–20 (*PIP5K1C*), and 2.0–10.0 (*JARID2*) in units of Tn5 insertion counts observed in each cell type.
See also Figure S6.

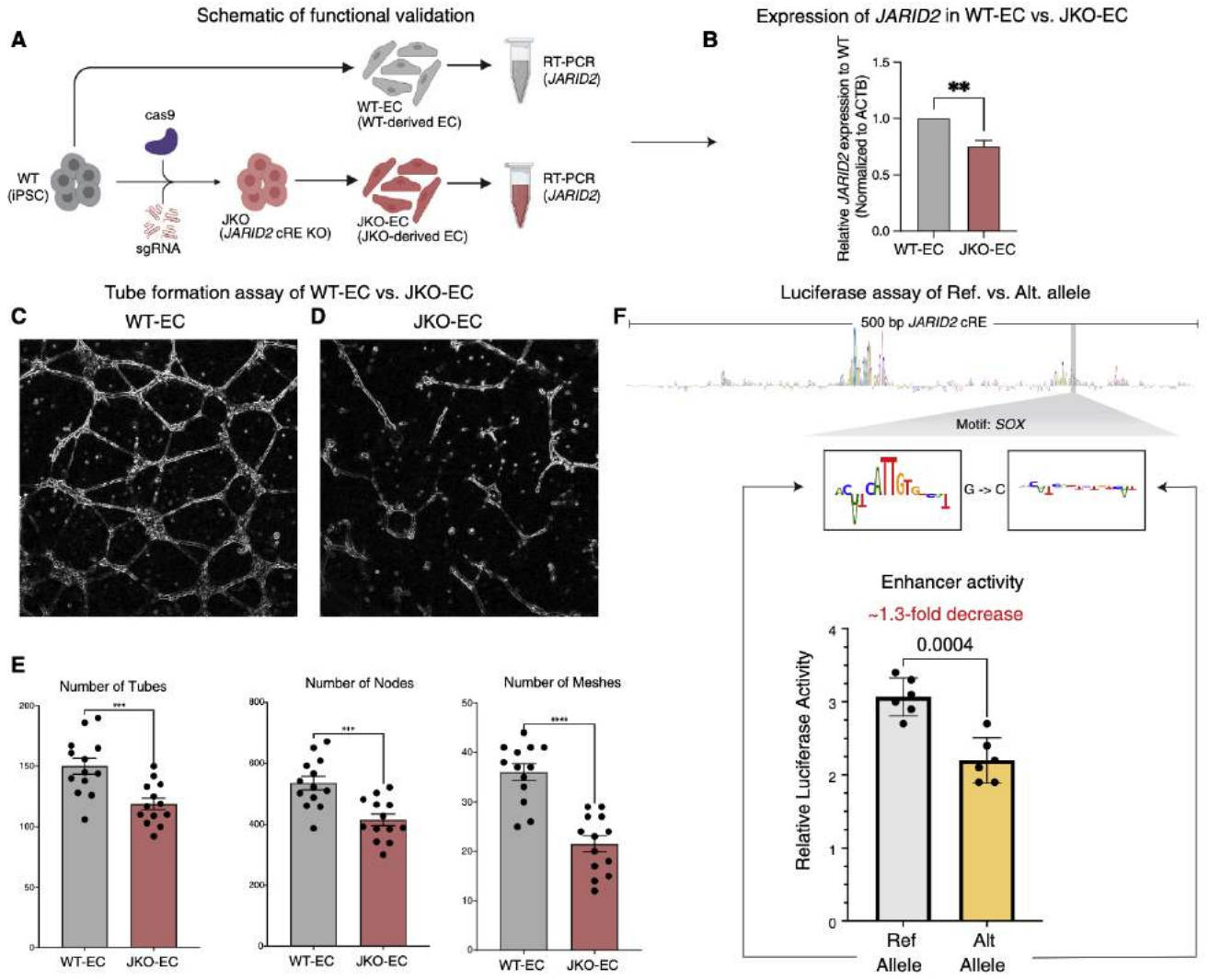


Figure 7. Functional validation of the prioritized mutation in *JARID2* cRE

(A) Schematic of *in vitro* differentiation of iPSCs to EC lineage, and comparison of *JARID2* expression in iPSC-derived EC with and without CRISPR-Cas9 deletion of cRE containing prioritized CHD mutation shown in Figure 6G.

(B) CRISPR-Cas9 deletion of cRE containing the prioritized CHD mutation from Figure 6G shows significant decrease (**$p < 0.001$, two-sided t test) in expression of *JARID2* gene expression in knockout vs. wild-type iPSC-derived ECs.

(C) Tube formation assay of wild-type EC.

(D) Tube formation assay of *JARID2* enhancer knockout cells. KO EC have severe depletion of tubes in the angiogenesis assay, as assessed by quantification of number of tubes, nodes, and meshes.

(E) (left to right). Comparison of number of tubes, nodes, and meshes of tube formation between WT-EC (gray) and enhancer knockout (red) (***p value = 0.0004, two sided t test, ****p value < 0.0001, two sided t test).

(F) Luciferase reporter activity of wild type and mutant variants for *JARID2* cRE (G-to-C) mutation in iPSC derived endothelial cells. The mutant construct shows a significant decrease (p value < 0.0004, two-sided t test) in luciferase activity of the construct with the prioritized mutation vs. wild-type sequence in iPSC-derived ECs. Data are shown as mean +/- SEM for all error bars in this figure.

See also Figure S6.

as well as the relevant cell types within the developmental window.

DISCUSSION

In this study, we present a resource elucidating regulatory dynamics of human cardiogenesis at single-cell resolution. By generating scATAC-seq experiments in fetal hearts at early and mid-gestational developmental timepoints, we reveal the coordinated landscapes of dynamic cREs and genes that define major cell types, lineages, and differentiation trajectories in the developing human heart. By training and interpreting deep learning models, we were able to decipher the cell-type-specific sequence syntax of active TF binding sites. By coupling these

dynamic TF motif activity maps with TF expression across the cell types, we defined putative *trans*-factors that bind to TF motifs encoded in specific cREs and orchestrate dynamic gene regulatory programs that define differentiation trajectories of the major cardiac cell types.

We identify several previously characterized TFs in mice that are important for cell fate determination of the terminally differentiated cell types. For example, we identified *SOX17* to be a TF with predicted dynamic binding in the late capillary (Figure S4K) and mid venous (Figure 3M) differentiation trajectory in open chromatin peaks near *APLNR* (Figure 3P). Consistent with these findings, *Sox17* knockout in mice have been shown to retard the differentiation of endocardial cells due to the downregulation of the NOTCH signaling pathway and promote defective heart development.[55] We also nominate putative regulatory TFs. For example, we observe *SOX18* expression and chromatin activity in the mid to late temporal regulation of arterial endothelial cells. This activity pattern is consistent with other data implicating this factor, along with *SOX17*, in regulating vascular endothelium development in mouse retina[56] (Figure S4N) and controlling the expression of *MEOX2*[57] and *CLDN5*—downstream master regulators of arterial development[58] (Figure S4K). We also identify other TFs that exhibit strong chromatin activity changes along developmental lineage trajectories (Figures 3F, 3M, and S4), implicating these factors as potentially important for lineage specification.

We observed that the EMT program drives substantial differences *in vitro* compared *to in vivo* epicardial-derived lineages. Based on this observation, we successfully optimized the differentiation protocol for iPSC-derived epicardial cells to diminish EMT, which resulted in *in vitro* differentiated epicardial cells with substantially greater epigenomic similarity to their *in vivo* counterparts. This case study serves as proof of principle that single cell molecular "benchmarking" against *in vivo* derived data can serve as a useful computational tool for optimizing *in vitro* differentiation protocols.

Finally, by using the deep learning models, we predict the impact of *de novo* non-coding mutations on cell-type-specific chromatin accessibility profiles and infer the active TF binding sites disrupted by high impact mutations. We also identify ranking of cell types whose cREs are enriched for prioritized CHD mutations. Our CRISPR/Cas9 luciferase and angiogenesis experiments in iECs showed the impact of ablating an endothelial lineage-specific enhancer harboring a predicted high impact *de novo* CHD mutation related to *JARID2*, a key CHD gene. These data provide evidence that prioritized cRE mutations likely impact enhancers with critical developmental functions that are relevant for CHD. Importantly, we show that overlapping mutations with cell-type-resolved cRE maps of fetal heart cell types is not sufficient to enrich CHD mutations over controls unless augmented by mutation impact scores from our cell-type-specific deep learning models, highlighting the utility of the single cell atlas and basepair neural network models.

Limitations of the study

While most developmental trajectories exhibited no substantial "gaps" in cell density, obtaining samples both earlier and later in development might allow us to more fully populate the extremes of these trajectories, extending our understanding of these developmental paradigms. Second, our analysis of regulatory landscapes has largely focused on activators, and not on repressors that are more challenging to nominate using correlation-based analysis. Third, we restrict our prioritization of *de novo* CHD mutations to those that fall in the immediate vicinity of observed scATAC-seq peaks in our fetal heart atlas and are likely to disrupt and decrease accessibility. While this strategy reduces the likelihood of false positives, it does bias our prioritization against mutations that might result in gain of accessibility. The reduced sensitivity of peak identification from scATAC-seq profiles in some rare cell types (e.g., neural crest cells) with sparse coverage may also result in a greater false negative rate and reduced enrichments for these cell types. Our study is restricted to point mutations and our chromatin-centric approach cannot predict functional impact of non-coding mutations via other key regulatory mechanisms (e.g., splicing, structural variants). The modest number of CHD cases that are confidently explained by our prioritization framework may be due to some of these limitations. Finally, while we have directly validated the impact of one candidate enhancer harboring a specific *de novo* CHD mutation toward expression of its predicted target gene and on downstream angiogenesis-related phenotypes, more extensive computational and experimental validation of the gene expression and phenotypic impact of prioritized mutations would further dissect the validation rate of the model.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Patient recruitment
- METHOD DETAILS
  - Experimental methods
  - Epicardial cell differentiation (old protocol)
  - Cardiac fibroblast differentiation
  - Computational methods
- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2022.11.028.

ACKNOWLEDGMENTS

We thank members of the Kundaje, Greenleaf, Quertermous, Wang, Wu, Engreitz and Karakikes labs for discussion and advice, especially J. Granja, R. Ma, and G. Marinov. All schematics were created with BioRender. Sequencing of scATAC-seq libraries was performed by the Stanford Functional Genomics Facility (supported by NIH grants S10OD025212 and 1S10OD021763).

Funding: This work was supported by grants from the NIH 1DP2GM123485 (A.K.), U01HG012069 (A.K.), R01 HL139478 (T.Q.), R01 HL145708 (T.Q.), R01 HL134817 (T.Q.), R01 HL151535 (T.Q.), R01 HL156846 (T.Q.), 1UM1 HG011972 (T.Q.), RM1-HG007735, (W.J.G.) UM1-HG009442 (W.J.G.), UM1-HG009436 (W.J.G.), R01- HG00990901 (W.J.G.), and U19- AI057266 (W.J.G.), R01 GM136737 (K.C.W.), R61 AR076815 (K.C.W.), a Human Cell Atlas grant from the Chan Zuckerberg Foundation (T.Q.), NIH R01 HL139679 and R01HL150414 (I.K); Stanford Maternal & Child Health Research Institute (I.K.) K08 HL119251 (K.D.W.), K99 HL135258 (M.G.); S10 OD018220 (Stanford Functional Genomics), NHGRI Genomic Innovator Award (R35HG011324 to J.M.E.); Gordon and Betty Moore and the BASE Research Initiative at the Lucile Packard Children's Hospital at Stanford University (J.M.E.); and the Stanford Maternal & Child Health Research Institute and Additional Ventures (to J.M.E.), NSF Graduate Research Fellowship Program (M.A.) and The Bio-X Bowes Fellowship (L.S.). K.C.W. is a New York Stem Cell Foundation–Robertson Investigator, and the Stephen Bechtel Endowed Faculty Scholar in Pediatric Translational Medicine, Stanford Maternal and Child Health Research Institute. This work was also supported by funding from the Rita Allen Foundation (W.J.G.), the Human Frontiers Science (RGY006S) (W.J.G.). W.J.G. is a Chan Zuckerberg Biohub investigator and acknowledges grants 2017-174468 and 2018-182817 from the Chan Zuckerberg Initiative and funding from Emerson Collective.

AUTHOR CONTRIBUTIONS

M.A., L.S., I.K., K.C.W. and A.K. conceived the project. L.S., M.A., T.Q, W.J.G and A.K. generated figures and wrote the manuscript with input from authors. M.A. designed and performed all experimental data generation for the manuscript with inputs from L.S., M.C., K.D.W, M.G., I.K., K.C.W., T.Q., A.K., and W.J.G. L.S designed and performed all computational analyses for the manuscript with inputs from M.A., A. Banerjee, S.K., S.N., A.S., A.V., N.V., A. Balsubramani, J.M.E., K.F., T.Q., W.J.G., and A.K.

DECLARATION OF INTERESTS

W.J.G. is named as an inventor on patents describing ATAC-seq methods. 10X Genomics has licensed intellectual property on which W.J.G. is listed as an inventor. W.J.G. holds options in 10X Genomics and is a consultant for Ultima Genomics and Guardant Health. W.J.G. is a scientific co-founder of Protillion Biosciences. A.S. is an employee of Insitro and is a consultant at Myokardia. A.K. is a consulting Fellow with Illumina, a member of the SAB of OpenTargets (GSK), PatchBio, SerImmune and a scientific co-founder of RavelBio. M.A., L.S., A. Banerjee, and K.F. are employees of Illumina. J.C.W. is a co-founder of Khloris Biosciences but has no competing interests, as the work presented here is completely independent. The other authors declare no competing interests.

Received: January 15, 2022
Revised: September 13, 2022
Accepted: November 23, 2022
Published: December 22, 2022

REFERENCES

1. Sylva, M., van den Hoff, M.J.B., and Moorman, A.F.M. (2014). Development of the human heart. Am. J. Med. Genet. *164A*, 1347–1371.
2. Meilhac, S.M., and Buckingham, M.E. (2018). The deployment of cell lineages that form the mammalian heart. Nat. Rev. Cardiol. *15*, 705–724.
3. Srivastava, D. (2006). Making or breaking the heart: from lineage determination to morphogenesis. Cell *126*, 1037–1048.
4. Suryawanshi, H., Clancy, R., Morozov, P., Halushka, M.K., Buyon, J.P., and Tuschl, T. (2020). Cell atlas of the foetal human heart and implications for autoimmune-mediated congenital heart block. Cardiovasc. Res. *116*, 1446–1457.
5. Asp, M., Giacomello, S., Larsson, L., Wu, C., Fürth, D., Qian, X., Wärdell, E., Custodio, J., Reimegård, J., Salmén, F., et al. (2019). A Spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. Cell *179*, 1647–1660.e19.
6. Miao, Y., Tian, L., Martin, M., Paige, S.L., Galdos, F.X., Li, J., Klein, A., Zhang, H., Ma, N., Wei, Y., et al. (2020). Intrinsic endocardial defects contribute to hypoplastic left heart syndrome. Cell Stem Cell *27*, 574–589.e8.
7. van der Linde, D., Konings, E.E.M., Slager, M.A., Witsenburg, M., Helbing, W.A., Takkenberg, J.J.M., and Roos-Hesselink, J.W. (2011). Birth prevalence of congenital heart disease worldwide: a systematic review and meta-analysis. J. Am. Coll. Cardiol. *58*, 2241–2247.
8. Zaidi, S., Choi, M., Wakimoto, H., Ma, L., Jiang, J., Overton, J.D., Romano-Adesman, A., Bjornson, R.D., Breitbart, R.E., Brown, K.K., et al. (2013). De novo mutations in histone-modifying genes in congenital heart disease. Nature *498*, 220–223.
9. Homsy, J., Zaidi, S., Shen, Y., Ware, J.S., Samocha, K.E., Karczewski, K.J., DePalma, S.R., McKean, D., Wakimoto, H., Gorham, J., et al. (2015). De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. Science *350*, 1262–1266.
10. Pediatric Cardiac Genomics Consortium, Gelb, B., Brueckner, M., Chung, W., Goldmuntz, E., Kaltman, J., Kaski, J.P., Kim, R., Kline, J., Mercer-Rosa, L., et al. (2013). The congenital heart disease genetic network study: rationale, design, and early results. Circ. Res. *112*, 698–706.
11. Jin, S.C., Homsy, J., Zaidi, S., Lu, Q., Morton, S., DePalma, S.R., Zeng, X., Qi, H., Chang, W., Sierant, M.C., et al. (2017). Contribution of rare inherited and de novo variants in 2, 871 congenital heart disease probands. Nat. Genet. *49*, 1593–1601.
12. Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., and Zeitlinger, J. (2021). Base-resolution models of transcription-factor binding reveal soft motif syntax. Nat. Genet. *53*, 354–366.
13. Trevino, A.E., Müller, F., Andersen, J., Sundaram, L., Kathiria, A., Shcherbina, A., Farh, K., Chang, H.Y., Pașca, A.M., Kundaje, A., et al. (2021). Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. Cell *184*, 5053–5069.e23. https://doi.org/10.1016/j.cell.2021.07.039.
14. Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. (2019). Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. Cell *176*, 1517.
15. Richter, F., Morton, S.U., Kim, S.W., Kitaygorodsky, A., Wasson, L.K., Chen, K.M., Zhou, J., Qi, H., Patel, N., DePalma, S.R., et al. (2020). Genomic analyses implicate noncoding de novo variants in congenital heart disease. Nat. Genet. *52*, 769–777.
16. Satpathy, A.T., Granja, J.M., Yost, K.E., Qi, Y., Meschi, F., McDermott, G.P., Olsen, B.N., Mumbach, M.R., Pierce, S.E., Corces, M.R., et al. (2019). Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nat. Biotechnol. *37*, 925–936. https://doi.org/10.1038/s41587-019-0206-z.
17. Granja, J.M., Corces, M.R., Pierce, S.E., Bagdatli, S.T., Choudhry, H., Chang, H.Y., and Greenleaf, W.J. (2021). ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nat. Genet. *53*, 403–411.
18. Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol. *37*, 38–44. https://doi.org/10.1038/nbt.4314.
19. McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: uniform manifold approximation and projection. J. Open Source Softw. *3*, 861. https://doi.org/10.21105/joss.00861.
20. Traag, V.A., Waltman, L., and van Eck, N.J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. *9*, 5233.

21. Cui, Y., Zheng, Y., Liu, X., Yan, L., Fan, X., Yong, J., Hu, Y., Dong, J., Li, Q., Wu, X., et al. (2019). Single-cell transcriptome analysis maps the developmental track of the human heart. Cell Rep. *26*, 1934–1950.e5.

22. Cusanovich, D.A., Hill, A.J., Aghamirzaie, D., Daza, R.M., Pliner, H.A., Berletch, J.B., Filippova, G.N., Huang, X., Christiansen, L., DeWitt, W.S., et al. (2018). A single-cell atlas of in vivo mammalian chromatin accessibility. Cell *174*, 1309–1324.e18.

23. Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences *70*, 3145–3153.

24. Lundberg, S.M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems, I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds. (Curran Associates, Inc.), pp. 4765–4774.

25. Sehnert, A.J., Huq, A., Weinstein, B.M., Walker, C., Fishman, M., and Stainier, D.Y.R. (2002). Cardiac troponin T is essential in sarcomere assembly and cardiac contractility. Nat. Genet. *31*, 106–110.

26. Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell *158*, 1431–1443.

27. Schep, A.N., Wu, B., Buenrostro, J.D., and Greenleaf, W.J. (2017). chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat. Methods *14*, 975–978.

28. Davis, C.A., Haberland, M., Arnold, M.A., Sutherland, L.B., McDonald, O.G., Richardson, J.A., Childs, G., Harris, S., Owens, G.K., and Olson, E.N. (2006). PRISM/PRDM6, a transcriptional repressor that promotes the proliferative gene program in smooth muscle cells. Mol. Cell Biol. *26*, 2626–2636.

29. Hellström, M., Kalén, M., Lindahl, P., Abramsson, A., and Betsholtz, C. (1999). Role of PDGF-B and PDGFR-beta in recruitment of vascular smooth muscle cells and pericytes during embryonic blood vessel formation in the mouse. Development *126*, 3047–3055.

30. Levéen, P., Pekny, M., Gebre-Medhin, S., Swolin, B., Larsson, E., and Betsholtz, C. (1994). Mice deficient for PDGF B show renal, cardiovascular, and hematological abnormalities. Genes Dev. *8*, 1875–1887.

31. Burridge, P.W., Matsa, E., Shukla, P., Lin, Z.C., Churko, J.M., Ebert, A.D., Lan, F., Diecke, S., Huber, B., Mordwinkin, N.M., et al. (2014). Chemically defined generation of human cardiomyocytes. Nat. Methods *11*, 855–860.

32. Lian, X., Hsiao, C., Wilson, G., Zhu, K., Hazeltine, L.B., Azarin, S.M., Raval, K.K., Zhang, J., Kamp, T.J., and Palecek, S.P. (2012). Robust cardiomyocyte differentiation from human pluripotent stem cells via temporal modulation of canonical Wnt signaling. Proc. Natl. Acad. Sci. USA *109*, E1848–E1857.

33. Cheung, C., Bernardo, A.S., Trotter, M.W.B., Pedersen, R.A., and Sinha, S. (2012). Generation of human vascular smooth muscle subtypes provides insight into embryological origin–dependent disease susceptibility. Nat. Biotechnol. *30*, 165–173.

34. Zhang, H., Tian, L., Shen, M., Tu, C., Wu, H., Gu, M., Paik, D.T., and Wu, J.C. (2019). Generation of Quiescent Cardiac Fibroblasts From Human Induced Pluripotent Stem Cells for In Vitro Modeling of Cardiac Fibrosis. Circ. Res. *125*, 552–566.

35. Paik, D.T., Tian, L., Lee, J., Sayed, N., Chen, I.Y., Rhee, S., Rhee, J.-W., Kim, Y., Wirka, R.C., Buikema, J.W., et al. (2018). Large-Scale Single-Cell RNA-Seq Reveals Molecular Signatures of Heterogeneous Populations of Human Induced Pluripotent Stem Cell-Derived Endothelial Cells. Circ. Res. *123*, 443–450.

36. Friedman, C.E., Nguyen, Q., Lukowski, S.W., Helfer, A., Chiu, H.S., Miklas, J., Levy, S., Suo, S., Han, J.-D.J., Osteil, P., et al. (2018). Single-cell transcriptomic analysis of cardiac differentiation from human PSCs reveals HOPX-dependent cardiomyocyte maturation. Cell Stem Cell *23*, 586–598.e8.

37. Churko, J.M., Garg, P., Treutlein, B., Venkatasubramanian, M., Wu, H., Lee, J., Wessells, Q.N., Chen, S.-Y., Chen, W.-Y., Chetal, K., et al. (2018). Defining human cardiac transcription factor hierarchies using integrated single-cell heterogeneity analysis. Nat. Commun. *9*, 4906.

38. Granja, J.M., Klemm, S., McGinnis, L.M., Kathiria, A.S., Mezger, A., Corces, M.R., Parks, B., Gars, E., Liedtke, M., Zheng, G.X.Y., et al. (2019). Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. Nat. Biotechnol. *37*, 1458–1465.

39. Vincentz, J.W., Barnes, R.M., Firulli, B.A., Conway, S.J., and Firulli, A.B. (2008). Cooperative interaction of Nkx2.5 and Mef2c transcription factors during heart development. Dev. Dyn. *237*, 3809–3819.

40. Quijada, P., Trembley, M.A., and Small, E.M. (2020). The Role of the Epicardium During Heart Development and Repair. Circ. Res. *126*, 377–394.

41. von Gise, A., and Pu, W.T. (2012). Endocardial and epicardial epithelial to mesenchymal transitions in heart development and disease. Circ. Res. *110*, 1628–1645.

42. Risebro, C.A., Vieira, J.M., and Riley, P.R. (2015). Characterisation of the human embryonic and foetal epicardium during heart development. Development *142*, 3630–3636. https://doi.org/10.1242/dev.127621.

43. Lamouille, S., Xu, J., and Derynck, R. (2014). Molecular mechanisms of epithelial–mesenchymal transition. Nat. Rev. Mol. Cell Biol. *15*, 178–196.

44. Eppig, J.T., Richardson, J.E., Kadin, J.A., Ringwald, M., Blake, J.A., and Bult, C.J. (2015). Mouse Genome Informatics (MGI): reflecting on 25 years. Mamm. Genome *26*, 272–284.

45. Conway, R.E., Petrovic, N., Li, Z., Heston, W., Wu, D., and Shapiro, L.H. (2006). Prostate-specific membrane antigen regulates angiogenesis by modulating integrin signal transduction. Mol. Cell Biol. *26*, 5310–5324.

46. Wang, Y., Lian, L., Golden, J.A., Morrisey, E.E., and Abrams, C.S. (2007). PIP5KI gamma is required for cardiovascular and neuronal development. Proc. Natl. Acad. Sci. USA *104*, 11748–11753.

47. Zhang, Y., Chen, W., Zeng, W., Lu, Z., and Zhou, X. (2020). Biallelic loss of function NEK3 mutations deacetylate α-tubulin and downregulate NUP205 that predispose individuals to cilia-related abnormal cardiac left-right patterning. Cell Death Dis. *11*, 1005.

48. Fedak, P.W.M., Smookler, D.S., Kassiri, Z., Ohno, N., Leco, K.J., Verma, S., Mickle, D.A.G., Watson, K.L., Hojilla, C.V., Cruz, W., et al. (2004). TIMP-3 deficiency leads to dilated cardiomyopathy. Circulation *110*, 2401–2409.

49. Kawamoto, H., Yasuda, O., Suzuki, T., Ozaki, T., Yotsui, T., Higuchi, M., Rakugi, H., Fukuo, K., Ogihara, T., and Maeda, N. (2006). Tissue inhibitor of metalloproteinase-3 plays important roles in the kidney following unilateral ureteral obstruction. Hypertens. Res. *29*, 285–294.

50. Mesbah, K., Harrelson, Z., Théveniau-Ruissy, M., Papaioannou, V.E., and Kelly, R.G. (2008). Tbx3 is required for outflow tract development. Circ. Res. *103*, 743–750.

51. Mysliwiec, M.R., Bresnick, E.H., and Lee, Y. (2011). Endothelial Jarid2/Jumonji is required for normal cardiac development and proper Notch1 expression. J. Biol. Chem. *286*, 17193–17204.

52. Cho, E., Mysliwiec, M.R., Carlson, C.D., Ansari, A., Schwartz, R.J., and Lee, Y. (2018). Cardiac-specific developmental and epigenetic functions of Jarid2 during embryonic development. J. Biol. Chem. *293*, 11659–11673.

53. Lee, Y., Song, A.J., Baker, R., Micales, B., Conway, S.J., and Lyons, G.E. (2000). Jumonji, a nuclear protein that is necessary for normal heart development. Circ. Res. *86*, 932–938.

54. Barth, J.L., Clark, C.D., Fresco, V.M., Knoll, E.P., Lee, B., Argraves, W.S., and Lee, K.-H. (2010). Jarid2 is among a set of genes differentially regulated by Nkx2.5 during outflow tract morphogenesis. Dev. Dyn. *239*, 2024–2033.

55. Saba, R., Kitajima, K., Rainbow, L., Engert, S., Uemura, M., Ishida, H., Kokkinopoulos, I., Shintani, Y., Miyagawa, S., Kanai, Y., et al. (2019). Endocardium differentiation through Sox17 expression in endocardium precursor cells regulates heart development in mice. Sci. Rep. *9*, 11953.

56. Zhou, Y., Williams, J., Smallwood, P.M., and Nathans, J. (2015). Sox7, Sox17, and Sox18 cooperatively regulate vascular development in the mouse retina. PLoS One *10*, e0143650.

57. Douville, J.M., Cheung, D.Y.C., Herbert, K.L., Moffatt, T., and Wigle, J.T. (2011). Mechanisms of MEOX1 and MEOX2 regulation of the cyclin dependent kinase inhibitors p21 and p16 in vascular endothelial cells. PLoS One *6*, e29099.

58. Fontijn, R.D., Volger, O.L., Fledderus, J.O., Reijerkerk, A., de Vries, H.E., and Horrevoets, A.J.G. (2008). SOX-18 controls endothelial-specific claudin-5 gene expression and barrier function. Am. J. Physiol. Heart Circ. Physiol. *294*, H891–H900.

59. Feyen, D.A.M., Perea-Gil, I., Maas, R.G.C., Harakalova, M., Gavidia, A.A., Arthur Ataam, J., Wu, T.-H., Vink, A., Pei, J., Vadgama, N., et al. (2021). Unfolded protein response as a compensatory mechanism and potential therapeutic target in PLN R14del cardiomyopathy. Circulation *144*, 382–392.

60. Bao, X., Lian, X., Qian, T., Bhute, V.J., Han, T., and Palecek, S.P. (2017). Directed differentiation and long-term maintenance of epicardial cells derived from human pluripotent stem cells under fully defined conditions. Nat. Protoc. *12*, 1890–1900.

61. Mikawa, T., and Gourdie, R.G. (1996). Pericardial mesoderm generates a population of coronary smooth muscle cells migrating into the heart along with ingrowth of the epicardial organ. Dev. Biol. *174*, 221–232.

62. Cai, C.-L., Martin, J.C., Sun, Y., Cui, L., Wang, L., Ouyang, K., Yang, L., Bu, L., Liang, X., Zhang, X., et al. (2008). A myocardial lineage derives from Tbx18 epicardial cells. Nature *454*, 104–108.

63. Muhl, L., Genové, G., Leptidis, S., Liu, J., He, L., Mocci, G., Sun, Y., Gustafsson, S., Buyandelger, B., Chivukula, I.V., et al. (2020). Single-cell analysis uncovers fibroblast heterogeneity and criteria for fibroblast and mural cell identification and discrimination. Nat. Commun. *11*, 4493.

64. Dobnikar, L., Taylor, A.L., Chappell, J., Oldach, P., Harman, J.L., Oerton, E., Dzierzak, E., Bennett, M.R., Spivakov, M., and Jørgensen, H.F. (2018). Disease-relevant transcriptional signatures identified in individual smooth muscle cells from healthy mouse vessels. Nat. Commun. *9*, 5401.

65. Pham, T.T.D., Park, S., Kolluri, K., Kawaguchi, R., Wang, L., Tran, D., Zhao, P., Carmichael, S.T., and Ardehali, R. (2021). Heart and brain pericytes exhibit a pro-fibrotic response after vascular injury. Circ. Res. *129*, e141–e143.

66. Wang, W.-D., Melville, D.B., Montero-Balaguer, M., Hatzopoulos, A.K., and Knapik, E.W. (2011). Tfap2a and Foxd3 regulate early steps in the development of the neural crest progenitor population. Dev. Biol. *360*, 173–185.

67. Aird, W.C. (2007). Phenotypic heterogeneity of the endothelium: I. Structure, function, and mechanisms. Circ. Res. *100*, 158–173.

68. Kalucka, J., de Rooij, L.P.M.H., Goveia, J., Rohlenova, K., Dumas, S.J., Meta, E., Conchinha, N.V., Taverna, F., Teuwen, L.-A., Veys, K., et al. (2020). Single-cell transcriptome atlas of murine endothelial cells. Cell *180*, 764–779.e20.

69. Vodyanik, M.A., Yu, J., Zhang, X., Tian, S., Stewart, R., Thomson, J.A., and Slukvin, I.I. (2010). A mesoderm-derived precursor for mesenchymal stem and endothelial cells. Cell Stem Cell *7*, 718–729.

70. Podgrabinska, S., Braun, P., Velasco, P., Kloos, B., Pepper, M.S., and Skobe, M. (2002). Molecular characterization of lymphatic endothelial cells. Proc. Natl. Acad. Sci. USA *99*, 16069–16074.

71. Acharya, A., Baek, S.T., Huang, G., Eskiocak, B., Goetsch, S., Sung, C.Y., Banfi, S., Sauer, M.F., Olsen, G.S., Duffield, J.S., et al. (2012). The bHLH transcription factor Tcf21 is required for lineage-specific EMT of cardiac fibroblast progenitors. Development *139*, 2139–2149.

72. Nurnberg, S.T., Cheng, K., Raiesdana, A., Kundu, R., Miller, C.L., Kim, J.B., Arora, K., Carcamo-Oribe, I., Xiong, Y., Tellakula, N., et al. (2015). Coronary artery disease associated transcription factor TCF21 regulates smooth muscle precursor cells that contribute to the fibrous cap. PLoS Genet. *11*, e1005155.

73. Wirka, R.C., Wagh, D., Paik, D.T., Pjanic, M., Nguyen, T., Miller, C.L., Kundu, R., Nagao, M., Coller, J., Koyano, T.K., et al. (2019). Atheroprotective roles of smooth muscle cell phenotypic modulation and the TCF21 disease gene as revealed by single-cell analysis. Nat. Med. *25*, 1280–1289.

74. Ma, S., Zhang, B., LaFave, L.M., Earl, A.S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V.K., Tay, T., et al. (2020). Chromatin potential identified by shared single-cell profiling of RNA and chromatin. Cell *183*, 1103–1116.e20.

75. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

76. Corces, M.R., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W., Silva, T.C., Groeneveld, C., Wong, C.K., Cho, S.W., et al. (2018). The chromatin accessibility landscape of primary human cancers. Science *362*, eaav1898. https://doi.org/10.1126/science.aav1898.

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

78. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. Cell *177*, 1888–1902.e21.

79. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-R., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods *16*, 1289–1296.

80. Wolf, K., Hu, H., Isaji, T., and Dardik, A. (2019). Molecular identity of arteries, veins, and lymphatics. J. Vasc. Surg. *69*, 253–262.

81. Ng, S.Y., Wong, C.K., and Tsang, S.Y. (2010). Differential gene expressions in atrial and ventricular myocytes: insights into the road of applying embryonic stem cell-derived cardiomyocytes for future therapies. Am. J. Physiol. Cell Physiol. *299*, C1234–C1249.

82. Bentsen, M., Goymann, P., Schultheis, H., Klee, K., Petrova, A., Wiegandt, R., Fust, A., Preussner, J., Kuenne, C., Braun, T., et al. (2020). ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. Nat. Commun. *11*, 4267.

83. Lundberg, S.M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st international conference on neural information processing systems, pp. 4768–4777.

84. Korhonen, J., Martinmäki, P., Pizzi, C., Rastas, P., and Ukkonen, E. (2009). MOODS: fast search for position weight matrix matches in DNA sequences. Bioinformatics *25*, 3181–3182.

85. Sharma, B., Ho, L., Ford, G.H., Chen, H.I., Goldstone, A.B., Woo, Y.J., Quertermous, T., Reversade, B., and Red-Horse, K. (2017). Alternative progenitor cells compensate to rebuild the coronary vasculature in elabela- and Apj-deficient hearts. Dev. Cell *42*, 655–666.e3.

86. Kang, Y., Kim, J., Anderson, J.P., Wu, J., Gleim, S.R., Kundu, R.K., McLean, D.L., Kim, J.-D., Park, H., Jin, S.-W., et al. (2013). Apelin-APJ signaling is a critical regulator of endothelial MEF2 activation in cardiovascular development. Circ. Res. *113*, 22–31.

87. Inui, M., Fukui, A., Ito, Y., and Asashima, M. (2006). Xapelin and Xmsr are required for cardiovascular development in Xenopus laevis. Dev. Biol. *298*, 188–200.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
CD144 (VE-Cadherin) MicroBeads	MiltenyiBiotec	130-097-857
Anti-WT1	Abcam	ab89901/RRID:AB_2043201
Anti-ZO1	Thermo Fisher Scientific	33-9100/RRID:AB_2533147
Bacterial and virus strains		
Escherichia coli DH5a competent cells	Zymo Research	T3007
Biological samples		
Human fetal heart samples	Stanford University	N/A
Chemicals, peptides, and recombinant proteins		
Essential 8 Medium	Gibco	A1517001
RPMI 1640 Medium	Gibco	11875093
RPMI 1640 Medium, minus glucose	Gibco	11879020
DMEM, high glucose	Gibco	11965118
HBSS, calcium, magnesium, no phenol red	Gibco	14025092
TrypLE Select Enzyme (10X)	Gibco	A1217703
KnockOut Serum Replacement	Gibco	10828028
Advanced DMEM/F-12	Gibco	12634028
Ham's F-12 Nutrient Mix	Gibco	11765-054
IMDM	Gibco	12440-053
Opti-MEM I Reduced Serum Media	Gibco	11058021
DPBS without calcium and magnesium	Gibco	14190250
Chemically defined lipid concentrate	Gibco	11905-031
Glutamax	Gibco	35050-061
UltraPure 0.5M EDTA, pH 8.0	Invitrogen	15575-020
Retinoic acid	Sigma-Aldrich	R2625
L-Ascorbic acid 2-phosphate sesquimagnesium salt hydrate	MilliporeSigma	A8960
Accutase solution	MilliporeSigma	A6964
Gelatin solution, Type B	Sigma-Aldrich	G1393
Liberase TM	Sigma-Aldrich	5401127001
DNase I	Worthington	LK003172
Matrigel Basement Membrane Matrix	Corning	354234
Y-27632 2HCl (ROCK Inhibitor)	Selleck Chemicals	S1049
CHIR-99021 (CT99021) HCl 5mg	Selleck Chemicals	S2924
IWR-1	Selleck Chemicals	S7086
C59	Selleck Chemicals	S7037
LY294002	Selleck Chemicals	S1105
SB431542	Selleck Chemicals	S1067
B-27 Supplement, minus insulin	Thermo Fisher Scientific	A1895601
B-27 Supplement, serum free	Thermo Fisher Scientific	17504044
Recombinant Human FGF-2	PeproTech	100-18B
Human BMP4	PeproTech	120-05ET
Recombinant Human VEGF	R&D Systems	293-VE-010/CF

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
EGM-2 Endothelial Cell Growth Medium-2 Bullet Kit	Lonza	CC-3162
FGM-2 Fibroblast Growth Medium-2 Bullet Kit	Lonza	CC-3132
TRIzol Reagent	Thermo Fisher Scientific	15596026
MACS BSA Stock Solution	Miltenyi Biotec	130-091-376
Digitonin	Thermo Fisher Scientific	BN2006
Tris-HCl	Invitrogen	15568025
NaCl	Invitrogen	AM9759G
MgCl2	Invitrogen	AM9530G
Tween 20	Sigma-Aldrich	11332465001
NP40	Sigma-Aldrich	11332473001
Activin A	PeproTech	120-14E
PrimeSTAR GXL DNA Polymerase	Takara	R050B
Lipofectamine™ Stem Transfection	Thermo Fisher Scientific	STEM00001
Polyvinyl alcohol	MilliporeSigma	P8136
Transferrin	MilliporeSigma	T8158
Monothioglycerol	MilliporeSigma	M6145
Dimethyl sulfoxide	MilliporeSigma	D2650
SpCas9 2NLS Nuclease	Synthego	N/A
Gelatin solution	Sigma-Aldrich	G1393
Critical commercial assays		
Chromium Next GEM Single Cell ATAC Reagent Kits v1.1	10X Genomics	1000175
CytoTuneTM-iPS 2.0 Sendai Reprogramming Kit	Thermo Fisher Scientific	A16517
Direct-zol RNA Micro-Prep	Zymo Research	R2053
iQ SYBR Green Supermix	Bio-Rad	1708882
iScript cDNA Synthesis Kit	Bio-Rad	170-8891
Dual-Glo® Luciferase Assay System	Promega	E2920
Deposited data		
Data files for scATAC-seq	NCBI GEO	GEO: GSE181346
Experimental models: Cell lines		
Human iPSC	SCVI Biobank	SCVI274
Oligonucleotides		
Human ACTB Primers	IDT	Hs.PT.39a.22214847
Human JARID2 Primers	IDT	Hs.PT.58.20087641
Recombinant DNA		
pGL3-Promoter	Promega	E1761
pGL3-Control	Promega	E1741
pRL-CMV	Promega	E2261
Software and algorithms		
ImageJ	NIH	https://imagej-nih-gov.stanford.idm.oclc.org/ij/
Genome assembly	https://www.ncbi.nlm.nih.gov/grc/human	hg38/GRCh38
Cell Ranger	10x Genomics	CellRanger v3.1.0
Cell Ranger-ATAC	10x Genomics	Cell Ranger-ATAC v1.2.0
Seurat	https://satijalab.org/seurat/	Seurat v.3.1.4
MACS2	Zhang et al., 2008	MACS2 v2.1.1
ChromVAR	Schep et al., 2017	ChromVAR v.1.6

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
GraphPad Prism	GraphPad Software Inc	https://www.graphpad.com/scientific-software/prism/
ArchR	https://www.archrproject.com/	ArchR v0.9.4
KerasAC (BPNet code framework for ATAC-seq profiles)	https://zenodo.org/record/4248179#.Y1CRjHbMJmN	KerasAC v.2.5.1.
DeepLIFT	https://github.com/kundajelab/deeplift	DeepLIFT v0.6.13.0-alpha
Code repository for all analyses	https://github.com/kundajelab/Cardiogenesis_Repo	https://github.com/kundajelab/Cardiogenesis_Repo

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Thomas Quertermous (tomq1@stanford.edu).

Materials availability

This study did not generate unique reagents.

Data and code availability

- Aligned fragment files from single-cell chromatin assays are deposited in the Gene Expression Omnibus database with the Super-Series reference number GSE181346. The cell by gene accessibility scores matrices, along with cluster 5′ insertion bigWig tracks for the human heart samples are deposited to UCSC cell browser portal under reference url https://cardiogenesis-atac.cells.ucsc.edu to enable visualization of cell markers and genes. Reanalyzed scRNA Seurat objects are deposited to https://doi.org/10.5281/zenodo.7063223. The trained BPNet model weights are deposited to https://doi.org/10.5281/zenodo.6789181. Interactive HiGlass browser sessions with cell-type resolved tracks for measured base-resolution scATAC-seq coverage profiles and predicted base-resolution scATAC-seq coverage profiles from BPNet models as well as model-derived nucleotide-resolution contribution scores in peak regions could be found at: https://resgen.io/kundaje-lab/sundaram-2022/views/cardiogenesis.
- Code used for single cell analysis, training BPNet models and results for all figures can be found at: https://github.com/kundajelab/Cardiogenesis_Repo.
- Any additional information required to reanalyze the data and the raw data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Patient recruitment

Human subjects were enrolled in the study with informed consent approved by the Stanford Institutional Review Board and Stem Cell Research Oversight Committee. Human fetal heart tissues (day-42, day-56, and day-133 post-conception) were obtained from de-identified aborted fetuses in collaboration with the Stanford Family Planning Research Team, Department of Obstetrics and Gynecology, Division of Family Planning Services and Research, Stanford University School of Medicine. Human iPSCs were obtained from the Stanford CVI iPSC Biobank.

METHOD DETAILS

Experimental methods

Generation and culture of human induced pluripotent stem cells

Peripheral blood mononuclear cells (PBMCs) were reprogrammed to iPSCs using the CytoTune-iPS 2.0 Sendai Reprogramming Kit (Thermo Fisher Scientific) according to the manufacturer's instructions with modifications as previously described.[59] Stem cell-like colonies were manually picked about two weeks post-transduction and expanded in E8 stem cell media (Life Technologies). All iPSCs used for the subsequent studies were within passages 22 to 30. The genome integrity was assessed by a single nucleotide polymorphism-based karyotyping assay (Illumina, HumanOmniExpress-24 v1.1). The iPSCs were maintained in a defined E8 medium (Life Technologies) on cell culture plates coated with ESC-qualified Matrigel (BD Biosciences) in a hypoxic environment (8% O_2, 5%

CO_2) at 37°C. For routine passaging, iPSCs were dissociated with Gentle Cell Dissociation Reagent (StemCell Technologies) and cultured on with E8 medium supplemented with 5 μM Y-27632 (SelleckChem). The iPSCs were tested to be mycoplasma negative using the Mycoalert Mycoplasma testing kits (LT07-318, Lonza).

Cardiomyocyte differentiation

Cardiomyocytes were differentiated using a monolayer method as previously described.[59] The iPSCs were seeded in 6-wells at a density of 1.2 × 10^5 cells per well and grown for four days prior to differentiation. Differentiation was initiated by replacing the E8 media with RPMI supplemented with B27 without insulin (A1895601, Life Technologies) and 6 μM CHIR-99021 (CT99021, Selleckchem). Two days later the media was replaced with RPMI supplemented with B27 without insulin. Cultures were then treated with 3 μM IWR-1 (I0161, Sigma) in RPMI supplemented with B27 without insulin for two days. The cultures were then maintained in RPMI with B27 with insulin (17504-044, Life Technologies) and glucose starved for three days (using RPMI minus glucose). After glucose starvation, iPSC-CMs were maintained in RPMI with B27. Cells were collected at specific time points during differentiation, day 0 (iPSC), day 2 (i-Mes), day 5 (i-CP), day 15 (i-pCM), and day 30 (i-CM), The cells from three independent differentiation batches for each time point were collected and pooled for scATAC analysis.

Endothelial cell differentiation

The iPSCs were cultured as described above until reaching 80% confluence. The medium was switched to RPMI-B27 without insulin (Life Technologies) with 6 μM CHIR99021 for 2 days and then changed to 2 μM CHIR99021 for another 2 days. During differentiation, from days 4–12, the medium was changed to EGM2 (Lonza) supplemented with vascular endothelial growth factor (VEGF) (50 ng/mL) (PeproTech), bone morphogenetic protein 4 (BMP4) (20 ng/mL), and fibroblast growth factor 2 (FGF2) (20 ng/mL) (PeproTech). On day 12, cells were dissociated using TrypLE for 5 min and sorted using CD144-conjugated magnetic microbeads (Miltenyi Biotec) according to the manufacturer's instructions. CD144-positive cells were seeded on 0.2% gelatin-coated plates and maintained in EGM2 medium supplemented with 10 μM transforming growth factor β (TGFβ) inhibitor (SB431542). (Selleckchem). After passage 2, iPSC-ECs were cultured in EGM2. The iPSC-ECs were analyzed at passage 3 post differentiation.

Epicardial cell differentiation (old protocol)

EPCs were differentiated using a method as previously described [60]. The iPSCs were seeded in 6-wells at a density of 1.2 × 10^5 cells per well and grown for four days prior to differentiation. Differentiation was initiated by replacing the E8 media with RPMI supplemented with B27 without insulin (A1895601, Life Technologies) and 6 μM CHIR-99021 (CT99021, Selleckchem). Two days later the media was replaced with RPMI supplemented with B27 without insulin. Cultures were then treated with 5 μM IWR-1 (I0161, Sigma) in RPMI supplemented with B27 without insulin for two days. On day 5, human induced pluripotent stem cell-derived cardiac progenitor cells (iPSC-CPCs) were re-plated at a density of 20,000 cells/cm2 in advanced DMEM medium (12634028, Gibco, Life Technologies). On day 5 to day 8, cells were treated with 5 μM of CHIR99021 and 2 μM of retinoic acid (R2625, Sigma-Aldrich) for 3 days, and recovered in advanced DMEM for 4 days.

Epicardial cell differentiation (new protocol)

The iPSC-derived epicardial cells were differentiated in a chemically defined medium (CDM), which is composed of 50% IMDM, 50% Ham's F-12 Nutrient Mix, 1% chemically defined lipid concentrate, 2 mM Glutamax, 1 mg/mL PVA, 15 μg/mL transferrin, and 450 μM monothioglycerol. When iPSCs reached ~80% confluency they were dissociated with 1 mL of Accutase (Sigma) and re-plated a density of 1.5 x10^4 cells/cm^2 in 6-well plates and cultured in iPS-Brew medium (Miltenyi Biotech) supplemented with 10 μM Y27632. The next day (day 1), each well was washed with D-PBS, and epicardial cells differentiation was initiated by adding the mid-primitive streak induction medium (consisting of 10 ng/mL Activin A, 6 μM CHIR99021, 50 ng/mL BMP4, 20 ng/mL FGF2, and 2 μM LY294002 in CDM). On day 2, each well was refreshed with the lateral plate mesoderm induction medium (consisting of 1 μM A83-01, 30 ng/mL BMP4, and 1 μM C59 in CDM). On days 3-4, each well was refreshed with the splanchnic mesoderm induction medium (consisting of 1 μM A83-01, 30 ng/mL BMP4, 1 μM C59, 20 ng/mL FGF2, and 2 μM retinoic acid in CDM). On days 5-8, the media was refreshed with the septum transversum induction medium (consisting of 2 μM retinoic acid and 40 ng/mL BMP4 in CDM). On day 9, cells were dissociated using Accutase and sparsely seeded (10^4 cells/cm^2) on gelatin-coated 6-well plates in the proepicardium induction medium (consisting of 100 μg/mL ascorbic acid, 2 μM of retinoic acid, and 0.7 μg/mL insulin in CDM) for 2 days without medium change. Starting at day 11, each well was refreshed every other day with the epicardial cell induction/maintenance medium (consisting of 100 μg/mL ascorbic acid, 10 μM SB431542, and 0.7 μg/mL insulin in CDM). The iPSC-derived epicardial cells can preserve their cell type-specific markers (e.g., *TBX18*, *WT1*, and *TCF21*) for at least 18 passages in the epicardial cell induction/maintenance medium.

Cardiac fibroblast differentiation

To generate cardiac-specific fibroblasts, iPSC-derived epicardial cells were dissociated with Accutase and plated at a density of 10^4 cells/cm^2 in 6-well plates and cultured in fibroblast growth medium (Lonza) supplemented with 20 ng/mL FGF2 and 10 μM SB431542. The medium was refreshed every other day for 6 days. When the fibroblasts reached ~90% confluency, they were dissociated and split at a 1:3 ratio in fibroblast growth medium supplemented with 10 μM SB431542 for long-term maintenance. The differentiated fibroblasts exhibit a quiescent phenotype with negligible (<5%) α-SMA expression for at least five passages.

Smooth muscle cell differentiation

To generate cardiac-specific smooth muscle cells (SMCs), iPSC-derived epicardial cells were dissociated with Accutase and seeded at a density of 3×10^4 cells/cm^2 were seeded in the nascent SMC induction medium (consisting of 100 μg/mL ascorbic acid, 0.7 μg/mL insulin, 10 ng/mL Activin A, and 10 ng/mL PDGF-BB in CDM) for 2 days. The medium was refreshed every other day with Medium 231 supplemented with SMGS (ThermoFisher) for at least 14 days to allow the expression of SMC-specific markers (e.g., *TAGLN*, *CNN1*, *SMTNB*, and *MYH11*).

Single-cell ATAC-seq on iPSC-derived cardiac cells and human fetal heart

The iPSC-derived cardiac cells were dissociated using Tryple Express and resuspended in the RPMI medium. The human fetal hearts were minced and digested using Liberase (Sigma) for 10 min at 37°C, and resuspended in RPMI + B27 medium to stop the enzymatic reaction. The digested tissue was passed through a 70 μm filter before proceeding to single-nuclei sample preparation. Cells with viability >90% were washed in ice-cold ATAC-seq resuspension buffer (RSB, 10 mM Tris pH 7.4, 10 mM NaCl, 3 mM MgCl2), spun down, and resuspended in 100 μL ATAC-seq lysis buffer (RSB plus 0.1% NP-40 and 0.1% Tween 20). Lysis was allowed to proceed on ice for 5 min, then 900 μL RSB was added before spinning down again and resuspending in 50 μL 1X Nuclei Resuspension Buffer (10x Genomics). A sample of the nuclei was stained with Trypan Blue and inspected to confirm complete lysis. Nuclei were processed using a 10X chromium single-cell ATAC-seq kit (V1 version, 10X Genomics) at the Stanford Functional Genomics Facility (SFGF). All samples were sequenced using the Illumina HiSeq 4000 (150 bp paired-end).

CRISPR–Cas9-mediated genome editing of iPSCs

The genomic region (300-400bp) corresponding to *JARID2* cRE was deleted using CRISPR-Cas9 genome editing. Two guide RNAs (gRNAs) flanking the cRE upstream of *JARID2* were designed using a web-based tool (Benchling) and chosen based on a high score for on-target binding and the lowest off-target score. For cRE deletion, iPSCs ($3.5\text{x}10^5$) were nucleofected (1200 V, 20 ms, 1 pulse) with 60 pmol sgRNA (Synthego) and 20 pmol SpCas9 nuclease (Synthego) using the Neon Transfection System (ThermoFisher Scientific) and the 10 μL tip per the manufacturer's instructions). After electroporation, iPSCs were plated in E8 medium supplemented with 5 μM Y-27632 into a 12-well plate coated with Matrigel. After recovering (3 days post electroporation), the cells were dissociated with TrypLE Express and were plated in 6-well plates at a density of 2,000 cells per well. About 10 days after transfection, colonies were picked into 96-well plates and a small proportion of cells from each colony were used for DNA extraction using Quick Extract solution (Epicenter) and direct PCR with Prime STAR GXL DNA Polymerase (Clontech). PCR amplicons were sequenced by Sanger to verify the deletion (Figure S6M).

Tubular network formation assay

Tubular network formation was conducted in a 24-well plate format. Prior to experiments, 24-well plates were pre-chilled in −20°C. Then plates were coated with 250 μL growth factor-reduced Matrigel (Corning) per well and incubated at 37°C with 5% CO2 for 30 min iPSC-derived ECs at passage 2 were dissociated into single cells using accutase and resuspended in EMG-2 medium containing 5 ng/mL VEGF. A total of 100, 000 cells were seeded in each well and incubated at 37°C with 5% CO2. Bright-field images were taken 12 h after cell seeding with an inverted phase contrast SONY microscope using a 4× objective. Experiments were carried out in triplicates and repeated twice. Images were analyzed using a customized version of the "Angiogenesis Analyzer" developed for ImageJ .

Luciferase reporter vector construction

The luciferase reporter vectors pGL3-Promoter (E1761) and pGL3-Control (E1741) were purchased from Promega. JARID2 cRE with 500 bp in length harboring reference and variant alleles were synthesized by Twist. The cRE was cloned into the linearized pGL3-Promoter vector (cut by Xhol). The fusion product (pGL3-cRE) was subsequently transformed into Mix & Go Competent Cells Strain Zymo 5-α (Zymo Research, T3007). Clones were selected by Ampicillin and plasmids were prepared using the QIAprep Spin Miniprep Kit (Qiagen, 27,106).

Transfection and luciferase assays

i-ECs were transfected in a 24-well plate using the Lipofectamine Stem Transfection Reagent (Invitrogen, STEM00001) and Opti-MEM Reduced Serum medium (Invitrogen, 31,985- 070). On the day of transfection, cell density was 60-80% confluent. For each well, 500 ng of pGL3-enhancer, pGL3-control, or pGL3-promoter was co-transfected with 10 ng of pRL-CMV (Promega, E2261) as an internal control for the normalization of luciferase activity. Cells were incubated with DNA-lipid complex overnight and media was changed for another 2 days. The firefly and renilla luciferase activity were measured respectively using a Dual-Glo Luciferase Assay System (Promega, E2920). The ratio of firefly versus renilla luminescence was calculated and normalized to the control sample.

Computational methods

Fetal tissue - scATAC processing

Raw sequencing data were converted to FASTQ format using 'cellranger-atac mkfastq' (10x Genomics, v.1.2.0). 150 bp paired-end (PE) scATAC-seq reads were aligned to the GRCh38 (hg38) reference genome and quantified using '*cellranger-atac count*' (10x Genomics, v.1.2.0).

Fetal tissue - scATAC-seq quality control, dimensionality reduction, filtering and identification of cell types

Mapped Tn5 insertion sites (fragments.tsv files) from cellranger were read into the ArchR (v0.9.4) R package.[17] To ensure that each cell was both adequately sequenced and had a high signal-to-background ratio, we filtered cells with fewer than 1,000 unique fragments and enrichment at TSSs below 6. To calculate TSS enrichment, genome-wide Tn5-corrected insertions were aggregated ±2,000 bp relative (TSS-strand-corrected) to each unique TSS. This profile was normalized to the mean accessibility ±1,900–2,000 bp from the TSS, smoothed every 51 bp and the maximum smoothed value was reported as TSS enrichment in R (Figures S1A–S1F). Latent Semantic Indexing (LSI) dimensionality reduction was computed (*iterations = 4*, *variable features = 25,000*, *dim = 30*) by appending fragment files from all three timepoints together (Figures S1G and S1H). We did not observe any significant batch effects after the fourth iteration of iterative LSI. We computed chromatin-derived gene accessibility scores by aggregating scATAC-seq reads in each cell weighted by distance from each gene within its *cis*-regulatory domain.[17] A preliminary cell-type annotation was performed using these gene accessibility scores of known cell type markers (Figures 1C, 1D, S1I, and S1J, Table S1).

We observed two populations of cell types identified to be macrophages and immune cells (Figures S1H and S1I). Even though these sets of cell types are of interest from a biological standpoint, they do not directly contribute to the cardiogenesis process and hence were dropped from subsequent analysis. The final UMAP used in all subsequent analyses was generated by repeating the above mentioned iterative LSI with the same parameters as above after removing barcodes corresponding to the macrophage and immune cell clusters (Figure 1C). Final cell-type annotations for each cluster were assigned based on gene accessibility scores of marker genes of known cardiac cell types (Figures 1C, 1D, and S1J, Table S1).

Briefly, we identified cell types of the three major lineages and neural crest. Within the myocardial lineage, we found that *TNNT2*, *ACTN2*, and *NKX2-5* had high GA-scores across the early cardiomyocytes (eCM), ventricular cardiomyocytes (vCM), and atrial cardiomyocyte (aCM) clusters.[4–6,21] *TTN* and *HAND1* specifically marked the eCM and vCM cluster while *TBX10* and *NPPA* marked the aCM cluster (Figures 1D and S1J).

We observed diverse lineages within the epicardial derived cells. We discovered four cell types at PCW6: cardiac fibroblast progenitors (CFP) with high *WT1*, *TBX18*, and *TCF21* GA-scores, another set of similar cells with both *TBX18* and *TCF21* signal but lacking *WT1* which we called fibroblast-like cells (FB1), and the outflow tract (OFT)-like cells with high *PRDM6*[28] and *HOXA3* GA-scores (Figure S1J). We also found an undifferentiated epicardial cell cluster (EPC) with high *TBX18* and *WT1* GA-scores but lacking *TCF21*[61,62] (Figure S1J). We found different cardiac fibroblast cell populations (preCF and CF) that have high *TCF21* GA-scores but varying, low to high respectively, *DCN* and *LUM* GA-scores.[63] Another cluster of fibroblast like cells (FB2) with high *CNN1* and *COL9A2* GA-scores were also identified. We hypothesize that this cell type, along with FB1, may be related to valvular fibroblasts, but further studies are required to establish this potential relationship. Finally, we defined a cluster of pre-smooth muscle cells (preSMC) with high *MYH11*, *PDGFRB*, and *TAGLN* GA-scores but lacking *TCF21* activity,[64] a cluster of smooth muscle cells (SMC) exhibiting stronger GA-scores for *MYH11* and *PDGFRB* with major contributions from PCW19 and minor contributions from PCW8, and a cluster of pericytes (PC) with high GA-scores of *PDGFRB* and *ABCC9* (Figure S1J).[65] We also defined a cluster of neural crest (NC) cells with high *TFAP2A* GA-score (Figure S1J).[66]

The endocardial cell populations exhibited two distinct phenotypes: one with high *CDH11* GA-scores (Endo1) and a population that resembled endocardial-like transitioning cell types (Endo2).[67] Arterial endothelial cells (aEC) exhibited high *UNC5B* and *GJA5* GA-scores. Capillary cells (Cap) were marked by high *CA4*, *APLNR*, and *CD36* GA-scores (Figure S1J). Venous endothelial cells (vEC) were marked by high *SELE* and *SELP* GA-scores, amongst other markers.[68,69] In addition to these major endothelial cell types, we also found a sub-population of lymphatic endothelial cells (lEC) exhibiting high *LYVE* GA-score (Figure S1J).[70]

We also observed chromatin state changes consistent with promoter priming for genes in specific cell types that do not express the associated gene. For instance, the promoter of the developmental gene *TCF21* was accessible in cardiac fibroblast and SMC cell lineages but the gene was expressed only in cardiac fibroblasts and not in SMC[71,72] (Figures 1H and S2C). Interestingly, *TCF21* expression is known to be activated in SMC in adults in response to vascular stress,[73] promoting cell state changes such as proliferation and migration, consistent with a return to an embryonic-like phenotype for the SMC.[72] Thus, accessibility of the TSS at the *TCF21* gene may represent adaptive promoter priming[74] that allows the gene to rapidly respond to disease-related stress or cellular activation.

Fetal tissue - Peak calling in scATAC-seq datasets

Single-cell chromatin accessibility data were used to generate pseudobulk group coverages based on high-resolution cluster identities of scATAC-seq datasets before peak calling with MACS2 v2.1.1.20160309[75] using the *addReproduciblePeakSet()* in ArchR.[17] A background peak set controlling for total accessibility and GC-content was generated using *addBgdPeaks()*. Overlapping peaks were handled using an iterative removal procedure as previously described in.[76] First, the most significant (MACS2 *q*-value) extended peak summit is kept and any peak that directly overlaps with that significant peak is removed. This process reiterates to the next most significant peak until all peaks have either been kept or removed owing to direct overlap with a more significant peak. The most significant extended peak summits for each cluster were then merged and the previous iterative removal procedure was used. Lastly, we removed any peaks whose nucleotide content had any 'N' nucleotides and any peaks mapping to chrY.

Using the previously annotated clusters, we identified 215,163 putative cREs as scATAC-seq peak regions over all cell types and timepoints (Figure 1E). The clusters were enriched for expected gene ontology (GO) terms associated with cardiac development and cell-type specific attributes[77] (Figure 1E, Table S1).

Fetal tissue - scRNA processing

Raw sequencing data from two previous studies[6,21] corresponding to post-conception week (PCW) 6, 8 and 12, were converted to FASTQ format using the command '*cellranger mkfastq*' (10x Genomics, v.3.1.0). scRNA-seq reads were aligned to the GRCh38 (hg38) reference transcriptome (Ensembl 93) and quantified using '*cellranger count*' (10x Genomics, v.3.1.0). The filtered matrices from cell ranger count were combined with the filtered matrices of other datasets from Asp, et al.[5] and Suryawanshi, et al.[4] corresponding to PCW6 and 19 to create the scRNA object.

Count data were further processed using the 'Seurat' R package (v.3.1.4),[78] using GENCODE v.27 for gene identification. We removed cells with less than 250 expressed genes, cells with less than 300 reads, and cells with more than 40% read count corresponding to mitochondrial genes. Genes not contained in the GENCODE annotation were excluded from further analysis. Gene level read count data was scaled to 10,000 (TP/10k) and $\log_2$ transformed. We performed Principal Component Analysis (PCA) restricting to the 2,000 most variable genes as defined by Seurat. The top 30 principal components (PCs) were used for downstream clustering. Clusters were identified using Louvain clustering implemented in Seurat's *'FindClusters()'* function ('resolution = 1'). 2-dimensional representations were generated using uniform manifold approximation and projection (UMAP) (McInnes et al., 2020) as implemented in Seurat and the 'uwot' R packages (v.0.1.8; parameter settings: 'min.dist = 0.8', 'n.neighbors = 50', 'cosine' distance metric). We observed that the clustering was strongly influenced by sample of origin indicating significant batch effects (Figure S2A). To correct these batch effects, we used Harmony[79] with max_iters = 5 and other parameters set to their default values. We then reran Louvain clustering with higher resolution on the top 30 components from Harmony and generated a 2D UMAP for the corrected data with the same functions listed above. Post harmonization, clusters did not appear to be affected by the sample of origin (Figure S2A). Cell-type annotations for each cluster were assigned by resolving and merging clusters from higher resolutions based on the expression of known marker genes of cardiac cell types (Figures S2B and S2C, Table S1).

Fetal tissue - Matching cells from scRNA-seq and scATAC-seq data

Canonical correlation analysis (CCA) as implemented in Seurat[78] was used to align and match cells from the scRNA-seq and scATAC-seq experiments. For this purpose, we computed $\log_2$-transformed gene accessibility scores as surrogates for gene expression in the cells profiled by scATAC-seq. As integration features, we used the union of the 2,000 most variable genes in each modality as input to Seurat's *'FindTransferAnchors()'* function with reduction method 'cca' and parameter 'k.anchor = 10'. For each cell profiled by scRNA-seq, we identified the nearest neighbor cell in scATAC-seq by applying nearest-neighbor search in the joint CCA L2 space. Nearest neighbors were determined using the 'FNN' R package (https://rdrr.io/cran/FNN/) employing the 'kd_tree' algorithm with Euclidean distance. These nearest-neighbor-based cell matches from all gestational time points were concatenated to obtain dataset-wide cell matches across both modalities (Figures S2D and S2E).

We found high concordance (accuracy = 74.76%) between the cluster assignments for cells from the scATAC-seq and scRNA-seq data, further supporting our cell type annotations based on chromatin accessibility derived gene accessibility scores (Figure S2E). Examining a subset of cell-type specific marker genes, we found *TNNT2* marking the vCMs, *PECAM1* identifying endothelial cells, *CDH11* identifying endocardium, *MYH11* identifying SMC, and *DCN* identifying fibroblasts[80,81] (Figures 1D and 1G). We also observed a strong correlation (Table S1) between gene expression from the scRNA-seq data and the GA-scores from the scATAC-seq data across matched nearest-neighbor cells from the two complementary atlases (Figures 1D and 1G), further supporting our annotations.

BPNet deep learning models to predict base-resolution, cell-type resolved pseudobulk scATAC-seq profiles from DNA sequence

BPNet is a sequence-to-profile convolutional neural network that uses one-hot-encoded DNA sequence (A = [1,0,0,0], C = [0,1,0,0], G = [0,0,1,0], T = [0,0,0,1]) as input to predict single nucleotide-resolution read count profiles from assays of regulatory activity.[12,13] The models take in a sequence context of 2,114 bp around the summit of each ATAC-seq peak and predict cluster-specific scATAC-seq pseudobulk Tn5 insertion counts at each base pair for the central 1,000 bp. The BPNet model also uses an input Tn5 bias track which is concatenated to the pre-final layer as explained below. Our BPNet model is a higher capacity version of the architecture introduced in.[12] The model architecture consists of 8 dilated residual convolution layers, with 500 filters in each layer. At each layer, the Keras Cropping 1D layer is used to clip out the two edges of the sequence, to match the inputs concatenated to the output of each convolution, which naturally trims the 2,114 bp sequence to a final 1,000 bp profile. Each dilated convolutional layer has a kernel width of 21 and the dilation rate is doubled for every convolutional layer starting at 1. The model predicts the base-resolution 1,000 bp length Tn5 insertion count profile using two complementary outputs: (1) the total Tn5 insertion counts over the 1,000 bp region, and (2) a multinomial probability of Tn5 insertion counts at each position in the 1,000 bp sequence. The predicted (expected) count at a specific position is a multiplication of the predicted total counts and the multinomial probability at that position. To predict the total counts in the 1,000 bp window, the output from the last dilated convolutional layer is passed through a GlobalAveragePooling1D layer in Keras. We estimate the "tn5 bias" for the input sequence using the TOBIAS method.[82] This total

bias is concatenated with the output of the pooling layer and passed through a Dense layer with 1 neuron to predict total counts. To predict the per-base logits of the multinomial probability profile output, the output from the last dilated residual convolution is appended with per base TOBIAS "tn5 bias" and passed through a final convolution layer with a single kernel and a kernel width of 1 to predict the per-base logits. BPNet uses a composite loss function consisting of a linear combination of a mean squared error (MSE) loss on the log of the total counts and a multinomial negative log likelihood loss (MNLL) for the profile probability output. We use a weight of [4.9, 4.3, 18.5, 9.8, 8.9, 4.8, 4.6, 4.9, 12.4, 15.4, 4.3, 6.3, 1.4, 2.6, 7.6, 2.3, 16.3, 7.1 & 3.7] for the MSE loss for clusters c0–c20 (c15-c16 combined as one model), and a weight of 1 for the MNLL loss in the linear combination. The MSE loss weight is derived as the median of total counts across all peak regions for each cluster divided by a factor of 10^{12}. We used the ADAM optimizer with early stopping patience of 3 epochs.

A separate BPNet model was trained on pseudobulk scATAC-seq profiles from each scATAC-seq cluster. We used a 5-fold chromosome hold-out cross-validation framework for training, tuning, and test set performance evaluation. The training, evaluation, and test chromosomes used for each fold are as follows. Test chromosomes: fold 0: [chr1], fold 1: [chr19, chr2], fold 2: [chr3, chr20], fold 3: [chr13, chr6, chr22] & fold 4: [chr5, chr16]. Validation chromosomes: fold 0: [chr10, chr8], fold 1: [chr1], fold 2: [chr19, chr2], fold 3: [chr3, chr20] & fold 4: [chr13, chr6, chr22]. The model's performance was evaluated using two different metrics for the two output tasks separately. For the total counts predicted for the 1,000 bp region, the model's performance is computed with the Spearman correlation of predicted counts to actual counts. The profile prediction performance is evaluated using the Jensen-Shannon Distance, which computes the divergence between two probability distributions; in this case, the observed and predicted base-resolution probability profile over each 1,000 bp region.

For each cell type, BPNet models were trained, tuned, and evaluated on genomic windows consisting of 1 kb scATAC-seq profiles from (1) signal windows centered at summits of scATAC-seq peaks from the cell type and (2) background windows randomly sampled across the genome such that the number of background windows was 10% of the number of signal windows. The selected signal and background windows were further augmented with upto 10 random jitters (+/- 1000 bp). Code for training BPNet models is available at https://github.com/kundajelab/Cardiogenesis_Repo.

BPNet model-derived DeepLIFT/DeepSHAP nucleotide contribution scores of accessible cRE sequences

We used the DeepLIFT algorithm[23] to interrogate BPNet models and estimate the predictive contribution of each base in any query input sequence to the predicted total counts from the model. DeepLIFT backpropagates a score, analogous to gradients, which is based on comparing the activations of all the neurons in the network for the input sequence to those obtained from neutral 'reference' sequences. We use 20 dinucleotide-shuffled versions of each input sequence as reference sequences. We used the DeepSHAP implementation of DeepLIFT (https://github.com/slundberg/shap/blob/0.28.5/shap/explainers/deep/deep_tf.py) to obtain contribution scores for all observed bases in each sequence.[83] For each cell type, we obtained consolidated DeepLIFT/DeepSHAP contribution scores for each sequence from each of 5-folds of cross-validation and then averaged the scores per position from the 5-folds.

Annotation of PWM-based transcription factor motif instances in accessible cREs

We obtained position weight matrix (PWM) models of transcription factor (TF) sequence motifs from the ChromVAR motif catalog called 'human_pwms_v1',[27] which is collated from the Catalog of Inferred Sequence Binding Preferences (CIS-BP).[26]

We then annotated PWM-based motif instances in all cRE sequences from all cell types by scanning, scoring, and thresholding (p value < 5e-5) matches from all PWMs using the motifmatchr tool (https://github.com/GreenleafLab/motifmatchr) which uses the MOODSv.1.9.3 library.[84]

Annotation of cell-type specific active TF motif instances in accessible CREs with high contribution scores and motif mutagenesis scores

For each accessible cRE in each cell type, we defined active motif instances as a subset of PWM-based motif instances that have high DeepLIFT contribution scores or high motif mutagenesis scores from the corresponding cell-type specific BPNet models relative to a null background distribution of corresponding scores.

Motif instance contribution scores. We computed the contribution score of each PWM motif instance to accessibility in a specific cell type as the average of the consolidated DeepLIFT contribution scores from the cell-type specific BPNet models over all bases overlapping the motif instance.

Motif instance mutagenesis scores.

We also inferred mutagenesis scores (motif-ISM) for each PWM-motif instance in a cRE sequence with respect to accessibility in each cell type. To generate the motif-ISM scores for a PWM motif instance in a specific cell type,

1. We first used the fold-0 BPNet model of the specific cell type to predict the total scATAC-seq counts over a 1000 bp window (using a 2114 bp input sequence) centered at the motif instance.
2. We then generated 3 shuffled versions of the input sequence containing the motif instance such that we maintain di-nucleotide frequencies (dinucleotide shuffling).

3. We obtained 3 subsequences overlapping the positions of the original motif instance from the 3 shuffled dinucleotide shuffled sequences.
4. We replaced the subsequence of the motif instance in the original reference sequence with each of the 3 shuffled subsequences.
5. We then use the fold-0 BPNet model to once again predict the total scATAC-seq counts for each of these 3 disrupted sequences containing the shuffled versions of the motif instance.
6. We then computed the $\log_2$ ratio of the total predicted counts between the reference sequence from step 1. and each of the 3 disrupted sequences from step 5.
7. The motif-ISM score of the instance was computed as the average of the $\log_2$ ratio score from step 6. over all 3 disrupted sequences.

Empirical null distributions.
We generated empirical null distributions of motif-instance contribution scores as follows.

1. We constructed dinucleotide frequency preserving shuffled versions of all cREs from from chr4 and chr7.
2. We used the cell-type specific BPNet models from each of the 5-folds to compute DeepLIFT contribution scores over all randomized sequences from step 1. For each sequence, the contribution scores at each base were averaged over all 5-folds.
3. The contribution scores from all bases in all sequences from step 2. were used to derive an empirical null distribution of contribution scores.

We generated empirical null distributions of motif-instance ISM scores as follows.

1. We reused the predicted total scATAC-seq counts for each of these 3 disrupted sequences containing the shuffled versions of the motif instance from step 5. of the motif-ISM estimation process above. We computed the $\log_2$ ratio of the total predicted counts between each of the 3 pairs of disrupted sequences.
2. The empirical null distribution for motif-ISM scores was derived from the above computed scores over all motif instances in all cRE sequences in chr4 and chr7.

Active motif instances. Finally, to identify active motif instances in each cell type, we select PWM-based motif instances that have motif-instance contribution scores or motif-ISM scores that are above the 95th percentile or below the fifth percentile of corresponding empirical null distribution scores of that cell type. All other PWM-based instances were labeled as "inactive".

Enrichment of active motif instances and all PWM-motif instances in differential, cell-type specific scATAC-seq peaks

We identified differentially accessible, cell-type specific "marker peaks" for the ventricular cardiomyocyte cluster (vCM) relative to all other clusters using the *getMarkerFeatures()* function in ArchR,[17] which uses the Wilcoxon Rank-sum test to identify marker peaks while controlling for the TSS enrichment and $\log_{10}$(unique fragments) of cells when sampling the background set of cells. We then calculated the Fisher Exact test implemented in the *peakAnnoEnrichment()* function in ArchR to compute the enrichment of active motif instances of all TFs expressed in vCMs in vCM marker peaks relative to all vCM peaks. We compute analogous enrichments of all PWM-based motif instances. We compare the statistical significance of enrichments of active and all PWM instances in Figures 2H, 2I and S3A.

We observed a diverse set of TFs enriched across different cell types in our fetal atlas. Briefly, we found that *MEF2*, *TGIF1*, *NFI* motif families were highly enriched in vCMs and *TGIF* and *KLF* families in aCMs. The eCMs had similar TF motifs as the vCMs and aCMs, albeit with weaker enrichments, suggesting this cluster is the progenitor population for later cardiomyocyte subtypes. The CFPs and CFs had similar motif enrichment for *TCF21*/TCF, *MYOG, MSC*, with CF gaining enrichment for *TEAD* and *NFI* families and implicating a second set of TFs that become active during CF maturation. The other fibroblast-like clusters (FB1 and FB2) had lower *TCF21* enrichments than the cardiac fibroblast clusters, but stronger enrichment for *JUN*, *FOS* and *JDP* motif families. The OFT cells exhibited strong *RFX* and *TEAD* motif enrichments, while preSMC exhibited weaker enrichments for the *RFX* and *KLF* families and stronger enrichment for motifs associated with proliferation like *SP* and *RBPJ*. These enrichments became substantially stronger in the SMCs at PCW19, and with the gain of new TF enrichments such as the *MEF2* family, indicating a continuum of TF motif activity promoting the SMC cell fate trajectory. The PCW6 endocardial cells (Endo1) had stronger TF activity for *ETV* and *STAT* families and weaker enrichments for the *SOX* family. The capillary (Cap) cells, which are thought to derive from the endocardium, were strongly enriched for *SOX* family motifs. The aEC and Cap clusters, exhibited enrichments for *SOX*, *FOS* and *JUN* motifs and also retained endocardium TF motifs like *ELF* and *ETV*. vECs also had a motif landscape similar to the capillaries, with the addition of a few motifs, such as *STAT*.

ChromVAR motif deviation scores

To compute ChromVAR motif deviation scores for any peak set, a background peak set controlling for total accessibility and GC-content was generated using *addBgdPeaks()* for each cluster in ArchR. Chromvar[27] was run with *addDeviationsMatrix()* using active TF motif instances in both peak sets to calculate enrichment of chromatin accessibility over all active motif instances of each TF at single-cell resolution. We then computed the GC-bias-corrected deviation scores using the chromVAR 'deviationScores' function used in the *addDeviationsMatrix()* function in ArchR.

Defining cell transitions and trajectories from scATAC-seq data using optimal transport

Computing gene signatures. We created a cell by gene score matrix that was used for computing the gene signatures associated with cell cycle and apoptosis for optimal transport analysis. We used the list of curated genes for cell cycle and apoptosis as suggested in the original optimal transport paper.[14] We scored cells based on the chromatin derived gene accessibility scores[17] of genes in the curated gene signatures. We used the same procedure as in the original manuscript. For each cell, we compute the *Z score* of the gene accessibility scores for each gene in the set. We then clip these *z*-scores in the range of −5 to 5. We define the signature score of the cell to be the mean *Z score* over all genes in the gene set (Figures S3B and S3C). We estimated the initial growth rate with the same calculations as performed in the original method[14] with the cell cycle and apoptosis signal computed from the gene score matrix (Figure S3D).

Using gene score matrix for optimal transport calculation. We performed optimal transport-based trajectory analysis by following the original codebase (https://broadinstitute.github.io/wot/tutorial/).[14] The two changes between the original method and our implementation are the use of gene accessibility scores to compute the gene signatures and the use of the cell by gene accessibility score matrix for inferring the optimal transport maps as compared to the cell by gene expression used in the original method. The cell by gene accessibility score matrix was scaled to read per 10K and $\log_2$-transformed. The top 2000 variable genes based on Seurat (*FindVariableGenes()* method = "vst") were retained for further analysis. The coupling inference was obtained using parameters *e = 0.05; l1 = 1; l2 = 50; growth_iters = 3*.[14] We first computed the transport matrices between successive timepoints, inferred long-range temporal couplings and then computed the fate matrices to obtain the transition table (Figure 3B).

We observed 8 major differentiation trajectories within our single-cell atlas. Briefly, within the endocardium lineage, the endocardium-like cell clusters (Endo1/2) were predicted to give rise to the Cap cells, which in turn were predicted to transition into the vECs in PCW19. The aEC cluster was derived from Endo1/2 clusters as well as the PCW8 Cap cluster, suggesting that some terminal cell states can originate from different developmental origins (Figure 3B). We also identified cells that appeared to have already committed to their developmental fates based on their expression of lineage specific genes. For example, at PCW6, cells from the epicardial lineage (EPC, OFT, CFP and FB1) that expressed *TCF21* were predicted to transition into the cardiac fibroblasts at PCW8 (preCF) and PCW19 (CF) (Figure 3B). The OFT cluster which lacks *TCF21* expression was predicted to transition into SMC and PC clusters through the preSMC cluster. These observations are highly concordant with results from studies with lineage tracing in *TCF21* recombinase knock-in mice.[71] Finally, the FB1 cluster was predicted to transition into the FB2 cluster. For the myocardium cells, the eCM cluster was predicted to differentiate into vCM and aCM clusters.

Chromatin and gene expression dynamics across trajectories. For all the major trajectories identified using optimal transport, we identified the clusters that are predicted to be in the trajectory using the transition table (Figures 3C and S4). We provided these sets of cell clusters to ArchR's[17] *addTrajectory()* function and assigned cells pseudotime values. We then used the *plotTrajectory()* function to plot the chromatin peak dynamics associated with the identified trajectory. We estimated correlation between TF gene expression from scRNA-seq projected into the scATAC-seq subspace and TF ChromVAR deviation scores using *correlateMatrices()* in ArchR.[17] We defined correlated TFs for each trajectory as those who had correlation values >0.5.

In addition to the SMC trajectory, we would like to elaborate on one more main differentiation trajectory. The vEC differentiation trajectory captured cell state transitions from the Endo1/2 progenitor cells at PCW6 to vECs at PCW19 through the Cap cells in PCW8 (Figure 3K). Waves of TFs including *GATA2/3/4/6, NFATC2, SOX4, SOX17* and *MEOX1* with correlated expression and motif activity dynamics are predicted to regulate concordant cascades of dynamically accessible cREs targeting genes involved in different stages of angiogenesis (Figures 3L and 3M). We once again used cell-type specific BPNet models to decipher TFs that regulate dynamic cREs in the *cis*-regulatory domain of the *APLNR* gene, a primary marker of vECs,[85–87] which exhibited a coordinated and monotonic increase in gene expression, promoter accessibility and cumulative distal chromatin accessibility (gene accessibility scores) across the trajectory (Figure 3N). BPNet models trained on Endo1/2, Cap and vEC cells revealed *GATA3, SOX17* and *SP1* to specifically regulate three representative cREs in the *APLNR* locus with distinct temporal dynamics of chromatin accessibility based on cell-type specific predictive motif instances and concordant TF expression (Figures 3O–3Q).

iPSC derived in vitro cardiac cell types - scATAC-seq data processing, quality control, dimensionality reduction and motif annotations

Raw sequencing data were converted to FASTQ format using '*cellranger-atac mkfastq*' (10x Genomics, v.1.2.0). 150 bp paired-end (PE) scATAC-seq reads were aligned to the GRCh38 (hg38) reference genome and quantified using '*cellranger-atac count*' (10x Genomics, v.1.2.0). To ensure that each cell was both adequately sequenced and had a high signal-to-background ratio, we filtered cells with enrichment at TSSs below 6 and unique fragments (1,000-1,500) depending on the individual library (Figure S5).

Projecting iPSC derived in vitro cardiac cells based on scATAC-seq into the fetal heart scATAC-seq manifold

We projected the iPSC derived *in vitro* cardiac cells based on the scATAC-seq profiles into the scATAC-seq LSI subspace of fetal heart cells following the procedure described previously.[38] Briefly, when computing the TF-IDF transformation on the fetal samples, we stored the colSums, rowSums, and SVD. To project cells from additional samples into this subspace, we first zeroed out rows based on the initial TF-IDF rowSums. We next calculated the term frequency by dividing by the column sums and computed the inverse document frequency from the previous TF-IDF transformation. These were then used to compute the new TF-IDF. The resulting TF-IDF matrix was projected into the previously defined SVD of the fetal heart LSI.

Identifying scATAC-seq peaks across all in vivo and in vitro cardiac cells

To enable the comparison of epigenomic features between the *in vivo* and *in vitro* cells, we built a combined ArchR object of all post filtered cells from the three fetal heart samples and all the samples from the iPSC differentiation to major cardiac cell types. We performed peak calling on the combined data using ArchR, as described above. We used these peak calls from the combined object for all the downstream differential analyses between the *in vivo* and *in vitro* nearest cells identified by the projection analysis. PWM-based motif instances[26] were used to compute TF motif annotations and ChromVar deviations as described above.

Identifying differential scATAC-seq peaks and TF motif enrichments between matched in vivo and in vitro cardiac cell types

Differential peaks between *in vivo* and *in vitro* cell types were identified within the integrated peak set described in the above section. For each pair of match cell types, we obtained the integrated cell x peak matrix. We then computed row-wise two-sided *t*-tests for each peak and estimated the FDR using *p.adjust(method = "fdr")*. Peaks with absolute $\log_2$(fold changes) > 1 and FDR < 0.05 were labeled as differential.

To calibrate the magnitude of these differences, we also estimated differential peaks between two distant *in vivo* cell types, namely vCMs and excitatory neurons.[13] Reassuringly, the differences between *in vitro* and *in vivo* cardiac cells were substantially smaller than differences between vCMs and neurons (Figure 5A). We next identified the TF motifs enriched in up or down regulated differential peaks relative to all peaks for each pairwise comparison using *peakAnnoEnrichment()* in ArchR.

Predicting mutation impact scores of de novo non coding mutations from CHD cases and controls on cell-type resolved scATAC-seq profiles using neural network models

We obtained *de novo*, non coding mutations from CHD patients from the Pediatric Cardiac Genomics Consortium (PCGC) and from healthy controls (unaffected siblings) from the Simons simplex collection (SSC) from Richter, et al.[15] We restricted our analysis to single-nucleotide (point) mutations within these cohorts.

For each cell type, we used cell-type specific BPNet models to predict the allelic impact of all mutations that were found within 1000 bp windows around summits of scATAC-seq peaks in that cell type. For each mutation, we used the BPNet model to predict the base-resolution read count profile corresponding to the input sequence (2,114 bp) containing the reference allele of the mutation at its center. We then used the model to predict the 1 kb base-resolution read count profile (which is decomposed into total predicted counts over 1 kb and base-resolution read probability profile) corresponding to the input sequence (2,114 bp) containing the alternate allele of the mutation at its center. Using these predicted read probability profiles from the two alleles, we computed the impact score of the mutation as the $\log_2$ fold change in cumulative probability between the reference allele and the alternate allele, over a 100 bp window around the mutation using the formula:

$$\log\left(\sum_{j=i-50}^{i+50}\left(P_j^{ref}\right) \Big/ \sum_{j=i-50}^{i+50}\left(P_j^{alt}\right)\right)$$

where i = position of mutation

P^{ref}_j = predicted profile probability at position j for sequence containing reference allele

P^{alt}_j = predicted profile probability at position j for sequence containing alternate allele

For each mutation, the cell-type specific impact scores were computed and averaged over cluster-specific BPNet models trained on each of 5-folds.

We also computed an alternate mutation impact score based on the predicted cumulative read counts over the 100 bp window around the mutation, instead of the predicted cumulative read probability.

$$\log\left(\sum_{j=i-50}^{i+50}\left(Y_j^{ref}\right) \Big/ \sum_{j=i-50}^{i+50}\left(Y_j^{alt}\right)\right)$$

where i = position of mutation

Y^{ref}_j = predicted counts at position j for sequence containing reference allele

Y^{alt}_j = predicted counts at position j for sequence containing alternate allele

We found high concordance of cell type specific enrichments of high impact mutations in cases vs. controls for both scores (Figure S6B).

Thresholding mutation impact scores to define high impact prioritized mutations

Because we are investigating a cohort of children with CHD born to parents without CHD, our expectation is that the some of these cases will be caused by *de novo* mutations. On average, each individual has approximately 70 such mutations,[15] and because we assume mutations that lead to CHD are generally rare, we would expect just one would be a causal presentation and we would expect only a fraction of the cohort to have such causal mutations. Based on the expectation that a small proportion of mutations from CHD cases in cell type resolved scATAC-seq peaks will have a causal role, we prioritized high-impact mutations in each cell type, as those that have an impact score >95th percentile of the distribution of cell-type specific impact scores of all mutations from the CHD cohort that fall in 1kb scATAC-seq peak regions in that cell type. The same thresholds were used for mutation impact scores of control mutations as well to obtain enrichments as specified below.

Selection of prioritized mutations in aEC for deeper investigation

We further restricted deeper investigation into a subset of higher confidence CHD mutations prioritized by the arterial endothelial cells (aEC) BPNet model to those that were within 200 bp (+/− 100 bp) of summits of aEC scATAC-seq peaks that had >75 reads in a +/− 250 bp window around mutation. For each of these selected mutations, we obtained predicted profiles for sequences centered at the mutation for both alleles as well as the corresponding DeepLIFT scores and active motif instances. The gene closest to the mutation in linear genomic sequence was assigned as the putative target gene of the mutation.

Cell-type specific enrichment analysis of prioritized mutations in cases relative to controls

To compute the enrichment of case vs. control mutations in scATAC-seq peaks (cREs) of each cell type in the fetal heart, we computed a 2 x 2 contingency table. The first axis splits all *de novo* mutations based on whether they were found in cases versus controls. The second axis splits all *de novo* mutations based on whether they overlap a cluster-specific peak. The enrichment p value and odds ratio (OR) was computed using the Fisher Exact Test implemented in the SciPy package in Python.

We used a similar procedure to estimate enrichment of *de novo* mutations prioritized by cell-type specific models from cases versus control. In this case, the first axis of the 2 x 2 contingency table splits all *de novo* mutations based on whether they were found in cases versus controls. The second axis splits all *de novo* mutations based on whether they are predicted to have a high impact score (>95^{th} percentile) or not using a cell-type specific BPNet model. High impact score mutations are pre-filtered to those in peak regions in the cell type. This analysis was performed for each cell type separately and for the pseudobulk of all cell types separately.

Enrichments of case and control mutations using mutation impact scores from the HeartENN model

We obtained mutation impact scores as computed by the authors of the HeartENN model for all non-coding *de novo* mutations in the PCGC case and SSC unaffected controls.[15] We retained the *de novo* mutations that overlap 1 kb scATAC-seq peak regions in any of the fetal heart cell types. Finally, we performed Fisher's exact test for enrichment of high impact (scores $\geq$ 0.1 as recommended in Richter et al.[15]) mutations in peaks in cases vs controls.

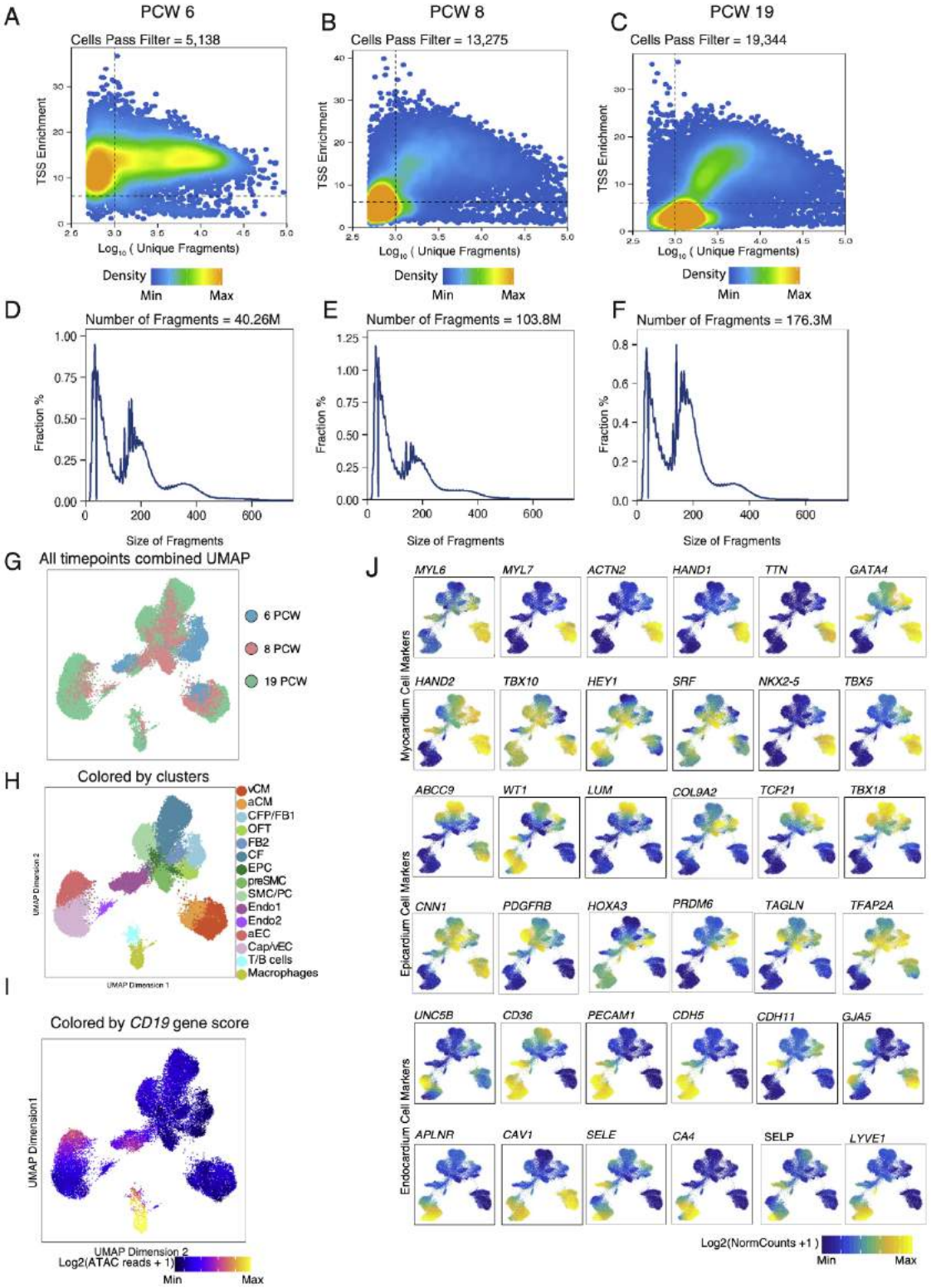
QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analyses were performed in R v4.0.5 or Python v3.8. Statistical tests are described in the relevant methods sections above.

ADDITIONAL RESOURCES

https://resgen.io/kundaje-lab/sundaram-2022/views/cardiogenesis.
https://cardiogenesis-atac.cells.ucsc.edu.

Supplemental figures



(legend on next page)

Figure S1. Quality control, clustering of cells and gene score of representative cell type markers for scATAC-seq data from fetal hearts at PCW 6 (left), PCW 8 (middle), and PCW19 (right), related to Figure 1

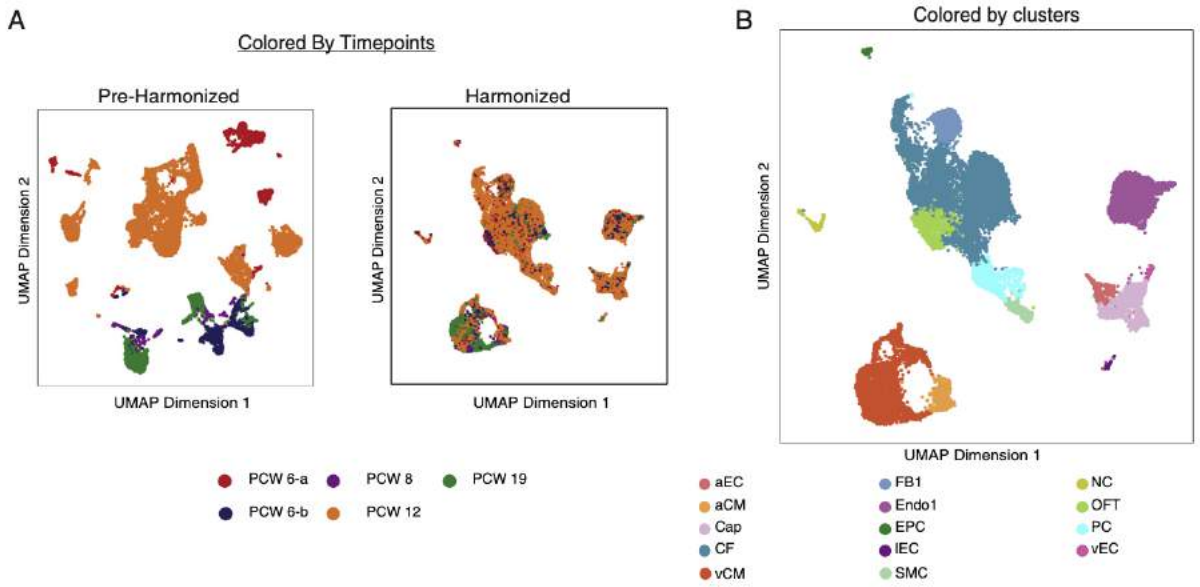
(A–C) Shown are the number of unique ATAC-seq nuclear fragments in each single cell (each dot) compared to TSS enrichment of all fragments in that cell. Dashed lines represent the thresholds for filtering cells (1,000 unique nuclear fragments and TSS score $\geq$ 6).

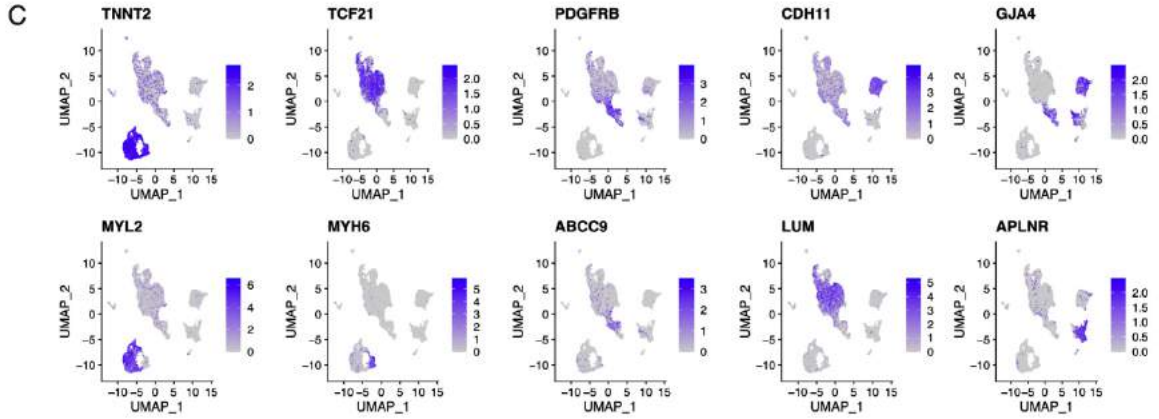
(D–F) The fragment length distribution for PCW 6 (left), PCW 8 (middle), and PCW19 (right).

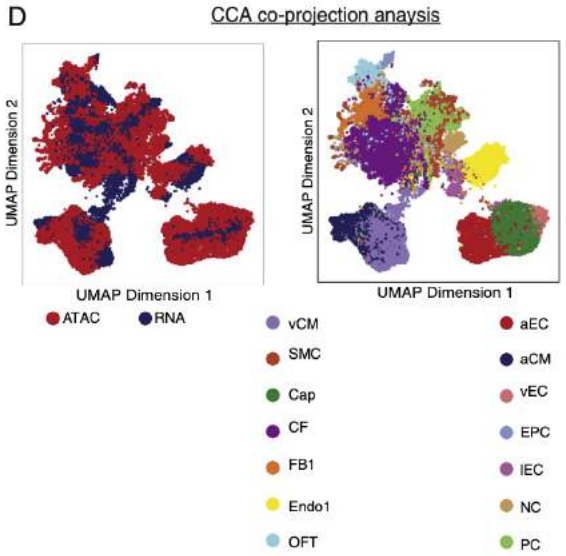
(G and H) UMAP of cells from three timepoints combined. Cells are colored according to (G) sample gestational time and (H) cluster membership.

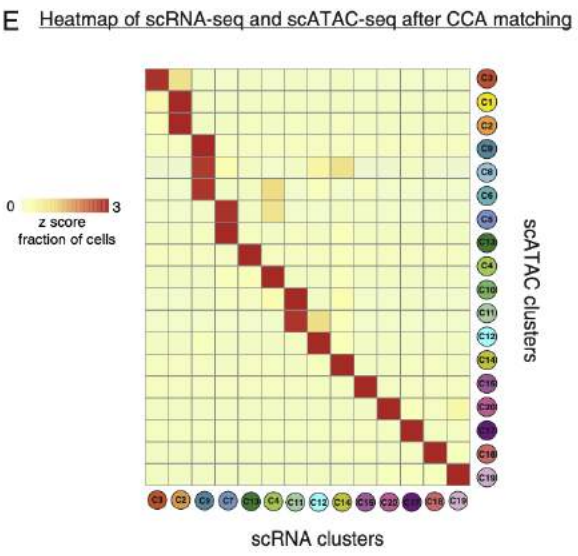
(I) scATAC-seq gene activity profiling of immune marker gene CD19.

(J) Units: $\log_2$(normalized ATAC gene-score). Scale: *MYL6* (min = 0.6,max = 1), *MYL7* (min = 0.25,max = 1.4), ACTN2 (min = 0.2,max = 1.2), *HAND1* (min = 0.4,max = 1.2),*TTN* (min = 0.4,max = 2.2),*GATA4* (min = 0.5,max = 1.6), *HAND2* (min = 0.5,max = 1.75),*TBX10* (min = 0.2,max = 0.8),*HEY1* (min = 0.8,max = 1.4), *SRF* (min = 1,max = 1.3), *NKX2-5* (min = 0.5,max = 2),*TBX5* (min = 0.2,max = 1), *ABCC9* (min = 0.15,max = 0.7), *WT1* (min = 0.4,max = 1),*LUM* (min = 0.05,max = 0.3), *COL9A2* (min = 0.2,max = 0.6), *TCF21* (min = 0.2,max = 1),*TBX18* (min = 0.2,max = 0.9), *CNN1* (min = 0.2,max = 0.6), *PDGFRB* (min = 0.4,max = 1.4), *HOXA3* (min = 0.2,max = 0.8), *PRDM6* (min = 0.2,max = 1), *TAGLN* (min = 0.2,max = 0.9),*TFAP2A* (min = 0.25,max = 0.7), *UNC5B* (min = 0.9,max = 1.4), *CD36* (min = 0.4,max = 1.2),*PECAM1* (min = 0.25,max = 1.25), *CDH5* (min = 0.3,max = 1.5), *CDH11* (min = 0.3,max = 1.5),*GJA5* (min = 0.2,max = 1), *APLNR* (min = 0.4,max = 1.5), *CAV1* (min = 0.2,max = 0.8),*SELE* (min = 0,max = 0.45), CA4 (min = 0.3,max = 1.1), *SELP* (min = 0,max = 0.25),*LYVE1* (min = 0.1,max = 0.6)









(legend on next page)

Figure S2. Integration of scRNA-seq and scATAC-seq data using canonical correlation analysis (CCA), related to Figure 1

(A) UMAP of cells from 5 scRNA-seq studies without (left) and with (right) batch effect correction and harmonization using Harmony (right). Cells are colored by the scRNA study of origin.

(B) Harmonized UMAP of scRNA-seq analysis used for downstream analysis. Cells are colored by clusters.

(C) Gene expression (Units: TP10K) of cell type specific and cluster specific markers in harmonized scRNA-seq UMAP.

(D) UMAPs of matched cells from scATAC-seq and scRNA-seq data modalities using the CCA subspace. On the left, cells are colored by their assay type and on the right, cells are colored by clusters from scRNA-seq.

(E) Heatmap showing the cluster–cluster mapping between scRNA-seq and scATAC-seq clusters after CCA matching.

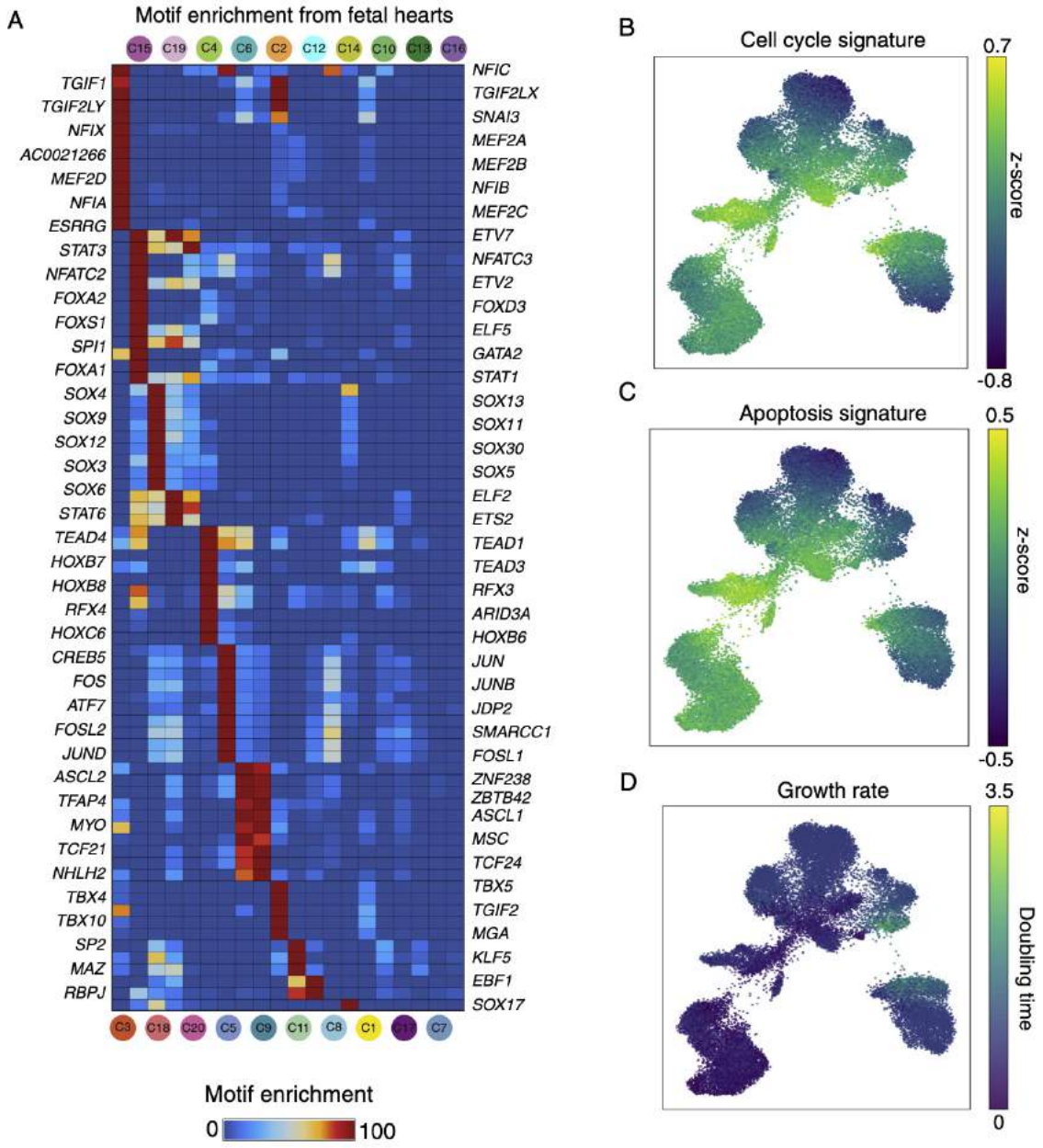
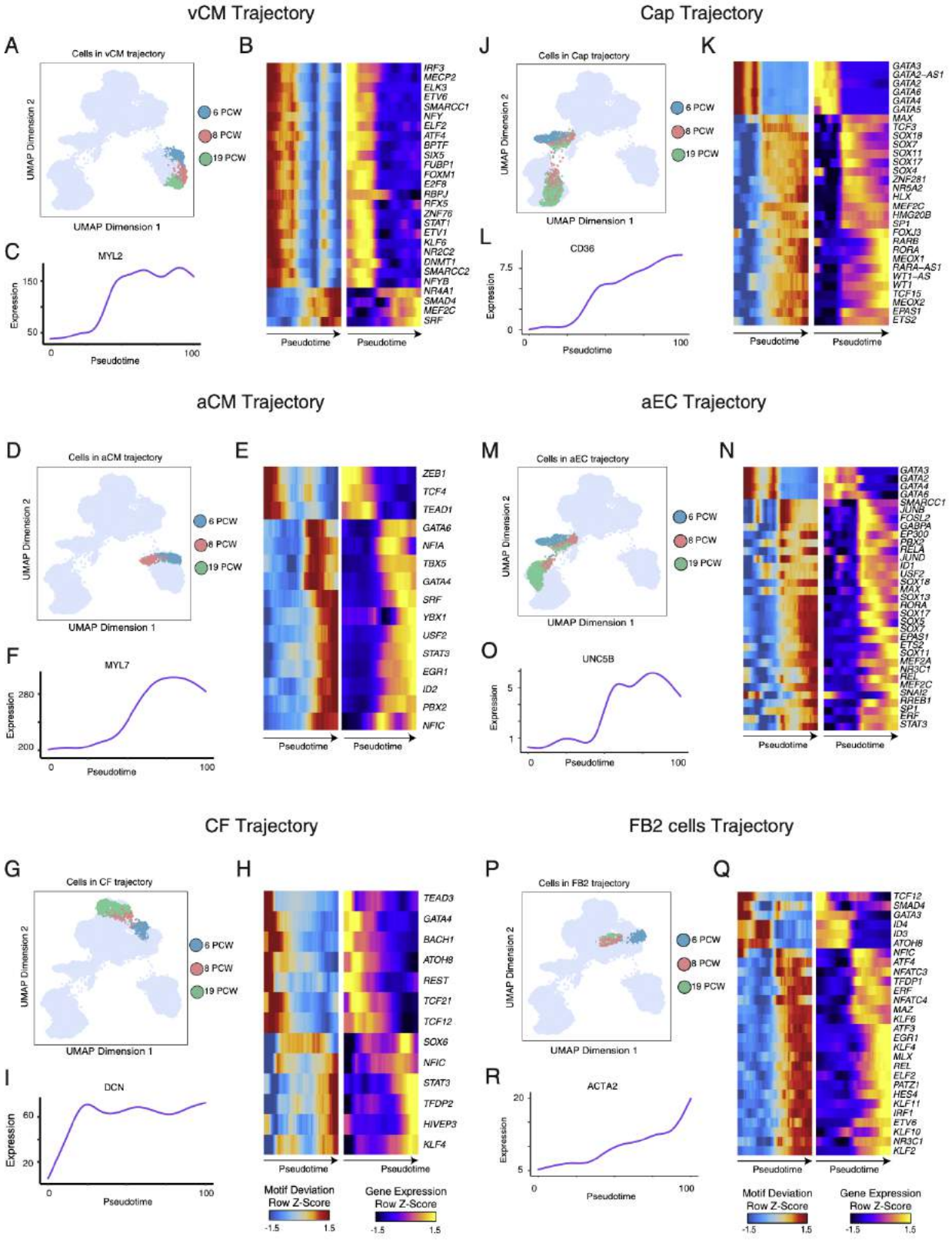


Figure S3. Overlap motif enrichment from fetal hearts and optimal transport cell signatures, related to Figures 2 and 3

(A) Overlap enrichment ($-\log_{10}$p adjusted) of position-weight matrix based motif instances in cell-type-specific marker scATAC-seq peaks of each cell type cluster from Figure 1E.

(B–D) UMAP of cells from scATAC-seq data showing (A) cell cycle signature *Z* scores, (B) apoptosis signature *Z* scores, and (C) growth rate estimates for optimal transport.



(legend on next page)

Figure S4. Optimal transport based developmental trajectories for vCM, aCM, CF, Cap, aEC, and FB2 cells using scATAC-seq, related to Figure 3.

(A) UMAPs of scATAC-seq cells in the ventricular cardiomyocyte (vCM) trajectory colored by the gestational sample time.

(B) Heatmaps showing *Z* score of ChromVAR motif deviation scores (left) and gene expression in units of $\log_2$(TP10K) (right) of TFs with correlated variable activity in cells identified to be in the vCM trajectory, as ordered by pseudotime.

(C) Expression dynamics of *MYL2*, an important marker gene for the vCM cell type.

(D–F) Trajectory analysis for atrial cardiomyocyte cluster (aCM), analysis as above.

(G–I) Trajectory analysis for cardiac fibroblast cluster (CF), as above.

(J–L) Trajectory analysis for capillary cells (Cap), as above.

(M–O) Trajectory analysis for arterial endothelial cell cluster (aEC), analysis as above.

(P–R) Trajectory analysis for Fibroblast like cells 2 (FB2), as above.

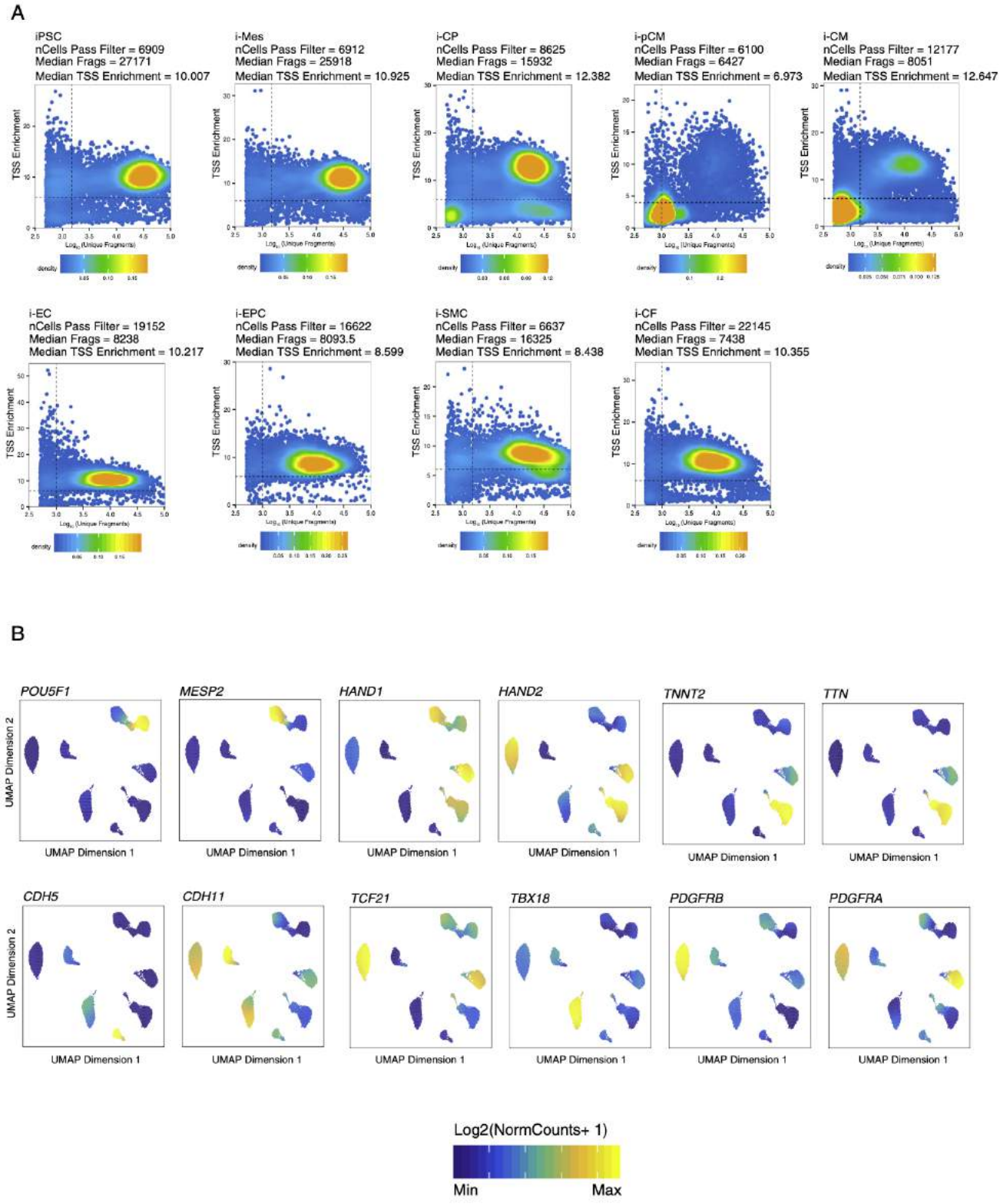


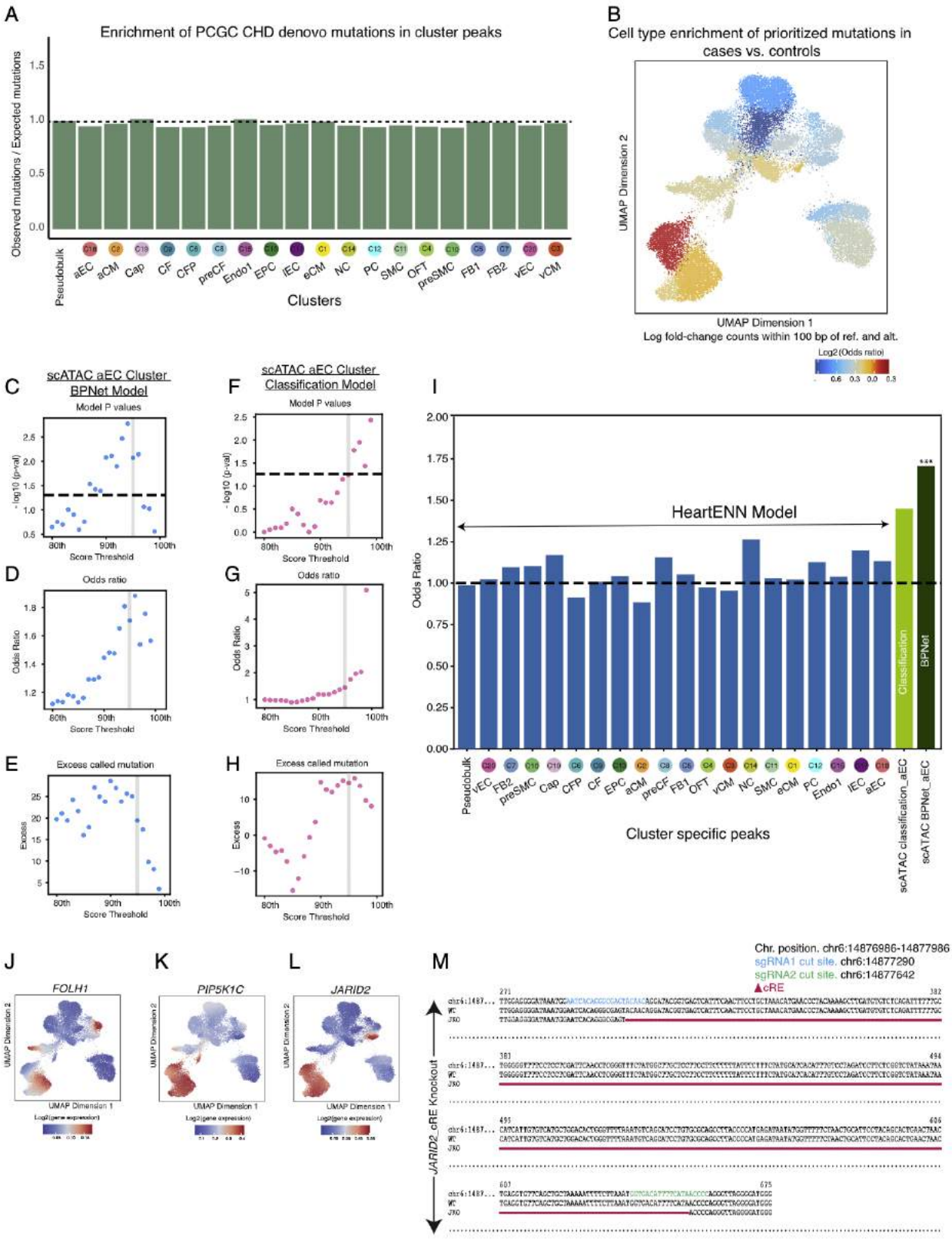
Figure S5. Quality control data and gene score of cell type markers for iPS derived cardiac cell types, related to Figures 4 and 5

(A) (Left to right, top to bottom) Representative scATAC-seq data quality control filters for Day 0, Day 2, Day 5, Day 15, Day 30 cardiomyocytes, Day 30 endothelial cells, Day 30 epicardial cells, Day 30 cardiac fibroblast cells, and Day 30 smooth muscle cells (top to bottom, left to right). Shown are the number of unique

(legend continued on next page)

ATAC-seq nuclear fragments in each single cell (each dot) compared to TSS enrichment of all fragments in that cell. Dashed lines represent the filters for high-quality single-cell data.

(B) UMAP plots showing gene scores of cell-type-specific and cluster specific markers. Units: log_2(normalized ATAC gene-score). Scale: *POU5F1* (min = 0,max = 0.7), *MESP2* (min = 0.25,max = 1.25), *HAND1* (min = 0.4,max = 1.6), *HAND2* (min = 0.8,max = 1.4), *TNNT2* (min = 0.25,max = 1.4), *TTN* (min = 0,max = 2), *CDH5* (min = 0.3,max = 1.5), *CDH11* (min = 0.4,max = 1.2), *TCF21* (min = 0.2,max = 0.9), *TBX18* (min = 0.4,max = 1), *PDGFRB* (min = 0.4,max = 1.2) & *PDGFRA* (min = 1.4,max = 2.2).



(legend on next page)

Figure S6. Prioritizing disease-associated non-coding variants using the cell-type-resolved scATAC-seq and predictive sequence models, related to Figures 6 and 7

(A) Enrichment of cases versus control mutations using naive overlap with cluster-specific ATAC-seq peaks, showing relevance of the deep learning model to capture pathogenic disruptions.

(B) Enrichment (log_2(OR) counts within +/− 50 bp, Fisher Exact Test) of prioritized mutations from each cell-type-specific BPNet model in CHD cases vs. controls plotted on the scATAC-seq UMAP of all fetal heart cells.

(C–E) Evaluation of robustness in disease prioritization of aEC model across different threshold values. (D) the -log_{10} (Fisher exact test p value), (E) the Fisher exact test odds ratio and (E) excess number of causal mutations observed in cases compared to controls are plotted across all threshold values.

(F–H) Similar metrics as (D–F) for a classification model with the same parameters as the BPNet model in aEC cluster.

(I) Barplot indicating the Fisher exact test odds ratio of the HeartENN model (Richter et al.[15]) subsetted to the *de novo* mutations in cases and controls overlapping cell type resolved peaksets (blue) scoring above 0.01 as recommended by (Richter et al.[15]) vs classification model in aEC cluster (light green) and BPNet model in aEC cluster (dark green). Stars indicate p values. (*** Fisher exact p value = 0.008).

(J–L) Gene expression of *FOLH1* (A), *PIP5K1C* (B) & *JARID2* (C) genes in UMAP of cells based on scATAC-seq data. Units: log_2(TP10K).

(M) Sanger sequencing confirms CRISPR/Cas9 targeted homozygous deletion in iPSC at the JARID2 cRE (red line).