Deep learning the dynamic regulatory sequence code of cardiac organoid differentiation

Eyal Metzl-Raz*[1], Ryan Zhao*[2], Salil Deshpande[3], Jackson Powell[1,5,10], Elizabeth G. Porter[1,5,10], Yassine Zouaghi[1,5,10], Betty B. Liu[6], Samuel H. Kim[7,8,9], Immanuel Abdi[1], Ivy Evergreen[1], Megha Agarwal[1,5,10], Maya U. Sheth[1,6,10], Joshua Rico[1,5,10,11], Matthew Miyamoto[1,5,10], Julie M. Sanchez[1,5,10,11], Jesse M. Engreitz[1,10,12,13,14], Anshul Kundaje[1,2], William J. Greenleaf[#1,15], Casey A. Gifford[#,1,5,10,11,12]

[1]Department of Genetics, Stanford University, Stanford, CA, USA.
[2]Department of Computer Science, Stanford University, Stanford, CA, USA.
[3]Department of Bioengineering, Stanford University, Stanford, CA, USA.
[4]Medical Scientist Training Program, Stanford University School of Medicine, Stanford, CA, USA.
[5]Division of Cardiology, Department of Pediatrics, Stanford University, Stanford, CA, USA.
[6]Department of Bioengineering, Stanford University, Stanford, CA, USA.
[7]Department of Biomedical Sciences, Stanford University, Stanford, CA, USA.
[8]Cancer Biology Program, Stanford University, Stanford, CA, USA.
[9]Medical Scientist Training Program, Stanford University, Stanford, CA, USA.
[10]Basic Science and Engineering (BASE) Initiative, Betty Irene Moore Children's Heart Center, Stanford Medicine Children's Health, Stanford, CA, USA.
[11]Institute for Stem Cell Biology and Regenerative Medicine, Stanford University, Stanford, CA, USA.
[12]Stanford Cardiovascular Institute, Stanford University, Stanford, CA, USA.
[13]Novo Nordisk Foundation Center for Genomic Mechanisms of Disease, Broad Institute of MIT and Harvard, Cambridge, MA, USA.
[14]Gene Regulation Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA.
[15]Department of Applied Physics, Stanford University, Stanford, CA, USA.
* These authors contributed equally to this work.
# Corresponding author

Abstract

Defining the temporal gene regulatory programs that drive human organogenesis is essential for understanding the origins of congenital disease. We combined a time-resolved, single-cell multi-omic atlas of human iPSC-derived cardiac organoids with deep learning models that predict chromatin accessibility from DNA sequence, enabling the discovery of the regulatory syntax underlying early heart development. This framework uncovered cell-state-specific rules of cardiogenesis, including context-dependent activities of TEAD, HAND, and TBX transcription factor families, and linked these motifs to their target genes. We identified distinct programs guiding lineage divergence, such as ventricular versus pacemaker cardiomyocytes, and validated predictions by perturbing Myocardin (*MYOCD*), establishing its essential role in ventricular specification. Integration of chromatin, transcriptional, and genetic data further highlighted regulatory regions and disease-associated variants that perturb differentiation state transitions, supporting evidence that suggests congenital heart disease emerges early in development. This work bridges developmental gene regulation with disease genetics, providing a foundation for mechanistic and therapeutic insights into congenital diseases.

Introduction

Organogenesis, the formation of functional organ systems, begins in the third week of human development and requires the coordinated differentiation of multiple cell lineages[1–3]. Dysregulation of the complex gene regulatory networks controlling this process can lead to congenital anomalies[4–6], with structural heart defects being the most common[7], affecting up to 3% of live births[8]. Despite being heritable, a significant proportion of congenital heart disease (CHD) cases remain genetically unsolved[9–12]. This failure of genetic attribution is likely because CHD arises from a complex combination of multiple coding and non-coding variants[7], and its origins lie in early developmental events[4,13,14], such as the critical paracrine signaling between nascent cell states[15–18], which are difficult to study *in vivo*.

To model the stepwise, multicellular events associated with organogenesis, human induced pluripotent stem cell (hiPSC) differentiation towards organoid models holds substantial promise[19]. These *in vitro* systems, which contain multiple organ-relevant cell types, allow for the study of the complex cellular interactions and paracrine signaling essential for organogenesis[20–22]. Compared with *in vivo* models, which are challenging to scale and genetically manipulate, and homotypic cell models, which lack multicellular interactions, organoids provide a tractable and scalable system. They enable the dissection of the temporal cis-regulatory architecture that governs cell identity and allow the functional testing of its role in human disease through targeted genetic perturbations[23,24].

Interpreting genetic perturbations at the level of gene regulation is currently challenging with existing experimental and computational approaches due to limited tools and a lack of resolution. Standard assays for mapping chromatin accessibility can identify active regulatory regions but lack the resolution to decipher discrete, complex regulatory grammar[25–28]. When paired with standard next-generation sequencing methods, the specific latent combinations, affinities, and spacing of transcription factor binding sites that dictate cell-type-specific gene expression patterns cannot be defined due to their clustered organization within the genome[29,30]. While techniques like ChIP-seq can infer TF occupancy[31–33], they are difficult to scale across all TFs and cellular contexts, and require prior knowledge of the TF of interest. However, deep learning models offer a powerful solution by effectively fusing the strengths of multiple assays[34–37]. They learn non-linear, salient predictive features at single-base-pair resolution using a corpus of genome or transcriptome-wide information[38–41]. This enables the de novo discovery of regulatory motifs and systematic interrogation of their syntax[14,41–43].

To identify complex TF motif grammar that governs the multicellular, nascent stages of human heart development and how its perturbation leads to CHD, here we

integrate advances in deep learning with a high-resolution cardiac organoid single-cell multi-modal atlas containing approximately 120,000 cells across six key differentiation time points. A base-pair resolution neural network learned complex regulatory motif grammar and then defined a compendium of syntax rules that guide cell fate, including a distinction between *HAND+TBX* dimers and *HAND+TEAD* dimers. To validate the model's predictions, we functionally tested and validated its nomination of *MYOCD* as a key driver of ventricular cardiomyocytes. This analysis also uncovered a non-autonomous effect that originates in cardiomyocytes but alters the ectoderm cell fate trajectory, highlighting how effective paracrine signaling is a crucial determinant of differentiation. We subsequently evaluated CHD-associated genes and variants using our *in vitro* and *in vivo* models, and observed that multiple intermediate, progenitor cell states are regulated and impacted by disease-associated gene regulatory networks. Ultimately, this work provides a powerful framework for decoding the molecular principles of early heart development and offers a valuable resource for investigating the complete genetic etiologies of congenital heart disease.

Results

Temporal sampling of cardioids reveals differentiation trajectories and lineage specification events

To map the regulatory landscape of early human heart development, we differentiated human induced pluripotent stem cells (hiPSCs) into cardiac organoids (Cardioids) by modulating key signaling pathways, including WNT, TGFβ, and retinoic acid[44,45] (Fig. 1A, S1A). The resulting cardioids successfully recapitulated key developmental milestones, transitioning from pluripotency (Day 0; *SOX2*, *OCT4*) through mesendoderm specification (Day 1; *TBXT*) and into early cardiac mesoderm (Day 3.5; *SRF*) and ectodermal states (Day 3.5; *NR2F2*). By day 5 and 10, the cardioids had developed into complex, multicellular structures containing cardiomyocytes (*MYL7*), endothelial cells (*PECAM1*), and non-neural ectoderm (*WNT5A*), among other lineages (Fig. 1A).

To resolve the cellular and molecular dynamics of this process, we generated a paired single-cell transcriptome and chromatin accessibility (scRNA/ATAC-seq) dataset from approximately 60,000 cells across six time points (days 0, 1, 2, 3, 5, and 10). We used Seurat[46] and DecontX[47,48] for preprocessing with RNA-contamination removal in the scRNA-seq data, followed by principal component analysis and UMAP embedding (Methods). Clustering analysis identified 20 distinct cell states, revealing a clear temporal progression of lineage specification (Fig. 1B-D, Fig. S1B-D). Early time points (days 0–2) were dominated by pluripotent (*NANOG*, *POU5F1*), epiblast (*NANOG*, *TBXT*) and mesendoderm (*TBXT*, *EOMES, MIXL1*) populations[44]. These progenitors

diversified at intermediate stages (days 3–5) into cardiac mesoderm (*NKX2-5*), non-neural ectoderm (*EYA1*), and hepatic progenitors (*HNF4A*). This transition from proliferative progenitors to differentiated cells (Fig. 1C) was further supported by EdU incorporation assays, which showed robust cell division at early stages that progressively decreased over time (Fig. S1E). By day 10, the atlas was composed primarily of diverse somatic-like cell types, including ventricular cardiomyocytes (vCMs; *MYL7*, *TTN*), pacemaker-like cardiomyocytes (pCMs; *SHOX2*, *TBX18*), endothelial cells (EC; *PECAM1*, *CDH5*), endoderm-derived hepatic-like cells (Endo; *AFP*, *HNF4A*), and non-neural ectoderm (Ecto; *TFAP2A*, *DLX5*) (Fig. S1F). The non-neural ectoderm state exhibited expression of the unique amniotic ectoderm marker *GABRP*, in addition to *VTCN1* and *TFAP2B*, which supports prior studies suggesting that the epiblast-derived non-neural ectoderm gives rise to amniotic ectoderm (AE)[49,50].

Having defined the cellular landscape, we next sought to connect cell identity to its underlying regulatory architecture. Performing k-means clustering of 628,378 consensus chromatin accessibility peaks (peaks shared across cell states out of 9,643,338 total cell-state-specific peaks) revealed dynamic, cell-state-specific patterns. We parsed this data into five main categories of chromatin accessibility: (i) consistently low accessibility in all cell types (n=40,850), (ii) TSS regions (n=11,689), (iiI) regions that are either increasing (n=36,270) or (iv) decreasing during differentiation regardless of the lineage (n=173,971), and (v) cell-type-specific accessibility (n=365,598) (Fig. 1E, left, Fig. S1G). We observed a qualitative correlation between the accessibility of these cis-regulatory elements (CREs) and the expression of nearby genes (Fig. 1E, right). This regulatory linkage was evident across a core set of 1,989 lineage-defining genes and transcription factors. The patterns included CREs that were accessible only in specific cell types, as well as those that were dynamic across all lineages during differentiation. Together, these analyses confirm that our cardioid system captures the rapid, lineage-specific regulatory events of early cardiogenesis, providing a high-resolution map connecting transcription factor activity, chromatin state, and gene expression.

A large, uncharacterized ZNF repertoire is dynamically expressed during cardiogenesis

The Zinc Finger (ZNF) protein superfamily, the most abundant class of transcription factors in the human genome, plays diverse roles as activators, repressors, and architectural proteins[51,52]. To systematically investigate their functions in cardiogenesis, we identified 502 ZNF genes with cell-state-specific chromatin accessibility across 14 developmental stages (Fig.1F, Fig. S1H, Table S1). A supervised AI text-mining approach (Note S5) identified a significant knowledge gap: 85% of these dynamically regulated ZNFs are poorly characterized or lack known functions, with fewer than 10%

being well understood (Fig. S1H). Structurally, many of these ZNFs contain domains consistent with transcriptional repression.

Despite the large number of uncharacterized factors, the mapping of ZNFs to cell states was validated by demonstrating the cell-state-specific activity of well-established cardiac regulators, such as the essential cardiac progenitor factor and *GATA4* cofactor *ZFPM2* (also called *FOG2*)[53] (Fig. S1K), which we assigned to early and mid-mesoderm clusters, marking its importance in very early mesoderm development. This framework enables the nomination of candidates for functional interrogation. Among these are *CASZ1*, a cardiac transcription factor essential for differentiation and proliferation, with mutations linked to congenital heart defects and dilated/arrhythmogenic cardiomyopathy[54–60], specifically assigned to the early developing vCMs (Fig. S1K). Another candidate for functional interrogation is the mid-mesoderm-expressed *ZNF516*, a KRAB-ZNF that suppresses EGFR signaling[61]. This under-characterized gene is a potential repressor that ensures cardiac lineage fidelity. These examples demonstrate the strength of our integrated approach in generating specific, data-driven hypotheses about the roles of previously uncharacterized factors in human cardiogenesis.

Out of the 502 ZNFs, 13 possess a functionally validated repressor domain[62] and exhibited cell-state-specific expression profiles (Fig. 1G, Fig. S1I-J): (i) ZNFs active during early differentiation stages (e.g., *ZNF7*, *ZIK1*), (ii) those with intermediate activity (e.g., *ZNF2*, *ZNF82*), and (iii) ZNFs active during later differentiation stages (e.g., *ZNF84*, *ZFP30*). Additionally, some ZNFs showed lineage-specific activity, such as *ZNF81*, mainly active in Pacemaker CMs; *ZNF34*, active during later stages of the Endoderm lineage and early differentiation; and *ZNF2*, *ZNF45*, *ZFP82*, which display lower expression in AE. Our analysis highlights these ZNFs, most of which have unknown functions (12 out of 13), as potential key regulators across different stages of differentiation.

Cardioids recapitulate human organogenesis and cardiac lineage specification trajectories

To demonstrate that the cardioid model broadly recapitulates early human development, we projected our scRNA-seq data onto a reference atlas of the human gastrula (Fig. 2A-B)[3,46]. This comparison revealed a strong alignment between our *in vitro* cell states and key *in vivo* stages of embryogenesis, including the epiblast, primitive streak, emergent mesoderm, and advanced mesoderm (Fig. S2A). Notably, the earliest time points of differentiation (days 0–3) showed the highest mapping scores to the gastrula reference, with the correlation decreasing at later stages as cells progressed toward more specified fates (Fig. 2B). These results establish that the

cardioid system is an accurate model for investigating the specific and often inaccessible timeframe of human gastrulation and nascent organogenesis.

Having confirmed the relevance of our model to early embryogenesis, we next sought to reconstruct the developmental trajectories that lead from pluripotent cells to differentiated lineages. To achieve this, we applied Waddington-OT (WOT), a computational approach that infers lineage relationships by calculating the most probable fate of each cell between successive time points[63]. This analysis quantitatively revealed differentiation paths into distinct lineages (Fig. 2C, S2B). For example, tracing a continuous progression from pluripotency through mesoderm (days 1-2), into pre-cardiac mesoderm (day 3), and ultimately to ventricular cardiomyocytes by day 10.

This lineage tracing demonstrated that lineage-specific gene expression profiles necessary for proper differentiation occur as early as the first day of differentiation, with a higher tendency toward mesodermal lineages, as expected (Fig. 2D, Fig. S2C). By day 2, distinct progenitor populations were already predicted to give rise to ventricular cardiomyocytes (vCMs), AE and hepatocyte-like cells. By day 3, many of these cells show strong commitment to their terminal fate (Fig. 2E-F, Fig. S2D), expressing increasingly more progressive gene programs (Fig. S2E). For pCMs, we observed shared cell states with the vCM trajectory (Mesoderm and Pre-cardiac Mesoderm) with a day 5 trajectory divergence (Fig. S2B-C). This illustrates a later differentiation divergence point from a common unknown progenitor state shared with the vCM state. The early lineage specification observed in our model aligns with gene expression signatures observed during human gastrulation *in vivo,* underscoring the model's capacity to accurately recapitulate gene expression programs that govern cell fate decisions.

Deep learning models decipher the regulatory sequence syntax of cardiac lineage commitment

To decipher the sequence syntax that mediates regulation of chromatin accessibility by TFs defining lineage commitment, we trained ChromBPNet deep learning models to predict base-pair resolution, cell-state-specific chromatin accessibility profiles directly from DNA sequence. ChromBPNet also corrects distortions in ATAC-seq profiles caused by sequence biases of the Tn5 enzyme used in the experiment[38]. Model interpretation techniques can be applied to assess the predictive contribution of every nucleotide in each regulatory element to its predicted accessibility profiles, revealing TF motifs and their syntax rules which govern cell-state-specific accessibility patterns[64,65]. This approach overcomes the limitations of traditional motif scanning by learning the quantitative effects of *de novo* discovered motifs on cell-state-resolved accessibility conditioned on their local genomic sequence context.

We first trained cell-state-resolved models for each of the clusters identified across cardioid differentiation (Fig. 3A). To assess the accuracy of our models, we computed the Pearson correlation between the observed and predicted total counts for each cell state across all peaks in that cell state. The model correlations ranged from 0.78 to 0.88, with a median of 0.85 (Fig. S3A-B). We also computed the Jensen-Shannon Divergence (JSD) between the observed and predicted base-resolution distribution of counts (profile shape) within each peak. The median JSD (smaller is better) across all peaks ranged from 0.47 to 0.73 across all models (Fig. S3A). As an illustrative example, we observe that our ventricular cardiomyocyte model accurately recapitulates the complex accessibility pattern at the promoter of *NKX2-5*, a key cardiac regulator (Fig. 3B, left).

We next used the DeepLIFT method to interrogate each model and estimate sequence contribution scores for all peaks in each cell state[64,66]. We then used the TF-MODISCO motif discovery method to infer all *de novo*, predictive sequence features, corresponding to TF motifs, learned by each of the models[65]. Unifying motifs learned by all models yielded a comprehensive regulatory lexicon of 418 individual DNA sequence patterns, consisting of 172 distinct motifs from 106 TF families (MoitfCompendium[67]). Motif families were defined as sets of motifs having very similar core sequence preferences with variable flanking bases (Methods). The models' cell-state specificity is evident, for instance, when inspecting the *NKX2-5* promoter, at which the predictive binding sites for vCM-specific factors like MEIS/TBX heterodimers and SRF are uniquely present in vCM models but absent in models from all other lineages (Fig. 3B, right).

We next used this unified motif lexicon to understand the dynamic activity of globally predictive motifs and their associated TFs across all cell states. As expected, analysis of cell-state-specific peaks confirmed the enrichment of known lineage-driving TFs: SMAD and TBX motifs were enriched in mesendoderm, MEF2 and NKX motifs in cardiac mesoderm, and HNF and TFAP2 motifs in endoderm and ectoderm, respectively (Fig. 3C, Fig. S3C).

Next, we used the cell-state-resolved sequence contributions to identify dynamic patterns of combinatorial TF occupancy at predictive motif instances mapped to the unified motif lexicon in all peaks across all cell types (Methods). Peaks encoded on average eight predictive motif instances, resulting in a more precise and specific annotation of motif syntax compared to traditional motif annotation methods that identified an excessive (90 per peak on average) number of motif instances (Fig. S3D). We further quantified the total number of motif instances across all peaks in each cell state and found both shared and cell-state-specific TF motifs (Fig. 3C). Thus, our models reveal local and global occupancy patterns of TFs through predictive motifs

syntax, mediating regulation of accessibility dynamics across the differentiation time course.

Integrating dynamic motif syntax and TF expression resolves context-dependent regulation of accessibility by distinct TF family members

The precise regulation of gene expression is governed by a complex interplay between TFs and cis-regulatory elements (CREs). A gene can be controlled by different TFs binding to multiple distinct CREs, or by different TFs acting on the same CRE in a cell-state-specific manner. The latter, which we term context-dependent cis-regulatory elements (cdCREs), are particularly challenging to decipher. Our integrative framework can resolve both types of regulatory logic. The cdCRE upstream of *DHRS3*, a gene involved in regulating retinoic acid levels during embryogenesis[68], serves as a clear example. While this single element is accessible in multiple lineages, our model predicts that distinct TF occupancy patterns contribute to its accessibility in each context, via pleiotropic motif syntax (Fig. 3E). In vCMs, NKX2 motifs contribute to the observed accessibility. In contrast, in AE, TFAP2 motifs contribute to the observed accessibility, while TEAD motifs contribute similarly in both cell states. Examining profiles of expressed TFs from the NKX family nominates more specific TF family members likely to bind the predicted motifs. In the vCMs, *NKX2-5* is the highest expressed gene in the NKX2 family, aligning with its known role as a critical regulator for proper heart development. In AE, other members of the NKX2 family are expressed (*NKX2-3*, *NKX2-6*, *NKX3-1*), although they are not predicted to contribute to the *DHRS3* regulatory element accessibility (Fig. S4).

We extended this analysis to examine a broader set of peaks accessible across multiple terminal lineage states (Fig. 3C, light orange fourth row). Despite exhibiting similar accessibility across these cell states, the most prevalent predicted drivers of these peaks were lineage-specific. For example, the NKX2 motif family was the most prevalent motif in these peaks in vCMs and was not a key driver in any other lineage. Similarly, TFAP2 was the largest driver in and specific to AE, POU:SOX was specific to hiPSCs and early states across all lineages, and FOX showed the highest presence in endoderm-derived cells (Fig. 3D, Fig. S3E). These motif dynamics are consistent with lineage-specific TF expression patterns (Fig. S4, Fig. S5). These results highlight the models' ability to disentangle activity from regulatory logic, suggesting that cdCREs with similar accessibility across cell states may encode pleiotropic motif syntax reflecting context-dependent regulation by distinct TFs.

Our framework can also resolve the ambiguity of multiple TFs from the same family binding to a common motif. The FOX family, which contains many members implicated in congenital heart disease, illustrates this challenge. While the canonical FOX motif is active in nearly all cell states, by integrating our model's FOX motif predictions (Fig. 3D) with single-cell gene expression (Fig. S5), we can nominate the specific FOX gene family members responsible for this activity in a cell-state-specific manner. For example, we find that the congenital heart disease-associated genes *FOXP1* and *FOXF1* are specifically expressed in vCMs, whereas other family members like *FOXM1* and *FOXO1* are transcriptionally inactive in the same cells (Fig. S5). These examples demonstrate our framework's ability to link transcription factor expression to motif activity in a cell-state-specific manner.

When comparing cell-state-specific motif proportional composition, different cell states might show similar composition of attributing motifs (Fig. S3F, top). When summarizing and focusing on our annotated motifs, we are able to identify specific cell states where the motifs are predicted to contribute to the observed accessibility (Fig. S3F, bottom). This suggests that the proportional usage of the regulatory vocabulary is conserved across developmental lineages. This finding points to a fundamental syntax pattern of the genome, where different cell states draw upon a shared toolbox of TFs, but in distinct levels and proportions, to achieve specialized functions.

in silico marginalization quantifies context-dependent motif accessibility

A central goal in developmental genomics is to resolve how individual TFs contribute to gene regulation within complex, multi-motif enhancers and across cell states. To address this, we used an *in silico* marginalization strategy[35,69]. This technique computes an estimate of a motif's impact on chromatin accessibility in each cell state by using the corresponding model to predict the accessibility of background genomic regions both with and without the motif present, thereby measuring the marginal effect of the motif itself. This approach measures the isolated effect of each motif, unconfounded by the effects of neighboring motifs present in accessible regions. Furthermore, this allows us to accurately predict the accessibility of rare or sparse motifs, which are often overlooked by conventional methods that rely on statistical enrichment.

The MEF2 family provides a clear case study. While MEF2 motifs are abundant in mesodermal peaks and several gene family members (*MEF2A*, *MEF2C*, *MEF2D*) are expressed in both vCM and AE lineages, our analysis revealed that MEF2 motif accessibility is restricted to the vCM lineage. We predict a wave of MEF2 accessibility that begins in pre-cardiac mesoderm and peaks in vCMs, which correlates with the

sequential expression of *MEF2A*, *MEF2D*, and *MEF2C* (Fig. 3F, upper panel). In contrast, our model does not predict MEF2 motif accessibility in the AE lineage despite detectable expression of MEF genes (Fig. 3F, lower panel). This demonstrates our approach's ability to discriminate between TFs that are expressed and those that are functionally active in shaping the chromatin landscape.

Modeling deciphers the context-dependent grammar of TF cooperativity

Extending the *in silico* marginalization strategy beyond individual TFs, our analysis can decipher the complex "regulatory grammar" that governs TF cooperativity. The model predicts, for instance, that a TBX homodimer motif exhibits strong synergistic chromatin activity when compared to a TBX monomer, while a tandem POU:SOX motif provides no increased chromatin activity over a monomer (Fig. 3G, Fig. S3G).

The TEAD motif provides an instructive case study in predicted context-dependent regulatory syntax. While the core TEAD motif is present across most cell states, its contribution to accessibility becomes highly specific when paired with other TFs' motifs (Fig. 3H, Fig. S6). The model predicts that the TEAD motif's chromatin activity is dictated by its partner motif: it cooperates with SRF motifs in cardiomyocytes, with GRHL in the ectoderm, and with HNF4 in the endoderm. The model's precision extends even to the base-pair spacing between motifs. For example, it is predicted that a HAND and TEAD motif pair has the highest chromatin activity when separated by exactly three base pairs, demonstrating a highly sensitive ability to resolve the subtle syntactical rules that define lineage-specific enhancer function (Fig. 3H). Other prominent TF families exhibit the same sequence-specific activity patterns (Fig. S6). This capacity to predict how subtle changes in motif syntax dictate cell-state-specific activity provides a powerful new lens for interpreting the non-coding genome.

A gene-centric framework links regulatory syntax to developmental cascades

To translate our understanding of motif syntax into gene-level regulation, we computationally linked active regulatory elements to their putative gene targets. We used the scE2G method[70] to assign our 628,378 consensus peaks to genes, which resulted in 1,256,939 scE2G individual peak-gene links across all cell states (Fig. S7A). We intersected the model's motif accessibility predictions with these region-to-gene links, calculated a cumulative contribution score for each motif to each gene and ranked the potential effects of motifs on gene regulation based on this score (Methods). This

analysis generates a cumulative contribution score for each motif on each gene in each identified cell state. (Fig. 4A, Fig. S7B).

We first validated this approach by using it to reconstruct the characterized regulatory cascade of vCM specification (Fig. 4B). To visualize general regulatory contributions of motifs during this differentiation, we average our gene-by-motif contribution across cell states matrix for all lineage-defining genes (Fig. 1E). The resulting average contributions of motif by cell state matrix accurately recapitulates known sequential regulatory changes over developmental time: from pluripotency factors (POU, SOX) in hiPSCs, specification-associated TFs (TBX, EOMES, SMAD) in epiblast-like cells, to mesendoderm-associated TFs (GATA, HAND) in early progenitors, and finally to the canonical cardiomyocyte TFs (SRF, NKX2-5, and later MEF2 and NFI) that orchestrate later stage differentiation *in vivo*. Therefore, the predicted TF activities we observe faithfully recapitulate the known principles of vertebrate heart development.

Expanding this analysis genome-wide and across differentiation trajectories, we generated a global map of motif-to-gene regulation across all cell states (Fig. 4C, Fig. S7B). This map captures the relative contributions to accessibility of both transcriptional activators and transcriptional repressors, such as *ZEB1/2* and *SNAI*. These repressors are specifically expressed in the mesodermic lineage, where our models assigned their motifs significantly lower contribution scores, consistent with their role in binding and compacting chromatin. The decrease in accessibility of ZEB/SNAI motifs in the mesodermal lineage is reflected in the Endothelial-to-Mesenchymal Transition (EndMT) genes, such as *CDH1*, *EPCAM*, *CRB3*, and *CLDN*, which were all downregulated in the mesoderm. Additionally, genes known to be activated by *ZEB/SANI*, such as *ACTA2*, *VIM* and *FN1*, were upregulated (Fig. S8A).

An integrative framework to define the regulatory programs of cardiac subtypes

Congenital heart disease can arise from the misregulation of highly specific cell subtypes, yet resolving their distinct regulatory programs has been a major challenge. Our framework provides a powerful solution, offering the ability to define the precise regulatory elements, from active enhancers to governing transcription factors, for any given cell state. By applying this framework to dissect the syntax of ventricular (vCM) and pacemaker (pCM) cardiomyocytes, we can directly investigate the programs controlling critical disease-associated genes. To identify the key TFs that drive this bifurcation, we focused on genes in their respective progenitor states that were both differentially expressed and linked to differentially accessible peaks (Fig. 4D), then

calculated the differential contributions of all TF motifs associated with this gene set (Fig. 4E).

Our analysis revealed that pCM progenitors exhibit high predicted contribution from the GATA, PDX/HOX, and LHX/SHOX motif families. These motifs were linked to key pacemaker identity genes, including *FBN1* and *TWIST1*, which are associated with congenital syndromes[71–74], and were highly accessible at the enhancers of other pCM-specific genes like *LHX2*, *TBX18* and the ion channel *KCNJ3* (Fig. 4F, Fig. S7D, Fig. S7F). In contrast, vCM progenitors exhibited strong contributions from MEF2, NKX2-5, and SRF motifs. These TF motifs were linked to vCM-specific targets like *BMP2*, as well as genes like *TTN*, *ANKRD1* and *MYH6*, which are linked to various cardiomyopathies, confirming that a distinct and clinically relevant regulatory syntax underlies the two progenitor states (Fig. 4G, Fig. S7E).

Functional validation reveals MYOCD is essential for ventricular cardiomyocyte specification

Based on our model's prediction that the MYOCD-SRF network is a key vCM-specific driver (Fig. 4B-C), we sought to functionally understand the dysregulation mechanism upon genetic perturbation. We used a CRISPR-interference (CRISPRi) system to knock down (KD) *MYOCD*, a lineage-specific coactivator for the ubiquitously expressed *SRF* (Fig. 5A-C; Fig. S9F). The perturbation resulted in a significant reduction in canonical *in vitro* vCMs by day 10, did not affect overall organoid size (Fig. 5B) and led to a non-significant decrease in cardioid overall cell numbers, suggesting that the perturbation induced a switch in cell fate rather than a general defect in cell viability or proliferation (Fig. S9A).

To identify the molecular dynamics affected by *MYOCD* KD, we generated single-cell multiomic data from control-matched knockdown cardioids at days 0-1, 1.5-2, 3, 5, and 10. We observed distinct alternative cell states, specifically in the vCM and AE lineages, where *MYOCD* KD resulted in alternative developing CM cells that did not cluster with the Control vCMs or any other cell state. In the AE lineage, we observed a slight shift in the Early AE state, resulting in a lack of the Late AE subcluster (Fig. 5D). To better identify the altered states, we projected control and knockdown cells onto a murine embryonic development atlas encompassing E8 to P0 embryos[75]. This revealed a qualitative divergence beginning on day 3 (Fig. 5E; Fig. S9D-E). While control mesodermal cells progressed toward a first heart field (FHF) ventricular state, *MYOCD* KD cells were arrested in a progenitor state most similar to the second heart field (SHF) (Fig. 5E; Fig. S9B). Additionally, projecting the control and *MYOCD* KD cells onto the mouse developmental atlas's fetal time confirmed that *MYOCD* KD CMs exclusively

mapped to the E8-E8.5 stage, whereas control CMs mapped mainly to the later developmental stages of E18.5-P0 (Fig. S9C). This confirms that *MYOCD* is essential for proper FHF-derived cardiomyocyte specification and that its loss causes a developmental stall, extending previous *in vivo* observations[76].

Differential gene expression analysis pinpointed the timing of the molecular effects induced by *MYOCD* KD. Through day 2, the only significantly changed gene was *MYOCD* itself (Fig. S9G, “mesoderm”), but by day 3, significant transcriptomic changes appeared, almost exclusively in the precardiac mesoderm (Fig.5F, left; Fig. S9G). These changes included a significant downregulation of genes associated with heart development and similar Gene Ontology (GO) terms (Fig. 5F, right), directly linking the perturbation to the molecular programs implicated in heart disease (Fig. 5G). Intriguingly, this analysis also suggested a non-cell autonomous effect of *MYOCD* KD on the amniotic ectoderm (AE) lineage, which exhibited significant transcriptomic changes, suggesting a previously unappreciated signaling axis between the developing mesoderm and ectoderm.

MYOCD loss perturbs cardiac and ectodermal regulatory programs

To identify the core regulatory networks disrupted by *MYOCD* KD, we analyzed motif enrichment within chromatin regions that lost or gained accessibility upon KD (Fig. 5H). For each cell state, we performed k-means clustering (k=3) based on the $\log_2$ fold change ($\log_2$FC) between Control and *MYOCD* KD in that specific cell type, aiming for three major chromatin accessibility profiles: (i) decreased accessibility with knockdown, (ii) unchanged accessibility, and (iii) increased accessibility with knockdown. Then, for each motif family, we conducted Fisher’s exact test to assess motif enrichment between the k-cluster showing increased accessibility in KD and the k-cluster showing decreased accessibility. Only active peaks (peaks with at least one motif within +/- 500bp) were considered for each cluster. This analysis revealed an apparent dichotomy in the cardiomyocyte lineage. Regions that *gained* accessibility were enriched for general promoter GC-rich motifs, likely reflecting indirect gene regulation of the observed stalled developmental state. In contrast, chromatin regions that *lost* accessibility were most significantly enriched for SRF and MEF2 motifs (Fig. 5H, grey vertical bars). This result provides direct evidence that the primary autonomous effect of the perturbation is the targeted disruption of the *MYOCD*-*SRF*/*MEF2* regulatory network, which is essential for cardiac development[76].

Next, we investigated the regulatory basis for the phenotype observed in the amniotic ectoderm (AE). Our analysis revealed that cardiac TFs did not directly drive the

disruption in this lineage, but instead, the most significant perturbation was in the activity of the PDX/HOX transcription factor family (Fig. 5H). The indirect effect of *MYOCD* KD on PDX/HOX motifs suggests that defects in cardiomyocyte lineage differentiation may lead to cell non-autonomous perturbed signaling in the AE lineage. Indeed, our data indicate that cardiomyocytes ectopically express the signaling molecule *BMP2*, while AE cells express its receptor, *BMPR1A* (Fig. S9H). Given that BMP signaling is a known inducer of HOX gene expression[77–79], these findings suggest a paracrine mechanism: aberrant BMP2 signaling from the perturbed mesoderm acts on the ectoderm, where it converges on the PDX/HOX regulatory network to mediate the non-autonomous effect.

Disrupted regulatory syntax at the *MYH7* locus aligns with independent experimental validation

To understand precisely how *MYOCD* KD leads to silencing of its target genes, we examined the regulatory syntax at the cardiomyocyte-specific gene *MYH7*. Our model revealed a significant disruption of the regulatory program at the onset of gene expression. In control cells, *MYH7* is driven by a coordinated activity of TFs (e.g., MEIS, MEF2, HAND and SRF). In knockdown cells, key activators like MEIS and MEF2 lose their predicted activity, though they remain expressed (Fig. 5I-J). This regulatory imbalance provides a mechanistic view of how a gene's regulatory landscape is altered upon *MYOCD* KD.

Independent experimental data validate model predictions

To validate our model's predictions of contributive motif instances to *MYOCD* regulation, we compared our findings to a recent *in vivo* reporter assay of human heart enhancers[80]. Applying our scE2G analysis on the *in vivo* tested enhancers, one enhancer ("HT2") was linked to the putative target gene *SSH2*, a phosphatase that regulates the cytoskeleton and is critical for cell migration and morphogenesis. Our model predicts that the primary downstream effect of *MYOCD* KD is a sharp reduction in the regulatory activity of the enhancer key cofactors, MEF2 and SRF (Fig. S10A).

The Kosicki et al.[80] study provides a direct experimental test of this prediction. They found that experimentally mutating the MEF2 or SRF motifs within the HT2 enhancer resulted in a "Full loss" or "Major loss" of reporter gene expression in the mouse heart. This provides an *in vivo* phenotypic readout for the molecular event our model predicts. Furthermore, their experiment showed that reducing MEF2 activity did not affect a nearby GATA motif, which mirrors our model's prediction that this GATA motif remains unperturbed following MYOCD loss (Fig. S10B). This concordance

provides strong external validation that our model accurately identifies the specific TF motifs driving the downstream effects of a perturbation at disease-relevant enhancers linked to critical developmental genes.

Noncoding variants associated with CHD genes have cell-state-specific predicted effects

Congenital heart disease (CHD) encompasses a broad spectrum of structural malformations present at birth, most of which originate during the first four weeks of embryogenesis. Genetic etiologies involving variants that perturb these early developmental processes remain difficult to define and experimentally validate owing to the limited availability of physiologically relevant models[9,81,82]. This challenge is especially pronounced for noncoding variants, which are predicted to exert effects in a highly cell state–specific manner[42,83,84].

To identify cell states disrupted in CHD, we evaluated the gene expression pattern of 410 genes associated with CHD through cohort studies[70]. We first combined them into expression modules via k-means clustering (Fig. 6A, Fig. S11A-B) and confirmed that 88% of genes are expressed across multiple lineages and trajectories in our data, including progenitor states. This result reiterates the need to expand the cell states used for variant adjudication and functional genomics studies, moving away from employing somatic cell states that fail to reflect the cause of disease.

To define state and/or lineage-specific gene regulatory logic for CHD genes expressed across multiple states, we next integrated cell state-specific motif hits derived from ChromBPNet models with scE2G links. This analysis identified 6,607 unique putative CREs linked to known disease genes, with a median of 12 linked peaks per gene (Fig. S11C). *NOTCH1* is a prominent CHD gene associated with myriad structural defects of disparate origins[85–87] and was expressed early in differentiation in all lineages (Fig. S11D). While most prominently studied in endothelial cells[88], *in vivo* studies have suggested *NOTCH1* is additionally required for second heart field development[89] and neural crest-derived smooth muscle differentiation[90]. *NOTCH1* is also required for the diversification of cells derived from the pharyngeal pouch endoderm[91], whose proper differentiation to FGF-secreting cells is critical for outflow tract and aortic arch development in the heart[92]. Indeed, this analysis identified 33 unique peaks predicted to regulate *NOTCH1* across all cell states, highlighting cell state-specific putative CREs (Fig. 6B). Given our aforementioned observations that peak level analyses obscure TF-driven regulatory logic, we next defined the cell state-specific *NOTCH1* motif syntax based on ChromBPNet nominations with scE2G-linked peaks (Fig. 6C, Fig. S11E). This analysis defined discrete roles for TF families along the differentiation trajectories; a POU-SOX role transitioned to GATA/HAND role in the mesoderm and cardiac

progenitors, and finally the MEIS TF family in cardiomyocytes, while nominating the FOX/TEAD/LHX families in the endoderm (Fig. 6C-D). To predict the effect of the nominated motifs, we next quantitatively evaluated the effect of mutating single base pairs that ChromBPNet identified as necessary for accessibility, which revealed cell-state-specific effects on accessibility, supporting the motif nominations (Fig. 6E). This result highlights that, unlike coding variants, which typically affect all isoforms of a gene, noncoding variants in CREs can perturb gene expression in a cell-type- or stage-specific manner. Consequently, they may give rise to phenotypes that diverge from those observed in individuals with coding variants in the same gene.

To identify noncoding variants associated with CHD that are likely to perturb differentiation, we next quantitatively evaluated the predicted effect on chromatin accessibility of various single-nucleotide polymorphisms encoded within a noncoding region predicted to regulate *TBX5*[93]. This analysis revealed a single rare variant in that region (rs377307764) that disrupts a contributive GATA motif (Fig. 6F). The presence of this variant reduced accessibility beginning at the mesendoderm stage of differentiation, prior to detectable *TBX5* expression (Fig. 6G). This result highlights how integration of scMultiome data with noncoding variant interpretation tools can uncover disease-relevant genetic mechanisms.

Discussion

Efforts to define genotype–phenotype relationships in CHD cohorts have been hampered by the complexity of noncoding variation, which likely acts through disruption of cis-regulatory elements and their associated gene regulatory networks. While fetal atlases have mapped later-stage cardiac cell states relevant to CHD pathogenesis, they lack resolution of the early post-gastrulation developmental transitions when lineage specification and regulatory architecture are first established—stages at which noncoding perturbations are most likely to exert pathogenic effects. In this study, we generated a high-resolution, single-cell multi-omic atlas of human cardiac organoid differentiation. We then leveraged this resource to train a base-pair-resolution deep learning framework that decodes the complex cis-regulatory syntax governing cell fate. Our models uncovered a latent layer of regulation encoded by cell-state-specific combinations of transcription factor motifs, a feature obscured by traditional genomic analyses. We then employed this approach to uncover gene regulatory networks disrupted by noncoding variants discovered in patient cohorts. This work presents a comprehensive framework for linking non-coding genetic variants to their mechanistic roles in congenital diseases.

Deep learning frameworks provide a powerful means to move beyond conventional motif scanning by inferring the regulatory grammar that orchestrates

developmental programs. While single-cell chromatin accessibility assays delineate the global landscape of regulatory potential, they lack the resolution to disentangle how specific combinations of transcription factor (TF) motifs, often embedded in close proximity, synergize to drive cell state transitions. This limitation is compounded by the rapid kinetics of differentiation, wherein successive states frequently exhibit only subtle differences in chromatin accessibility despite profound shifts in lineage identity. Our framework addresses this limitation by systematically evaluating the accessibility contributions of each nucleotide. This methodological innovation has enabled us to assess the dynamic, lineage-specific activities of key regulators such as MEF2, to delineate the syntax of cell-state-specific regulation within cdCREs at loci like DHRS3, which is broadly required for retinoic acid signaling. We extend previous results nominating the TEAD family to highlight how its activity is modulated by its interacting partners, exemplified by SRF in cardiomyocytes and GRHL in ectodermal tissues, as well as by discrete spacing between motifs. By linking cis-regulatory logic to their cognate target genes, we can reconstruct gene regulatory networks that chart the developmental trajectories underlying cardiac lineage diversification. This framework captures the nuanced regulatory differences that distinguish closely related cell types, such as ventricular versus pacemaker cardiomyocytes, which arise from a recent shared progenitor pool. Importantly, the predictive capacity of this model provides a substantial advance in our ability to interpret how combinations of transcription factor motifs encode the functional instructions within the noncoding genome, thereby enabling systematic dissection of the cis-regulatory architecture that drives cell fate specification.

While *in vivo* studies have highlighted essential roles of enhancers in heart development, identifying noncoding variants that disrupt enhancer activity and contribute to the causal genetic etiologies of CHD has faced many technical obstacles. Specifically, the large number of non-coding variants in the human genome makes it difficult to pinpoint the ones that are functionally significant. Furthermore, linking these variants to their specific target genes is challenging, as enhancers can regulate multiple genes over considerable genomic distances. Lastly, traditional genomic assays often lack the resolution and scale needed to address these issues, leaving the regulatory basis of many CHD cases unresolved. By employing base-pair resolution computational models to elucidate the regulatory grammar that is active during the early stages of cardiac differentiation, our research delineates lineage-specific regulatory cascades and identifies disease-associated noncoding variants that disrupt these pathways. Through the identification of variants predicted to regulate the initial phases of cardiogenesis, our study enhances the understanding of the precise developmental contexts in which noncoding variants may interfere with normal development, leading to the manifestation of CHD. Furthermore, our findings underscore the notion that noncoding variants residing within the regulatory landscape of a known disease gene may impact a limited

set of cell states, differing from those caused by coding variants in the same gene. Because cis-regulatory elements and enhancers can modulate the expression of multiple genes, a single noncoding variant within these regions has the potential to perturb several targets simultaneously, thereby broadening its pathogenic impact. Moreover, previous efforts to link noncoding variants to CHD have faced several key limitations, which may explain the lack of robust genetic etiologies. First, many studies have mainly focused on cardiomyocytes, overlooking the diverse cell types involved in early heart development. Second, they have often depended solely on transcriptomic data, which does not reveal the full regulatory landscape. Finally, these analyses are usually conducted on later-stage or postnatal tissues, missing the crucial post-gastrulation period when cardiac lineages are first established. Our finding that noncoding variants in CHD patients are predicted to affect chromatin accessibility across multiple lineages during these early developmental stages directly addresses these issues. These findings highlight the urgent need for experimental models that accurately represent relevant developmental stages to properly interpret the effects of genetic variants.

In summary, this work presents an integrated framework for transitioning from a descriptive atlas of human organogenesis to a predictive model of its underlying gene regulatory code, which can be utilized to identify the genetic etiologies of human congenital abnormalities. The power of deep learning models trained on high-resolution, rich sc-Multi-omic data lies in their potential to address a central challenge of modern genetics: interpreting the function of the vast number of non-coding variants linked to human disease. Our regulatory models can now be used as an *in silico* screening tool to systematically prioritize disease-associated variants that are most likely to disrupt cis-regulatory syntax and alter cell fate. This provides a direct path from correlative GWAS findings to testable hypotheses about the origins of congenital disease. Ultimately, by learning to read the grammar of the genome, we can improve our ability to decipher how its misspellings lead to disease.

Limitations of our study

While our cardioid model successfully recapitulates early developmental events, its present cell state composition presents limitations for modeling the later, more complex aspects of organogenesis and disease. The current system lacks key supportive cell types known to be critical *in vivo*, such as neural crest cells. The absence of these lineages and their associated paracrine signals likely limits the maturation potential of the cell types that are present. Therefore, while our framework is powerful for dissecting early cell fate decisions, future studies aimed at modeling complex disease phenotypes will require more comprehensive organoid systems. Incorporating additional growth

factors and mechanical cues, such as sheer stress, will be a critical next step in generating a more complete model of the human heart.

Methods

1. Cell culture

1.1. Cardioid generation

The cardioid generation protocol was adapted from [44,45]. Induced pluripotent stem cells (iPSCs) were seeded into 96-well plates at a density of 10,000 cells per well after achieving 70% confluency and dissociating with Accutase. Cells were resuspended in mTesR+ (StemCell Technologies, 100-0276) with ROCKi (Tocris, 1254) and incubated overnight. On day 0, media was changed to FlyAB(Ins), containing RPMI (Thermo Fisher, 11875119), B-27 with insulin supplement (Thermo Fisher, 17504044), FGF2 (Fisher, 23-3FB), BMP4 (Fisher, 314-BP), LY294002 (Fisher, 1130), CHIR (Tocris, 4423), and Activin A (R&D, 338-AC) in specified concentrations. Subsequent media changes were performed as follows: on day 1, with a refresh of FlyAB(Ins); after 36-40 hours, media was changed to BWIIFRa, which includes RPMI, B-27 with insulin supplement, BMP4, FGF2, IWP2 (Tocris, 3533), and retinoic acid (Tocris, 0695). Media was refreshed every 24 hours for days 2-4 with BWIIFRa. On days 5/6, media was changed to BFI containing RPMI, B-27 with insulin supplement, BMP4, and FGF2. On day 7, the media was switched to maintenance media with RPMI and B-27 , refreshed every 48 hours.

1.2. Cardioid dissociation and cryopreservation

At every time point (Days 0, 1, 2, 3, 5, 10), cardioids were dissociated into single cells with >90% live cells before freezing. TrypLE 10x Gibco™ A1217701) was warmed to 37°C, and the centrifuge was cooled to 4°C. Cardioids were washed with DPBS on a bi-directional 37 µm filter (Stemcell, 27215) and then dissociated using prewarmed TrypLE 10x in a thermomixer set at 1000 RPM at 37°C for 20 minutes with gentle pipetting every 5 minutes. When cardioids were fully dissociated, cells were centrifuged at 300g for 5 minutes at 4°C, resuspended in PBS, and viability checks were performed, aiming for >95% live cells. Cells were thoroughly resuspended in a freezing medium (Bambanker, NIPPON Genetics, BB05) and cryopreserved using a slow rate-controlled cooling process. Cryovials were then transferred to liquid nitrogen for long-term storage.

1.3. Cardioid Thawing

Cryopreserved dissociated cardioids were transferred from liquid nitrogen to a bead bath until approximately half-thawed. The cell suspension was then immediately mixed with 1 mL of cold RPMI and transferred to a low-bind Eppendorf tube. The cells underwent two subsequent wash steps: they were centrifuged at 300g for 5 minutes at 4°C, and then gently resuspended in 1 mL of 1X PBS (Thermofisher, 10010023) with 0.04% BSA (Thermofisher, 15260037) using a wide-bore tip. After the final wash, the cell pellet was resuspended in an appropriate volume of 1X PBS with 0.04% BSA. To remove any remaining aggregates or debris, the final cell suspension was passed through a 70 µm cell strainer. The recovered cells were kept on ice, and their viability was assessed via trypan blue staining, with a quality control goal of greater than 95% viability.

2. Immunofluorescence and microscopy

Organoids were fixed in 4% paraformaldehyde (PFA) and embedded in Optimal Cutting Temperature (OCT) compound (Tissue-Tek, 4583). Samples were cryosectioned at 10 µm using a cryomicrotome (Thermo Scientific; Microm HM 550), and adjacent sections were collected on coated glass slides (VWR, VWRU48311-703). Sections from the central region of each organoid were selected for analysis. For immunofluorescence staining, sections were fixed again in 4% PFA for 20 minutes and washed with PBS. Samples were blocked with 5% goat serum (BioAbChem, 72-0480) in PBS containing 0.1% Triton X-100 (MP Biochemicals, 194854) for 1 hour at room temperature, then incubated overnight at 4°C with primary antibodies. After three PBS washes, sections were incubated with fluorescently conjugated secondary antibodies for 2 hours at room temperature and washed again. Finally, slides were mounted with DAPI Fluoromount-G (Southern Biotech, 0100-20) and imaged using a Zeiss LSM 990 confocal microscope.

3. Single-cell processing

3.1. SHARE-seq

SHARE-seq protocol was adopted from [94,95], and a detailed protocol is provided in Note S1.

3.2. Library Sequencing

All DNA libraries were sequenced on a NextSeq 500/550 High Output Kit v2.5 (150 Cycles, 20024907), NovaSeq X Series 25B Reagent Kit (300 cycles, 20104706), NovaSeq 6000 S4 Reagent Kit v1.5 (300 cycles, 20028312) with XP workflow. Paired-end sequencing was run with a 96-99-8-96 configuration (Read1-Index1-Index2-Read2). We quantified DNA libraries using Qubit and tapestation, then prepared library pools at 4nM concentration. Sequencing was performed at the Stanford Genome

Technology Center (NovaSeq6000), NovoGen (NovaSeqX) and in-house instrument (NextSeq550).

4. SHARE-seq data preprocessing

Raw sequencing data was preprocessed using a customized pipeline suited to SHARE-seq https://github.com/GreenleafLab/shareseq-pipeline (stable release v1.0.0) [40].

5. Data analysis

All analysis scripts can be found here https://github.com/GreenleafLab/Cardioid. A brief description of the analysis steps is provided in the following sections.

5.1. SHARE-seq data QC and filtering

We performed per-sample QC filtering by manually inspecting and thresholding the following metrics: 1) TSS enrichment ratio and number of fragments for ATAC fragment files, 2) number of UMIs, number of genes and percentage of mitochondrial reads for RNA sparse matrices, 3) ratio of RNA UMIs vs ATAC fragments to remove cells with low quality in one modality. All sample filtering thresholds are summarized in Table S2.

5.2. RNA normalization, ambient RNA removal, dimensionality reduction, and clustering

We used Seurat (v4.3.0)[46] in R (v4.1.2) to process filtered RNA sparse matrices into Seurat objects. We adopted an iterative dimensionality reduction and clustering workflow to sequentially annotate cell types. For each iteration, we first performed SCTransform v2 and variable feature selection on RNA raw counts of each sample, then selected the top 3,000 consensus variable features across samples using the SelectIntegrationFeatures function from Seurat, excluding mitochondrial genes, sex chromosome genes, and cell cycle genes to minimize batch effects. We merged the raw RNA counts from per-sample objects into a single matrix, performed SCTransform v2 using consensus features, and used the DecontX function from the celda (v1.6.1) package [47] to remove ambient RNA contamination per cell. The decontaminated counts were then scaled to 10,000 UMIs per cell and log-normalized. Similar to the process mentioned above, we selected a list of the top 3,000 consensus variable features. PCA was performed on the merged object with the consensus features, followed by cell clustering using the Louvain algorithm at a resolution of 0.3 with 50 PCs and UMAP embedding. We then inspected each cluster and removed any low-quality clusters with significantly lower UMIs than other clusters. After removing cells in the low-quality clusters, we repeated the processing steps starting from RNA raw counts for each sample. This process was repeated until no more low-quality clusters were identified, which required three iterations. Cells in

the final set of clusters passing this iterative QC were considered "whitelisted". For each cell type cluster, marker genes were identified in a one-vs-all Wilcoxon rank-sum test versus all other clusters. Top marker genes were filtered to genes with a $\log_2$ fold change greater than 1.

All final cluster annotations are included in Table S3. All cluster markers are summarized in Table S4.

5.3. scATAC data processing and label transfer

We used ArchR (v1.0.2)[96] to process filtered ATAC fragment files into a single ArchR project. After filtering to the final whitelisted cell barcodes from the iterative RNA processing workflow and transferring the clustering and cell type annotations, we called peaks per cluster using Macs2 (v2.2.7.1)[97], merged peaks into a single reproducible peak set using ArchR's iterative overlap strategy, and created a cell-by-peak matrix of fragment counts. We identified marker peaks per cluster using a Wilcoxon rank-sum test. For final cluster annotation, we use the well-defined RNA cluster labels and store them as metadata in the ArchR object.

5.4. scE2G

We used the scE2G^{Multiome} model to predict enhancer-gene connections in each cell type (https://github.com/EngreitzLab/scE2G/tree/v1.2)[70]. Briefly, scE2G first defines a set of candidate element-gene pairs where candidate elements are derived from peaks called on the pseudobulk ATAC-seq data from each cell type and are paired with genes with a transcription start site within 5 Mb. We used the model's default TSS reference file with some manual adjustments to select the appropriate TSS for each gene. Next, it annotates each pair with seven features calculated from the multiome data, including 1) a pseudobulk Activity-By-Contact (ABC) score, where 3D contact is estimated by an inverse function of genomic distance; 2) the Kendall correlation across single cells between element accessibility and gene expression; 3) whether the gene is "ubiquitously expressed", and 4) several other measures of genomic distance and chromatin accessibility around the element and promoter. Then, the features are integrated to assign each candidate element-gene pair a score representing the likelihood that the element regulates expression of the gene. Regulatory enhancer-gene interactions were defined as element-gene pairs with a score greater than 0.177, which is as the score yielding 70% recall when evaluating predictions in K562 cells against CRISPRi-validated enhancer-gene pairs[98].

5.5. Azimuth

To compare our *in vitro* cell states to an *in vivo* reference, we performed a projection analysis using Azimuth[46] (v2.1.0). The decontaminated scRNA-seq counts matrix from the integrated cardioid dataset was used as the query. We ran the RunAzimuth function to map these query cells onto the human gastrula reference atlas ("hGas"). The mapping was performed using pre-computed reference data, and cell annotations were transferred at two levels: celltype.l1 (broad developmental lineages) and celltype.l2 (specific cell identities). The resulting object, containing the projected UMAP coordinates and cell predictions (predicted.celltype.l1, predicted.celltype.l2, and mapping.score), was saved for downstream analysis. The metadata, including query cluster labels and all predicted annotations and scores, are available in Table S8.

5.6. Optimal transport

We applied Waddington-OT[63] to model single-cell trajectories for each modality, with default pipeline settings. A cell at a time t is represented by the gene-expression vector $x \in R^{G}$, where G is the number of genes. For each adjacent pair of time points (t_i, t_{i+1}), we first carried out a pairwise principal-component analysis, retaining the top d components shared by both time points. Coordinates in this R^{d} subspace served as input to an entropic, unbalanced optimal-transport problem [99]. Only the row sums were constrained, allowing the column sums to vary so that proliferation or death could be accommodated. Solving this optimisation yielded a coupling matrix $\pi_{t_i \to t_{i+1}} \in R^{N_{ti} \times N_{ti+1}}$. Matrix multiplication of the cell embedding matrix at t_i with $\pi_{t_i \to t_{i+1}}$ denotes the mass transport between cells in t_i to t_{i+1}. Assuming Markovian dynamics, long-range couplings were approximated by successive matrix multiplication: $\gamma_{t_i \to t_j} = \pi_{t_i \to t_{i+1}} \pi_{t_{i+1} \to t_{i+2}} \ldots \pi_{t_{j-1} \to t_j}$. For each starting cell, the probability mass in γ arriving in user-defined final clusters was summed to create an aggregate fate matrix $F \in R^{N \times K}$, where N is the total number of cells and K is the number of fates. The full pipeline was executed three times—once on RNA expression, once on ATAC gene scores, and once on a concatenated RNA + ATAC feature matrix. Within each modality, we trained separate models on control cells, on knockdown cells, and on the combined dataset, enabling direct comparison of transport maps and fate probabilities across perturbations.

5.7. ChromBPNet

5.7.1. Training

For each cardioid cell state, ATAC-seq fragments from cells constituting that cell state were aggregated. These aggregated fragments were fed into the ENCODE peak calling pipeline to identify peaks in each cell state. The per-cell state aggregated fragments and peaks were then used to train ChromBPNet (https://pubmed.ncbi.nlm.nih.gov/39829783/) models.

The ChromBPNet models were trained using the framework provided in the ChromBPNet Python package (https://github.com/kundajelab/chrombpnet). For each cell state, 5 individual ChromBPNet models were trained by following the 5-fold cross-validation scheme used in the ChromBPNet package, in which each chromosome is present in the test set of at least one cross-validation fold. The chromosomes used per fold are shown in Table S9. All training was done using the hg38 reference genome. All of the ChromBPNet models were trained using a Tn5 bias model trained on the background regions from the fold 0 chromosomes of the Ventricular Cardiomyocytes cell state. The per-fold background regions were defined by the ChromBPNet package. A bias threshold factor of 0.6 was used.

5.7.2. Per-cell state predictions

For every cell state, once the 5-fold models were trained, predictions for that cell state were made by averaging individual predictions from the 5-fold models. The log counts prediction for the cell state was the average of the log counts predictions of all the individual folds. The profile logits prediction for the cell state was the average of the profile logits predictions of all the individual folds. The profile logits prediction for the cell state was then softmaxed to produce the profile prediction for the cell state. The log counts prediction for the cell state was exponentiated and then multiplied by the profile prediction for the cell state to produce the counts-scaled profile prediction for the cell state. This averaging procedure was used for both the with-bias and de-biased model predictions.

5.7.3. Generating contribution scores

For any genomic sequence and for each model, we used the DeepSHAP[66] implementation of the DeepLIFT algorithm[64] to compute the contribution score that each base is predicted to have towards the total chromatin accessibility over that sequence. The contribution scores in each cell state are computed as the average

of the contribution scores from each of the 5-fold models. Contribution scores were calculated on the de-biased models.

5.7.4. Identifying a motif lexicon

In each cell state, we identified the motifs of transcription factors that are active in that cell state and contribute towards chromatin accessibility predictions. To do this, for each cell state, we began by computing contribution scores for all 2114 bp sequences centered at every peak summit identified in that cell state. Then, TF-MoDISco [65] (https://github.com/kundajelab/tfmodisco) was run on the per-cell-state contribution scores to identify motifs present in that cell state. The *modisco motifs* command was run with 1,000,000 seqlets (the *-n 1000000* option).

Once motifs were identified in each cell state, the motifs were clustered to create a non-redundant set of motifs. The 2169 motifs identified across all 59 control and knockdown cell states were clustered using the MoitfCompendium[67] (https://github.com/kundajelab/MotifCompendium) Python package to yield a set of 418 distinct motifs. Clustering was done at the 0.95 similarity threshold. The 418 distinct motifs were further classified into 172 motif groups by automated annotation (MoitfCompendium) followed by manual curation. The 172 groups were next clustered into 106 distinct motif families (Note S3).

5.7.5. Marginal footprinting

For each motif and for each cell state, an *in silico* marginal footprint score was computed to quantify the effect that motif has in that cell state. First, for each model, 1000 of the training background sequences identified during ChromBPNet model training were sampled. Then, the predicted counts from the de-biased model were computed (in this case, the predicted counts were computed as the average of the predicted counts of the forward and reverse sequence). This served as a baseline for the amount of accessibility latent in these background regions. Then, for each motif, a consensus DNA sequence was determined by taking the maximum contributing base pair at every position. For each motif, that motif's consensus sequence was then inserted into the center of the 1000 background sequences. Predictions from the de-biased models were made on these sequences with *in silico* motif insertions in them; the predicted counts of the forward and reverse sequences were averaged to produce the predicted counts of these sequences. The log2 fold change between the predicted counts of

the motif-inserted-sequence and the predicted counts of the without-motif-sequence was then computed per sequence. This log2 fold change was then averaged across all 1000 sequences to produce an effect size for the marginal effect of the motif in that model. These effect sizes were then averaged across the five-fold models to give cell-state-specific marginal footprint scores of the motif. This process was done per motif to give marginal footprint scores for each motif in each footprint.

These marginal footprint scores were plotted as a line plot across the differentiation trajectory to produce Figure 3F, S6 and Note S3. These marginal footprint scores were row-normalized and then plotted in a heatmap to produce Figure 3G and S6. Figure 3H shows rows of Figure 3G, which have been subset to highlight differences in the marginal footprint effect sizes of different motifs in the same family.

5.7.6. Annotating motifs in each cell state's regulatory landscape

We performed a comprehensive identification of motifs present in the regulatory regions of every cell state. For each cell state, we used the FiNeMo (https://github.com/kundajelab/Fi-NeMo) Python package to identify all occurrences of each motif in the non-redundant motif set within the summit-centered peak sequences of that cell state. The result was an annotation of all motif instances of every motif within the peak sequences of every cell state. The *finemo call-hits* command was run with default options.

5.7.7. Motif enrichment in peaks

We computed motif enrichment within cell states across peak clusters (Figure S3C). To do this, we first began by selecting two peak clusters of interest in a cell state. Then, we counted the number of motif instances of each type of motif present in that cell state in each peak in the two peak clusters. Then, for each motif, we created a contingency table where the top left was the number of instances of that motif in all peaks in the first peak cluster, the top right was the number of motif instances of that motif in all peaks in the second peak cluster, the bottom left was the total number of other motif instances in all peaks in the first peak cluster, and the bottom right was the total number of other motif instances in all peaks in the second peak cluster. Then, an odds ratio and p-value were computed for each contingency table using Fisher's exact test.

5.7.8. Gene-motif linking analysis

The first step of the gene-motif linking analysis was to filter genes. First, for each gene, the maximum expression of the gene across all cell states in normalized units of logTPM was computed. Then, any gene whose maximum expression is < 1% of the maximum expression (in units of logTPM) of any gene in any cell state is removed.

Then, the next step is to identify linked peaks for each gene. For each gene and each cell state, all scE2G links to that gene in that cell state were considered. If any of those links overlapped a peak in that cell state, that peak was added to the total set of peaks linked to that gene. This was done across all cell states to yield a comprehensive set of scE2G-linked peak regions that are linked to that gene in any cell state.

Then, for each cell state, gene, and motif, all instances of that motif identified in that cell state that overlapped any scE2G-linked peak region were considered. For each one of those hits, the sum of the contribution scores of that motif CWM is weighed by a function of the distance of the hit from the gene TSS. The distance weighting function is *e^-(distanceToTSS/5000)+1/e*. The distance-weighted contribution scores of all hits of a motif are summed to compute a score for how much a motif contributes to a gene in a cell state. This is done across all motifs, genes, and cell states.

5.8. Consensus non-overlapping peakset

A consensus, non-overlapping peakset was constructed by merging peaks from every cell state. For each cell state, each peak set was transformed into a list of 500 bp bins centered around each peak summit. Then, these summit-centered bin lists were combined across all control and knockdown cell states, leading to a list of 9,643,338 500 bp regions. The peak calling pipeline, which uses macs3, annotates each summit with a p-value indicating how likely that summit is really a peak summit. The combined summit-centered bin list was then sorted by macs3 p-value, with the bins with the lowest p-value (most likely to be a peak summit) at the top and bins with the highest p-value (least likely to be a peak summit) at the bottom. At this point, the consensus peakset is an empty set and contains no summit-centered bins. Then, each summit-centered bin in the sorted list was considered, and the non-overlapping peakset was built with the following rule: if the bin had any overlap with any bin already in the peakset, the bin was not retrained in the final peakset; however, if the bin did not overlap with any bin that was already in the peakset, it was added

to the peakset. Through this procedure, a non-overlapping summit-centered peakset of 628,378 500 bp bins was created.

5.9. Peak clusters

For each of the 628,378 peaks in the consensus peakset, its total accessibility is computed per cell state as the sum of the total number of Tn5 insertions in the peak region in that cell state. This gives a peak by cell state counts matrix. This raw counts matrix is then normalized by the read depth in each cell state so that each element in the matrix, which corresponds to a specific peak in a specific cell state is divided by the read depth in that cell state. Then, a copy of the normalized counts matrix is made which is then column-wise z-score normalized. Within each column (corresponding to a specific peak), this z-score normalization is done by first computing the column-wise mean (the mean normalized accessibility across all cell states) and then subtracting this mean from all values in the column, and then computing the column-wise standard deviation (the standard deviation of normalized accessibility across cell states) and then dividing the column values by this standard deviation. Then, the normalized counts matrix and the z-score-normalized counts matrix are concatenated along the cell state axis so that each peak is now represented by a vector of the normalized counts value across all cell states, concatenated with the z-score-normalized values across all cell states. Then, the peaks are clustered using scikt-learn's Mini-Batch K-Means clustering with k-20 (*sklearn.cluster.MiniBatchKMeans(n_clusters=20, random_state=0, batch_size=1000, n_init=50)*). Three pairs of clusters with similar dynamics were merged (original clusters 6 and 16 were merged, original clusters 1 and 11 were merged, and original clusters 8 and 19 were merged). Finally, the clusters were manually re-ordered. This led to 17 peak clusters.

5.10. Motif changes in knockdown

Every control cell state has a corresponding knockdown cell state. For every control/knockdown cell state pair, for each peak in the consensus peakset, we compute the $\log_2$ fold change between the normalized counts per million in the control and knockdown cell states; a positive log2 fold change indicates that the peak was more accessible in the control than the knockdown, and a negative value indicates that the peak was more accessible in the knockdown than the control. Peaks that had no counts in neither control nor knockdown were removed from this analysis for that cell state pair. Then, the remaining peaks were clustered by their log2 fold change value using k-means clustering with k=3. For each cell state pair,

this separated the peaks into 3 clusters: a cluster containing peaks more accessible in control than knockdown, a cluster containing peaks more accessible in knockdown than control, and a cluster containing peaks with similar accessibility across both control and knockdown cell states.

Then, for each motif, we created a contingency table where the top left was the number of peaks in the control-accessible peakset that contained any instance of that motif family, the top right was the number of peaks in the knockdown-accessible peakset that contained any instances of that motif family, the bottom left was the number of peaks in the control-accessible peakset that did not contain any instances of the motif family, and the bottom right was the number of peaks in the knockdown-accessible peakset that did not contain any instances of the motif family. Then, an odds ratio and p-value were computed for each contingency table using Fisher's exact test. The odds ratio and p-value of differential motif enrichment between control and knockdown are shown for each motif family and each control/knockdown cell state pair in Figure 5H.

5.11. Variant Scoring.

We made *in silico* variant effect predictions of noncoding variants using our ChromBPNet models. Each variant was specified by a genomic coordinate, a reference allele (which matched the hg38 reference genome), and an alternate allele. For each variant, we took the 2114 bp genomic sequence centered around the variant to be the "reference sequence". We substituted the center reference allele for the alternate allele to create the "alternate sequence". To predict the effect of a variant in a model, we compared the model's prediction on the reference sequence to the prediction on the alternate sequence. All predictions were made using the de-biased models.

When comparing the reference and alternate predictions, a variety of scores are computed to quantify the effect of the variant. These include the log2 fold change (LFC) of the total predicted coverage (total counts) over the two sequences, the absolute value of the LFC, the Jensen Shannon Distance (JSD) between the base-resolution predicted probability profiles, and the active allele quantile (AAQ), which is a measure of how the predicted counts in the reference and alternate sequences compare to the distribution of total counts in all peak regions used to train the models. For each of these variant effect scores, an empirical p-value is computed by comparing the variant effect against a null distribution generated from the scores of variants placed in dinucleotide shuffled sequences. Variant scores and p-values were

computed using the variant-scorer Python package (https://github.com/kundajelab/variant-scorer). To compute cell state scores, the 5-fold model scores were averaged; all scores were averaged using an arithmetic mean, while the p-values were averaged using a geometric mean. The heatmap in Figure 6E and the line plot in Figure 6G show the predicted fold-averaged LFC across all cell states.

Data accessibility

Code availability

All analysis code is available at https://github.com/GreenleafLab/Cardioid.

Acknowledgements

We thank Nick Addleman, John Coller, and the Stanford Genome Technology Center for their assistance with sequencing libraries, and members of the Greenleaf, Gifford, and Kundaje labs for their helpful discussions. This work was supported by RM1HG007735, R01NS128028, R01HL171611, DP1HG013599, R01HG013317, UM1 HG012660, UM1HG011972 (to W.J.G); support from the Stanford Cardiovascular Institute, Saving Tiny Hearts Society and Additional Ventures (to C.A.G); support from NHGRI Impact of Genomic Variation on Consortium (UM1HG011972); the Novo Nordisk Foundation Center for Genomic Mechanisms of Disease (NNF21SA0072102); NHLBI R01HL159176; the Applebaum Foundation; Gordon and Betty Moore; and the BASE Research Initiative at the Lucile Packard Children's Hospital at Stanford University (to J.M.E). W.J.G. is an Arc Innovation Investigator and Director of the Stanford RNA Medicine Program. C.A.G. is the Akiko Yamazaki and Jerry Yang Faculty Scholar in Pediatric Translational Medicine, Stanford Maternal & Child Health Research Institute and Director of the Mann Charities Biobank at the Betty Irene Moore Children's Heart Center, Lucile Packard Children's Hospital, Stanford. M.U.S. acknowledges the support of an NSF Graduate Research Fellowship (DGE-1656518) and a graduate fellowship award from Knight-Hennessy Scholars at Stanford University. M.A. acknowledges the support from the AHA postdoctoral fellowship. E.G.P acknowleges the support of the*Training Grant in Myocardial Biology at Stanford (NIH 5T32HL094274-13)*. E.M acknowledges the support from the HFSP postdoctoral fellowship.

Competing 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. is a consultant for Ultima Genomics and Guardant Health. W.J.G. is a scientific cofounder of Protillion Biosciences. J.M.E. has received materials from 10x Genomics unrelated to this study and has received speaking honoraria from GSK plc, Roche Genentech, and Amgen. A.K. is on the scientific

advisory board of SerImmune, TensorBio, AINovo, is a consultant with Arcadia Science, Inari, Precede Biosciences, was a consultant with Illumina and PatchBio and has a financial stake in DeepGenomics, Immunai and Freenome. All other authors declare no competing interests.

Author Contributions

E.M. and R.Z. contributed equally as co-first authors. E.M., W.J.G., and C.A.G. conceived the study; E.M., J.P., E.G.P., J.M.S and J.R. optimized cardioid cultures and performed sample collection, with input from C.A.G.; E.M. conducted SHARE-seq experiments; E.M. programmed and performed data preprocessing, with input from B.B.L. and S.H.K.; E.M., E.G.P. and M.M. annotated all cell types, with input from C.A.G.; R.Z. and S.D. trained ChromBPNet models and performed analyses related to the motif lexicon and transcription factor cooperativity, with input from A.K.; M.U.S performed scE2G processing and analysis, with input from J.M.E.; I.A. programmed and performed optimal transport analysis, with input from E.M.; M.A. performed immunofluorescence imaging, with input from C.A.G.; Y.Z., I.E., S.D. and C.A.G. performed causal variant analysis, with input from A.K.; E.M., R.Z., S.D., W.J.G., A.K., and C.A.G. wrote the manuscript, with input from all authors.; A.K., W.J.G., and C.A.G. jointly supervised the work.

Figures

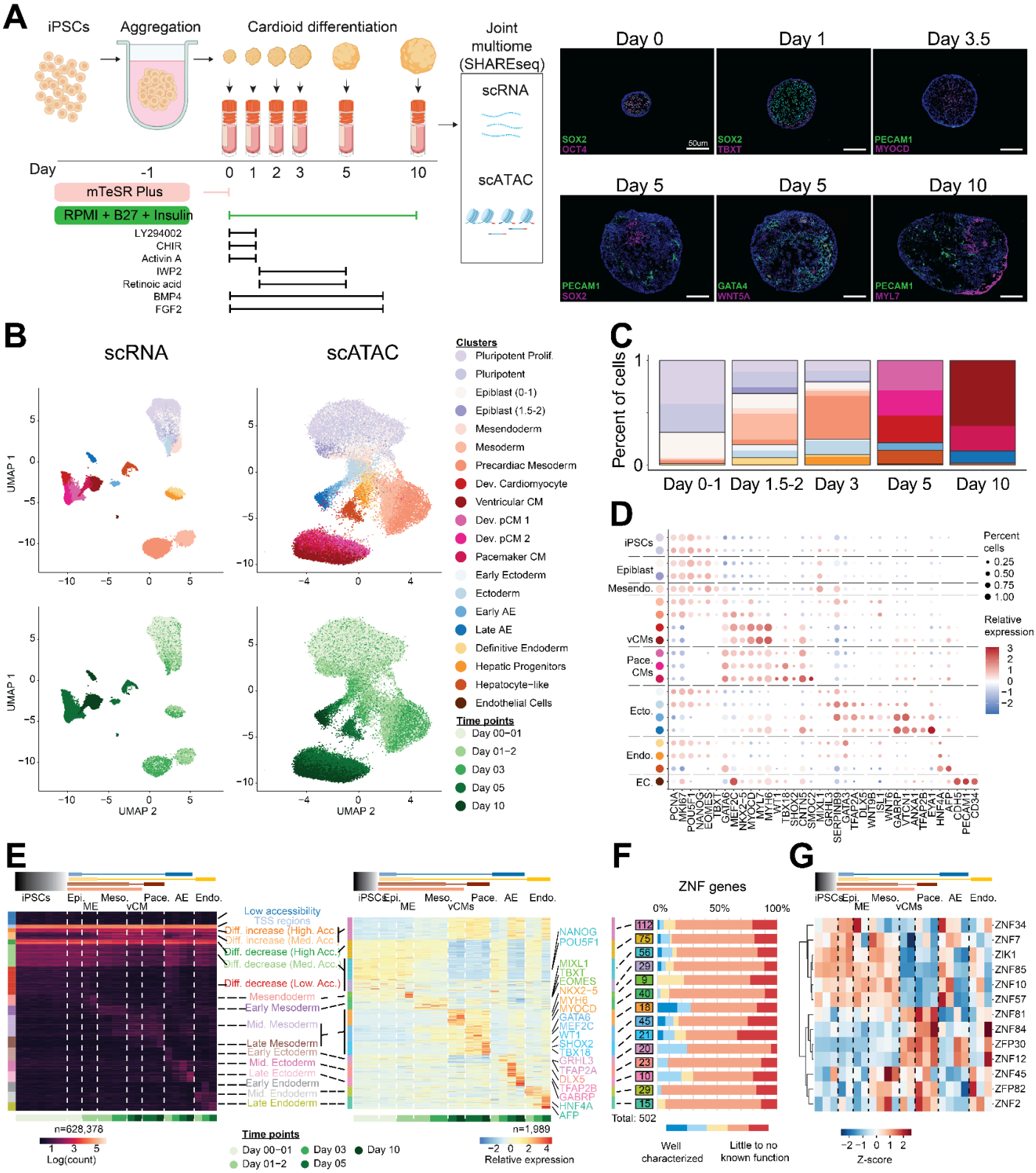
A
iPSCs
Aggregation
Cardioid differentiation
Joint multiome (SHAREseq)
scRNA
scATAC
Day
-1
0
1
2
3
5
10
mTeSR Plus
RPMI + B27 + Insulin
LY294002
CHIR
Activin A
IWP2
Retinoic acid
BMP4
FGF2
Day 0
Day 1
Day 3.5
Day 5
Day 5
Day 10
SOX2
OCT4
50um
SOX2
TBXT
PECAM1
MYOCD
PECAM1
SOX2
GATA4
WNT5A
PECAM1
MYL7
B
scRNA
scATAC
UMAP 1
UMAP 2
Clusters
Pluripotent Prolif.
Pluripotent
Epiblast (0-1)
Epiblast (1.5-2)
Mesendoderm
Mesoderm
Precardiac Mesoderm
Dev. Cardiomyocyte
Ventricular CM
Dev. pCM 1
Dev. pCM 2
Pacemaker CM
Early Ectoderm
Ectoderm
Early AE
Late AE
Definitive Endoderm
Hepatic Progenitors
Hepatocyte-like
Endothelial Cells
Time points
Day 00-01
Day 01-2
Day 03
Day 05
Day 10
C
Percent of cells
Day 0-1
Day 1.5-2
Day 3
Day 5
Day 10
D
iPSCs
Epiblast
Mesendo.
vCMs
Pace. CMs
Ecto.
Endo.
EC.
Percent cells
0.25
0.50
0.75
1.00
Relative expression
E
iPSCs
Epi.
ME
Meso.
vCM
Pace.
AE
Endo.
Low accessibility
TSS regions
Diff. increase (High. Acc.)
Diff. increase (Med. Acc.)
Diff. decrease (High Acc.)
Diff. decrease (Med. Acc.)
Diff. decrease (Low. Acc.)
Mesendoderm
Early Mesoderm
Mid. Mesoderm
Late Mesoderm
Early Ectoderm
Mid. Ectoderm
Late Ectoderm
Early Endoderm
Mid. Endoderm
Late Endoderm
n=628,378
Log(count)
Time points
Day 00-01
Day 01-2
Day 03
Day 05
Day 10
n=1,989
Relative expression
F
ZNF genes
0%
50%
100%
Total: 502
Well characterized
Little to no known function
G
ZNF34
ZNF7
ZIK1
ZNF85
ZNF10
ZNF57
ZNF81
ZNF84
ZFP30
ZNF12
ZNF45
ZFP82
ZNF2
Z-score

Figure 1. (A) Left: Experimental workflow. Right: Immunofluorescence staining at multiple differentiation time points. (B) scRNA and scATAC modalities from SHARE-seq multiome processing of ~60,000 cells throughout cardioid differentiation. (C) Cell type compositions in each sampled day, colors correspond to (B). (D) Manually curated gene markers verify numerous heart-relevant cell types. Colors correspond to (B), z-score normalized per gene. (E) kmeans clustering of consensus peaks across all cell types in all time points in the differentiating cardioid (~630K, left) and k-means clustering of all expressed TFs merged with the top 50 variable genes in each cluster (~2000 genes). Filled color boxes on top of heatmaps designate trajectories; clusters under the dashed lines are not part of the respective trajectory. (F). Known and predicted function of 502 ZNFs with cell-state specificity as defined in (E) Right. (G) Z-score normalized expression for 13 ZNFs with validated repressor domains.

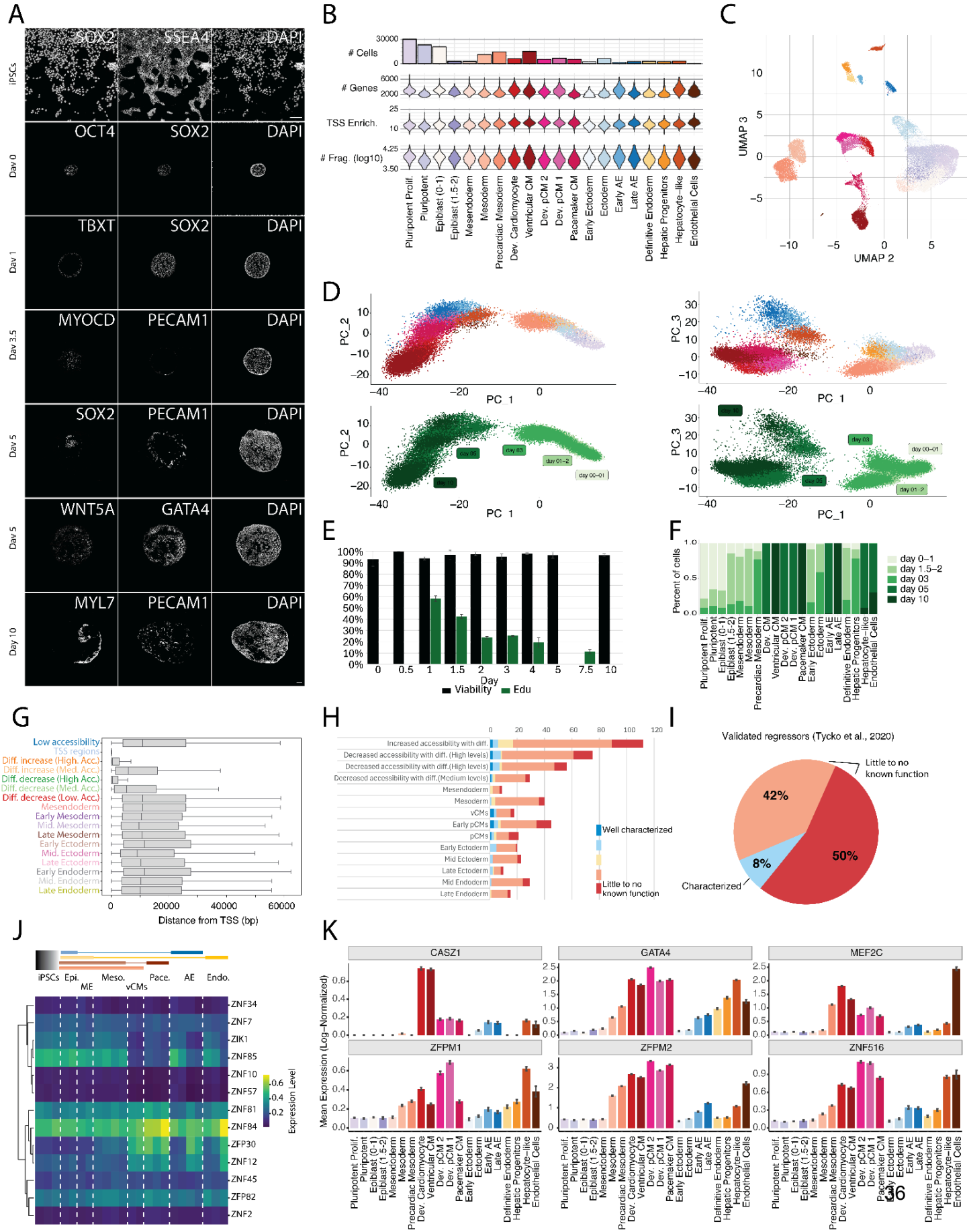
A
iPSCs
SOX2
SSEA4
DAPI
Day 0
OCT4
SOX2
DAPI
Day 1
TBXT
SOX2
DAPI
Day 3.5
MYOCD
PECAM1
DAPI
Day 5
SOX2
PECAM1
DAPI
Day 5
WNT5A
GATA4
DAPI
Day 10
MYL7
PECAM1
DAPI
B
Cells
Genes
TSS Enrich.
Frag. (log10)
Pluripotent Prolif.
Pluripotent
Epiblast (0-1)
Epiblast (1.5-2)
Mesendoderm
Mesoderm
Precardiac Mesoderm
Dev. Cardiomyocyte
Ventricular CM
Dev. pCM 2
Dev. pCM 1
Pacemaker CM
Early Ectoderm
Ectoderm
Early AE
Late AE
Definitive Endoderm
Hepatic Progenitors
Hepatocyte-like
Endothelial Cells
C
UMAP 3
UMAP 2
D
PC_1
PC_2
PC_3
day 10
day 03
day 05
day 01-2
day 00-01
E
Viability
Edu
Day
F
Percent of cells
day 0-1
day 1.5-2
day 03
day 05
day 10
G
Low accessibility
TSS regions
Diff. increase (High. Acc.)
Diff. increase (Med. Acc.)
Diff. decrease (High Acc.)
Diff. decrease (Med. Acc.)
Diff. decrease (Low. Acc.)
Mesendoderm
Early Mesoderm
Mid. Mesoderm
Late Mesoderm
Early Ectoderm
Mid. Ectoderm
Late Ectoderm
Early Endoderm
Mid. Endoderm
Late Endoderm
Distance from TSS (bp)
H
Increased accessibility with diff.
Decreased accessibility with diff. (High levels)
Decreased accessibility with diff.(High levels)
Decreased accessibility with diff.(Medium levels)
Mesendoderm
Mesoderm
vCMs
Early pCMs
pCMs
Early Ectoderm
Mid Ectoderm
Late Ectoderm
Mid Endoderm
Late Endoderm
Well characterized
Little to no known function
I
Validated regressors (Tycko et al., 2020)
Little to no known function
42%
50%
8%
Characterized
J
iPSCs
Epi.
Meso.
Pace.
AE
Endo.
ME
vCMs
ZNF34
ZNF7
ZIK1
ZNF85
ZNF10
ZNF57
ZNF81
ZNF84
ZFP30
ZNF12
ZNF45
ZFP82
ZNF2
Expression Level
K
CASZ1
GATA4
MEF2C
ZFPM1
ZFPM2
ZNF516
Mean Expression (Log-Normalized)

Figure S1. (A) Raw channel intensities. Ploripotency: *SOX2*, *SSEA4*, *OCT4*; Early mesendoderm specification: *TBXT*; Endothelial cells: *PECAM1*; Early ectoderm: *SOX2* on day 5; Mesoderm: *GATA4*; Ectoderm: *WNT5A*. (B) single-cell QCs per cell type. (C) (D) PCA plots. Left - PC1 vs PC2 captures the general time component of cardioid differentiation. Right - PC3 captures the variance between the lineages. The top panels are annotated by cell type, and the bottom panels are annotated by time point. (E) Black bars: Viability of harvested cells during cardioid differentiation. Green bars: EdU staining. Proliferation drastically decreases in the first few days of differentiation. (F) The time point composition of each cell type shows that later time points are more prevalent in more mature cell types. (G) Distance of peaks in the peakset from the nearest TSS. (H) Known and predicted function of 502 ZNFs with cell-state specificity, raw numbers for Fig. 1F. (I) Fraction of known function for the 13 ZNFs with validated repressor domain. Out of the 13 ZNFs, only *ZNF57* was annotated as “Characterized”, mainly because of its known activity in methylation imprinting, also in embryonic development. (J) log Normalized expression, corresponding to Fig. 1G. (K) log Normalized expression across cell states.

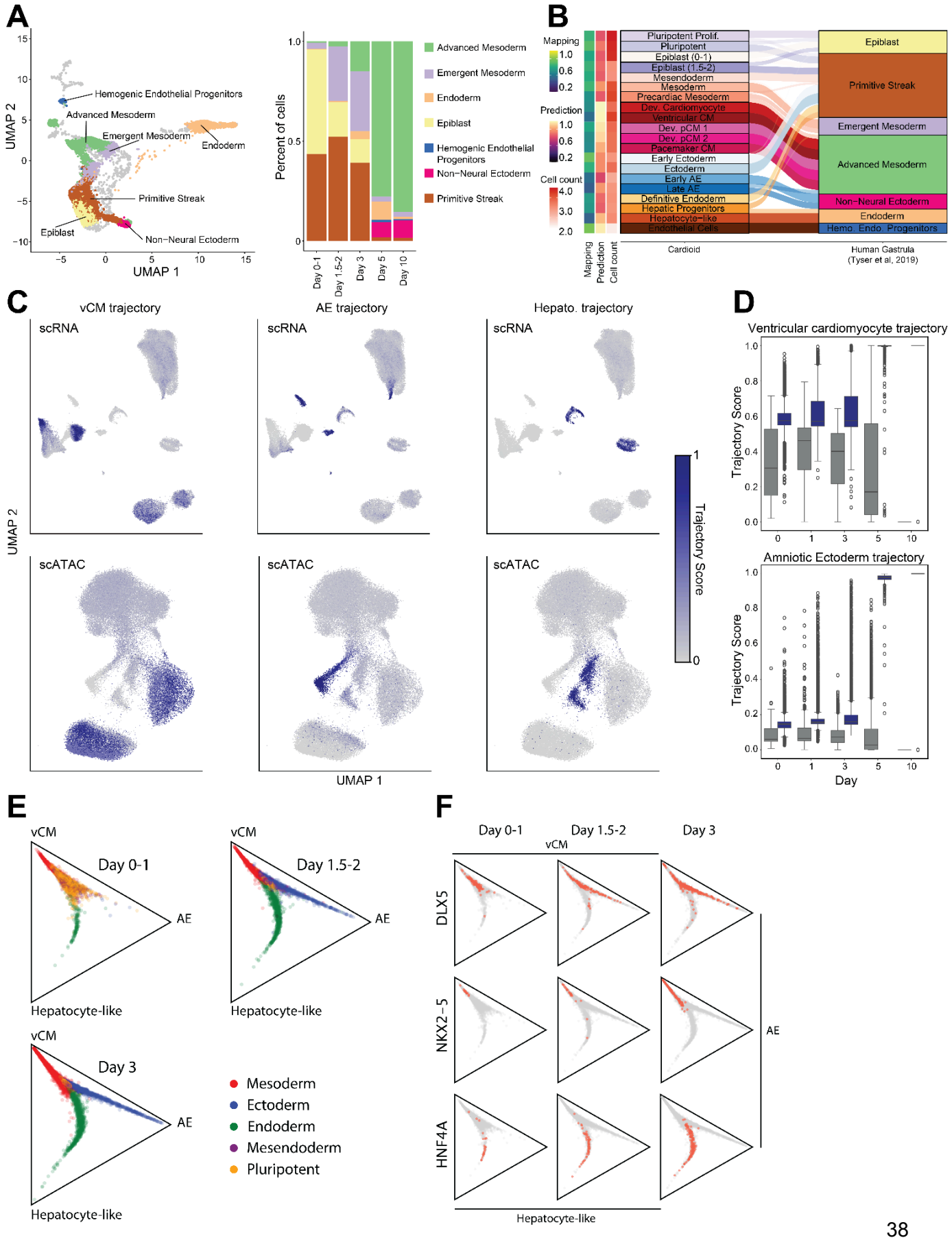
A
UMAP 2
UMAP 1
Hemogenic Endothelial Progenitors
Advanced Mesoderm
Emergent Mesoderm
Endoderm
Primitive Streak
Epiblast
Non−Neural Ectoderm
Percent of cells
Day 0-1
Day 1.5-2
Day 3
Day 5
Day 10
Advanced Mesoderm
Emergent Mesoderm
Endoderm
Epiblast
Hemogenic Endothelial Progenitors
Non−Neural Ectoderm
Primitive Streak
B
Mapping
Prediction
Cell count
Pluripotent Prolif.
Pluripotent
Epiblast (0-1)
Epiblast (1.5-2)
Mesendoderm
Mesoderm
Precardiac Mesoderm
Dev. Cardiomyocyte
Ventricular CM
Dev. pCM 1
Dev. pCM 2
Pacemaker CM
Early Ectoderm
Ectoderm
Early AE
Late AE
Definitive Endoderm
Hepatic Progenitors
Hepatocyte−like
Endothelial Cells
Epiblast
Primitive Streak
Emergent Mesoderm
Advanced Mesoderm
Non−Neural Ectoderm
Endoderm
Hemo. Endo. Progenitors
Cardioid
Human Gastrula
(Tyser et al, 2019)
C
vCM trajectory
AE trajectory
Hepato. trajectory
scRNA
scATAC
UMAP 2
UMAP 1
Trajectory Score
D
Ventricular cardiomyocyte trajectory
Amniotic Ectoderm trajectory
Trajectory Score
Day
E
vCM
Day 0-1
Day 1.5-2
Day 3
AE
Hepatocyte-like
Mesoderm
Ectoderm
Endoderm
Mesendoderm
Pluripotent
F
Day 0-1
Day 1.5-2
Day 3
vCM
DLX5
NKX2−5
HNF4A
AE
Hepatocyte-like

Figure 2. (A) Azimuth projection of the cardioid scRNA data to a human gastrula dataset. Left - Mapping the cardioid cells to the human gastrula, UMAP shows overlapping cell states and labels from the human gastrula. Right - earlier cardioid time points' cells are mapped to earlier human gastrula developmental stages. (B) The projection breakdown shows the mapping of individual cells to the reference. Less mature cell states in the cardioid map have higher Mapping and lower Prediction scores, while more mature cardioid cell states have higher Prediction scores and lower Mapping scores. (C) Optimal Transport Trajectory analysis identifies distinct cells predicted to differentiate to a more mature state (Black arrow). (D) Trajectory scores obtained through OT analysis within predefined trajectories (blue), as shown in Fig.1D, are significantly higher than those outside these trajectories (grey). (E) The Optimal Transport Fate by differentiation day displays the fate probabilities for each cell, highlighting the early commitment of cells toward either of the final states (Late AE, vCMs, Hepatocytes). (F) Optimal Transport Fate. Expression of cell state markers, shown are the top 90% of cells expressing these markers.

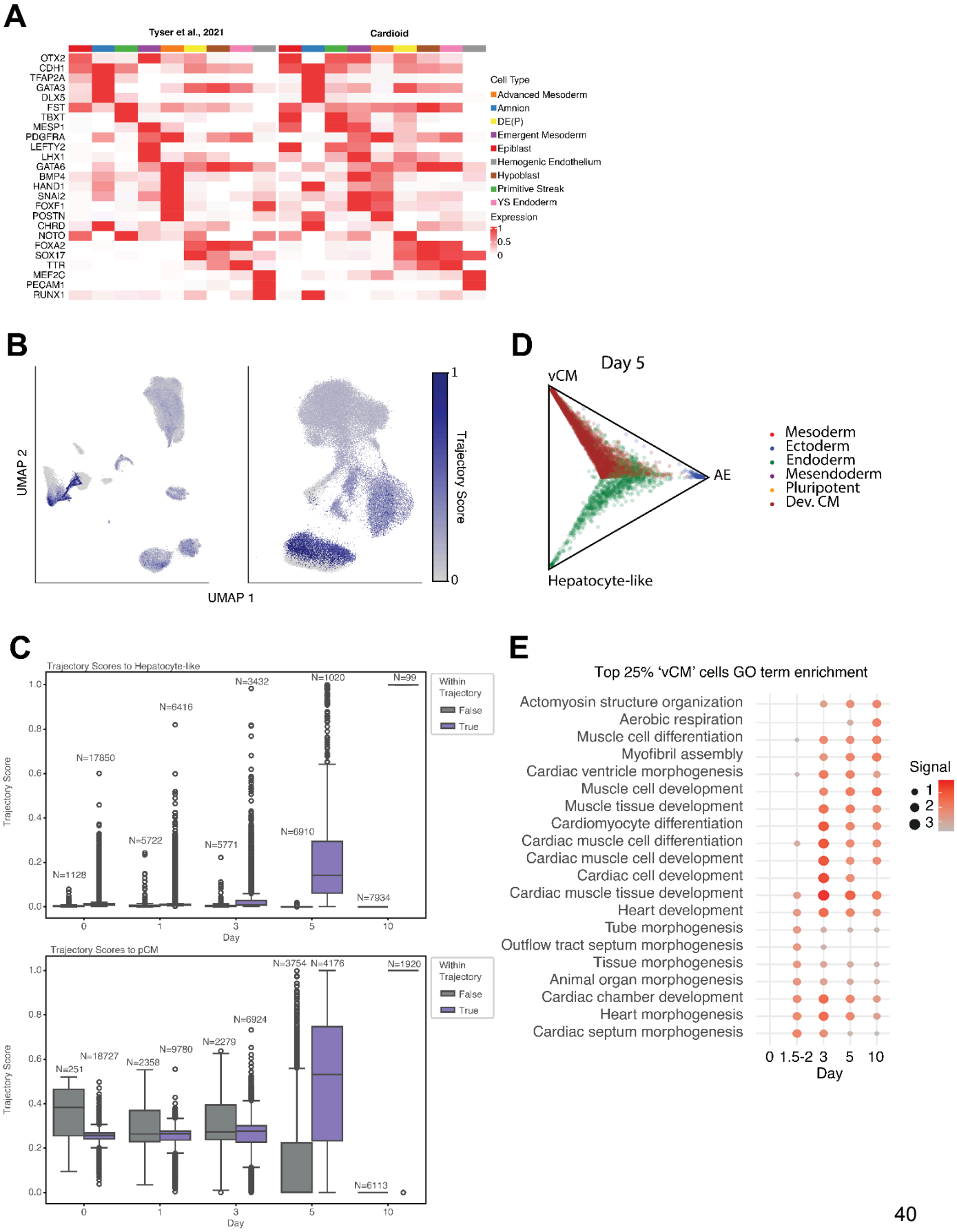
A
Tyser et al., 2021
Cardioid
OTX2
CDH1
TFAP2A
GATA3
DLX5
FST
TBXT
MESP1
PDGFRA
LEFTY2
LHX1
GATA6
BMP4
HAND1
SNAI2
FOXF1
POSTN
CHRD
NOTO
FOXA2
SOX17
TTR
MEF2C
PECAM1
RUNX1
Cell Type
Advanced Mesoderm
Amnion
DE(P)
Emergent Mesoderm
Epiblast
Hemogenic Endothelium
Hypoblast
Primitive Streak
YS Endoderm
Expression
1
0.5
0
B
UMAP 2
UMAP 1
Trajectory Score
1
0
D
Day 5
vCM
AE
Hepatocyte-like
Mesoderm
Ectoderm
Endoderm
Mesendoderm
Pluripotent
Dev. CM
C
Trajectory Scores to Hepatocyte-like
Trajectory Score
Day
Within Trajectory
False
True
N=1128
N=17850
N=5722
N=6416
N=5771
N=3432
N=6910
N=1020
N=7934
N=99
Trajectory Scores to pCM
N=251
N=18727
N=2358
N=9780
N=2279
N=6924
N=3754
N=4176
N=6113
N=1920
E
Top 25% 'vCM' cells GO term enrichment
Actomyosin structure organization
Aerobic respiration
Muscle cell differentiation
Myofibril assembly
Cardiac ventricle morphogenesis
Muscle cell development
Muscle tissue development
Cardiomyocyte differentiation
Cardiac muscle cell differentiation
Cardiac muscle cell development
Cardiac cell development
Cardiac muscle tissue development
Heart development
Tube morphogenesis
Outflow tract septum morphogenesis
Tissue morphogenesis
Animal organ morphogenesis
Cardiac chamber development
Heart morphogenesis
Cardiac septum morphogenesis
0 1.5-2 3 5 10
Day
Signal
1
2
3

Figure S2. (A) Human gastrula cell type markers expression from Tysler et al.[3], Extended Figure 2A and mapped cardioid cells gene expression. Left and right, respectively. (B) Optimal Transport Trajectory analysis identifies distinct cells predicted to differentiate to Pacemaker CM state (Black arrow). (C) Same as Fig. 2D but for Hepatocyte and Pacemaker CM trajectories. (D) The Optimal Transport Fate Matrix by Day displays the Fate probabilities for each cell at Day 5 of differentiation toward either of the final states (Late AE, vCMs, Hepatocytes). (E) GO term enrichments during cardioid development for the top 25% of cells with the highest probability of becoming vCMs.

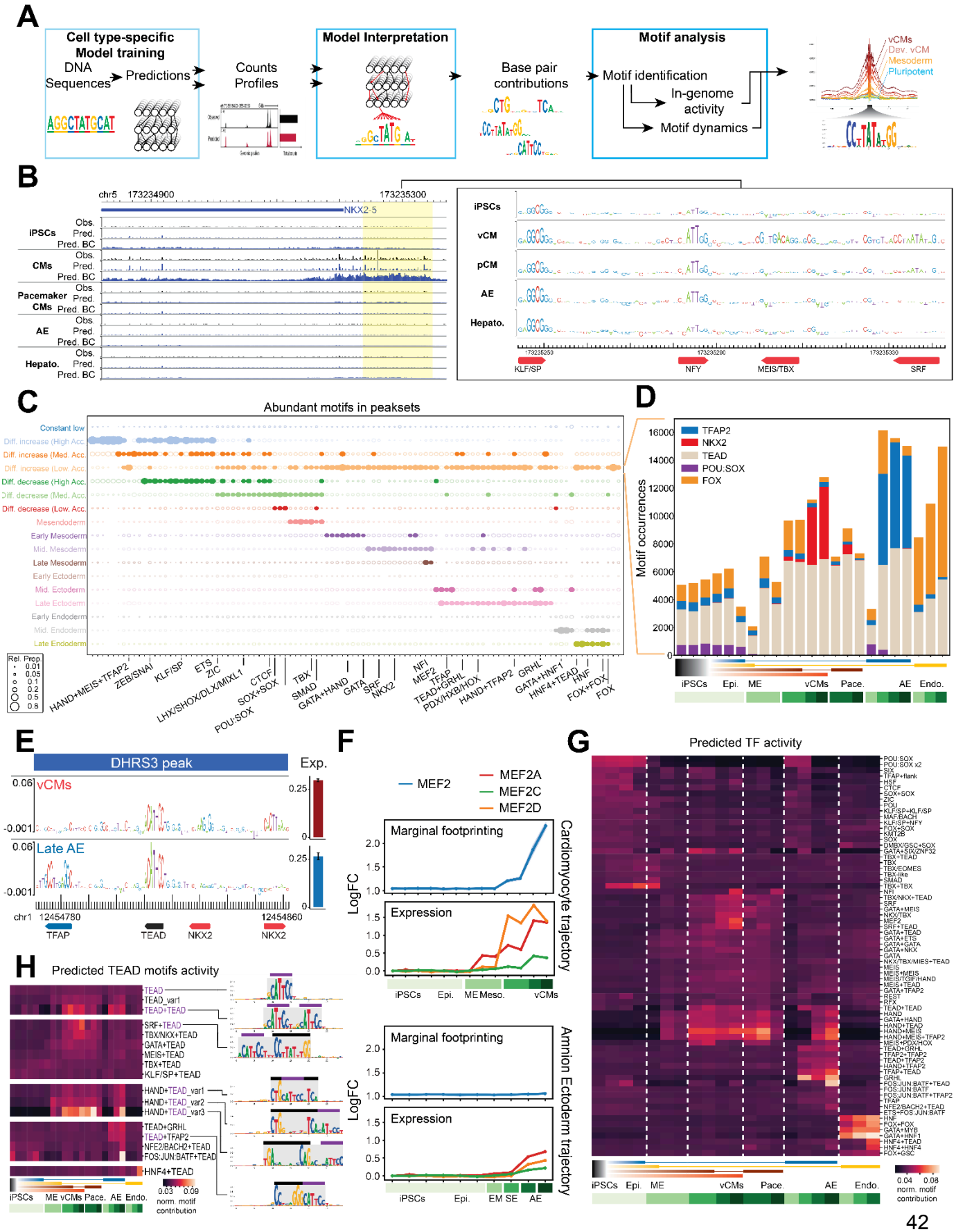
A
Cell type-specific Model training
DNA Sequences
Predictions
Counts Profiles
Model Interpretation
Base pair contributions
Motif analysis
Motif identification
In-genome activity
Motif dynamics
vCMs
Dev. vCM
Mesoderm
Pluripotent
B
chr5 173234900
173235300
NKX2-5
iPSCs
CMs
Pacemaker CMs
AE
Hepato.
Obs.
Pred.
Pred. BC
vCM
pCM
KLF/SP
NFY
MEIS/TBX
SRF
C
Abundant motifs in peaksets
Constant low
Diff. increase (High Acc.)
Diff. increase (Med. Acc.)
Diff. increase (Low. Acc.)
Diff. decrease (High Acc.)
Diff. decrease (Med. Acc.)
Diff. decrease (Low. Acc.)
Mesendoderm
Early Mesoderm
Mid. Mesoderm
Late Mesoderm
Early Ectoderm
Mid. Ectoderm
Late Ectoderm
Early Endoderm
Mid. Endoderm
Late Endoderm
Rel. Prop.
HAND+MEIS+TFAP2
ZEB/SNAI
KLF/SP
ETS
ZIC
LHX/SHOX/DLX/MIXL1
CTCF
SOX+SOX
POU:SOX
TBX
SMAD
GATA+HAND
GATA
SRF
NKX2
NFI
MEF2
TFAP
TEAD+GRHL
PDX/HXB/HOX
HAND+TFAP2
GRHL
GATA+HNF1
HNF4+TEAD
HNF
FOX+FOX
FOX
D
TFAP2
NKX2
TEAD
POU:SOX
FOX
Motif occurrences
iPSCs
Epi.
ME
vCMs
Pace.
AE
Endo.
E
DHRS3 peak
Exp.
vCMs
Late AE
chr1 12454780
12454860
TFAP
TEAD
NKX2
NKX2
F
MEF2
MEF2A
MEF2C
MEF2D
Marginal footprinting
Expression
LogFC
Cardiomyocyte trajectory
Amnion Ectoderm trajectory
iPSCs
Epi.
ME
Meso.
vCMs
EM
SE
AE
G
Predicted TF activity
norm. motif contribution
H
Predicted TEAD motifs activity
TEAD
TEAD_var1
TEAD+TEAD
SRF+TEAD
TBX/NKX+TEAD
GATA+TEAD
MEIS+TEAD
TBX+TEAD
KLF/SP+TEAD
HAND+TEAD_var1
HAND+TEAD_var2
HAND+TEAD_var3
TEAD+GRHL
TEAD+TFAP2
NFE2/BACH2+TEAD
FOS:JUN:BATF+TEAD
HNF4+TEAD
norm. motif contribution

Figure 3. (A) chromBPNet is a basepair-resolution deep learning model that predicts cell state-specific TF activity. (B) Observed (black) and predicted (before and after bias correction, top and bottom blue tracks, respectively) accessibility in the *NKX2-5* locus for several cell types (left panel) shows increased accessibility and a high correlation with predicted accessibility in mature vCMs. The TSS and proximal promoter regions (highlighted, right panel) are predicted to have high contributions from KLF/SP, NFY, MEIS/TBX and SRF motifs. (C) Performing Motif abundance enrichment in peak sets (Fig. 1F) using chromBPNet predictions allows us to identify specific cell-state profiles of accessible motifs. (D) An allegedly canonical peak set (accessibility increases with differentiation in all lineages) has, in fact, cell-type-specific compositions of different motifs. Whereas TEAD has the same activity dynamics in all lineages in these peaks, other TFs show cell-state-specific activity. (E) Cell-state-specific motifs differentially regulate a *DHRS3*-related peak: Left - TEAD is active in both cell types, but TFAP and NKX2 motifs are cell-state-specific. Right - *DHRS3* log-norm expression. (F) marginal footprinting for MEF2 motif (blue) in two lineages (upper versus lower panels) and corresponding expression dynamics (color lines). In the vCM lineage, MEF2 motif activity is explained by sequential expression patterns of the MEF2 genes. Although MEF2 is lowly expressed in the AE lineage, it is not predicted to have binding activity. (G) *in silico*, row-normalized cell-type-specific motif contribution (Marginal footprinting) shows various predicted motif activities throughout lineages. (H) The TEAD motif family *in silico*, row-normalized contributions, reveals differential predicted activity of TEAD variant motifs and TEAD composite motifs. Purple bar - TEAD motif; Black bar - partner motif.

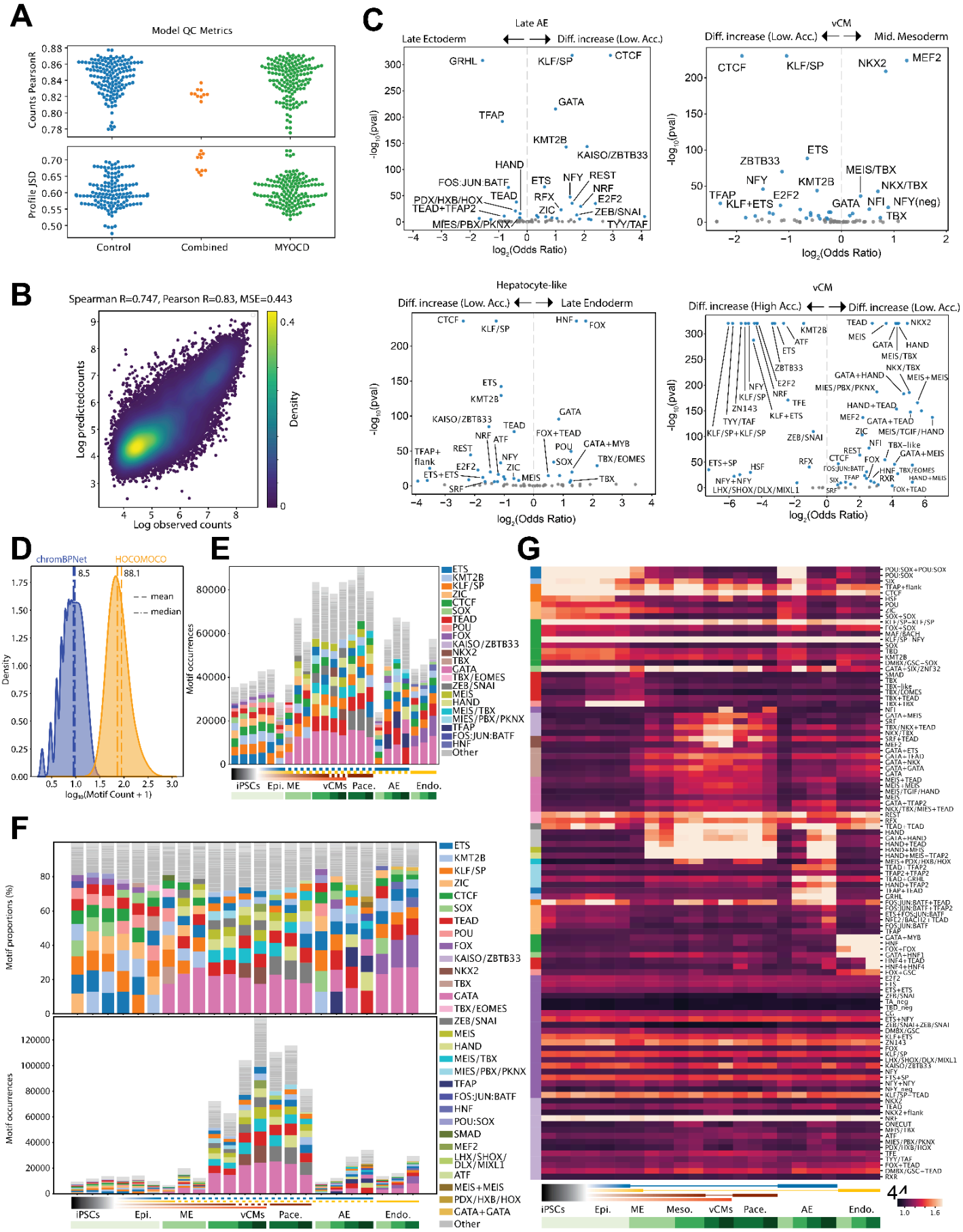
A
Model QC Metrics
Counts PearsonR
Profile JSD
Control
Combined
MYOCD
B
Spearman R=0.747, Pearson R=0.83, MSE=0.443
Log predicted counts
Log observed counts
Density
C
Late AE
Late Ectoderm
Diff. increase (Low. Acc.)
vCM
Mid. Mesoderm
Hepatocyte-like
Late Endoderm
Diff. increase (High Acc.)
-log10(pval)
log2(Odds Ratio)
D
chromBPNet
HOCOMOCO
mean
median
Density
log10(Motif Count + 1)
E
Motif occurrences
iPSCs
Epi.
ME
vCMs
Pace.
AE
Endo.
F
Motif proportions (%)
Motif occurrences
G
Meso.

Figure S3. (A) chromBPnet QCs. Counts Pearson R of the control and MYOCD KD cardioids. For the Endothelial cells, we merged the fragments to train the model ("Combined"). All models achieved correlations ranging from 0.78 to 0.88, with a median of 0.85. We also quantified the extent to which the base pair-resolution predictions of the model were accurate by computing the Jensen-Shannon Distance (JSD) between the observed and predicted distributions of counts across bases per peak. For each cell-type-specific model, the median JSD among all peaks in that cell-type was an average of 0.61. (B) Correlation of predicted vs observed counts for control vCM cluster. (C) log Odd-Ratio of motif abundance between different peak sets in Fig. 3C. (D) Number of motifs per peak for chromBPNet and HOCOMOCO V12 (blue with mean=8.5; median=8 and orange with mean=88.1; median=72, respectively) (E) Same as Fig. 3D but with all motifs. The top 10 abundant motifs in each cell state are annotated. (F) Motif proportion and occurrences for Fig. 3C, peak cluster "Mid. Mesoderm" (G) Log_2FC values for Fig. 3G.

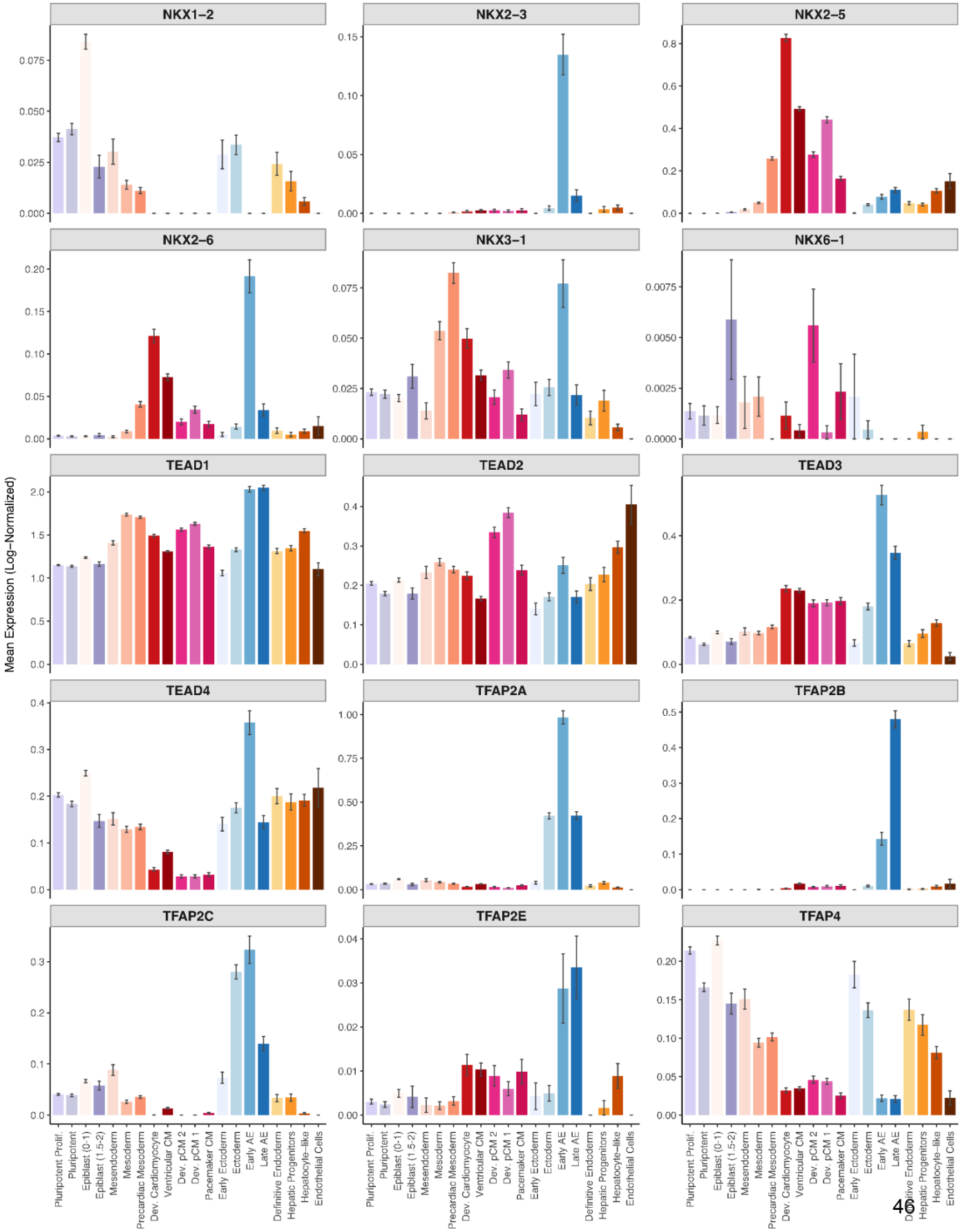
NKX1−2
NKX2−3
NKX2−5
NKX2−6
NKX3−1
NKX6−1
TEAD1
TEAD2
TEAD3
TEAD4
TFAP2A
TFAP2B
TFAP2C
TFAP2E
TFAP4
Mean Expression (Log−Normalized)
Pluripotent Prolif.
Pluripotent
Epiblast (0-1)
Epiblast (1.5-2)
Mesendoderm
Mesoderm
Precardiac Mesoderm
Dev. Cardiomyocyte
Ventricular CM
Dev. pCM 2
Dev. pCM 1
Pacemaker CM
Early Ectoderm
Ectoderm
Early AE
Late AE
Definitive Endoderm
Hepatic Progenitors
Hepatocyte−like
Endothelial Cells

Figure S4. NKX, TEAD and TFAP gene families' expression levels in our cardioid cell states. Genes with counts below 0.001 were excluded.

A

FOXA1
FOXA2
FOXA3
FOXB1
FOXC1
FOXC2
FOXD3
FOXE3
FOXF1
FOXF2
FOXH1
FOXI3
FOXJ1
FOXJ2
FOXJ3
FOXK1
FOXK2
FOXL1
FOXM1
FOXN2
FOXN3
FOXN4
FOXO1
FOXO3
FOXO4
FOXO6
FOXP1
FOXP2
FOXP3
FOXP4
FOXQ1
Mean Expression (Log−Normalized)
Pluripotent Prolif.
Pluripotent
Epiblast (0-1)
Epiblast (1.5-2)
Mesendoderm
Mesoderm
Precardiac Mesoderm
Dev. Cardiomyocyte
Ventricular CM
Dev. pCM 2
Dev. pCM 1
Pacemaker CM
Early Ectoderm
Ectoderm
Early AE
Late AE
Definitive Endoderm
Hepatic Progenitors
Hepatocyte−like
Endothelial Cells

Figure S5. FOX gene family expression levels in our cardioid cell states. Genes with counts below 0.001 were excluded.

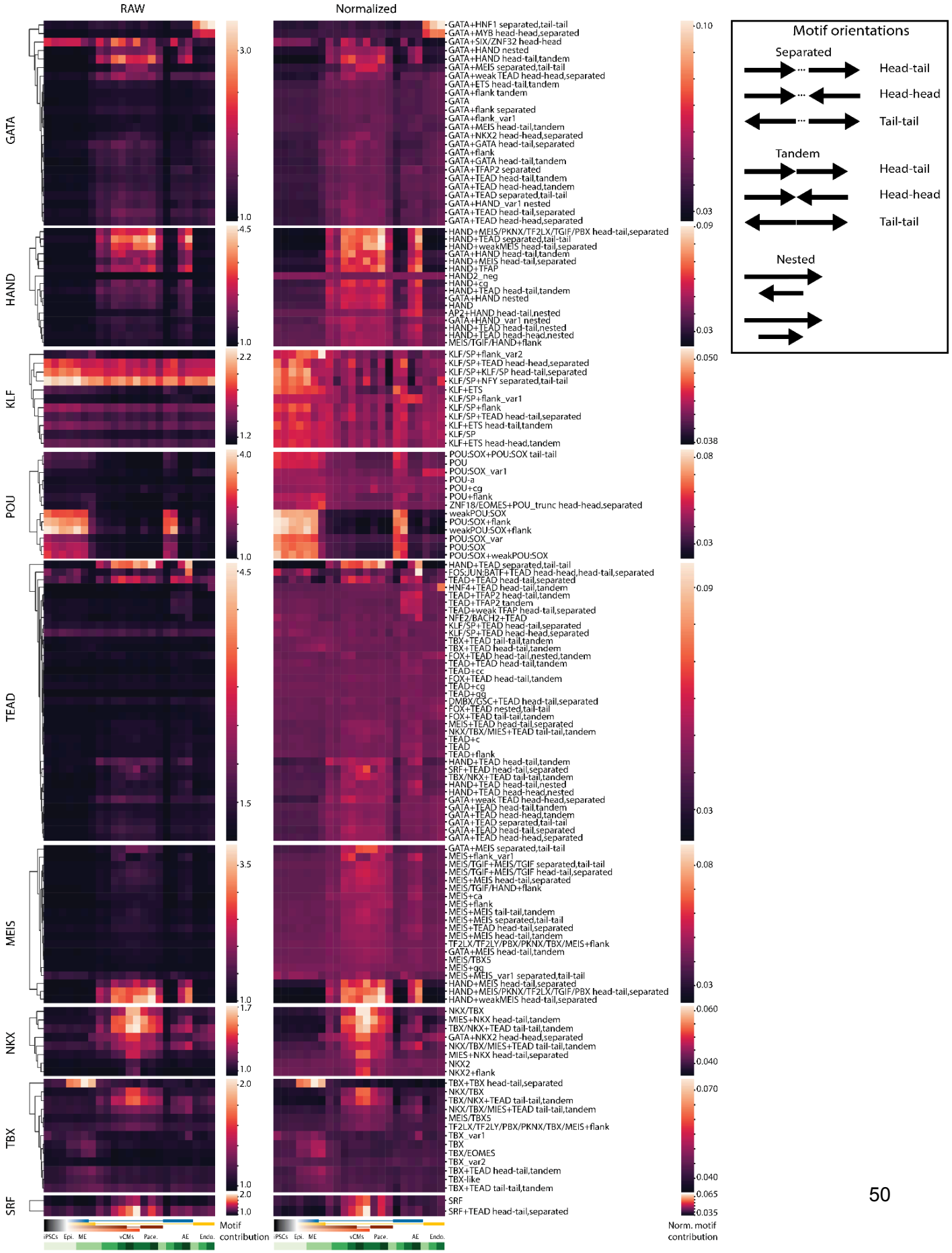
RAW
Normalized
GATA
HAND
KLF
POU
TEAD
MEIS
NKX
TBX
SRF
GATA+HNF1 separated,tail-tail
GATA+MYB head-head,separated
GATA+SIX/ZNF32 head-head
GATA+HAND nested
GATA+HAND head-tail,tandem
GATA+MEIS separated,tail-tail
GATA+weak TEAD head-head,separated
GATA+ETS head-tail,tandem
GATA+flank tandem
GATA
GATA+flank separated
GATA+flank_var1
GATA+MEIS head-tail,tandem
GATA+NKX2 head-head,separated
GATA+GATA head-tail,separated
GATA+flank
GATA+GATA head-tail,tandem
GATA+TFAP2 separated
GATA+TEAD head-tail,tandem
GATA+TEAD head-head,tandem
GATA+TEAD separated,tail-tail
GATA+HAND_var1 nested
GATA+TEAD head-tail,separated
GATA+TEAD head-head,separated
HAND+MEIS/PKNX/TF2LX/TGIF/PBX head-tail,separated
HAND+TEAD separated,tail-tail
HAND+weakMEIS head-tail,separated
GATA+HAND head-tail,tandem
HAND+MEIS head-tail,separated
HAND+TFAP
HAND2_neg
HAND+cg
HAND+TEAD head-tail,tandem
GATA+HAND nested
HAND
AP2+HAND head-tail,nested
GATA+HAND_var1 nested
HAND+TEAD head-tail,nested
HAND+TEAD head-head,nested
MEIS/TGIF/HAND+flank
KLF/SP+flank_var2
KLF/SP+TEAD head-head,separated
KLF/SP+KLF/SP head-tail,separated
KLF/SP+NFY separated,tail-tail
KLF+ETS
KLF/SP+flank_var1
KLF/SP+flank
KLF/SP+TEAD head-tail,separated
KLF+ETS head-tail,tandem
KLF/SP
KLF+ETS head-head,tandem
POU:SOX+POU:SOX tail-tail
POU
POU:SOX_var1
POU-a
POU+cg
POU+flank
ZNF18/EOMES+POU_trunc head-head,separated
weakPOU:SOX
POU:SOX+flank
weakPOU:SOX+flank
POU:SOX_var
POU:SOX
POU:SOX+weakPOU:SOX
HAND+TEAD separated,tail-tail
FOS:JUN:BATF+TEAD head-head,head-tail,separated
TEAD+TEAD head-tail,separated
HNF4+TEAD head-tail,tandem
TEAD+TFAP2 head-tail,tandem
TEAD+TFAP2 tandem
TEAD+weak TFAP head-tail,separated
NFE2/BACH2+TEAD
KLF/SP+TEAD head-tail,separated
KLF/SP+TEAD head-head,separated
TBX+TEAD tail-tail,tandem
TBX+TEAD head-tail,tandem
FOX+TEAD head-tail,nested,tandem
TEAD+TEAD head-tail,tandem
TEAD+cc
FOX+TEAD head-tail,tandem
TEAD+cg
TEAD+gg
DMBX/GSC+TEAD head-tail,separated
FOX+TEAD nested,tail-tail
FOX+TEAD tail-tail,tandem
MEIS+TEAD head-tail,separated
NKX/TBX/MIES+TEAD tail-tail,tandem
TEAD+c
TEAD
TEAD+flank
HAND+TEAD head-tail,tandem
SRF+TEAD head-tail,separated
TBX/NKX+TEAD tail-tail,tandem
HAND+TEAD head-tail,nested
HAND+TEAD head-head,nested
GATA+weak TEAD head-head,separated
GATA+TEAD head-tail,tandem
GATA+TEAD head-head,tandem
GATA+TEAD separated,tail-tail
GATA+TEAD head-tail,separated
GATA+TEAD head-head,separated
GATA+MEIS separated,tail-tail
MEIS+flank_var1
MEIS/TGIF+MEIS/TGIF separated,tail-tail
MEIS/TGIF+MEIS/TGIF head-tail,separated
MEIS+MEIS head-tail,separated
MEIS/TGIF/HAND+flank
MEIS+ca
MEIS+flank
MEIS+MEIS tail-tail,tandem
MEIS+MEIS separated,tail-tail
MEIS+TEAD head-tail,separated
MEIS+MEIS head-tail,tandem
TF2LX/TF2LY/PBX/PKNX/TBX/MEIS+flank
GATA+MEIS head-tail,tandem
MEIS/TBX5
MEIS+gg
MEIS+MEIS_var1 separated,tail-tail
HAND+MEIS head-tail,separated
HAND+MEIS/PKNX/TF2LX/TGIF/PBX head-tail,separated
HAND+weakMEIS head-tail,separated
NKX/TBX
MIES+NKX head-tail,tandem
TBX/NKX+TEAD tail-tail,tandem
GATA+NKX2 head-head,separated
NKX/TBX/MIES+TEAD tail-tail,tandem
MIES+NKX head-tail,separated
NKX2
NKX2+flank
TBX+TBX head-tail,separated
NKX/TBX
TBX/NKX+TEAD tail-tail,tandem
NKX/TBX/MIES+TEAD tail-tail,tandem
MEIS/TBX5
TF2LX/TF2LY/PBX/PKNX/TBX/MEIS+flank
TBX_var1
TBX
TBX/EOMES
TBX_var2
TBX+TEAD head-tail,tandem
TBX-like
TBX+TEAD tail-tail,tandem
SRF
SRF+TEAD head-tail,separated
iPSCs Epi. ME vCMs Pace. AE Endo.
Motif contribution
Norm. motif contribution
Motif orientations
Separated
Head-tail
Head-head
Tail-tail
Tandem
Head-tail
Head-head
Tail-tail
Nested

Figure S6. *in silico* marginalization for fine-grained motif families. Row normalized.

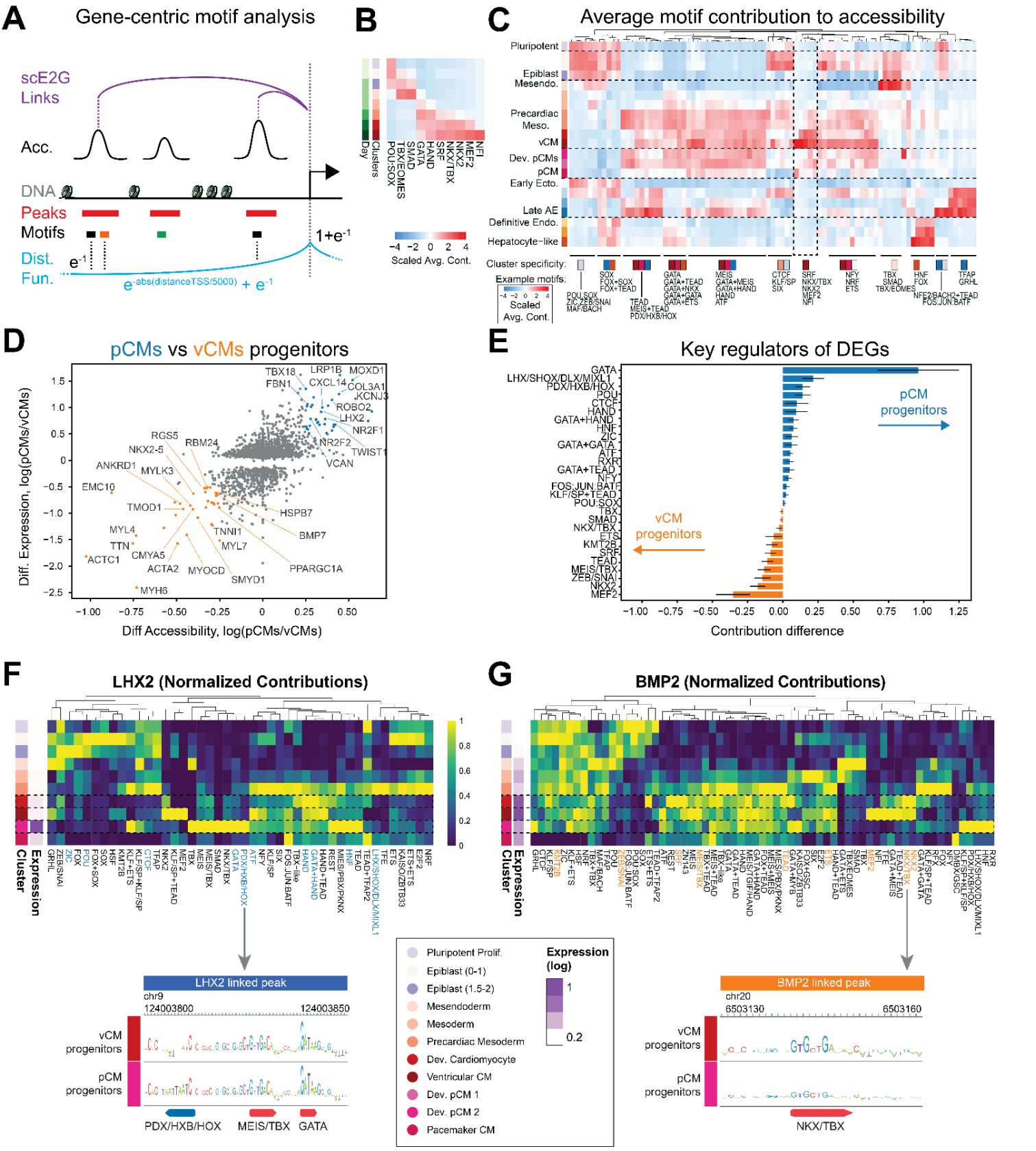
A
Gene-centric motif analysis
scE2G Links
Acc.
DNA
Peaks
Motifs
Dist. Fun.
1+e^-1
e^-1
e^-abs(distanceTSS/5000) + e^-1
B
Day
Clusters
POU:SOX
TBX/EOMES
GATA
SMAD
HAND
SRF
NKX/TBX
NKX2
MEF2
NFI
Scaled Avg. Cont.
C
Average motif contribution to accessibility
Pluripotent
Epiblast
Mesendo.
Precardiac Meso.
vCM
Dev. pCMs
pCM
Early Ecto.
Late AE
Definitive Endo.
Hepatocyte-like
Cluster specificity:
Example motifs:
Scaled Avg. Cont.
D
pCMs vs vCMs progenitors
Diff. Expression, log(pCMs/vCMs)
Diff Accessibility, log(pCMs/vCMs)
TBX18
LRP1B
MOXD1
FBN1
CXCL14
COL3A1
KCNJ3
ROBO2
LHX2
NR2F1
NR2F2
TWIST1
VCAN
RGS5
RBM24
NKX2-5
ANKRD1
MYLK3
EMC10
TMOD1
HSPB7
MYL4
TNNI1
BMP7
TTN
MYL7
ACTC1
CMYA5
ACTA2
MYOCD
PPARGC1A
SMYD1
MYH6
E
Key regulators of DEGs
pCM progenitors
vCM progenitors
Contribution difference
F
LHX2 (Normalized Contributions)
Cluster
Expression
LHX2 linked peak
chr9
124003800
124003850
vCM progenitors
pCM progenitors
PDX/HXB/HOX
MEIS/TBX
GATA
Pluripotent Prolif.
Epiblast (0-1)
Epiblast (1.5-2)
Mesendoderm
Mesoderm
Precardiac Mesoderm
Dev. Cardiomyocyte
Ventricular CM
Dev. pCM 1
Dev. pCM 2
Pacemaker CM
Expression (log)
G
BMP2 (Normalized Contributions)
BMP2 linked peak
chr20
6503130
6503160
NKX/TBX

Figure 4. (A) The gene-centric approach links active motifs in a cell-state-specific manner. Motifs in called peaks are filtered by consensus scE2G peak-gene linkages and are weighted down by a distance function before summing their contributions. (B) Average in genome contribution for the SRF regulome in the vCM lineage. (C) Average in genome contribution for the entire motif compendium across all cell states. The sums of contributions (intensity) for each motif family were averaged and scaled across all expressed genes (approximately 12,000 genes) in each cell state. The major cell-state specificity is indicated at the bottom, alongside motif examples (Fig.S4B for all annotations). (D) Comparison of differential genes (y-axis) versus differential accessibility (x-axis) between vCMs and pCMs progenitors. Marked genes represent the top differentials in either trajectory. (E) Differences in average contributions for each motif between vCM and pCM progenitors. Positive values indicate higher contributions in pCMs, while negative values indicate higher contributions in vCM progenitors. Error bars represent standard deviation (SD). (F-G) Motif syntaxes for *LHX2* (F) and *BMP2* (G), which are differentially expressed genes in pCMs and vCMs, respectively. Dashed boxes highlight vCM progenitors (top) and pCM progenitors (bottom), and the blue/orange annotations indicate the key regulators of the lineages shown in (E). Panels at the bottom display example peaks and motif contributions for each gene.

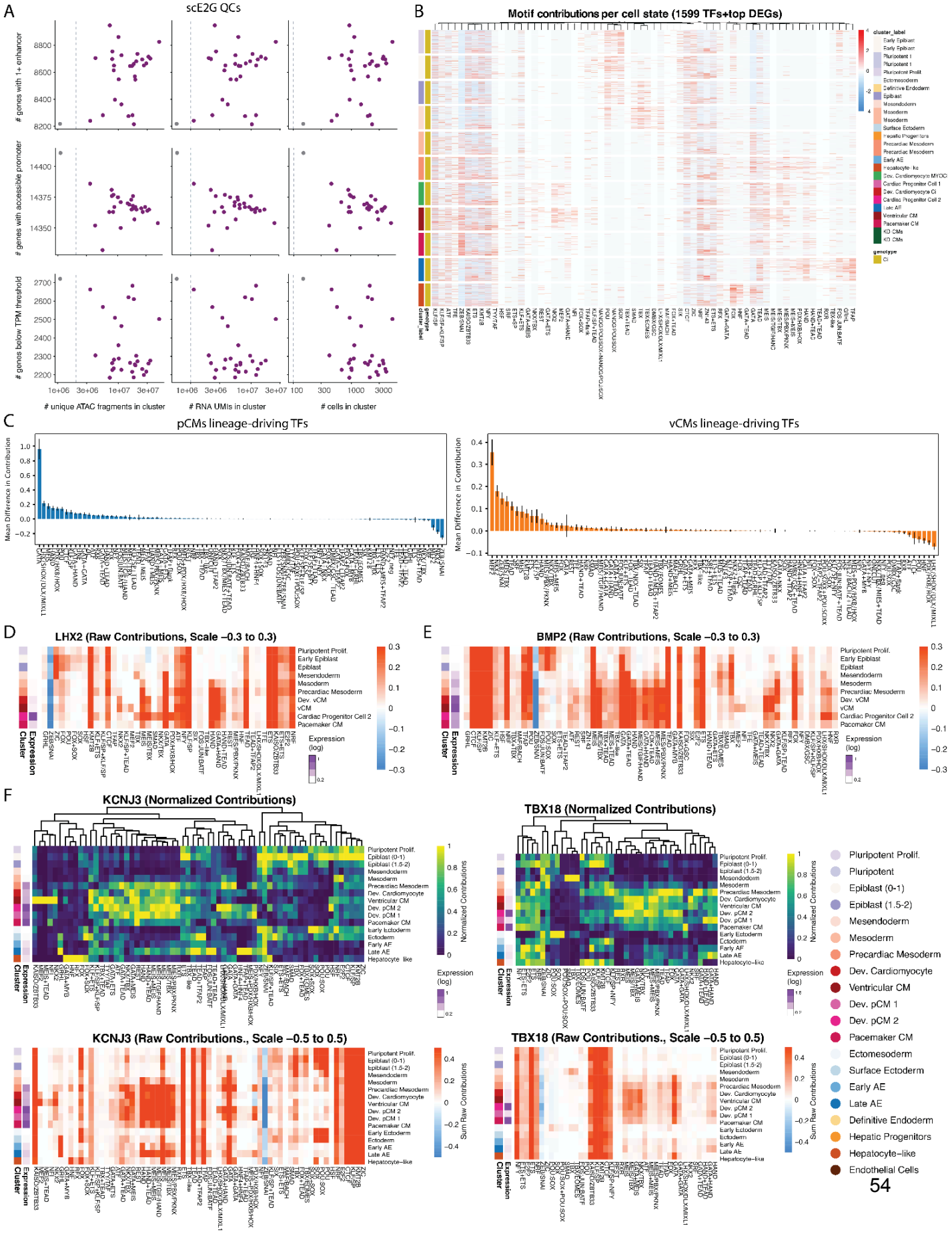
A
scE2G QCs
genes with 1+ enhancer
genes with accessible promoter
genes below TPM threshold
unique ATAC fragments in cluster
RNA UMIs in cluster
cells in cluster
B
Motif contributions per cell state (1599 TFs+top DEGs)
C
pCMs lineage-driving TFs
vCMs lineage-driving TFs
Mean Difference in Contribution
D
LHX2 (Raw Contributions, Scale −0.3 to 0.3)
E
BMP2 (Raw Contributions, Scale −0.3 to 0.3)
F
KCNJ3 (Normalized Contributions)
TBX18 (Normalized Contributions)
KCNJ3 (Raw Contributions., Scale −0.5 to 0.5)
TBX18 (Raw Contributions., Scale −0.5 to 0.5)
Pluripotent Prolif.
Pluripotent
Epiblast (0-1)
Epiblast (1.5-2)
Mesendoderm
Mesoderm
Precardiac Mesoderm
Dev. Cardiomyocyte
Ventricular CM
Dev. pCM 1
Dev. pCM 2
Pacemaker CM
Ectomesoderm
Surface Ectoderm
Early AE
Late AE
Definitive Endoderm
Hepatic Progenitors
Hepatocyte-like
Endothelial Cells

Figure S7. (A) scE2G QCs. All cell states passed the QC threshold, except for the Endothelial cells (grey dot), which likely resulted from very low cell numbers, resulting in UMIs and fragment numbers below the threshold. (B) in genome contribution for the entire motif compendium across all cell states. The sums of contributions (intensity) for each motif family are column-scaled across all expressed genes (approximately 12,000 genes) in each cell state. (C) Average contributions for each motif in pCM and VCM progenitors (left and right, respectively). Positive values indicate higher contributions in pCMs, while negative values indicate higher contributions in vCM progenitors. Error bars represent standard deviation (SD). (D-E) Raw sum of contributions for motif syntaxes for *LHX2* (D) and *BMP2* (E). (F) Normalized and Raw sum of contributions for motif syntaxes for *KCNJ3* and *TBX18*.

A
ACTA2
ANKRD1
ATF3
BMP4
CD34
CDH1
CDH2
CDH3
CLDN1
CLDN4
CLDN7
CRB3
DDIT3
ELK1
ELK3
EPCAM
FN1
FOXE3
ID2
ITGB4
KRT18
LLGL2
NIN
OCLN
PARD6A
PARD6B
PARD6G
SMAD3
SMAD4
SNAI1
SNAI2
VIM
ZEB1
ZEB2
Mean Expression (Log−Normalized)
Pluripotent Prolif.
Pluripotent
Epiblast (0-1)
Epiblast (1.5-2)
Mesendoderm
Mesoderm
Precardiac Mesoderm
Dev. Cardiomyocyte
Ventricular CM
Dev. pCM 2
Dev. pCM 1
Pacemaker CM
Early Ectoderm
Ectoderm
Early AE
Late AE
Definitive Endoderm
Hepatic Progenitors
Hepatocyte−like
Endothelial Cells

Figure S8. ZEB/SNAI genes and associated targets expression levels in our cardioid cell states. Genes with counts below 0.001 were excluded.

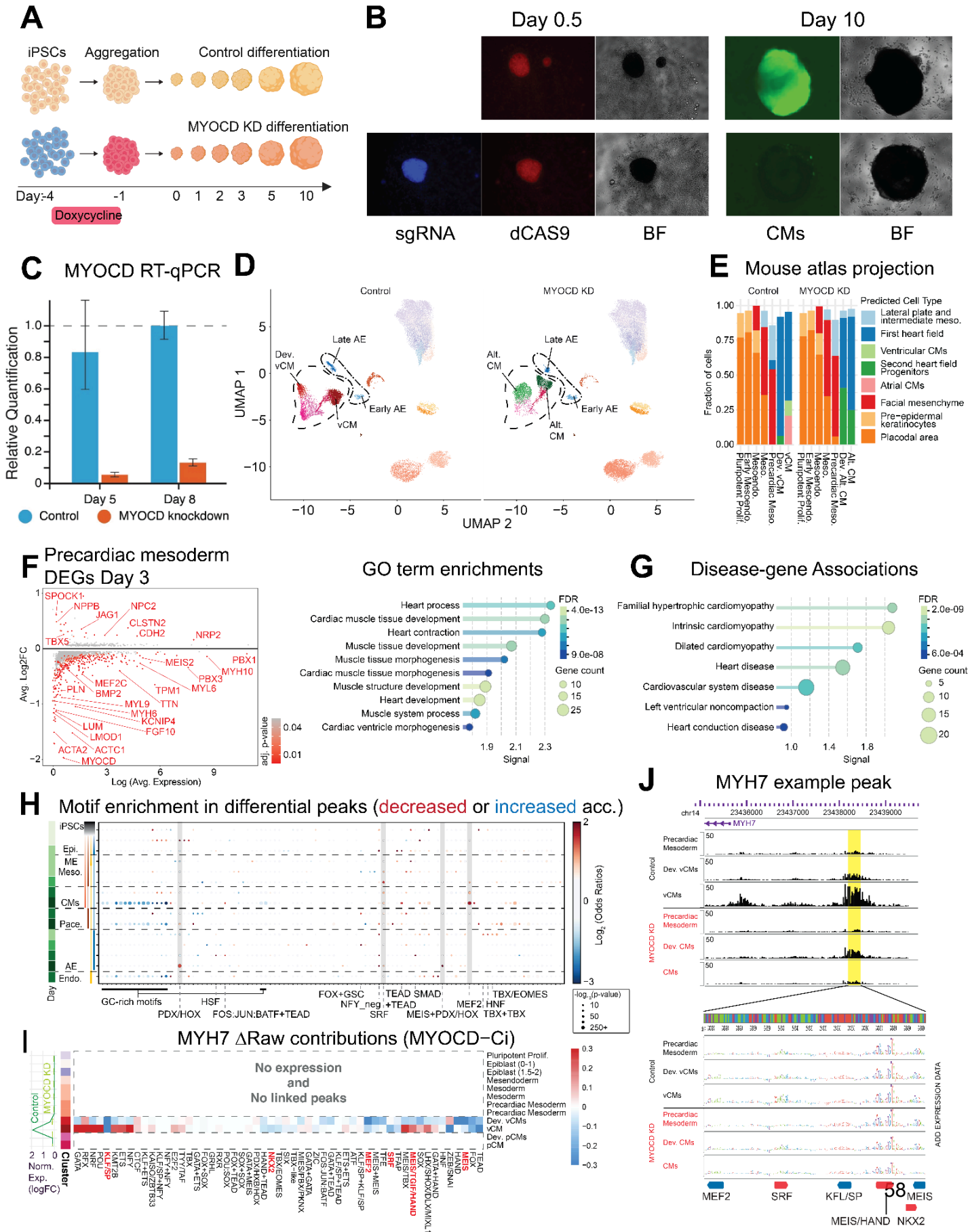
A
iPSCs
Aggregation
Control differentiation
MYOCD KD differentiation
Day:-4
-1
0 1 2 3 5 10
Doxycycline
B
Day 0.5
Day 10
sgRNA
dCAS9
BF
CMs
BF
C
MYOCD RT-qPCR
Relative Quantification
Day 5
Day 8
Control
MYOCD knockdown
D
Control
MYOCD KD
Late AE
Dev. vCM
Early AE
vCM
Alt. CM
UMAP 1
UMAP 2
E
Mouse atlas projection
Control
MYOCD KD
Fraction of cells
Predicted Cell Type
Lateral plate and intermediate meso.
First heart field
Ventricular CMs
Second heart field Progenitors
Atrial CMs
Facial mesenchyme
Pre-epidermal keratinocytes
Placodal area
F
Precardiac mesoderm DEGs Day 3
SPOCK1
NPPB
JAG1
NPC2
CLSTN2
CDH2
NRP2
TBX5
MEIS2
PBX1
MYH10
PBX3
MYL6
MEF2C
BMP2
PLN
TPM1
MYL9
TTN
MYH6
KCNIP4
FGF10
LUM
LMOD1
ACTA2
ACTC1
MYOCD
Avg. Log2FC
Log (Avg. Expression)
adj. p-value
GO term enrichments
Heart process
Cardiac muscle tissue development
Heart contraction
Muscle tissue development
Muscle tissue morphogenesis
Cardiac muscle tissue morphogenesis
Muscle structure development
Heart development
Muscle system process
Cardiac ventricle morphogenesis
FDR
Gene count
Signal
G
Disease-gene Associations
Familial hypertrophic cardiomyopathy
Intrinsic cardiomyopathy
Dilated cardiomyopathy
Heart disease
Cardiovascular system disease
Left ventricular noncompaction
Heart conduction disease
FDR
Gene count
Signal
H
Motif enrichment in differential peaks (decreased or increased acc.)
iPSCs
Epi.
ME Meso.
CMs
Pace.
AE
Endo.
Day
GC-rich motifs
HSF
PDX/HOX
FOS:JUN:BATF+TEAD
FOX+GSC
NFY_neg
SRF
TEAD SMAD
+TEAD
MEIS+PDX/HOX
MEF2
TBX/EOMES
HNF
TBX+TBX
Log2 (Odds Ratios)
-log10(p-value)
I
MYH7 ΔRaw contributions (MYOCD−Ci)
No expression
and
No linked peaks
Control
MYOCD KD
Norm. Exp. (logFC)
Cluster
Pluripotent Prolif.
Epiblast (0-1)
Epiblast (1.5-2)
Mesendoderm
Mesoderm
Mesoderm
Precardiac Mesoderm
Precardiac Mesoderm
Dev. vCMs
vCM
Dev. pCMs
pCM
J
MYH7 example peak
chr14
23436000
23437000
23438000
23439000
MYH7
Precardiac Mesoderm
Dev. vCMs
vCMs
Dev. CMs
CMs
Control
MYOCD KD
ADD EXPRESSION DATA
MEF2
SRF
KFL/SP
MEIS
MEIS/HAND
NKX2

Figure 5. (A) *MYOCD* KD experimental scheme. (B) *MYOCD* KD images taken 12 hours after aggregation and on Day 10 of differentiation. On Day 0.5, both the sgRNA (BFP) and the dCAS9 (mCherry) are still visible. By Day 10 of differentiation, the control cardioid displays robust beating vCMs (*MYL7*-GFP), while the *MYOCD* KD predominantly lacks them. (C) RT-qPCR relative expression of *MYOCD* in control and *MYOCD* KD cardioids on days 5 and 8 indicates ~90% KD efficiency. (D) scRNA UMAPs show the depletion of specific cell types (vCMs and AE lineages), whereas other clusters are shared across conditions. (E) Differential prediction of cell types, achieved by projecting the cardioid scRNA data onto a mouse developmental atlas, reveals that alternative vCMs are more closely aligned with an SHF state than with the mature states of CMs. (F) Differential gene expression at Day 3 of cardioid differentiation between *MYOCD* KD and control (left panel) and the associated GO term enrichments (right panel). (G) Disease association of the differential genes determined by DISEASES enrichment analysis. (H) Motif enrichment in differential peaks following *MYOCD* KD. Motifs enriched in peaks with decreased accessibility show positive log(OR), and those in peaks with increased accessibility demonstrate negative log(OR) following *MYOCD* KD. Grey vertical bars highlight motif examples with high cell-state-specificity. (I) The raw sum of the contribution difference between *MYOCD* KD and control illustrates how the *MYH7* regulatory syntax is impacted following *MYOCD* KD. The dashed box indicates no significant expression and no linked peaks via scE2G in specific cell states. Expression (Dark/light green for Control/*MYOCD* KD, respectively) was normalized to the iPSCs' basal expression levels. Red motif annotations correspond to motifs identified in the peak example (J). (J) Upstream MYH7-linked peak demonstrates motif-altered dynamics upon *MYOCD* KD.

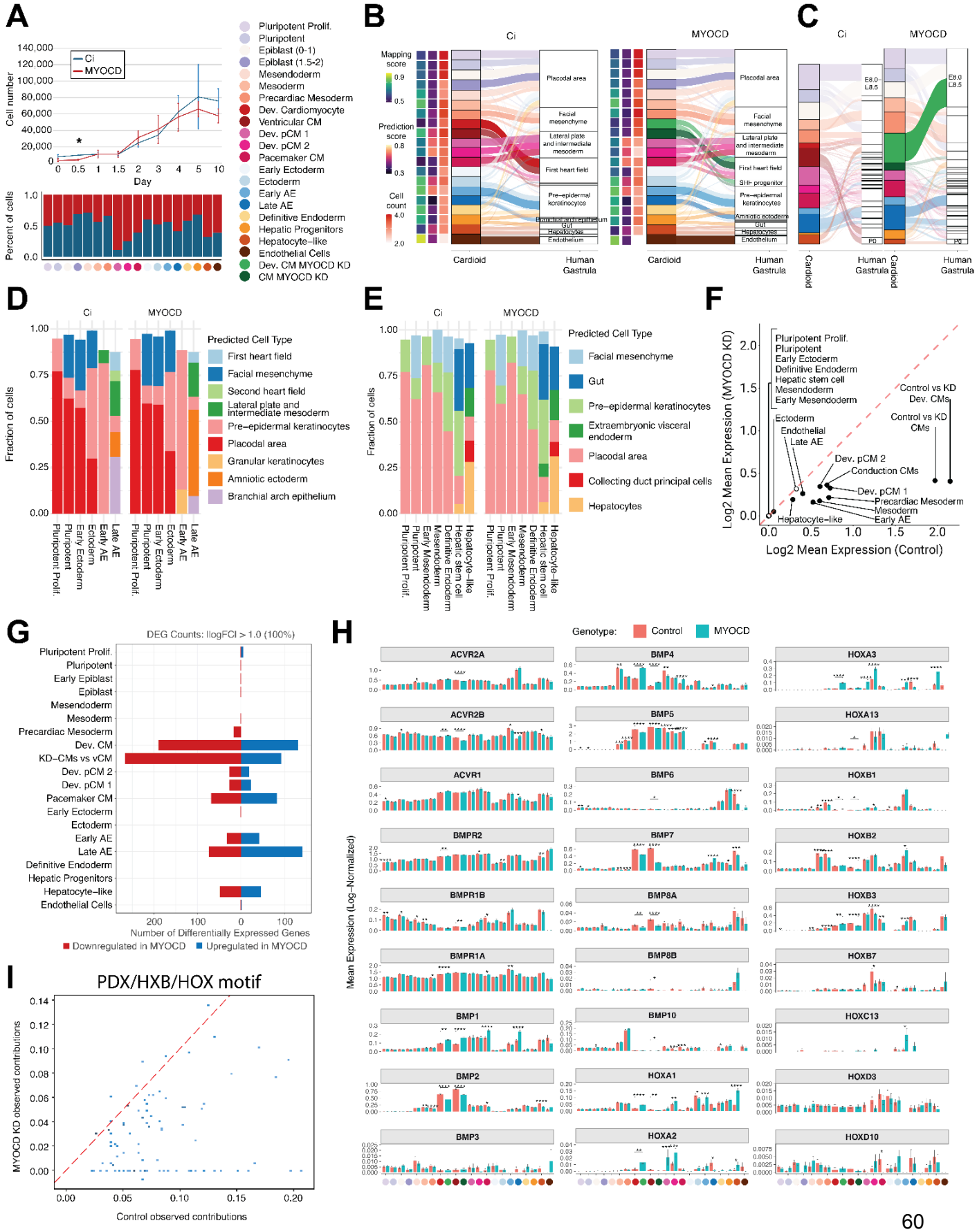
A
Cell number
Ci
MYOCD
Day
Percent of cells
Pluripotent Prolif.
Pluripotent
Epiblast (0-1)
Epiblast (1.5-2)
Mesendoderm
Mesoderm
Precardiac Mesoderm
Dev. Cardiomyocyte
Ventricular CM
Dev. pCM 1
Dev. pCM 2
Pacemaker CM
Early Ectoderm
Ectoderm
Early AE
Late AE
Definitive Endoderm
Hepatic Progenitors
Hepatocyte–like
Endothelial Cells
Dev. CM MYOCD KD
CM MYOCD KD
B
Mapping score
Prediction score
Cell count
Placodal area
Facial mesenchyme
Lateral plate and intermediate mesoderm
First heart field
Pre–epidermal keratinocytes
Gut
Hepatocytes
Endothelium
Cardioid
Human Gastrula
C
E8.0–L8.5
P0
D
Fraction of cells
Predicted Cell Type
First heart field
Facial mesenchyme
Second heart field
Lateral plate and intermediate mesoderm
Pre–epidermal keratinocytes
Placodal area
Granular keratinocytes
Amniotic ectoderm
Branchial arch epithelium
E
Facial mesenchyme
Gut
Pre–epidermal keratinocytes
Extraembryonic visceral endoderm
Placodal area
Collecting duct principal cells
Hepatocytes
F
Log2 Mean Expression (MYOCD KD)
Log2 Mean Expression (Control)
Pluripotent Prolif.
Pluripotent
Early Ectoderm
Definitive Endoderm
Hepatic stem cell
Mesendoderm
Early Mesendoderm
Ectoderm
Endothelial
Late AE
Control vs KD Dev. CMs
Control vs KD CMs
Dev. pCM 2
Conduction CMs
Dev. pCM 1
Precardiac Mesoderm
Mesoderm
Early AE
Hepatocyte–like
G
DEG Counts: |logFC| > 1.0 (100%)
Number of Differentially Expressed Genes
Downregulated in MYOCD
Upregulated in MYOCD
Dev. CM
KD–CMs vs vCM
H
Genotype:
Control
MYOCD
Mean Expression (Log–Normalized)
ACVR2A
ACVR2B
ACVR1
BMPR2
BMPR1B
BMPR1A
BMP1
BMP2
BMP3
BMP4
BMP5
BMP6
BMP7
BMP8A
BMP8B
BMP10
HOXA1
HOXA2
HOXA3
HOXA13
HOXB1
HOXB2
HOXB3
HOXB7
HOXC13
HOXD3
HOXD10
I
PDX/HXB/HOX motif
MYOCD KD observed contributions
Control observed contributions

Figure S9. (A) Top - cardioid cell number during differentiation. Astrix marks p-value < 0.05; two-sample two-tailed t-test. Bottom - cell number fraction of Control and *MYOCD* KD cardioids in shared cell-state annotations. (B) The projection breakdown shows the mapping of individual cells to the reference cell states. (C) The projection breakdown shows the mapping of individual cells to the reference developmental stage. (D) Same as Figure 5E but for Late AE. (E) Same as Figure 5E but for Hepatocyte-like cells. (F) *MYOCD* differential expression in the various cell states. (G) Number of DEGs between the control and *MYOCD* KD in all cell states. (H) BMP, BMPR and HOX gene expression. Significance by Wilcox test. (I) HOX/PDX/HXB in genome motif contributions in *MYOCD* KD vs Control for all control motif locations.

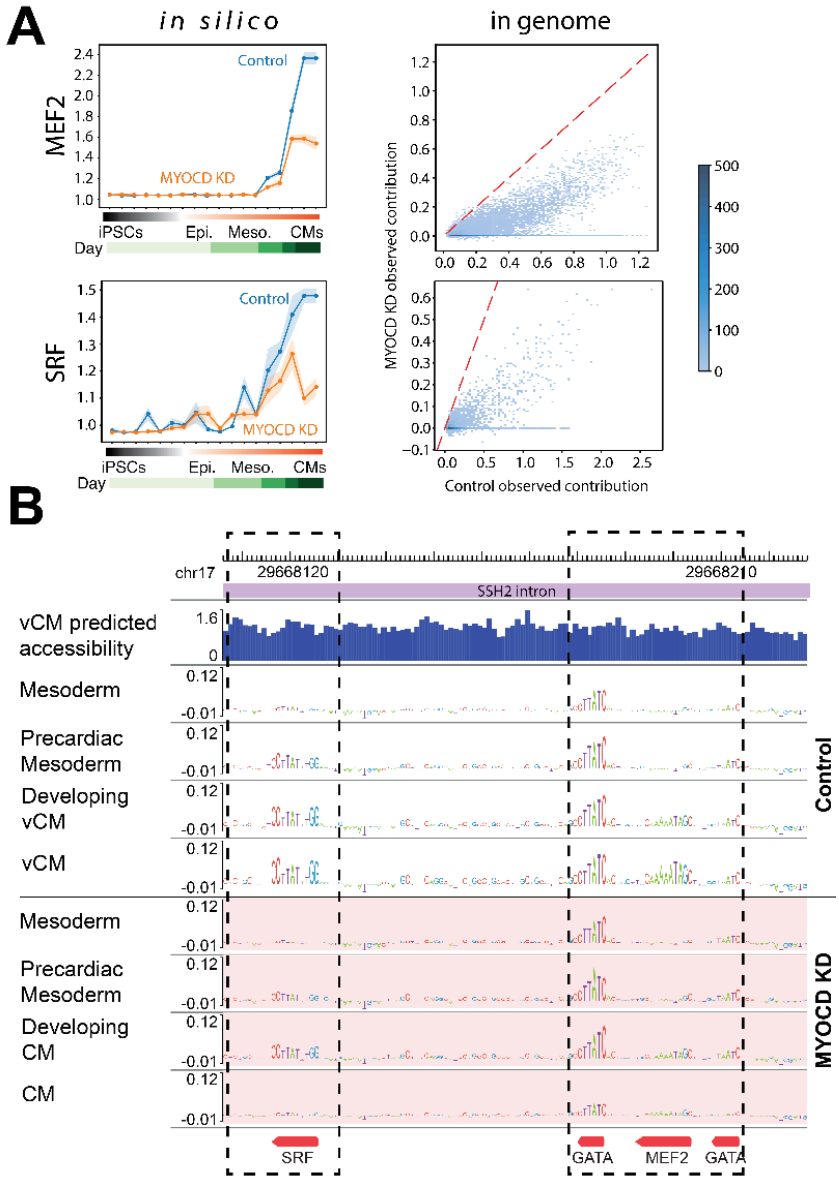
A
in silico
in genome
MEF2
Control
MYOCD KD
iPSCs
Epi.
Meso.
CMs
Day
SRF
MYOCD KD observed contribution
Control observed contribution
B
chr17
29668120
29668210
SSH2 intron
vCM predicted accessibility
Mesoderm
Precardiac Mesoderm
Developing vCM
vCM
Developing CM
CM
Control
MYOCD KD
SRF
GATA
MEF2
GATA

Figure S10. (A) chromBPNet predicted motifs' contributions. Left panels - *in silico* marginalization of MEF2 and SRF motifs along vCM differentiation. Right - *in genome* motif contributions in *MYOCD* KD vs Control for all control motif locations. (B) Comparison of motif contributions between Kosicki et al. [80], and cardioid data. HT2 enhancer (top panel) and mutation tracks (insert). Adopted from Kosicki et al. (C) Control and *MYOCD* KD contribution scores along CM differentiation.

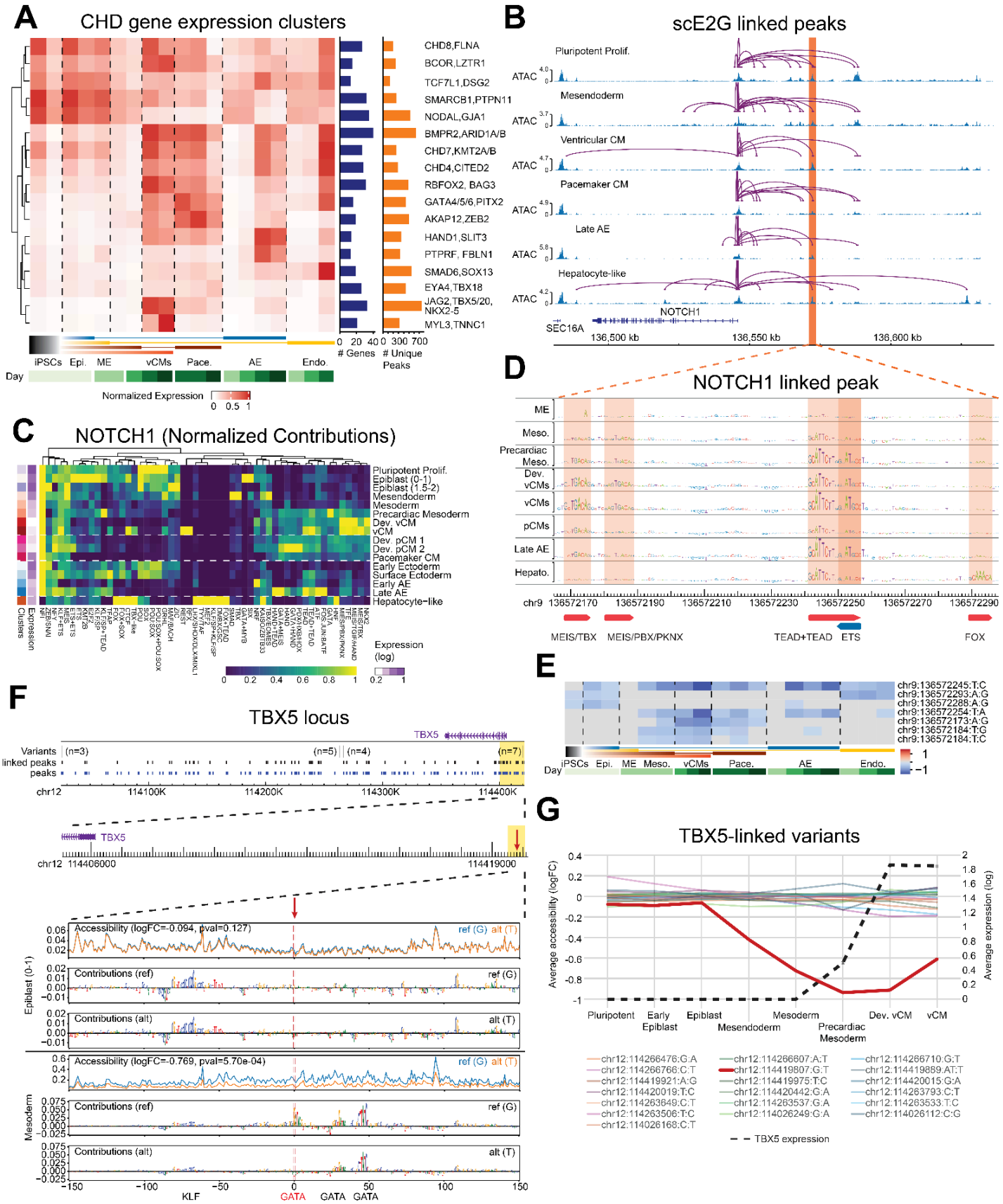
A
CHD gene expression clusters
CHD8,FLNA
BCOR,LZTR1
TCF7L1,DSG2
SMARCB1,PTPN11
NODAL,GJA1
BMPR2,ARID1A/B
CHD7,KMT2A/B
CHD4,CITED2
RBFOX2, BAG3
GATA4/5/6,PITX2
AKAP12,ZEB2
HAND1,SLIT3
PTPRF, FBLN1
SMAD6,SOX13
EYA4,TBX18
JAG2,TBX5/20, NKX2-5
MYL3,TNNC1
Genes
Unique Peaks
iPSCs Epi. ME vCMs Pace. AE Endo.
Day
Normalized Expression
B
scE2G linked peaks
Pluripotent Prolif.
Mesendoderm
Ventricular CM
Pacemaker CM
Late AE
Hepatocyte-like
ATAC
NOTCH1
SEC16A
136,500 kb
136,550 kb
136,600 kb
C
NOTCH1 (Normalized Contributions)
Pluripotent Prolif.
Epiblast (0-1)
Epiblast (1.5-2)
Mesendoderm
Mesoderm
Precardiac Mesoderm
Dev. vCM
vCM
Dev. pCM 1
Dev. pCM 2
Pacemaker CM
Early Ectoderm
Surface Ectoderm
Early AE
Late AE
Hepatocyte-like
Clusters
Expression
Expression (log)
D
NOTCH1 linked peak
ME
Meso.
Precardiac Meso.
Dev. vCMs
vCMs
pCMs
Late AE
Hepato.
chr9 136572170 136572190 136572210 136572230 136572250 136572270 136572290
MEIS/TBX
MEIS/PBX/PKNX
TEAD+TEAD
ETS
FOX
E
chr9:136572245:T:C
chr9:136572293:A:G
chr9:136572288:A:G
chr9:136572254:T:A
chr9:136572173:A:G
chr9:136572184:T:G
chr9:136572184:T:C
iPSCs Epi. ME Meso. vCMs Pace. AE Endo.
Day
F
TBX5 locus
TBX5
Variants
linked peaks
peaks
(n=3)
(n=5)
(n=4)
(n=7)
chr12 114100K 114200K 114300K 114400K
chr12 114406000 114419000
Accessibility (logFC=-0.094, pval=0.127)
ref (G) alt (T)
Contributions (ref)
Contributions (alt)
Epiblast (0-1)
Accessibility (logFC=-0.769, pval=5.70e-04)
Mesoderm
KLF GATA GATA GATA
G
TBX5-linked variants
Average accessibility (logFC)
Average expression (log)
Pluripotent
Early Epiblast
Epiblast
Mesendoderm
Mesoderm
Precardiac Mesoderm
Dev. vCM
vCM
chr12:114266476:G:A
chr12:114266607:A:T
chr12:114266710:G:T
chr12:114266766:C:T
chr12:114419807:G:T
chr12:114419889:AT:T
chr12:114419921:A:G
chr12:114419975:T:C
chr12:114420015:G:A
chr12:114420019:T:C
chr12:114420442:G:A
chr12:114263793:C:T
chr12:114263649:C:T
chr12:114263537:G:A
chr12:114263533:T:C
chr12:114263506:T:C
chr12:114026249:G:A
chr12:114026112:C:G
chr12:114026168:C:T
TBX5 expression

Figure 6. CHD gene mapping. (A) CHD gene norm expression. Blue bars are the number of genes in the cluster. Kmeans clustered (n=17). Orange bars represent the number of unique linked peaks for each CHD gene cluster. (B) *NOTCH1*-linked peaks in selected cell states. For each cell state, shown are linked peaks in purple and the corresponding ATAC tracks. Orange highlights a cdCRE element zoomed in (D). (C) *NOTCH1* motif contributions. Column normalized. (D) *NOTCH1* linked cdCRE peak (highlighted in (B)) with highlighted accessibility-contributing motifs. (E) effect of single-basepair polymorphism on accessibility across cell states of the seven highlighted motifs in (D). (F) Top - *TBX5* locus and known variants; Bottom - logFC effect on Accessibility and basepair contributions of the rs377307764 *TBX5* G>C variant in Early Epiblast (no effect) and in Mesoderm (false discovery rate < 0.01, aaq > 0.05). Red arrow shows the variant location. (G) rs377307764 *TBX5* G>C variant accessibility logFC change in the vCM lineage compared to control sequence (left axis). *TBX5* expression (dashed line, right axis). rs377307764 variant in red.

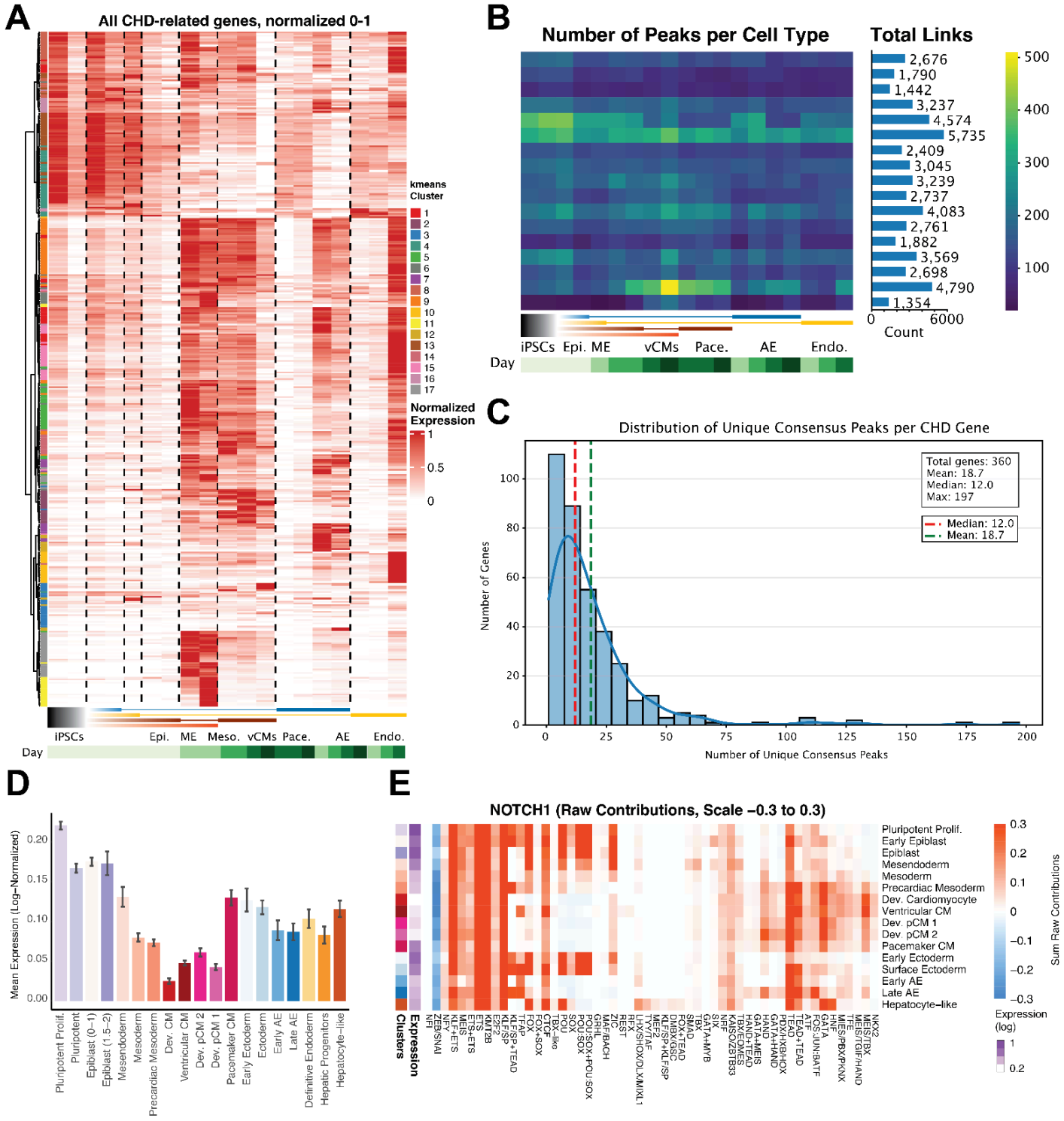
A
All CHD-related genes, normalized 0-1
kmeans Cluster
Normalized Expression
iPSCs
Epi.
ME
Meso.
vCMs
Pace.
AE
Endo.
Day
B
Number of Peaks per Cell Type
Total Links
2,676
1,790
1,442
3,237
4,574
5,735
2,409
3,045
3,239
2,737
4,083
2,761
1,882
3,569
2,698
4,790
1,354
Count
C
Distribution of Unique Consensus Peaks per CHD Gene
Total genes: 360
Mean: 18.7
Median: 12.0
Max: 197
Median: 12.0
Mean: 18.7
Number of Genes
Number of Unique Consensus Peaks
D
Mean Expression (Log-Normalized)
Pluripotent Prolif.
Pluripotent
Epiblast (0–1)
Epiblast (1.5–2)
Mesendoderm
Mesoderm
Precardiac Mesoderm
Dev. CM
Ventricular CM
Dev. pCM 2
Dev. pCM 1
Pacemaker CM
Early Ectoderm
Ectoderm
Early AE
Late AE
Definitive Endoderm
Hepatic Progenitors
Hepatocyte–like
E
NOTCH1 (Raw Contributions, Scale −0.3 to 0.3)
Clusters
Expression
Sum Raw Contributions
Expression (log)

Figure S11. (A) k-means cluster breakdown to all CHD genes in Fig. 6A. Hierarchical clustering shows a good correspondence with the k-means clustering (left column). (B) Number of peaks per cell state and the corresponding total number of scE2G links (blue bars) per cell state. (C) Distribution of unique consensus peaks per CHD gene across cell states. (D) *NOTCH1* log-norm expression with error bars. (E) Raw sum of contributions for motif syntaxes for *NOTCH1*.

References

1. Xu, Y., Zhang, T., Zhou, Q., Hu, M., Qi, Y., Xue, Y., Nie, Y., Wang, L., Bao, Z., and Shi, W. (2023). A single-cell transcriptome atlas profiles early organogenesis in human embryos. Nature Cell Biology *25*, 604–615.

2. Wang, Z., Wu, Z., Wang, H., Feng, R., Wang, G., Li, M., Wang, S.-Y., Chen, X., Su, Y., Wang, J., et al. (2023). An immune cell atlas reveals the dynamics of human macrophage specification during prenatal development. Cell *186*, 4454–4471.e19.

3. Tyser, R.C.V., Mahammadov, E., Nakanoh, S., Vallier, L., Scialdone, A., and Srinivas, S. (2021). Single-cell transcriptomic characterization of a gastrulating human embryo. Nature *600*, 285–289.

4. Bruneau, B.G., Nemer, G., Schmitt, J.P., Charron, F., Robitaille, L., Caron, S., Conner, D.A., Gessler, M., Nemer, M., Seidman, C.E., et al. (2001). A murine model of Holt-Oram syndrome defines roles of the T-box transcription factor Tbx5 in cardiogenesis and disease. Cell *106*, 709–721.

5. Sugathan, A., Biagioli, M., Golzio, C., Erdin, S., Blumenthal, I., Manavalan, P., Ragavendran, A., Brand, H., Lucente, D., Miles, J., et al. (2014). CHD8 regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors. Proc Natl Acad Sci U S A *111*, E4468–E4477.

6. Banerjee, K.K., Saxena, M., Kumar, N., Chen, L., Cavazza, A., Toke, N.H., O'Neill, N.K., Madha, S., Jadhav, U., Verzi, M.P., et al. (2018). Enhancer, transcriptional, and cell fate plasticity precedes intestinal determination during endoderm development. Genes Dev *32*, 1430–1442.

7. Morton, S.U., Quiat, D., Seidman, J.G., and Seidman, C.E. (2022). Genomic frontiers in congenital heart disease. Nat. Rev. Cardiol. *19*, 26–42.

8. Bakker, M.K., Bergman, J.E.H., Krikov, S., Amar, E., Cocchi, G., Cragan, J., de Walle, H.E.K., Gatt, M., Groisman, B., Liu, S., et al. (2019). Prenatal diagnosis and prevalence of critical congenital heart defects: an international retrospective cohort study. BMJ Open *9*, e028139.

9. Xiao, F., Zhang, X., Morton, S.U., Kim, S.W., Fan, Y., Gorham, J.M., Zhang, H., Berkson, P.J., Mazumdar, N., Cao, Y., et al. (2024). Functional dissection of human cardiac enhancers and noncoding de novo variants in congenital heart disease. Nat. Genet. *56*, 420–430.

10. 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.

11. Broberg, M., Ampuja, M., Jones, S., Ojala, T., Rahkonen, O., Kivelä, R., Priest, J., FinnGen, Palotie, A., Ollila, H.M., et al. (2024). Genome-wide association studies highlight novel risk loci for septal defects and left-sided congenital heart defects. BMC Genomics *25*, 256.

12. Lahm, H., Jia, M., Dreßen, M., Wirth, F., Puluca, N., Gilsbach, R., Keavney, B.D., Cleuziou,

J., Beck, N., Bondareva, O., et al. (2021). Congenital heart disease risk loci identified by genome-wide association study in European patients. J. Clin. Invest. *131*. https://doi.org/10.1172/JCI141837.

13. Yamaguchi, N., Chang, E.W., Lin, Z., Shekhar, A., Bu, L., Khodadadi-Jamayran, A., Tsirigos, A., Cen, Y., Phoon, C.K.L., Moskowitz, I.P., et al. (2023). An Anterior Second Heart Field Enhancer Regulates the Gene Regulatory Network of the Cardiac Outflow Tract. Circulation *148*, 1705–1722.

14. Ameen, M., Sundaram, L., Shen, M., Banerjee, A., Kundu, S., Nair, S., Shcherbina, A., Gu, M., Wilson, K.D., Varadarajan, A., et al. (2022). Integrative single-cell analysis of cardiogenesis identifies developmental trajectories and non-coding mutations in congenital heart disease. Cell *185*, 4937–4953.e23.

15. Rasouli, S.J., and Stainier, D.Y.R. (2017). Regulation of cardiomyocyte behavior in zebrafish trabeculation by Neuregulin 2a signaling. Nat Commun *8*, 15281.

16. Lavine, K.J., Yu, K., White, A.C., Zhang, X., Smith, C., Partanen, J., and Ornitz, D.M. (2005). Endocardial and epicardial derived FGF signals regulate myocardial proliferation and differentiation in vivo. Dev Cell *8*, 85–95.

17. Ma, L., Lu, M.-F., Schwartz, R.J., and Martin, J.F. (2005). Bmp2 is essential for cardiac cushion epithelial-mesenchymal transition and myocardial patterning. Development *132*, 5601–5611.

18. Plein, A., Calmont, A., Fantin, A., Denti, L., Anderson, N.A., Scambler, P.J., and Ruhrberg, C. (2015). Neural crest-derived SEMA3C activates endothelial NRP1 for cardiac outflow tract septation. J. Clin. Invest. *125*, 2661–2676.

19. Kim, J., Koo, B.-K., and Knoblich, J.A. (2020). Human organoids: model systems for human biology and medicine. Nature Reviews Molecular Cell Biology *21*, 571–584.

20. Xu, Q., Halle, L., Hediyeh-Zadeh, S., Kuijs, M., Riedweg, R., Kilik, U., Recaldin, T., Yu, Q., Rall, I., Frum, T., et al. (2025). An integrated transcriptomic cell atlas of human endoderm-derived organoids. Nat Genet *57*, 1201–1212.

21. Nikolova, M.T., He, Z., Seimiya, M., Jonsson, G., Cao, W., Okuda, R., Wimmer, R.A., Okamoto, R., Penninger, J.M., Camp, J.G., et al. (2025). Fate and state transitions during human blood vessel organoid development. Cell *188*, 3329–3348.e31.

22. Birey, F., Li, M.-Y., Gordon, A., Thete, M.V., Valencia, A.M., Revah, O., Paşca, A.M., Geschwind, D.H., and Paşca, S.P. (2022). Dissecting the molecular basis of human interneuron migration in forebrain assembloids from Timothy syndrome. Cell Stem Cell *29*, 248–264.e7.

23. Jain, A., Gut, G., Sanchis-Calleja, F., Tschannen, R., He, Z., Luginbühl, N., Zenk, F., Chrisnandy, A., Streib, S., Harmel, C., et al. (2025). Morphodynamics of human early brain organoid development. Nature. https://doi.org/10.1038/s41586-025-09151-3.

24. Lo, Y.-H., Horn, H.T., Huang, M.-F., Yu, W.-C., Young, C.-M., Liu, Q., Tomaske, M., Towers, M., Dominguez, A., Bassik, M.C., et al. (2025). Large-scale CRISPR screening in primary human 3D gastric organoids enables comprehensive dissection of gene-drug interactions.

Nat Commun *16*, 7566.

25. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–1218.

26. Schones, D.E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., Wei, G., and Zhao, K. (2008). Dynamic regulation of nucleosome positioning in the human genome. Cell *132*, 887–898.

27. Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

28. Kaya-Okur, H.S., Wu, S.J., Codomo, C.A., Pledger, E.S., Bryson, T.D., Henikoff, J.G., Ahmad, K., and Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. Nat Commun *10*, 1930.

29. Farley, E.K., Olson, K.M., Zhang, W., Rokhsar, D.S., and Levine, M.S. (2016). Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. Proc Natl Acad Sci U S A *113*, 6508–6513.

30. Smith, R.P., Taher, L., Patwardhan, R.P., Kim, M.J., Inoue, F., Shendure, J., Ovcharenko, I., and Ahituv, N. (2013). Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. Nat Genet *45*, 1021–1028.

31. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

32. Lai, W.K.M., Mariani, L., Rothschild, G., Smith, E.R., Venters, B.J., Blanda, T.R., Kuntala, P.K., Bocklund, K., Mairose, J., Dweikat, S.N., et al. (2021). A ChIP-exo screen of 887 Protein Capture Reagents Program transcription factor antibodies in human cells. Genome Res *31*, 1663–1679.

33. Luna-Zurita, L., Stirnimann, C.U., Glatt, S., Kaynak, B.L., Thomas, S., Baudin, F., Samee, M.A.H., He, D., Small, E.M., Mileikovsky, M., et al. (2016). Complex Interdependence Regulates Heterotypic Transcription Factor Distribution and Coordinates Cardiogenesis. Cell *164*, 999–1014.

34. Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D.R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods *18*, 1196–1203.

35. Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., et al. (2021). Base-resolution models of transcription-factor binding reveal soft motif syntax. Nat Genet *53*, 354–366.

36. Chen, K.M., Wong, A.K., Troyanskaya, O.G., and Zhou, J. (2022). A sequence-based global map of regulatory activity for deciphering human genetics. Nat Genet *54*, 940–949.

37. de Almeida, B.P., Reiter, F., Pagani, M., and Stark, A. (2022). DeepSTARR predicts

enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. Nat Genet *54*, 613–624.

38. Pampari, A., Shcherbina, A., Kvon, E.Z., Kosicki, M., Nair, S., Kundu, S., Kathiria, A.S., Risca, V.I., Kuningas, K., Alasoo, K., et al. (2024). ChromBPNet: bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants. bioRxiv. https://doi.org/10.1101/2024.12.25.630221.

39. Linder, J., Srivastava, D., Yuan, H., Agarwal, V., and Kelley, D.R. (2025). Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. Nat Genet *57*, 949–961.

40. Liu, B.B., Jessa, S., Kim, S.H., Ng, Y.T., Higashino, S.I., Marinov, G.K., Chen, D.C., Parks, B.E., Li, L., Nguyen, T.C., et al. (2025). Dissecting regulatory syntax in human development with scalable multiomics and deep learning. bioRxiv, 2025.04.30.651381. https://doi.org/10.1101/2025.04.30.651381.

41. Ma, X.R., Conley, S.D., Kosicki, M., Bredikhin, D., Cui, R., Tran, S., Sheth, M.U., Qiu, W.-L., Chen, S., Kundu, S., et al. (2024). Molecular convergence of risk variants for congenital heart defects leveraging a regulatory map of the human fetal heart. medRxiv. https://doi.org/10.1101/2024.11.20.24317557.

42. Marderstein, A.R., Kundu, S., Padhi, E.M., Deshpande, S., Wang, A., Robb, E., Sun, Y., Yun, C.M., Pomales-Matos, D., Xie, Y., et al. (2025). Mapping the regulatory effects of common and rare non-coding variants across cellular and developmental contexts in the brain and heart. bioRxiv. https://doi.org/10.1101/2025.02.18.638922.

43. 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.

44. Hofbauer, P., Jahnel, S.M., Papai, N., Giesshammer, M., Deyett, A., Schmidt, C., Penc, M., Tavernini, K., Grdseloff, N., Meledeth, C., et al. (2021). Cardioids reveal self-organizing principles of human cardiogenesis. Cell *184*, 3299–3317.e22.

45. Schmidt, C., Deyett, A., Ilmer, T., Haendeler, S., Torres Caballero, A., Novatchkova, M., Netzer, M.A., Ceci Ginistrelli, L., Mancheno Juncosa, E., Bhattacharya, T., et al. (2023). Multi-chamber cardioids unravel human heart development and cardiac defects. Cell *186*, 5587–5605.e27.

46. Integrated analysis of multimodal single-cell data (2021). Cell *184*, 3573–3587.e29.

47. Yang, S., Corbett, S.E., Koga, Y., Wang, Z., Johnson, W.E., Yajima, M., and Campbell, J.D. (2020). Decontamination of ambient RNA in single-cell RNA-seq with DecontX. Genome Biol. *21*, 57.

48. Janssen, P., Kliesmete, Z., Vieth, B., Adiconis, X., Simmons, S., Marshall, J., McCabe, C., Heyn, H., Levin, J.Z., Enard, W., et al. (2023). The effect of background noise and its removal on the analysis of single-cell expression data. Genome Biol. *24*, 140.

49. Rostovskaya, M., Andrews, S., Reik, W., and Rugg-Gunn, P.J. (2022). Amniogenesis occurs in two independent waves in primates. Cell Stem Cell *29*, 744–759.e6.

50. Nakanoh, S., Sham, K., Ghimire, S., Mohorianu, I., Rayon, T., and Vallier, L. (2024). Human surface ectoderm and amniotic ectoderm are sequentially specified according to cellular density. Sci. Adv. *10*, eadh7748.

51. Cassandri, M., Smirnov, A., Novelli, F., Pitolli, C., Agostini, M., Malewicz, M., Melino, G., and Raschellà, G. (2017). Zinc-finger proteins in health and disease. Cell Death Discovery *3*, 1–12.

52. Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The Human Transcription Factors. Cell *175*, 598–599.

53. Tevosian, S.G., Deconinck, A.E., Tanaka, M., Schinke, M., Litovsky, S.H., Izumo, S., Fujiwara, Y., and Orkin, S.H. (2000). FOG-2, a cofactor for GATA transcription factors, is essential for heart morphogenesis and development of coronary vessels from epicardium. Cell *101*, 729–739.

54. Dorr, K.M., Amin, N.M., Kuchenbrod, L.M., Labiner, H., Charpentier, M.S., Pevny, L.H., Wessels, A., and Conlon, F.L. (2015). Casz1 is required for cardiomyocyte G1-to-S phase progression during mammalian cardiac development. Development *142*, 2037–2047.

55. Liu, Z., Li, W., Ma, X., Ding, N., Spallotta, F., Southon, E., Tessarollo, L., Gaetano, C., Mukouyama, Y.-S., and Thiele, C.J. (2014). Essential role of the zinc finger transcription factor Casz1 for mammalian cardiac morphogenesis and development. J. Biol. Chem. *289*, 29801–29816.

56. Orlova, A., Guseva, D., and Ryzhkova, O. (2022). Identification of a novel de Novo variant in the CASZ1 causing a rare type of dilated cardiomyopathy. Int. J. Mol. Sci. *23*, 12506.

57. Jian, H., and Poetsch, A. (2023). CASZ1: Current implications in cardiovascular diseases and cancers. Biomedicines *11*, 2079.

58. Huang, R.-T., Xue, S., Wang, J., Gu, J.-Y., Xu, J.-H., Li, Y.-J., Li, N., Yang, X.-X., Liu, H., Zhang, X.-D., et al. (2016). CASZ1 loss-of-function mutation associated with congenital heart disease. Gene *595*, 62–68.

59. Sojka, S., Amin, N.M., Gibbs, D., Christine, K.S., Charpentier, M.S., and Conlon, F.L. (2014). Congenital heart disease protein 5 associates with CASZ1 to maintain myocardial tissue integrity. Development *141*, 3040–3049.

60. Liu, T., Li, T., and Ke, S. (2023). Role of the CASZ1 transcription factor in tissue development and disease. Eur. J. Med. Res. *28*, 562.

61. Li, L., Liu, X., He, L., Yang, J., Pei, F., Li, W., Liu, S., Chen, Z., Xie, G., Xu, B., et al. (2017). ZNF516 suppresses EGFR by targeting the CtBP/LSD1/CoREST complex to chromatin. Nature Communications *8*, 1–17.

62. Tycko, J., DelRosso, N., Hess, G.T., Aradhana, Banerjee, A., Mukund, A., Van, M.V., Ego, B.K., Yao, D., Spees, K., et al. (2020). High-throughput discovery and characterization of

human transcriptional effectors. Cell *183*, 2020–2035.e16.

63. 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.

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

65. Shrikumar, A., Tian, K., Avsec, Ž., Shcherbina, A., Banerjee, A., Sharmin, M., Nair, S., and Kundaje, A. (2018). Technical note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5. arXiv [cs.LG].

66. Lundberg, S.M., and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems *30*.

67. Deshpande, S., Yun, C.M., Ramalingam, V., Hecht, V., Patel, A., Pampari, A., Jessa, S., Zhao, R., Wang, A., and Kundaje, A. (2025). A unified lexicon of predictive DNA sequence motifs from ENCODE transcription factor binding and chromatin accessibility assays. Preprint at Zenodo, https://doi.org/10.5281/ZENODO.17179111 https://doi.org/10.5281/ZENODO.17179111.

68. Billings, S.E., Pierzchalski, K., Butler Tjaden, N.E., Pang, X.-Y., Trainor, P.A., Kane, M.A., and Moise, A.R. (2013). The retinaldehyde reductase DHRS3 is essential for preventing the formation of excess retinoic acid during embryonic development. FASEB J *27*, 4877–4889.

69. Kim, D.S., Risca, V.I., Reynolds, D.L., Chappell, J., Rubin, A.J., Jung, N., Donohue, L.K.H., Lopez-Pajares, V., Kathiria, A., Shi, M., et al. (2021). The dynamic, combinatorial cis-regulatory lexicon of epidermal differentiation. Nat. Genet. *53*, 1564–1576.

70. Sheth, M.U., Qiu, W.-L., Rosa Ma, X., Gschwind, A.R., Jagoda, E., Tan, A.S., Einarsson, H., Gorissen, B.L., Dubocanin, D., McGinnis, C.S., et al. (2024). Mapping enhancer-gene regulatory interactions from single-cell data. bioRxivorg. https://doi.org/10.1101/2024.11.23.624931.

71. Sakai, L.Y., Keene, D.R., Renard, M., and De Backer, J. (2016). FBN1: The disease-causing gene for Marfan syndrome and other genetic disorders. Gene *591*, 279–291.

72. Arbustini, E., Grasso, M., Ansaldi, S., Malattia, C., Pilotto, A., Porcu, E., Disabella, E., Marziliano, N., Pisani, A., Lanzarini, L., et al. (2005). Identification of sixty-two novel and twelve known FBN1 mutations in eighty-one unrelated probands with Marfan syndrome and other fibrillinopathies. Hum. Mutat. *26*, 494.

73. Vrljicak, P., Cullum, R., Xu, E., Chang, A.C.Y., Wederell, E.D., Bilenky, M., Jones, S.J.M., Marra, M.A., Karsan, A., and Hoodless, P.A. (2012). Twist1 transcriptional targets in the developing atrio-ventricular canal of the mouse. PLoS One *7*, e40815.

74. Lee, M.P., and Yutzey, K.E. (2011). Twist1 directly regulates genes that promote cell proliferation and migration in developing heart valves. PLoS One *6*, e29758.

75. Qiu, C., Martin, B.K., Welsh, I.C., Daza, R.M., Le, T.-M., Huang, X., Nichols, E.K., Taylor,

M.L., Fulton, O., O'Day, D.R., et al. (2024). A single-cell time-lapse of mouse prenatal development from gastrula to birth. Nature *626*, 1084–1093.

76. Wang, D., Chang, P.S., Wang, Z., Sutherland, L., Richardson, J.A., Small, E., Krieg, P.A., and Olson, E.N. (2001). Activation of cardiac gene expression by myocardin, a transcriptional cofactor for serum response factor. Cell *105*, 851–862.

77. Li, X., and Cao, X. (2003). BMP signaling and HOX transcription factors in limb development. Front. Biosci. *8*, s805–s812.

78. Lengerke, C., McKinney-Freeman, S., Wang, Y., Jang, I.-H., Green, J., and Daley, G.Q. (2006). BMP signaling via the Cdx-Hox pathway allocates mesoderm to hematopoietic vs cardiac fates. Blood *108*, 4183–4183.

79. Genikhovich, G., Fried, P., Prünster, M.M., Schinko, J.B., Gilles, A.F., Fredman, D., Meier, K., Iber, D., and Technau, U. (2015). Axis patterning by BMPs: Cnidarian network reveals evolutionary constraints. Cell Rep. *10*, 1646–1654.

80. Kosicki, M., Zhang, B., Hecht, V., Pampari, A., Cook, L.E., Slaven, N., Akiyama, J.A., Plajzer-Frick, I., Novak, C.S., Kato, M., et al. (2025). In vivo mapping of mutagenesis sensitivity of human enhancers. Nature *643*, 839–846.

81. Lim, F., Solvason, J.J., Ryan, G.E., Le, S.H., Jindal, G.A., Steffen, P., Jandu, S.K., and Farley, E.K. (2024). Affinity-optimizing enhancer variants disrupt development. Nature *626*, 151–159.

82. Jindal, G.A., Bantle, A.T., Solvason, J.J., Grudzien, J.L., D'Antonio-Chronowska, A., Lim, F., Le, S.H., Song, B.P., Ragsac, M.F., Klie, A., et al. (2023). Single-nucleotide variants within heart enhancers increase binding affinity and disrupt heart development. Dev Cell *58*, 2206–2216.e5.

83. Pushkarev, O., van Mierlo, G., Kribelbauer, J.F., Saelens, W., Gardeux, V., and Deplancke, B. (2024). Non-coding variants impact cis-regulatory coordination in a cell type-specific manner. Genome Biol *25*, 190.

84. Xiao, Y., Wang, J., Li, J., Zhang, P., Li, J., Zhou, Y., Zhou, Q., Chen, M., Sheng, X., Liu, Z., et al. (2023). An analytical framework for decoding cell type-specific genetic variation of gene regulation. Nat Commun *14*, 3884.

85. Garg, V., Muth, A.N., Ransom, J.F., Schluterman, M.K., Barnes, R., King, I.N., Grossfeld, P.D., and Srivastava, D. (2005). Mutations in NOTCH1 cause aortic valve disease. Nature *437*, 270–274.

86. McBride, K.L., Riley, M.F., Zender, G.A., Fitzgerald-Butt, S.M., Towbin, J.A., Belmont, J.W., and Cole, S.E. (2008). NOTCH1 mutations in individuals with left ventricular outflow tract malformations reduce ligand-induced signaling. Hum Mol Genet *17*, 2886–2893.

87. Page, D.J., Miossec, M.J., Williams, S.G., Monaghan, R.M., Fotiou, E., Cordell, H.J., Sutcliffe, L., Topf, A., Bourgey, M., Bourque, G., et al. (2019). Whole Exome Sequencing Reveals the Major Genetic Contributors to Nonsyndromic Tetralogy of Fallot. Circ Res *124*, 553–563.

88. Krebs, L.T., Xue, Y., Norton, C.R., Shutter, J.R., Maguire, M., Sundberg, J.P., Gallahan, D., Closson, V., Kitajewski, J., Callahan, R., et al. (2000). Notch signaling is essential for vascular morphogenesis in mice. Genes Dev *14*, 1343–1352.

89. High, F.A., Jain, R., Stoller, J.Z., Antonucci, N.B., Lu, M.M., Loomes, K.M., Kaestner, K.H., Pear, W.S., and Epstein, J.A. (2009). Murine Jagged1/Notch signaling in the second heart field orchestrates Fgf8 expression and tissue-tissue interactions during outflow tract development. J Clin Invest *119*, 1986–1996.

90. Mead, T.J., and Yutzey, K.E. (2012). Notch pathway regulation of neural crest cell development in vivo. Dev Dyn *241*, 376–389.

91. Li, J., Gordon, J., Chen, E.L.Y., Xiao, S., Wu, L., Zúñiga-Pflücker, J.C., and Manley, N.R. (2020). NOTCH1 signaling establishes the medullary thymic epithelial cell progenitor pool during mouse fetal development. Development *147*. https://doi.org/10.1242/dev.178988.

92. Frank, D.U., Fotheringham, L.K., Brewer, J.A., Muglia, L.J., Tristani-Firouzi, M., Capecchi, M.R., and Moon, A.M. (2002). An Fgf8 mouse mutant phenocopies human 22q11 deletion syndrome. Development *129*, 4591–4603.

93. Smemo, S., Campos, L.C., Moskowitz, I.P., Krieger, J.E., Pereira, A.C., and Nobrega, M.A. (2012). Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. Hum Mol Genet *21*, 3255–3263.

94. 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.

95. Kim, S.H., Marinov, G.K., Bagdatli, S.T., Higashino, S.I., Shipony, Z., Kundaje, A., and Greenleaf, W.J. (2023). Simultaneous single-cell profiling of the transcriptome and accessible chromatin using SHARE-seq. Methods Mol. Biol. *2611*, 187–230.

96. 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.

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

98. Gschwind, A.R., Mualim, K.S., Karbalayghareh, A., Sheth, M.U., Dey, K.K., Jagoda, E., Nurtdinov, R.N., Xi, W., Tan, A.S., Jones, H., et al. (2023). An encyclopedia of enhancer-gene regulatory interactions in the human genome. bioRxivorg, 2023.11.09.563812. https://doi.org/10.1101/2023.11.09.563812.

99. Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. (2018). Scaling algorithms for unbalanced optimal transport problems. Math. Comput. *87*, 2563–2609.