HHS Public Access

Author manuscript

Nat Cell Biol. Author manuscript; available in PMC 2020 September 15.

Published in final edited form as:

Nat Cell Biol. 2019 June ; 21(6): 674–686. doi:10.1038/s41556-019-0336-z.

A single cell transcriptional roadmap for cardiopharyngeal fate diversification

Wei Wang[1,#], **Xiang Niu**[2,3,5,#], **Tim Stuart**[3], **Estelle Jullian**[4], **William Mauck**[3], **Robert G. Kelly**[4], **Rahul Satija**[2,3,*], **Lionel Christiaen**[1,*]

[1]Center for Developmental Genetics, Department of Biology, New York University, New York, NY, USA

[2]Center for Genomics and Systems Biology, Department of Biology, New York University, New York, NY, USA

[3]New York Genome Center, New York, NY, USA

[4]Aix-Marseille University, CNRS UMR 7288, Developmental Biology Institute of Marseille, Campus De Luminy Case 907, 13288 Marseille Cedex 9, France

[5]present address: Tri-Institutional Program in Computational Biology and Medicine, Weill Cornell Medical College, New York, New York, USA

Abstract

In vertebrates, multipotent progenitors located in the pharyngeal mesoderm form cardiomyocytes and branchiomeric head muscles, but the dynamic gene expression programs and mechanisms underlying cardiopharyngeal multipotency and heart vs. head muscle fate choices remain elusive. Here, we used single cell genomics in the simple chordate model Ciona, to reconstruct developmental trajectories forming first and second heart lineages, and pharyngeal muscle precursors, and characterize the molecular underpinnings of cardiopharyngeal fate choices. We show that FGF-MAPK signaling maintains multipotency and promotes the pharyngeal muscle fate, whereas signal termination permits the deployment of a pan-cardiac program, shared by the first and second lineages, to define heart identity. In the second heart lineage, a Tbx1/10-Dach pathway actively suppresses the first heart lineage program, conditioning later cell diversity in the beating heart. Finally, cross-species comparisons between Ciona and the mouse evoke the deep evolutionary origins of cardiopharyngeal networks in chordates.

*: corresponding authors: lc121@nyu.edu (L.C.), rsatija@nygenome.org (R.S.).

#:contributed equally to the work

AUTHORS' CONTRIBUTIONS

W.W. performed the *Ciona* experiments. E.J. performed the mouse experiments. X.N., T.S., W.W. and R.S. performed computational analyses. W.M.M., W.W., and R.S. performed the single cell RNA-seq experiments. W.W., X.N., R.K., R.S., and L.C. designed the experiments and analyses. W.W., X.N., R.S. and L.C. wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Distinct cell types form multicellular animals and execute specialized functions within defined organs and systems, implying that individual cells within progenitor fields must acquire both organ-level and cell-type-specific identities. The mammalian heart comprises chamber-specific cardiomyocytes, various endocardial cell types, fibroblasts and smooth muscles[1], and despite their specialized features, these cells share a cardiac identity. Popular models posit that heart cells emerge from multipotent cardiovascular progenitors, implying that multipotent progenitors are first imbued with a cardiac identity, before producing a diversity of cell types. Consistent with this model, mammalian heart cells emerge primarily from *Mesp1*+ mesodermal progenitors[2-4]. However, lineage tracing and clonal analyses indicated that distinct compartments arise from separate progenitor pools, referred to as the first and second heart fields[5-8]. In addition, most early cardiac progenitors produce only one cell type[3], and cell type segregation occurs early, possibly prior to commitment to a heart identity[9]. Moreover, derivatives of the second heart field (e.g. cardiomyocytes of the right ventricle and outflow tract) share a common origin with branchiomeric head muscles, in the cardiopharyngeal mesoderm[3,10-16]. The characteristics of multipotent cardiopharyngeal progenitors, and the mechanisms underlying early heart vs. pharyngeal/branchiomeric muscle fate choices remain largely elusive, and studies are partially hindered by the complexity of vertebrate embryos[16].

The tunicate Ciona emerged as an innovative chordate model to study cardiopharyngeal development with unprecedented spatio-temporal resolution. In Ciona, invariant cell divisions produce distinct first and second heart lineages, and pharyngeal muscle precursors from defined multipotent cardiopharyngeal progenitors[17,18] (Fig. 1a). Multipotent progenitors exhibit multilineage transcriptional priming, whereby conserved fate-specific determinants are transiently co-expressed, before regulatory cross-antagonisms partition the heart and pharyngeal muscle programs to their corresponding fate-restricted precursors[17-19]. Here, we characterised, with single cell resolution, the genome-wide characteristics and regulatory mechanisms governing cardiopharyngeal multipotency, early fate choices, and the establishment of cell diversity in the beating heart.

RESULTS

Single cell transcriptome profiling of early cardiopharyngeal lineages

To characterize gene expression changes underlying the transitions from multipotent progenitors to distinct fate-restricted precursors, we performed plate-based single cell RNA sequencing (scRNA-seq) with SMART-Seq2[20] on cardiopharyngeal-lineage cells FACS-purified from synchronously developing embryos and larvae (Fig. 1a). We obtained 848 high-quality single cell transcriptomes from 5 time points covering early cardiopharyngeal development (Fig. 1a, Supplementary Fig. 1a). Using an unsupervised strategy[21], we clustered single cell transcriptomes from each time point, and identified clusters according to known markers and previously established lineage information (Fig. 1b, Supplementary Fig. 1b-c). Focusing on fate-restricted cells isolated from post-hatching larvae (18 and 20 hours post-fertilization (hpf), FABA stages 26-28; Supplementary Table 1), we identified clusters of *Gata4/5/6*+ first and second heart precursors, and *Ebf*+ atrial siphon muscle (ASM) precursors[18,19,22] (Fig. 1b, Supplementary Fig. 1c). Differential expression analyses

identified (1) ASM/pharyngeal muscle vs. pan-cardiac specific markers and (2) first vs. second heart precursor-specific markers (Fig. 1 c, Supplementary Table 2). The top 111 predicted pan-cardiac genes comprised established cardiac determinants, including *Gata4/5/6*, *Nk4/Nkx2-5*, and *Hand*, and we confirmed heart-specific expression by fluorescent *in situ* hybridization (FISH) for 19 candidate markers, including *Meis* and *Lrp4/8* (Fig. 1d, Supplementary Fig. 2a; Supplementary Table 2, 3). The pan-cardiac vs. pharyngeal muscle contrast dominated late cellular heterogeneity, but first and second heart precursor populations also segregated (Fig. 1b, Supplementary Fig. 1c), revealing 18 and 7 first- and second-lineage-specific markers, respectively (e.g. *Mmp21* and *Dach*; Fig. 1c-d, Supplementary Fig. 3a; Supplementary Table 2,3). Our analyses thus uncovered specific programs activated in fate-committed progenitors, including both shared ('pan-cardiac') and first- vs. second-lineage-specific signatures for heart precursors.

To characterize gene expression dynamics, we ordered single cell transcriptomes from successive time points on pseudotemporal developmental projections[23]. Using the whole dataset while ignoring established clonality, we identified multipotent progenitors and separate cardiac and pharyngeal muscles branches (Supplementary Fig. 4a). However, this unsupervised analysis failed to correctly distinguish the first and second heart lineages, likely because the shared pan-cardiac program dominates lineage-specific signatures (Fig. 1c; Supplementary Table 2). Taking advantage of the invariant lineage (Fig. 1a), we combined cells corresponding to each branch, and created three unidirectional trajectories representing first- and second heart, and pharyngeal muscle lineages (Fig. 2a-b). The distribution of cells along pseudotime axes corresponded to the time points of origin (Fig. 2c; average Pearson Correlation Coefficient, PCC = 0.889), while providing higher-resolution insights into the gene expression dynamics. Validating this approach, *in silico* trajectories captured known lineage-specific expression changes of cardiopharyngeal regulators[18,19,24] (Supplementary Fig. 4c-e).

Developmental trajectories suggest a continuous process marked by gradual changes in gene expression. However, the latter occur preferentially in defined 'pseudotime' windows for multiple genes (Supplementary Fig. 4c-e), consistent with more abrupt biological transitions, such as cell divisions[24]. To identify possible switch-like discontinuities, we determined cell-to-cell cross-correlations along lineage-specific trajectories. Using constrained hierarchical clustering[25], we identified 10 putative discrete regulatory states across the cardiopharyngeal trajectories, including two multipotent states, and eight successive transitions towards fate restriction (Fig. 2d, Supplementary Fig. 4b).

These successive transitions revealed underlying lineage-specific transcriptional dynamics. For example, the multipotent cardiopharyngeal progenitor state (aka TVC) differed markedly from subsequent cardiac states along the first cardiac trajectory (Fig. 2d-e), and gene expression mapping distinguished 'primed' and '*de novo*'-expressed heart markers, such as *Gata4/5/6* and *Slit1/2/3*, respectively (Fig. 2f). Conversely, primed pharyngeal muscle markers, like *Hand-r*, were downregulated along cardiac trajectories, as expected[26] (Fig. 2f; Supplementary Fig. 4c-e). Multilineage transcriptional priming is a hallmark of cardiopharyngeal multipotency, but remained to be characterized globally[19]. Here, we estimated that 41% (73/176) of the pharyngeal-muscle-specific and 53% (59/111) of the

pan-cardiac markers are already expressed in multipotent progenitors, indicating that lineage-specific maintenance of primed genes is a major determinant of cell-type-specific transcriptomes in the cardiopharyngeal lineage. Nevertheless, 88% (3,504/3,982) of late-expressed transcripts, were already detected in multipotent progenitors (Fig. 2g), indicating extensive stability of the global transcriptome and thus showing that *de novo* cell-type-specific gene activation contributes significantly to cell-type-specific programs (i.e. the fractions of *de novo*-expressed genes among cell-type-specific markers are greater than expected by chance, Fisher's exact test, $P < 2.2 \times 10^{-16}$ for both the pan-cardiac- and ASM-specific gene sets, respectively).

We further explored the molecular basis for progression through regulatory states. As a proof of concept, we first focused on the pharyngeal muscle trajectory, for which the key regulators Hand-r, Tbx1/10 and Ebf have been characterized[18,19,22,24,27]. The first two regulatory states corresponded to successive generations of multipotent cardiopharyngeal progenitors (aka TVCs and STVCs, Fig. 2d-e), confirming that asymmetric cell divisions provide the biological basis for these first transitions (Fig. 1a). To our surprise, the majority of newborn pharyngeal muscle precursors isolated from 16 hpf larvae clustered with multipotent progenitors isolated from 14 hpf embryos (aka STVCs, Fig. 2d-e), although they already expressed *Ebf* (Fig. 3a), as previously observed by FISH[18,24]. This indicates that, although newborn pharyngeal muscle progenitors already express a key determinant, their transcriptome remains similar to their multipotent mother cells. Indeed, the pharyngeal muscle transcriptome is progressively remodelled as cells transition through successive states, involving both downregulation of primed cardiac markers and '*de novo*' activation of pharyngeal muscle markers (Fig. 3a-b, Supplementary Fig. 5a-b, d). Moreover, systematic comparison with expression profiles following perturbations of Ebf function[19] indicated that candidate Ebf target genes, including *Myogenic regulatory factor* (Mrf) (the *MyoD/Myf5* homolog), and *Myosin heavy chain 3* (*Mhc3*), are activated at later time points, consistent with Ebf-dependent transitions to committed pharyngeal muscle states[24] (Fig. 3a, c-f, Supplementary Fig. 5c).

Termination of FGF-MAPK signaling launches a pan-cardiac program for heart identity

Next we investigated gene expression changes underlying state transitions during cardiac specification. We focused on the first heart trajectory, which provided the largest pseudotemporal range, to characterize the pan-cardiac program. Activation of '*de novo*' pan-cardiac markers and down-regulation of primed pharyngeal muscle and multipotent-specific markers accounted for most gene expression changes (Fig. 4a, Supplementary Fig. 5e-g). These coordinated gene expression changes explained major transitions along the cardiac trajectories. For example, we used logistic regression to determine that *Slit1/2/3* is activated at the transition between the multipotent and FHP1 states, which we confirmed by FISH (Fig. 2f, Supplementary Fig. 2b). We identified a principal component (PC1), which correlated highly with pseudotime (PCC=0.96; Fig. 4b), and used the PC1-loading of each gene to estimate the relative contribution of each class of markers to discrete regulatory states (Fig. 4c). This suggested that the multipotent state is primarily determined by the combined expression of TVC-specific genes, and primed cardiac and pharyngeal muscle markers. The TVC-to-FHP1 transition is marked by a sharp decline in TVC-specific gene

expression, accompanied by down-regulation of primed ASM genes, upregulation of primed pan-cardiac genes, and activation of *de novo*-expressed pan-cardiac genes. In this regard, the FHP1 state may be considered a "transition state" between a multipotent TVC state, and the FHP2 state[28]. The latter is defined by the virtual absence of TVC-specific and primed ASM-specific transcripts, and high levels of both primed and *de novo*-expressed pan-cardiac markers, thus probably corresponding to a heart-specific state, whereas activation of cell-type/lineage-specific genes underlie the FHP2-to-FHP3 transition, and their expression helps define the first-lineage-specific state, FHP3, as is the case for *Mmp21* (Fig. 4d-e).

A true pan-cardiac program should unfold following similar dynamics in the first and second heart lineage, reflecting shared regulatory logics. Accordingly, we observed a striking agreement between the ordered activation pattern of individual genes along each trajectory (Fig. 4f), suggesting a remarkably conserved developmental program. Notably, the onset of each gene was consistently delayed in the second heart trajectory, starting with the STVC-to-SHP1 transition, as second heart precursors are born from a second generation of multipotent progenitors, ~2 hours later than first heart precursors (Fig. 1a, 4f-g). Therefore, the *de novo* pan-cardiac program is tightly regulated and deployed in a reproducible cascade, whose onset is independently induced in the first and second heart lineages as they arise from multipotent progenitors.

We then sought to identify regulatory switches triggering the full pan-cardiac program in heart lineages. The FGF-MAPK signaling pathway is active and maintained specifically in multipotent cardiopharyngeal progenitors (TVCs and STVCs), and in early pharyngeal muscle precursors (ASMF), where it promotes the expression of *Hand-r*, *Tbx1/10*, and *Ebf*. By contrast, signaling is terminated in newborn first and second heart precursors[24]. We integrated sc- and bulk RNA-seq performed on FACS-purified cardiopharyngeal lineage cells following defined perturbations, and determined that FGF-MAPK signaling opposed pan-cardiac gene expression, while promoting the pharyngeal muscle program in swimming larvae (18 and 20 hpf, Fig. 5a-b, Supplementary Fig. 5h, 6a-d, Supplementary Table 6). At earlier stages (12 and 15 hpf), FGF-MAPK perturbations generally did not affect the expression of primed pan-cardiac genes in multipotent progenitors, whereas *de novo*-expressed genes were upregulated upon signaling inhibition by $Fgfr^{DN}$ misexpression (Fig. 4a-b, Supplementary Figs. 5i-j, 6e-h; Supplementary Table 6). FISH assays further demonstrated that the *de novo*-expressed pan-cardiac marker *Lrp4/8* was upregulated in TVCs upon misexpression of $Fgfr^{DN}$ (Fig. 5c). These analyses indicated that heart-lineage-specific termination of FGF-MAPK signaling permits the activation of *de novo*-expressed pan-cardiac genes, and subsequent heart fate specification, whereas ongoing FGF-MAPK signaling in cardiopharyngeal progenitors promotes multipotency both by maintaining the primed pharyngeal muscle program and by inhibiting the full deployment of the heart-specific program (summary Fig. 5d).

Differences between cardiac lineages foster cellular diversity in the beating heart

A shared pan-cardiac gene program progressively defines the heart identity, but distinct precursors nevertheless clustered separately, revealing significant differences between the first and second heart lineages (Fig. 1b-d; Supplementary Table 2). We mined the second

heart trajectory to explore the development and evolution of the vertebrate second heart field (SHF), since the ascidian and vertebrate SHFs share regulatory inputs from *Nk4/Nkx2-5* and *Tbx1/10* orthologs[8,10,18,29-31]. Examining our list of markers distinguishing second and first heart precursors, we identified the *dachshund* homolog *Dach* as the only known transcription regulator[32] (Fig. 6a, Supplementary Fig. 3a; Supplementary Table 2), and its upregulation as cells transitioned from a multipotent state suggested a role in specifying the second cardiac identity (Fig. 6a-b).

Dach1 and −2 have not been previously implicated in mammalian SHF development, but they belong to the conserved "retinal network"[33], which comprises homologs of Six and Eya transcription factors that contribute to cardiopharyngeal development in the mouse[34,35]. Lineage-specific CRISPR/Cas9-mediated loss-of-function[36,37], followed by gene expression assays, indicated that *Dach* is neither required for activation of the pharyngeal muscle marker *Ebf*, nor for its exclusion from the SHPs (Fig. 6c). By contrast, loss of Dach function caused ectopic expression of the FHP marker *Mmp21* in the second heart precursors (Fig. 6d). A CRISPR-resistant $Dach^{PAMmut}$ cDNA rescued the ectopic *Mmp21* expression in the second heart lineage, but did not abolish endogenous *Mmp21* expression in the first heart lineage (Fig. 6d). Dach is thus both a marker and a key regulator of second heart lineage specification, which is required, but not sufficient, to prevent activation of the first-lineage-specific marker *Mmp21*.

Since the second heart lineage emerges from *Tbx1/10*+ multipotent progenitors[18], we tested whether Tbx1/10 regulates *Dach* expression. CRISPR/Cas9-mediated lineage-specific loss of Tbx1/10 function[27] inhibited *Dach* expression (Fig. 6e), and caused ectopic activation of *Mmp21* (Fig. 6d), indicating that Tbx1/10 promotes second heart lineage specification, in part by regulating *Dach* activation, in addition to its role in pharyngeal myogenesis.

Indeed, Tbx1/10 is also necessary in parallel to FGF-MAPK activity to activate *Ebf* and promote the pharyngeal muscle program[18,24,27], in a manner similar to Tbx1 function in vertebrate branchiomeric myogenesis[38,39]. To explore the mechanism distinguishing between Tbx1/10 dual functions, we used the MEK/MAPKK inhibitor U0126, which inhibits *Ebf* expression[24], and caused ectopic *Dach* activation in the lateral-most cardiopharyngeal cells that normally form *Ebf*+ pharyngeal muscle precursors (Fig. 6f). Moreover, Tbx1/10 misexpression and MEK/MAPKK inhibition synergized to cause precocious and ectopic *Dach* activation in cardiopharyngeal progenitors (Fig. 6g). Taken together, these data indicate that termination of FGF-MAPK signaling in *Tbx1/10*+ cardiopharyngeal progenitors suffice to activate *Dach* expression and promote the second heart lineage identity, and demonstrate how distinct signaling environments can promote divergent regulatory programs in concert with Tbx1/10 expression.

First and second heart precursors share a common pan-cardiac signature, but initial molecular differences open the possibility that each lineage contributes differently to cardiogenesis. The beating *Ciona* heart is demonstrably simpler than its vertebrate counterpart; yet, diverse cell types form its single U-shaped compartment[40]. In post-metamorphic juveniles, the heart already beats, and double labeling with a *Mesp>nls::lacZ* reporter and the cardiac-specific *myosin heavy chain 2* (*Mhc2/Myh6*) marker showed that

beta-galactosidase+; *Mhc2/Myh6*– cells surround *Mhc2/Myh6*+ cardiomyocytes[18,22]. Lineage tracing using the photoconvertible reporter Kaede[19] indicated that first and second heart precursors derivatives remain within largely separate domains in juvenile hearts (Fig. 7a). Specifically, first-lineage-derived cells form the inner layer of *Mhc2+* cardiomyocytes, whereas second-lineage-derived cells contribute to the outer layer of *Mhc2*− cells (Fig. 7b; ; Supplementary Movie S1) . Triple labeling and cell quantification using pan-cardiopharyngeal and second-heart-lineage-specific reporters, and the *Mhc2* probe, indicated that most *Mhc2*+ cardiomyocytes were located in the first-lineage-derived inner layer, whereas only ~17% of second-lineage-derived cells express *Mhc2* in control juveniles (Fig. 7c). Thus, the first and second heart lineages contribute primarily *Mhc2/Myh6*+ cardiomyocytes and *Mhc2/Myh6*– cells to the beating juvenile heart, respectively.

To further characterize cellular diversity in the juvenile heart, we performed scRNA-seq on 386 FACS-purified cardiopharyngeal lineage cells dissociated from stage 38 juveniles, and identified clusters corresponding to pharyngeal muscle and cardiac lineages, including *Mhc2/Myh6*+ cardiomyocytes and an *Mhc2/Myh6*– population that expressed the second heart lineage markers *Dach* and *Matn* (Fig. 7d-f). Triple labeling with *Mesp* and *Tbx1/10* reporters indicated that these *Dach+;Matn+* cells derived principally from the second heart precursors and formed the outer layer of the juvenile heart (Fig.7g), suggesting that cellular diversity in the beating heart emerges from the initial segregation of the first and second heart lineages. Consistent with this hypothesis, CRISPR/Cas9-mediated loss of Dach and Tbx1/10 early functions increased the proportions of *Mhc2/Myh6*+ cells in the SHP progeny (Fig. 7c, Supplementary Fig. 3b), demonstrating that early inhibition of the *Mmp21*+ first-lineage-specific program by the Tbx1/10-Dach pathway limits the potential of second heart lineage derivatives to differentiate into *Mhc2/Myh6*+ cardiomyocytes during organogenesis.

Conserved cardiopharyngeal transcriptional signatures in chordates

Finally, we asked whether molecular features of cardiopharyngeal development are shared between Ciona and vertebrates. Recent scRNA-seq analysis of early mesodermal lineages in mice identified a population of pharyngeal mesoderm marked by high levels of both *Tbx1* and *Dach1* expression[41] (Fig. 8a). Multicolor immunohistochemical staining revealed that Dach1 expression starts broadly in the pharyngeal mesoderm, and becomes restricted to second heart field cells in the dorsal pericardial wall, and to a defined population of outflow tract cells, both of which also express Isl1 (Fig. 8b, Supplementary Fig. 7). Dach1 expression was excluded from the Nkx2.5+ ventricle, and absent from the Isl1+ skeletal muscle progenitor cells in the core mesoderm of the first and second pharyngeal arches (Fig. 8b, Supplementary Fig. 7), in a manner reminiscent of *Dach* exclusion from the pharyngeal muscles in *Ciona* (Fig. 6e-f, Supplementary Fig. 3a).

We extended the Ciona-to-mouse comparison of cardiopharyngeal transcriptomes using published scRNA-seq datasets[9,41,42]. We used canonical correlation analysis[43] to identify genes that separated cardiac and pharyngeal mesoderm cells in both Ciona and E8.25 mouse embryos (Fig. 8c-e, Supplementary Table 4). We then used only the 30 best correlated genes to re-cluster scRNA-seq data independently from each species, and found that these markers sufficed to distinguish cardiac and pharyngeal muscle cells in either species, revealing a

shared transcriptional program (Fig. 8c-e). We repeated this analysis using mouse datasets from earlier embryonic stages[9,41], and consistently identified genes that separated cardiac and pharyngeal cells in both species, and were enriched in transcription factor- and DNA binding protein-coding genes (Supplementary Fig. 7c-d, 8). For instance, *Gata4* and *Ebf1* homologs were identified in all three comparisons as discriminating markers that separated cardiac and pharyngeal cells (Fig. 8c-e, Supplementary Fig. 8, Supplementary Table 4). Overall, this analysis suggests that an evolutionary conserved transcriptional program, comprising homologs of *Ebf1*, *Gata4* and other regulatory genes, govern the heart vs. pharyngeal muscle fate choice in cardiopharyngeal mesoderm.

DISCUSSION

Here, we present an extensive analysis of the transcriptome dynamics underlying early cardiopharyngeal development in a tractable chordate model. Using established clonal relationships to inform the reconstruction of developmental trajectories, we characterized essential features of the transcriptome dynamics underlying cardiopharyngeal multipotency and early fate specification. Multipotent cardiopharyngeal progenitors exhibit extensive multilineage transcriptional priming. Together with the identification of heart vs. pharyngeal muscle-specific expression of E3 ubiquitin ligases and RNA binding proteins, this extensive multilineage priming opens the possibility that post-transcriptional regulatory mechanisms contribute to cell-type-specific expression profiles by clearing primed gene products in a lineage-specific manner. Nevertheless, *de novo* gene activation significantly contributes to cell-type-specific transcriptomes, highlighting the importance of transcriptional regulation in early heart vs. pharyngeal muscle fate choices.

Both first- and second-heart-lineage cells acquire a cardiac identity as they down-regulate multipotent progenitors markers and primed pharyngeal muscle genes, and deploy the full pan-cardiac program, entering this ‘transition state’[28] upon termination of FGF-MAPK signaling. We propose that the dual functions of FGF-MAPK signaling, as observed during early cardiac specification in Ciona, are conserved in vertebrates considering that FGF-MAPK inputs are necessary to induce multipotent progenitors[44-46], whereas signal termination is required for subsequent commitment to a heart fate and cardiomyocyte differentiation[47-51].

Following commitment to a cardiac identity, first heart progenitors transition to an *Mmp21*+ state that precedes differentiation into *Mhc2/Myh6*+ cardiomyocytes in the beating heart. By contrast, second heart progenitors activate *Dach* in response to Tbx1/10 inputs inherited from their distinct multipotent mother cells. This Tbx1/10-Dach pathway inhibits *Mmp21* expression and the *Mhc2/Myh6*+ cardiomyocyte potential to foster heart cell diversity. As is the case in vertebrates[30,31,38,39,52,53], Tbx1/10 plays a dual roles in branchiomeric/pharyngeal myogenesis[18,24,27] and second heart lineage development, thus acting as a *bona fide* regulator of cardiopharyngeal multipotency. Moreover, the first and second heart precursors share a common cardiac identity but differ because they emerge from successive multipotent progenitors before and after the onset of *Tbx1/10* expression. Cellular diversity in the Ciona heart thus emerges by temporal patterning, as is the case for neuronal fates in *Drosophila*[54]. The proposal that first and second heart lineages form a coherent cardiac

developmental unit, whereas distinct lineages contribute to cellular diversity within the heart, reconciles different views about the significance of the second heart field.

Finally, by leveraging recent computational methods for cross-species comparisons of single cell RNA-seq datasets, we identified shared markers of cardiopharyngeal regulatory states, while highlighting differences in expression dynamics, as seen for *Dach* homologs, thus illustrating the plasticity of gene regulatory networks controlling conserved developmental programs.

In conclusion, this study revealed that the first and second lineages of heart progenitors acquire a cardiac identity following the deployment of a shared and potentially ancestral program, whereas intra-cardiac cell diversity emerges from specific molecular differences established in distinct multipotent progenitors for the first and second heart lineages. This model for a modular control of heart cell identity potentially reconciles prevalent but somewhat antagonistic views about the significance of heart fields in chordates.

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for reagents may be directed to Lionel Christiaen (lc121@nyu.edu) and Rahul Satija (rsatija@nygenome.org).

METHODS

Animals

All animal care and experiments were carried out in accord with current NIH guidelines. *Ciona robusta* adults were purchased from M-Rep (San Diego, CA), maintained in artificial seawater with constant illumination, and used for experiment within one week after arrival. CD1 mice were crossed to generate embryos that were staged using the day of the copulation plug as embryonic day (E) 0.5.

Isolation of gametes, fertilization, dechorionation, electroporation, and development

Ciona eggs and sperm were collected from at least two individual adult animals and kept separately in filtered artificial seawater (FASW) until fertilization. Eggs were mixed with activated sperm and incubated in FASW at room temperature (18-20 °C) for 5 min. Chorion and surrounding follicle cells were chemically removed with a pronase solution (FASW, 7.5mg/L sodium thioglycolate, 0.05% pronase, 0.042N NaOH), as described[56]. Fertilized and dechorionated eggs were electroporated as described[56], and cultured in FASW in agarose coated plastic Petri dishes at 18°C. The amount of fluorescent reporter DNA (*Mesp>nls::lacZ*, *Mesp>hCD4::mCherry*, *Mesp>tagRFP*, *MyoD905>EGFP* and *Hand-r>tagBFP*) for electroporation was typically 50 μg, except for *Tbx1/10*^{3XT12}*>H2B::mCherry* is 30 μg. In order to increase the survival rate of juveniles, healthy and fluorescent positive larvae were selected at St. 26-27 using Leica M205 FA fluorescence stereo microscopes and transferred to uncoated Petri dishes with fresh FASW. The larvae were cultured at 18 °C until harvest.

Dissociation and FACS

Sample dissociation and FACS were performed essentially as described[19,56,57]. Embryos and larvae were harvested at 12, 14, 16, 18 and 20 hpf in 5 mL borosilicate glass tubes (Fisher Scientific, Waltham, MA. Cat.No. 14-961-26) and washed with 2 mL calcium- and magnesium-free artificial seawater (CMF-ASW: 449 mM NaCl, 33 mM Na_2SO4, 9 mM KCl, 2.15 mM $NaHCO_3$, 10 mM Tris-Cl pH 8.2, 2.5 mM EGTA). The tunique of juveniles was peeled off using BD PrecisionGlide™ Needles (REF 305115) under the dissecting microscope to facilitate dissociation. Embryos and larvae were dissociated in 2 mL 0.2% trypsin (w/v, Sigma, T- 4799) CMF-ASW by pipetting with glass Pasteur pipettes. The dissociation was stopped by adding 2 mL filtered ice cold 0.05% BSA CMF-ASW. Dissociated cells were passed through 40 μm cell-strainer and collected in 5 mL polystyrene round-bottom tube (Corning Life Sciences, Oneonta, New York. REF 352235). Cells were collected by centrifugation at 800 g for 3 min at 4 °C, followed by two washes with ice cold 0.05% BSA CMF-ASW. Cell suspensions were filtered again through a 40 μm cell-strainer and stored on ice. Following dissociation, cell suspensions were used for sorting within 1 hour.

B7.5 lineage cells were labeled by *Mesp>tagRFP* reporter. Contaminating B-line mesenchyme cells were counter-selected using *MyoD905>EGFP* as described[18,19,56]. The TVC-specific *Hand-r>tagBFP* reporter was used in a 3-color FACS scheme for positive co-selection of TVC-derived cells, in order to minimize the effects of mosaicism. Dissociated cell were loaded in a BD FACS Aria™ cell sorter[57]. 488 nm laser, FITC filter was used for EGFP; 407 nm laser, 561 nm laser, DsRed filter was used for tagRFP and Pacific BlueTM filter was used for tagBFP. The nozzle size was 100 μm. tagRFP +, tagBFP + and EGFP – cells were collected for downstream RNA sequencing analysis.

Fluorescent *In Situ* Hybridization-Immunohistochemistry (FISH-IHC) in *Ciona* Larvae

FISH-IHC were performed essentially as described[18,19,56]. Embryos were harvested and fixed at desired developmental stages for 2 hrs in 4% MEM-PFA and stored in 75 % ethanol at −20 °C. Antisense RNA probes were synthesized using either Gateway gene collections or amplified fragments of desired genes as templates (Supplementary Table 3). *In vitro* antisense RNA synthesis was performed using T7 RNA Polymerase (Roche, Cat. No. 10881767001) and DIG RNA Labeling Mix (Roche, Cat. No. 11277073910). Anti-Digoxigenin-POD Fab fragment (Roche, IN) was first used to detect the hybridized probes, then the signal were revealed using the Tyramide Signal Amplification (TSA) with Fluorescein TSA Plus Kits (Perkin Elmer, MA). Anti–β-galactosidase monoclonal mouse antibody (Promega) was co-incubated with anti-mCherry polyclonal rabbit antibody (Bio Vision, Cat. No. 5993-100) for immunodetection of *Mesp>nls::lacZ* and *Mesp>hCD4::mCherry* products respectively. Goat anti-mouse secondary antibodies coupled with AlexaFluor-555 and AlexaFluor-633 were used to detect β-galactosidase-bound mouse antibodies and mCherry-bound rabbit antibodies after the TSA reaction. FISH samples were mounted in ProLong® Gold Antifade Mountant (ThermoFisher Scientific, Waltham, MA. Catalog number P36930).

Multicolor Immunohistochemical Staining in Mouse Embryo

After dissection, embryos were fixed for 30 minutes (E8.5 and E9.5) to 1 hour (E7.5 in decidua) in 4 % paraformaldehyde, dehydrated and embedded in paraffin prior to sectioning at 10 μm. Immunofluorescence was performed using standard protocols. Briefly, after rehydration, sections were treated for 15 minutes with antigen unmasking solution (H-3300, Vector laboratories). Slides were washed twice in 1 x PBS Tween (0.05 %) and incubated for one hour in TNB buffer (0.1 M Tris-HCl pH 7.5, 0.15 M NaCl, 0.5% Blocking reagent (Roche 11096176001)). Sections were incubated with primary antibodies for 36 hours in TNB using the following dilutions: Dach1 (1/100, Proteintech 10914-1-AP), Nkx2-5 (1/100, Santa Cruz sc-8697), Islet1 (1/100, DSHB 39.4D5 and 40.2D6). After three 5 minutes washes in 1 x PBS Tween (0.05 %) sections were incubated with secondary antibodies for one hour using Alexa 488, 568 and 647 (1/500, Invitrogen). Sections were counterstained with Hoechst (Sigma 33258), mounted using Fluoromount (Southern Biotech 0100-001) and imaged using a Zeiss Axio Imager Z1 with an Apotome module.

CRISPR/Cas9-Mediated Gene Knock-Down

Two guide RNAs targeting the third and the fifth exon of *Dach* with Fusi scores (http://crispor.tefor.net) 63 (sgDach1: AAAAGATTAAGCATCGCCC) and 64 (sgDach2: GAGCATTGCCATTGACGTG), respectively, were designed to mutagenize the *Dach* locus in the B7.5 lineage using CRISPR/Cas9 as described[37]. Two guide RNAs described by Tolkin et al[27]. (sgTbx1.6 TGCGGCTTCGGCTCCGTGG; sgTbx1.8 AACGAAAGATTGGTGGCCG), were used to mutagenize the *Tbx1/10* coding region. The efficiency of guide RNAs were evaluated using the peakshift method[37]. Guide RNAs were expressed using the *Ciona robusta U6* promoter[36]. For each gene, two guide RNAs were used in combination with 25 μg of each expression plasmid. 30 μg of *Mesp>nls::Cas9::nls* plasmid were co-electroporated with guide RNA expression plasmids for B7.5 lineage-specific CRISPR/Cas9-mediated mutagenesis. Rescue of the *Dach* loss-of-function was achieved by TVC-specific overexpression of $Dach^{PAMmut}$ driven by a *Foxf* enhancer[58]. Point mutations (G303A and C852A) were introduced to the PAMs of sgDach1 and sgDach2 using an optimized QuikChange™ Site-directed Mutagenesis protocol. Two pairs of mutagenesis primers were designed using PrimerX website (http://www.bioinformatics.org/primerx/cgi-bin/DNA_1.cgi) as sgDACH_1_G303A_F: GATTAAGCATCGCCCCAGTCGTGTGCAACGTTG; sgDACH_1_G303A_R: CAACGTTGCACACGACTGGGGCGATGCTTAATC and sgDACH_2_C852A_F: CGTCGGGAATTCCACCCACGTCAATG; sgDACH_2_C852A_R: CATTGACGTGGGTGGAATTCCCGACG. The PCR mixture was prepared as 20 μl 5X HF buffer, 71 μl H2O, 3 μl 10 mM dNTP, 1.75 μl WT Dach plasmid at 15ng/μl, 2 μl 125 ng/μl top mutagenesis primer, 2 μl 125 ng/μl bottom mutagenesis primer, 1 μl Phusion® High-Fidelity DNA Polymerase (NEB M0530). PCR mixture were evenly distributed into 8 PCR tube-strip and PCR was performed with a denaturation at 95 °C for 4 min, followed by 18 cycles of (95 °C for 30 s, 50 to 72 °C (gradient) for 1 min, and 72 °C for 1 min) and a final extension at 72° for 5 min. The PCR products were pooled and 4 μl of DpnI was added directly to the tube, followed by the incubation at 37 °C for 2 hours. After the purification using QIAquick PCR Purification Kit (QIAGEN), the eluate are transformed into TOP10 cells. Successful mutagenesis was confirmed by sequencing.

Photoconversion and Lineage Tracing

Photoconversion and lineage tracing were performed as described[19]. Fertilized eggs were electroporated with 50 μg Mesp>Kaede:nls to label the B7.5 lineage. Embryos were raised on agarose coated plastic Petri dishes in ASW at 18°C and transferred individually into Nunc™ MicroWell™ 96-Well Optical-Bottom plates (ThermoFisher Scientific, Waltham, MA. Supplier No. 164588) at 15 hpf. Photoconversions were performed using the HC PL FLUOTAR 20×/0.50 objective on Leica Microsystems inverted TCS SP8 X confocal microscope, by shedding 405 nm UV light on ROI continuously for 2 min. Stack scanning of whole TVC lineage were documented at 16, 22.5, 40, 48 and 65 hpf.

Confocal Microscopy

Images were acquired with an inverted Leica TCS SP8 X confocal microscope, using HC PL APO 63×/1.30 objective. Z-stacks were acquired with 1 μm z-steps. Maximum projections were processed with maximum projection tools from the LEICA software LAS-AF.

Image Processing and Quantification

Confocal z stacks were processed using Imaris x64 8.4.1 (BitPlane). A region containing the TVC progeny was segmented first. For nuclei detection, the expected nucleus diameter was set at 2.5 μm, Nucleus Threshold(Absolute Intensity) were calculated automatically by Imaris. The cell segmentation was carried out using “Detect Cell Boundary from Cell Membrane” function, the “Cell Smallest Diameter” was set as 5 μm. The transcripts signal within the cell boundary was detected using “Vesicles Detection” function, the estimated diameter of dots was set as 1.44μm. To count the number of (Mhc2+, Matn+ or Dach+) cells in juvenile heart, a region of juvenile heart was segmented first, then the nuclei detection was performed as described above. A surface was created from the channel of FISH detention. Then the “Find Spots Close To Surface” function was used to define the nucleus with transcripts (the surface) closing to it using a 2 μm threshold.

Bulk RNA-seq Library Preparation, Sequencing and Analyses

200 to 800 cells were directly sorted in 100 μL lysis buffer from the RNAqueous®-Micro Total RNA Isolation Kit (Ambion). For each condition, samples were obtained in biological duplicates. The total RNA extraction was performed following the manufacturer’s instruction. The quality and quantity of total RNA was checked using Agilent RNA 6000 Pico Kit (Agilent) with Agilent 2100 Bioanalyzer. RNA samples with RNA Integrity Number (RIN)>8 were kept for downstream cDNA synthesis. 250-2000 pg of total RNA were loaded as template for cDNA synthesis using the SMART-Seq v4 Ultra Low Input RNA Kit (Clontech) with template switching technology. RNA-Seq Libraries were prepared and barcoded using Ovation® Ultralow System V2 1–16 (NuGen). Up to 6 barcoded samples were pooled in one lane of the flow cell and sequenced by Illumina Hi-Seq 2500. One direction and 50 bp length reads were obtained from all the bulk RNA-seq libraries.

Sequencing reads were mapped to the *Ciona robusta* genome (Joined-scaffold (KH), http://ghost.zool.kyotou.ac.jp/datas/JoinedScaffold.zip) using TopHat 2.0.12[59,60] with parameter: --no-coverage-search. Cufflinks 2.2.0 [60] was used to calculate the Fragments Per Kilobase of transcript per Million mapped reads (FPKM). We used edgeR[61] to analyze differential gene

expression in pairwise comparisons. Detailed summary statistics are provided in Supplementary Table 6.

Single Cell RNA-seq Library Preparation and Sequencing

Reverse transcription and cDNA amplification were carried out using modified Smart-seq2 protocol[20]. Single cells were sorted by FACS as described above into 96-well plates and collected in 3.4 μl RT buffer (0.5 μL 10 μM 3' RT Primer (5' - AAG CAG TGG TAT CAA CGC AGA GTA C T30 VN - 3'), 0.5 μL 10 μM dNTP Mix, 0.5 μL 4 U/μL RNase Inhibitor, 1 μL Maxima RT Buffer, 0.9 μL nuclease-free water) in each well. Plates were either stored at −80°C or processed immediately. Plates were incubated at 72°C for 3min and chilled on ice to denature the template RNA. 2μL RT reaction mixture (0.5 μL 10 μM TSO primer (5′-AGACGTGTGCTCTTCCGATCTNNNNNrGrGrG-3′), 0.925 μL 5 M Betaine, 0.4 μL 100 mM $MgCl_2$, 0.125 μL 40 U/μL RNase inhibitor, 0.05 μL 200 U/μL Maxima H Minus Reverse Transcriptase) were added to each well. Reverse transcription was carried out by incubating the plate at 42 °C for 90 min, followed by 10 cycles of (50 °C for 2 min, 42 °C for 2 min) and heat inactivation at 70 °C for 15 min. 7 μL PCR amplification mixture (0.25 μL 10 μM PCR primer (5′AGACGTGTGCTCTTCCGATCT-3′), 6.25 μL KAPA HIFI ReadyMix, 0.5 μL nuclease-free water) were added to each well. PCR amplification was carried out with a denaturation at 98 °C for 3 min, followed by 21 cycles of (98 °C for 15 s, 67 °C for 20 s, and 72 °C for 6 min) and a final extension at 72 °C for 5 min. PCR products were purified by adding 10 μL (0.8×) Agencourt AMPureXP SPRI beads (Beckman-Coulter) to each well, followed by 5 min incubation and two washes with 100 μL freshly prepared 70 % ethanol at room temperature. Purified cDNA were eluted in 20 μL TE buffer. The concentration of amplified cDNA was measured across the entire plate using Picogreen assays. The concentration of amplified cDNA was in a 0.5–2 ng/μL range. Fragment size distributions were checked for randomly selected wells with High-Sensitivity Bioanalyzer Chip (Agilent), the expected size average should be ~2 kb. For each sample, the amplified cDNA were normalized to a working concentration ranging from 0.1 to 0.2 ng/μL with TE buffer. 1.25 μL of diluted cDNA from each well were used for library preparation. Single cell libraries were prepared using the Nextera XT DNA Sample Kit (Illumina) according to manufacturer's instructions. After library amplification, 2.5 μL from each well were pooled into a single 1.5 mL microcentrifuge tube, purified using Agencourt AMPure XP beads and eluted with 30 μL TE buffer. 1 μL purified library was used to measure the fragment size distribution using the Agilent HS DNA BioAnalyzer chip and another 1 μL of the purified library was loaded into Qubit fluorometer to estimate library concentration according to the manufacturer's instructions. Libraries were sequenced on an Illumina HiSeq 2500 sequencer to obtain paired-end 50 bp reads.

QUANTIFICATION AND STATISTICAL ANALYSIS

Read alignment and generation of gene expression matrix

For each demultiplexed bulk and single cell RNA-seq library, sequencing reads were mapped to the *Ciona robusta* genome (Joined-scaffold (KH), http://ghost.zool.kyotou.ac.jp/datas/JoinedScaffold.zip) using TopHat 2.0.12[59,60] with parameter: --no-coverage-search.

Cufflinks 2.2.0[60] was used to calculate the Fragments Per Kilobase of transcript per Million mapped reads (FPKM).

Preprocessing and batch effect removal

We adopted multiple quality control criteria to filter out low quality single cell transcriptomes. First, we only retained single cells that had more than 2,000 and less than 6,000 expressed genes, and genes that were detected in more than 3 cells. 1,182 out of 1,796 single cells and 14,864 out of 15,287 genes were retained. We used total reads and overall read mapping rates from TopHat output files to assess the quality of scRNA-seq. Cells with mapping rates less than 30% and total reads more than 2-million were removed (see Supplementary Note). 1,138 out of 1,182 cells passed the quality controls and were retained for downstream analyses.

Batch effects were identified by principal component analysis (PCA) using all detected genes. Principal components 2, 5 and 7 were dominated either by ribosomal genes or unannotated genes that showed strong expressions only in certain batches (Supplementary Note). These PC's were considered as batch effects created by sequencing and library preparation. For each gene j, its expression level y_j was fitted by a linear mixed model of the total sum of latent batch effects (x_i) and its real biological expression level (ϵ_j) as the formula $y_j = \sum_i a_i x_i + \epsilon_j$, where a_i denotes coefficient of batch effect x_i. In our case, PCA rotation matrices of PC2, PC5 and PC7 served as batch effects and were regressed out by the above model (Supplementary Note).

Contaminating subpopulations were discovered upon clustering. 59 *Twist1+* mesenchymal cells and 198 cells without previously identified lineage markers were detected in cluster 8, 9, 11 and 12 (Supplementary Note). These contaminating non-cardiopharyngeal lineage cells were removed before downstream analysis. 881 out of 1,138 single cells were retained for clustering and trajectory analysis.

Identification of variable genes and dimensional reduction analysis.

All scRNA-seq analyses were performed on each time point data individually. Downstream analysis followed the procedures of 'Seurat' R package v1.2 [21]; http://satijalab.org/seurat).

We first identified the set of genes that was most variable in 12, 14 and 20 hpf single-cell data. We calculated the mean and dispersion (variance/mean) for each gene across all single cells, and placed genes into 20 bins based on their average expression. Within each bin, we then z-normalized the dispersion measure of all genes within the bin to identify genes whose expression values were highly variable even when compared to genes with similar average expression. We used a z-score cutoff of 2 for dispersion and average expression cutoff of 4 to identify highly variable genes. We then used those highly variable genes as input to the PCA to identify the primary data structures in 12, 14 and 20hpf data. For intermediate stages 16hpf and 18hpf, because of the known cell type similarity, we used cell type specific markers from 14hpf and 20hpf as input to PCA to obtain more robust dimensional reduction.

We extended the results of PCA analysis globally by projecting the PCA rotation matrix across the entire transcriptome. This additional projection allows us to identify other genes

with strong PCA loadings that may not be included in our variable gene list. Statistically significant PCs were identified using a permutation test and independently confirmed using a modified resampling procedure[62] encoded in Seurat's 'jackStraw' function. Significant and biological meaningful PCs were retained for clustering and visualization. To visualize single cell data, we projected individual cells based on their PC scores onto a single two-dimensional map using t-distributed Stochastic Neighbor Embedding (t-SNE)[63].

Single cell clustering and differential gene expression

Clustering of single cells was performed using the weighted shared nearest neighbor (SNN) graph-based clustering method[64]. To validate the legitimacy of clusters, we used 'ValidateClusters' function in Seurat, where we selected top 30 genes from significant PC's as defined above and utilized them to build a linear kernel SVM. The predictive accuracy of the SVM was assessed by repeated 5-fold cross validation. The accuracy cutoffs of 0.8 and 0.85 were used, and the merging of clusters was done based on the minimal connectivity from the SNN graph with a threshold of 0.001. Subsequently, TVC, STVC, FHP, SHP and ASM cells were identified from each time point data based on both known and candidate markers (Supplementary Table 2). Specifically, for 12hpf data, no significant PC was identified due to the small and homogeneous TVC population. In 18hpf data, we also identified a cluster of 33 cells that expressed noticeable degree of both cardiac and ASM markers (Supplementary Note). We inferred this cluster of cells is possibly due to insufficient tissue-dissociation or sorting and sequencing errors. We removed these cells from further analysis.

The dataset can be mined through an online tool (ShinyApp) available at: (https://christiaenlab.shinyapps.io/tvc-lineage/) (e.g. test using the pan-cardiac and ASM markers GATA4/5/6 and EBF1/2/3/4)

To find markers differentially expressed among clusters, we used the same approach as in [65]. We used the binary classifier with ROC curve that was incorporated in Seurat's 'find.markers' function with parameters: test.use = 'roc', thresh.use = 1 and min.pct = 0.5, which selects genes that are expressed in more than 50% of single cells in the given cluster and with average expression larger than 1 $\log_2$(FPKM) for differential expression analysis. The selected genes were ranked based on AUCs from 0 to 1. The higher the AUC or power value the more differentially expressed the gene is for the given cluster. AUCs of 0.5 or below have limited predictive power.

Single cell trajectory and transition state

We retrieved scRNA-seq data for each of the FHP, SHP, ASM trajectories by subsetting the master Seurat object containing all the single cell data. We adopted nonlinear dimensionality reduction technique of diffusion map[66,67], which reduces dimensionality through a random walking process, to identify developmental trajectory. We used the markers (power>0.3) of all cell types in each trajectory to calculate a cell-to-cell pairwise Euclidean distance matrix and used this matrix as input to diffusion map ('diffuse' function from 'diffusionMap' package). We retained only the first two diffusion map components as developmental trajectory for pseudotime analysis. Every cell was assigned to a pseudotime coordinate by

fitting a principal curve[68] to the first two diffusion map components. Pseudotime was determined by the unit-speed arc-length parameterization of each cell on principal curve and normalized to [0,1] range.

After identification of pseudotime, we selected genes that are dynamically expressed across the pseudotime using 'aic' function in the 'locfit' package. Genes expressed in more than 50% of single cells with mean expression level greater than 2 log(FPKM) were considered as expressed in each cell type. For every expressed gene, we built two local polynomial models: a null model with degree 0 that assumes the gene expression stays constant along pseudotime and an alternative model with degree 2 that assumes gene expression changes along pseudotime[41]. We evaluated these two model using Akaike Information Criterion (AIC) to calculate the AIC score differences as AIC(degree=2) – AIC(degree = 0). Genes with AIC score differences lower than −5 were considered to favor the alternative model to be dynamically expressed in pseudotime space.

We then used these dynamic genes to subdivide the pseudotime space into distinct regulatory states separated by discrete transitions. First, we built a cell-to-cell cross-correlation matrix based on dynamic gene expressions for each trajectory. Constrained hierarchical clustering tree, which maintains pseudotime ordering, was built with the CONISS algorithm[25] using 'chclust' function from 'rioja' package based on the cross-correlation matrix. We used gap statistics to empirically determine the number of clusters (k) considered as the transition states along pseudotime. Briefly, we start with k = 1 (no transition), and increase it to 10. If the k^{th} gap statistics Gap_k increases less than 10% of the previous k-1^{th} gap statistics Gap_{k-1} then we consider the current k as the suitable number of transition states.

Primed and *de novo* gene expression

Before defining primed and *de novo* genes, we first unbiasedly clustered the temporal gene expression patterns of both pan cardiac and ASM markers across all three trajectories. With k=2 of 'kmeans' clustering, we identified two groups of gene expression patterns in both cardiac and ASM genes that mimicked the primed and *de novo* patterns. Then we performed more stringent selection to define primed and *de novo* cardiac/ASM marker genes. We first identified all pan-cardiac and ASM markers (power>0.5) using the 18hpf and 20hpf datasets. Then we defined primed genes as genes that were expressed in more than 50% of single cells in both the multipotent progenitors (12 hpf TVCs) and the fate restricted cells (18/20 hpf FHPs/SHPs/iASMs/oASMs). Similarly, we defined *de novo*-expressed genes as expressed in less than 25% of single cells in the multipotent progenitors (12 hpf TVCs) but expressed in more than 50% of single cells in the fate restricted cells (18/20 hpf FHPs/SHPs/iASMs/oASMs). The genes expressed between 25%-50% of single cells were classified as ambiguous. The progenitor genes were defined as genes that were expressed in more than 50% of single cells in 12 hpf TVCs but less than 25% of single cells in any of 18/20 hpf FHP/SHP/iASM/oASM clusters. FHP/SHP specific markers were defined as genes that only belonged to FHP/SHP group but not to pan-cardiac or ASM gene set.

To visualize the smoothed pseudotime expression pattern and predict the induction time of a given gene, we smoothed the expression profiles along the pseudotime axis using local polynomial fit ('loess' function) with degree of smoothing equals to 0.75. Gene induction

time was predicted based on the smoothed pseudotime expression profile using a logistic regression model. Gene expressions were first normalized to [0,1] range. Normalized expression values that were smaller than 0.5 were considered in 'off' state and bigger or equal to 0.5 were considered in 'on' state. We used this binary state notation and pseudotime coordinates to train a logistic model ('glm' function with family=binomial(link='logit')) to predict the on/off state of a given gene. The induction time was determined as the closest pseudotime coordinate to 0.5. Genes were then subdivided into two groups, either turning on or turning off, and sorted by their induction pseudotime.

To quantify the relative contribution of multipotent progenitor, primed ASM/cardiac, *de novo* ASM/cardiac and FHP/SHP specific genes, we performed principal component analysis using these groups of genes defined using above criteria on FHP and SHP trajectories separately. We observed that principal component 1 (PC1) strongly correlated with the defined pseudotime (PCCs>0.9). This allowed us to used PC1 loadings of each gene to calculate the contribution of each group as a scaled score using the formula: $G = X_g^T * PCA^{rot}_{[\bullet,1]}$. The X_g represents scaled expression matrix of group g where each column is a single cell and each row is a gene from group g. The $PCA^{rot}_{[\bullet,1]}$ represents the PC1 variable loadings which is the first column of the loading matrix. The G is the scaled score of gene group g. For heatmap visualization, we smoothed this score along the pseudotime axis using local polynomial fit ('loess' function) with degree of smoothing equals to 0.75.

Mouse single cell RNA-seq data

Mouse single cell data shown in Fig. 8a were retrieved from the following website: http://gastrulation.stemcells.cam.ac.uk/scialdone2016 [41]. The count matrix were transformed into log Transcripts per Million (logTPM). The variable genes, PCA and tSNE analysis were performed as described above. The visualization was based on the clustering results of the original paper. The mouse single cell data used for canonical correlation analysis were retrieved from the following papers and websites: http://singlecell.stemcells.cam.ac.uk/mesp1[9] and https://marionilab.cruk.cam.ac.uk/organogenesis [42].

Canonical correlation analysis

Ciona and mouse genes were subset to those with known orthologs (https://www.aniseed.cnrs.fr/aniseed/) and the best blast hits (using the cutoff as e-value<0.01 and qcovs>30) in both species. In cases where one gene in Ciona was duplicated in mouse, the corresponding gene was duplicated in the ciona gene expression matrix (Supplementary Table 5). Each dataset was first independently clustered and marker genes identified that were expressed in a subset of the clusters using a Wilcoxon rank sum test. Scaled and centered log-normalized expression values for common marker genes between both species were then used as the input gene set for a canonical correlation analysis between the two species[43]. The top canonical correlation vector that separated cardiac from pharyngeal muscle cells across species was then identified, and the top genes 30 that contributed most to this canonical correlation vector identified. The scaled and centered expression values for these 30 genes were then used to computed a 2-dimensional tSNE embedding for each species separately[69]. Clusters in this 30-gene space were identified using an unsupervised graph-based approach, as described previously[64]. Gene set enrichment analysis was

performed on a non-redundant list of top CC genes for each mouse experiment using Panther[70], with the set of all gene orthologs between Ciona and mouse used as the background gene set.

Statistics and reproducibility

Experimental perturbations, except those with drug treatment, were performed in biological replicates. For the drug treatment, larger numbers of embryos were treated in one batch. At least 10 animals were randomly selected for each condition for analysis. The analysis, statistical tests, measurement, definition of error bars, and the batches of independent experiments are indicated in the figure legends. All of the statistical analysis for the single cell RNA-seq and bulk RNA-seq are described in detail in the methods section.

Code availability

The code/Rmarkdown files for the analyses reported in this paper are available at https://github.com/ChristiaenLab/single-cell-ciona.

Data availability

Sequencing data that support the findings of this study have been deposited in the Gene Expression Omnibus (GEO) under accession code GSE99846. Previously published microarray data that were re-analysed here are available under accession codes GSE54746. Source data for figures are provided in Supplementary Table 6. All other data supporting the findings of this study are available from the corresponding author on reasonable request.

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We are grateful to Florian Razy-Krajka for discussions and sharing reagents prior to publication. We thank Ashley Powers for help processing the single cell samples in the early phase of this study, Christopher Hafemeister and Andrew Butler for discussion and help on computational analyses. This project was funded by NIH/NHLBI R01 award HL108643 to L.C., trans-Atlantic network of excellence award 15CVD01 from the Leducq Foundation to R.K. and L.C., and an NIH New Innovator Award (DP2-HG-009623) to R.S.

REFERENCES

1. Pinto AR et al. Revisiting Cardiac Cellular Composition. Circ. Res 118, 400–409 (2016). [PubMed: 26635390]
2. Saga Y et al. MesP1 is expressed in the heart precursor cells and required for the formation of a single heart tube. Development 126, 3437–3447 (1999). [PubMed: 10393122]
3. Lescroart F et al. Early lineage restriction in temporally distinct populations of Mesp1 progenitors during mammalian heart development. Nat. Cell Biol 16, 829–840 (2014). [PubMed: 25150979]

4. Devine WP, Wythe JD, George M, Koshiba-Takeuchi K & Bruneau BG Early patterning and specification of cardiac progenitors in gastrulating mesoderm. Elife 3, (2014).
5. Meilhac SM, Esner M, Kelly RG, Nicolas J-F & Buckingham ME The clonal origin of myocardial cells in different regions of the embryonic mouse heart. Dev. Cell 6, 685–698 (2004). [PubMed: 15130493]
6. Kelly RG, Brown NA & Buckingham ME The arterial pole of the mouse heart forms from Fgf10-expressing cells in pharyngeal mesoderm. Dev. Cell 1, 435–440 (2001). [PubMed: 11702954]
7. Mosimann C et al. Chamber identity programs drive early functional partitioning of the heart. Nat. Commun 6, 8146 (2015). [PubMed: 26306682]
8. Nevis K et al. Tbx1 is required for second heart field proliferation in zebrafish. Dev. Dyn 242, 550–559 (2013). [PubMed: 23335360]
9. Lescroart F et al. Defining the earliest step of cardiovascular lineage segregation by single-cell RNA-seq. Science 359, 1177–1181 (2018). [PubMed: 29371425]
10. Diogo R et al. A new heart for a new head in vertebrate cardiopharyngeal evolution. Nature 520, 466–473 (2015). [PubMed: 25903628]
11. Nathan E et al. The contribution of Islet1-expressing splanchnic mesoderm cells to distinct branchiomeric muscles reveals significant heterogeneity in head muscle development. Development 135, 647–657 (2008). [PubMed: 18184728]
12. Harel I et al. Pharyngeal mesoderm regulatory network controls cardiac and head muscle morphogenesis. Proc. Natl. Acad. Sci. U. S. A 109, 18839–18844 (2012). [PubMed: 23112163]
13. Tirosh-Finkel L, Elhanany H, Rinon A & Tzahor E Mesoderm progenitor cells of common origin contribute to the head musculature and the cardiac outflow tract. Development 133, 1943–1953 (2006). [PubMed: 16624859]
14. Lescroart F et al. Clonal analysis reveals common lineage relationships between head muscles and second heart field derivatives in the mouse embryo. Development 137, 3269–3279 (2010). [PubMed: 20823066]
15. Gopalakrishnan S et al. A Cranial Mesoderm Origin for Esophagus Striated Muscles. Dev. Cell 34, 694–704 (2015). [PubMed: 26387456]
16. Mandal A, Holowiecki A, Song YC & Waxman JS Wnt signaling balances specification of the cardiac and pharyngeal muscle fields. Mech. Dev 143, 32–41 (2017). [PubMed: 28087459]
17. Kaplan N, Razy-Krajka F & Christiaen L Regulation and evolution of cardiopharyngeal cell identity and behavior: insights from simple chordates. Curr. Opin. Genet. Dev 32, 119–128 (2015). [PubMed: 25819888]
18. Wang W, Razy-Krajka F, Siu E, Ketcham A & Christiaen L NK4 antagonizes Tbx1/10 to promote cardiac versus pharyngeal muscle fate in the ascidian second heart field. PLoS Biol. 11, e1001725 (2013). [PubMed: 24311985]
19. Razy-Krajka F et al. Collier/OLF/EBF-dependent transcriptional dynamics control pharyngeal muscle specification from primed cardiopharyngeal progenitors. Dev. Cell 29, 263–276 (2014). [PubMed: 24794633]
20. Picelli S et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat. Methods 10, 1096–1098 (2013). [PubMed: 24056875]
21. Satija R, Farrell JA, Gennert D, Schier AF & Regev A Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol 33, 495–502 (2015). [PubMed: 25867923]
22. Stolfi A et al. Early chordate origins of the vertebrate second heart field. Science 329, 565–568 (2010). [PubMed: 20671188]
23. Trapnell C et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol 32, 381–386 (2014). [PubMed: 24658644]
24. Razy-Krajka F et al. An FGF-driven feed-forward circuit patterns the cardiopharyngeal mesoderm in space and time. Elife 7, (2018).
25. Grimm EC CONISS: a FORTRAN 77 program for stratigraphically constrained cluster analysis by the method of incremental sum of squares. Comput. Geosci 13, 13–35 (1987).
26. Nimmo RA, May GE & Enver T Primed and ready: understanding lineage commitment through single cell analysis. Trends Cell Biol. 25, 459–467 (2015). [PubMed: 26004869]

27. Tolkin T & Christiaen L Rewiring of an ancestral Tbx1/10-Ebf-Mrf network for pharyngeal muscle specification in distinct embryonic lineages. Development 143, 3852–3862 (2016). [PubMed: 27802138]
28. Moris N, Pina C & Arias AM Transition states and cell fate decisions in epigenetic landscapes. Nat. Rev. Genet 17, 693–703 (2016). [PubMed: 27616569]
29. Zhang L et al. Mesodermal Nkx2.5 is necessary and sufficient for early second heart field development. Dev. Biol 390, 68–79 (2014). [PubMed: 24613616]
30. Chen L, Fulcoli FG, Tang S & Baldini A Tbx1 regulates proliferation and differentiation of multipotent heart progenitors. Circ. Res 105, 842–851 (2009). [PubMed: 19745164]
31. Liao J et al. Identification of downstream genetic pathways of Tbx1 in the second heart field. Dev. Biol 316, 524–537 (2008). [PubMed: 18328475]
32. Davis RJ, Shen W, Heanue TA & Mardon G Mouse Dach , a homologue of Drosophila dachshund , is expressed in the developing retina, brain and limbs. Dev. Genes Evol 209, 526–536 (1999). [PubMed: 10502109]
33. Kumar JP The molecular circuitry governing retinal determination. Biochim. Biophys. Acta 1789, 306–314 (2009). [PubMed: 19013263]
34. Guo C et al. A Tbx1-Six1/Eya1-Fgf8 genetic pathway controls mammalian cardiovascular and craniofacial morphogenesis. J. Clin. Invest 121, 1585–1595 (2011). [PubMed: 21364285]
35. Zhou Z et al. Temporally Distinct Six2-Positive Second Heart Field Progenitors Regulate Mammalian Heart Development and Disease. Cell Rep. 18, 1019–1032 (2017). [PubMed: 28122228]
36. Stolfi A, Gandhi S, Salek F & Christiaen L Tissue-specific genome editing in Ciona embryos by CRISPR/Cas9. Development 141, 4115–4120 (2014). [PubMed: 25336740]
37. Gandhi S, Haeussler M, Razy-Krajka F, Christiaen L & Stolfi A Evaluation and rational design of guide RNAs for efficient CRISPR/Cas9-mediated mutagenesis in Ciona. Dev. Biol (2017). doi:10.1016/j.ydbio.2017.03.003
38. Kelly RG, Jerome-Majewska LA & Papaioannou VE The del22q11.2 candidate gene Tbx1 regulates branchiomeric myogenesis. Hum. Mol. Genet 13, 2829–2840 (2004). [PubMed: 15385444]
39. Kong P et al. Tbx1 is required autonomously for cell survival and fate in the pharyngeal core mesoderm to form the muscles of mastication. Hum. Mol. Genet 23, 4215–4231 (2014). [PubMed: 24705356]
40. Anderson HE & Christiaen L Ciona as a Simple Chordate Model for Heart Development and Regeneration. J Cardiovasc Dev Dis 3, (2016).
41. Scialdone A et al. Resolving early mesoderm diversification through single-cell expression profiling. Nature 535, 289–293 (2016). [PubMed: 27383781]
42. Ibarra-Soria X et al. Defining murine organogenesis at single-cell resolution reveals a role for the leukotriene pathway in regulating blood progenitor formation. Nat. Cell Biol 20, 127–134 (2018). [PubMed: 29311656]
43. Butler A, Hoffman P, Smibert P, Papalexi E & Satija R Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol 36, 411–420 (2018). [PubMed: 29608179]
44. Abu-Issa R, Smyth G, Smoak I, Yamamura K-I & Meyers EN Fgf8 is required for pharyngeal arch and cardiovascular development in the mouse. Development 129, 4613–4625 (2002). [PubMed: 12223417]
45. Barron M, Gao M & Lough J Requirement for BMP and FGF signaling during cardiogenic induction in non-precardiac mesoderm is specific, transient, and cooperative. Dev. Dyn 218, 383–393 (2000). [PubMed: 10842364]
46. Reifers F, Walsh EC, Léger S, Stainier DY & Brand M Induction and differentiation of the zebrafish heart requires fibroblast growth factor 8 (fgf8/acerebellar). Development 127, 225–235 (2000). [PubMed: 10603341]
47. Tirosh-Finkel L et al. BMP-mediated inhibition of FGF signaling promotes cardiomyocyte differentiation of anterior heart field progenitors. Development 137, 2989–3000 (2010). [PubMed: 20702560]

48. Hutson MR et al. Arterial pole progenitors interpret opposing FGF/BMP signals to proliferate or differentiate. Development 137, 3001–3011 (2010). [PubMed: 20702561]
49. Marques SR, Lee Y, Poss KD & Yelon D Reiterative roles for FGF signaling in the establishment of size and proportion of the zebrafish heart. Dev. Biol 321, 397–406 (2008). [PubMed: 18639539]
50. van Wijk B et al. Epicardium and myocardium separate from a common precursor pool by crosstalk between bone morphogenetic protein- and fibroblast growth factor-signaling pathways. Circ. Res 105, 431–441 (2009). [PubMed: 19628790]
51. Zhang J et al. Frs2alpha-deficiency in cardiac progenitors disrupts a subset of FGF signals required for outflow tract morphogenesis. Development 135, 3611–3622 (2008). [PubMed: 18832393]
52. Vitelli F, Morishima M, Taddei I, Lindsay EA & Baldini A Tbx1 mutation causes multiple cardiovascular defects and disrupts neural crest and cranial nerve migratory pathways. Hum. Mol. Genet 11, 915–922 (2002). [PubMed: 11971873]
53. Zhang Z, Huynh T & Baldini A Mesodermal expression of Tbx1 is necessary and sufficient for pharyngeal arch and cardiac outflow tract development. Development 133, 3587–3595 (2006). [PubMed: 16914493]
54. Li X et al. Temporal patterning of Drosophila medulla neuroblasts controls neural fates. Nature 498, 456–462 (2013). [PubMed: 23783517]
55. Hotta K et al. A web-based interactive developmental table for the ascidian Ciona intestinalis, including 3D real-image embryo reconstructions: I. From fertilized egg to hatching larva. Dev. Dyn 236, 1790–1805 (2007). [PubMed: 17557317]

METHODS REFERENCES

56. Christiaen L, Wagner E, Shi W & Levine M The sea squirt Ciona intestinalis. Cold Spring Harb. Protoc 2009, db.emo138 (2009).
57. Wang W, Racioppi C, Gravez B & Christiaen L Purification of Fluorescent Labeled Cells from Dissociated Ciona Embryos. in Advances in Experimental Medicine and Biology 101–107 (2018).
58. Beh J, Shi W, Levine M, Davidson B & Christiaen L FoxF is essential for FGF-induced migration of heart progenitor cells in the ascidian Ciona intestinalis. Development 134, 3297–3305 (2007). [PubMed: 17720694]
59. Kim D et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36 (2013). [PubMed: 23618408]
60. Trapnell C et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc 7, 562–578 (2012). [PubMed: 22383036]
61. Robinson MD, McCarthy DJ & Smyth GK edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140 (2010). [PubMed: 19910308]
62. Chung NC & Storey JD Statistical significance of variables driving systematic variation in high-dimensional data. Bioinformatics 31, 545–554 (2015). [PubMed: 25336500]
63. van der Maaten L & Hinton G Visualizing Data using t-SNE. J. Mach. Learn. Res 9, 2579–2605 (2008).
64. Villani A-C et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science 356, (2017).
65. Macosko EZ et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161, 1202–1214 (2015). [PubMed: 26000488]
66. Coifman RR & Lafon S Diffusion maps. Appl. Comput. Harmon. Anal 21, 5–30 (2006).
67. Haghverdi L, Buettner F & Theis FJ Diffusion maps for high-dimensional single-cell analysis of differentiation data. Bioinformatics 31, 2989–2998 (2015). [PubMed: 26002886]
68. Hastie T & Stuetzle W Principal Curves. J. Am. Stat. Assoc 84, 502–516 (1989).
69. Van Der Maaten L Barnes-hut-sne. arXiv.org (2013).
70. Mi H, Muruganujan A & Thomas PD PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. 41, D377–86 (2013). [PubMed: 23193289]

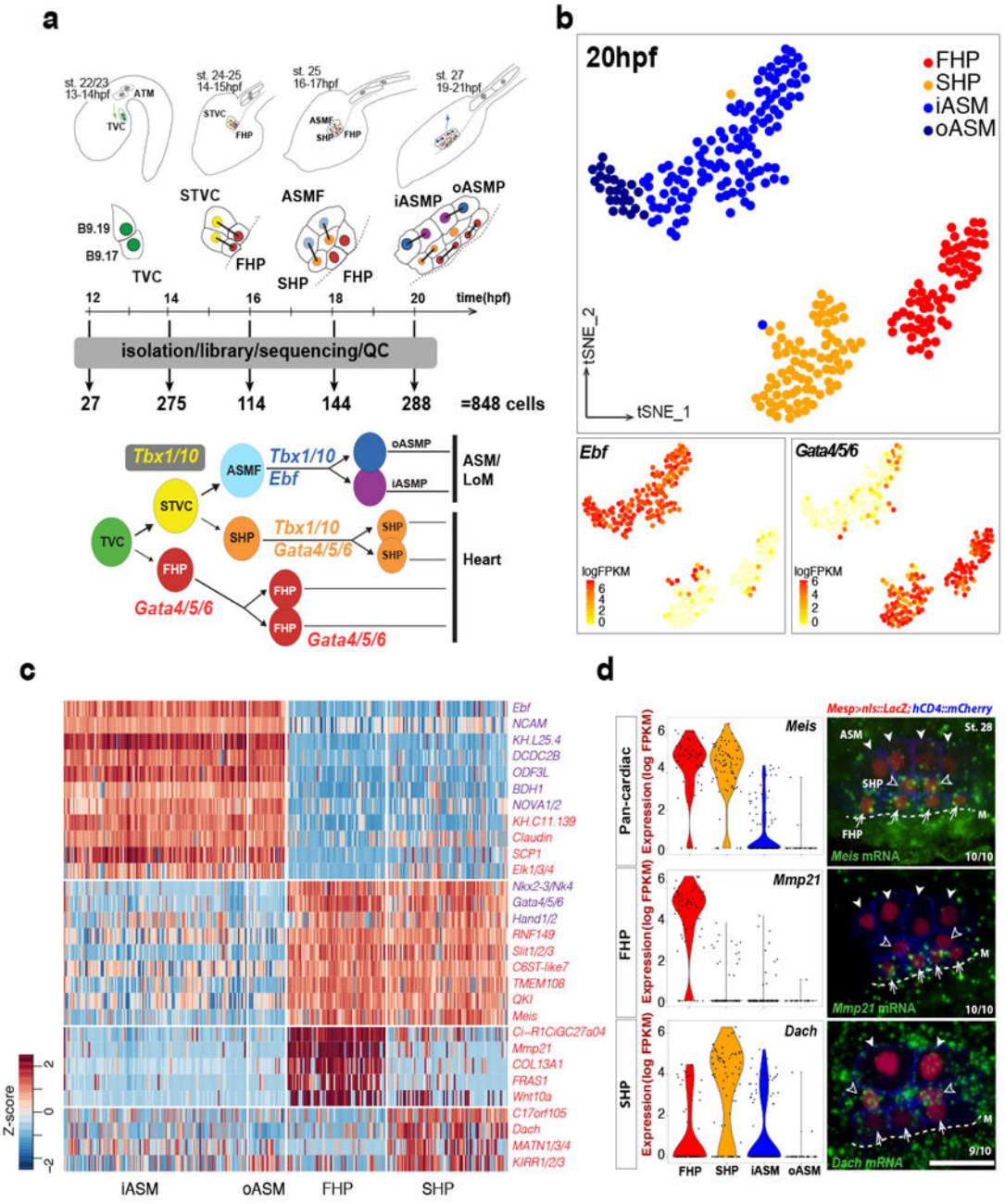

Figure 1|. Cell clustering and cell-type-specific markers.

(**a**) Early cardiopharyngeal development in *Ciona*, and sampling stages and established lineage tree. Cardiopharyngeal lineage cells are shown for only one side and known cell-type-specific marker genes are indicated. st., FABA stage[55]; hpf, hours post-fertilization; TVC, trunk ventral cell; STVC, second trunk ventral cell; FHP, first heart precursor; ASMF, atrial siphon muscle founder cells; SHP, second heart precursor; iASMP, inner atrial siphon muscle precursor; oASMP, outer atrial siphon muscle precursor; LoM, longitudinal muscles; QC, quality control. Dotted line: midline. (**b**) t-distributed Stochastic Neighbor Embedding (t-SNE) plots of 20 hpf scRNA-seq data (n=288 cells) showing distinct clusters of progenitor

subtypes: FHP (red), SHP (orange), iASMP (blue) and oASMP (dark blue). Color-coded marker gene expression levels are shown on corresponding clusters. (**c**) Expression heatmap of 20 hpf single cell transcriptomes showing top predicted differentially expressed marker genes across different cell types. Blue: previously known ASM and heart markers, red: candidate markers. (**d**) Violin plots and FISH validations of candidate cell-type-specific markers in St. 28 embryos. mRNAs visualized by whole mount fluorescent *in situ* hybridization (green). Cardiopharyngeal nuclei marked by Mesp>nls::LacZ revealed by anti beta-galactosidase antibody (red). Mesp>hCD4::mCherry, revealed by anti-mCherry antibody, marks cell membranes (blue). Anterior to the left. Scale bar, 10 μm. Solid arrowheads, ASM; open arrowheads, SHPs; arrows, FHPs; M, midline (dotted line). The numbers of observed embryos and those showing the illustrated gene expression pattern are indicated at the right bottom corner of each image. Violin plots are to visualize the distributions of the expression (log FPKM) of the indicated genes. The wide of the violin indicates the frequency of cells with indicated gene expression level. The number of cells in each cell cluster is summarized in Supplementary Table 6 (Data sheet: cell identity and number).

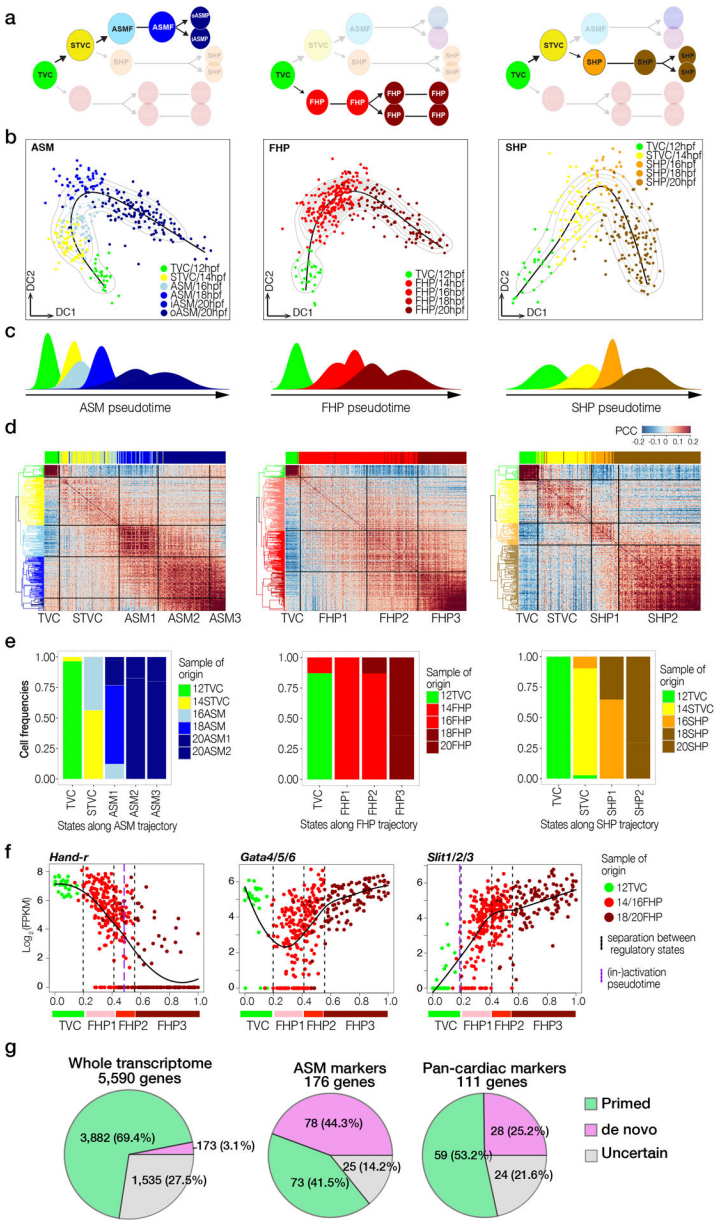

Figure 2|. Reconstruction of cardiopharyngeal developmental trajectories.

(**a**) Cell lineages used to reconstruct three unidirectional cardiopharyngeal trajectories. (**b**) Diffusion maps showing the cardiopharyngeal trajectories. Color-coded cell identities as defined by unsupervised clustering from larvae dissociated at indicated time points (Supplementary Fig. 1a). Black lines: principal curve; light gray contours: single cell density distribution. Color codes correspond to assigned cell identities following clustering at each time point. hpf, hours post-fertilization. DC: Diffusion Coordinate. (**c**) Distribution of identified cell types isolated at defined time points along the trajectories, showing the general agreement between the time series and developmental progression, but also that cells isolated from a given time point are not all at the same developmental “pseudotime”. (**d**) Cross-correlation heatmaps to infer regulatory states along the trajectories. Dendrogram

(left) obtained from constrained hierarchical clustering. Top bars indicate the sample of origin with color codes as in (c). PCC, Pearson Correlation Coefficient. (**e**) Relative cell identity composition for each regulatory states identified on the trajectories. Note the 16ASM cells clustering with the 'STVC' state in the ASM trajectory, indicating that these cells retain most STVC characteristics and have not yet activated the ASM-specific program. (**f**) Pseudotemporal expression profiles of indicated genes along the FHP trajectory. X-axis: normalized pseudotime as defined in (b), Y-axes: relative expression level. Black lines indicate the smoothed expression. Black dashed lines indicate the transitions between predicted regulatory states as defined in (d) and color-coded below. Purple dashed lines indicate calculated activation or inactivation pseudotime. Dot colors refer to the sample of origin as indicated in (c). (**g**) Proportions of primed vs. *de novo*-expressed genes among defined categories of marker genes.

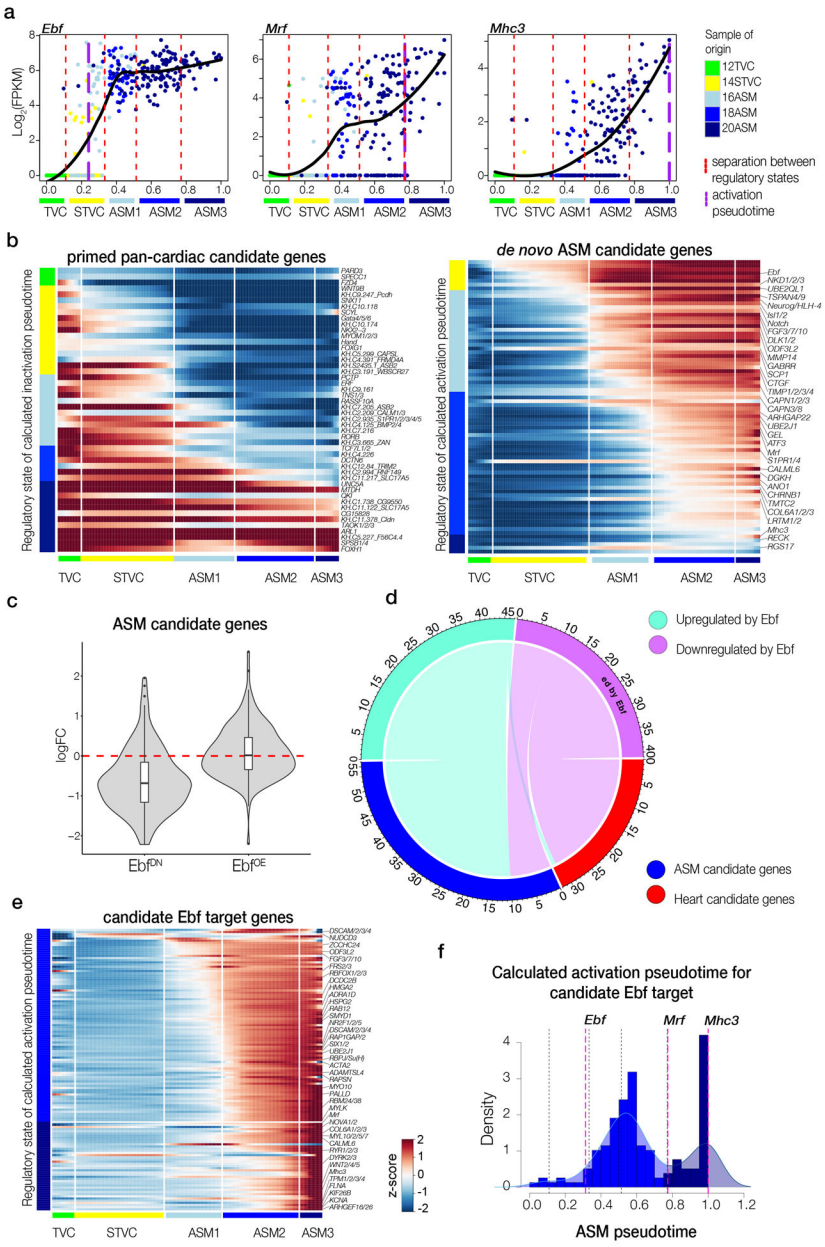

Figure 3l. Transcriptional regulation of ASM fate specification.

(**a**) Pseudotemporal expression profiles of indicated genes along the ASM trajectory. X-axis: normalized pseudotime as defined in Fig.2. Y-axes: relative expression levels. Black lines indicate the smoothed expression. Red dashed lines indicate the transitions between predicted regulatory states, indicated as in Fig. 2d, and purple dashed lines indicate calculated activation pseudotime. (**b**) Heatmap of smoothed expression profiles along the ASM trajectory showing primed pan-cardiac genes (inhibited, left) and candidate *de novo* ASM genes (activated, right). White vertical lines mark transitions between indicated regulatory states along the ASM trajectory. Colored bars on the left indicate the regulatory state of calculated activation pseudotime. (**c**) Violin plots showing the $\log_2$(fold-change) of candidate ASM-specific genes (n=159) in response to indicated perturbations of Ebf

function, a dominant-negative (Ebf^{DN}) and Ebf over-expression (Ebf^{OE}) as in Razy-Krajka et al.[4]. The white bars indicate the interquartile range. The black whiskers extended from the bars represent the upper (max) and lower (min) adjacent values in the data. The black lines in the middle of the bars show the median values. (**d**) Chord diagram showing mutual enrichment of ASM vs. Cardiac genes among candidate target genes activated or inhibited by Ebf, respectively. Ebf is predicted to downregulate a few ASM candidate genes, which are primed and quickly downregulated after ASM specification (e.g. *Hand-r*[4]). (**e**) Heatmap of smoothed expression profiles for candidate ASM-specific Ebf target genes defined in Razy-Krajka et al.[4], showing activation pseudotimes in regulatory states ASM2 and ASM3 (left blue bars). White vertical lines mark transitions between indicated regulatory states. (**f**) Predicted induction pseudotime of candidate ASM-specific Ebf target genes (black dashed lines separate corresponding ASM regulatory states).

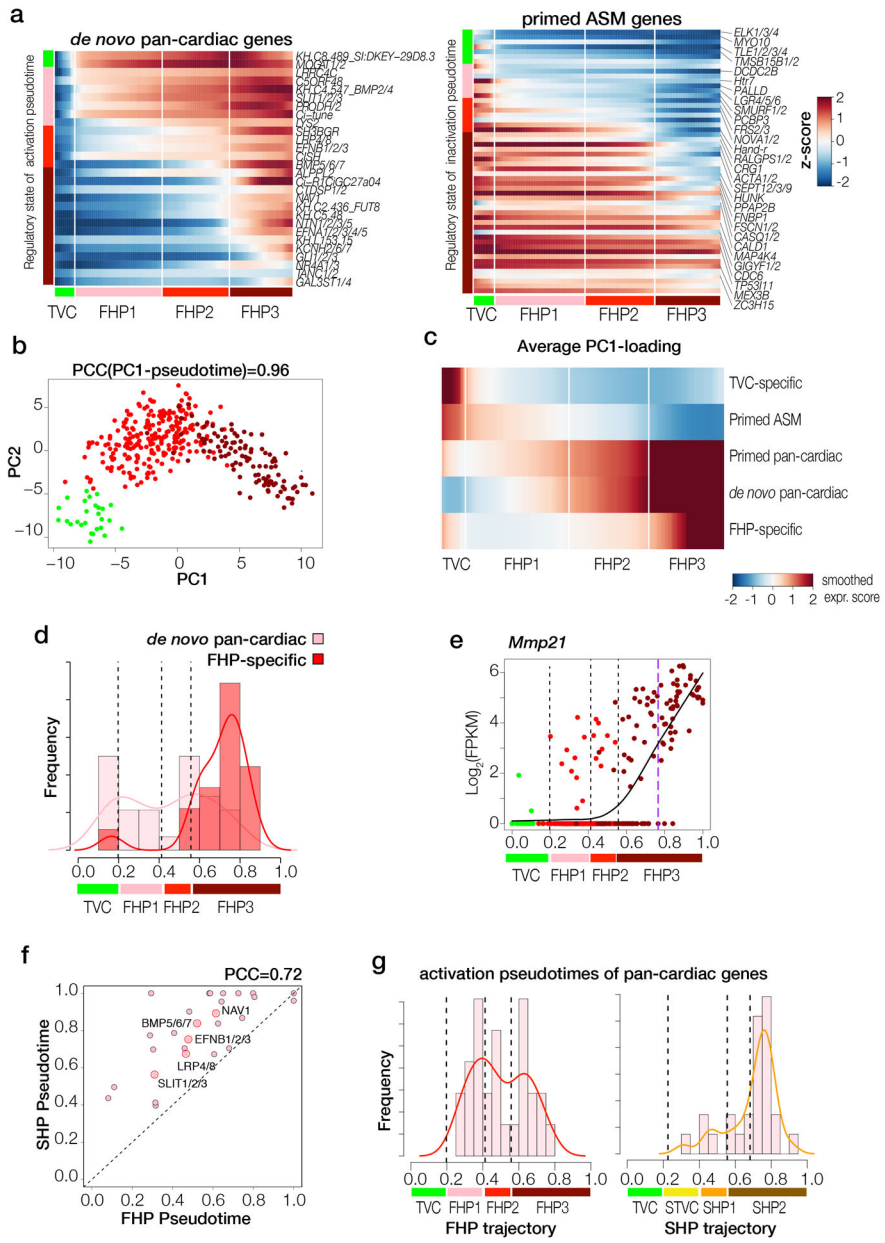

Figure 4l. A pan-cardiac program for heart fate specification.

(**a**) Smoothed gene expression along FHP pseudotime for *de novo*-expressed pan-cardiac genes (activated) and primed ASM genes (down-regulated). White vertical lines: transitions between predicted regulatory states. (**b**) Principal component (PC1) correlates with pseudotime. PCC: Pearson's Correlation Coefficient. Sample size (cells on the FHP trajectory) n=379. (**c**) Average PC1-loading scores for indicated gene category, mapped onto the FHP trajectory. (**d**) Proportions of *de novo* pan-cardiac and FHP-specific genes with calculated activation pseudotime in binned pseudotime windows along FHP trajectory. (**e**) Expression profiles of *Mmp21* along the FHP trajectory. Purple dashed line: calculated activation pseudotime. Dots colors: samples of origin as in Fig. 2b. (**f**) Activation pseudotimes for *de novo*-expressed pan-cardiac genes along first and second heart lineages

pseudotime axes. (**g**) Proportions of *de novo*-expressed pan-cardiac genes with calculated activation pseudotime in binned pseudotime windows along FHP and SHP trajectories. (d-e, g) X-axis: normalized pseudotime. Black dashed lines: transitions between regulatory states.

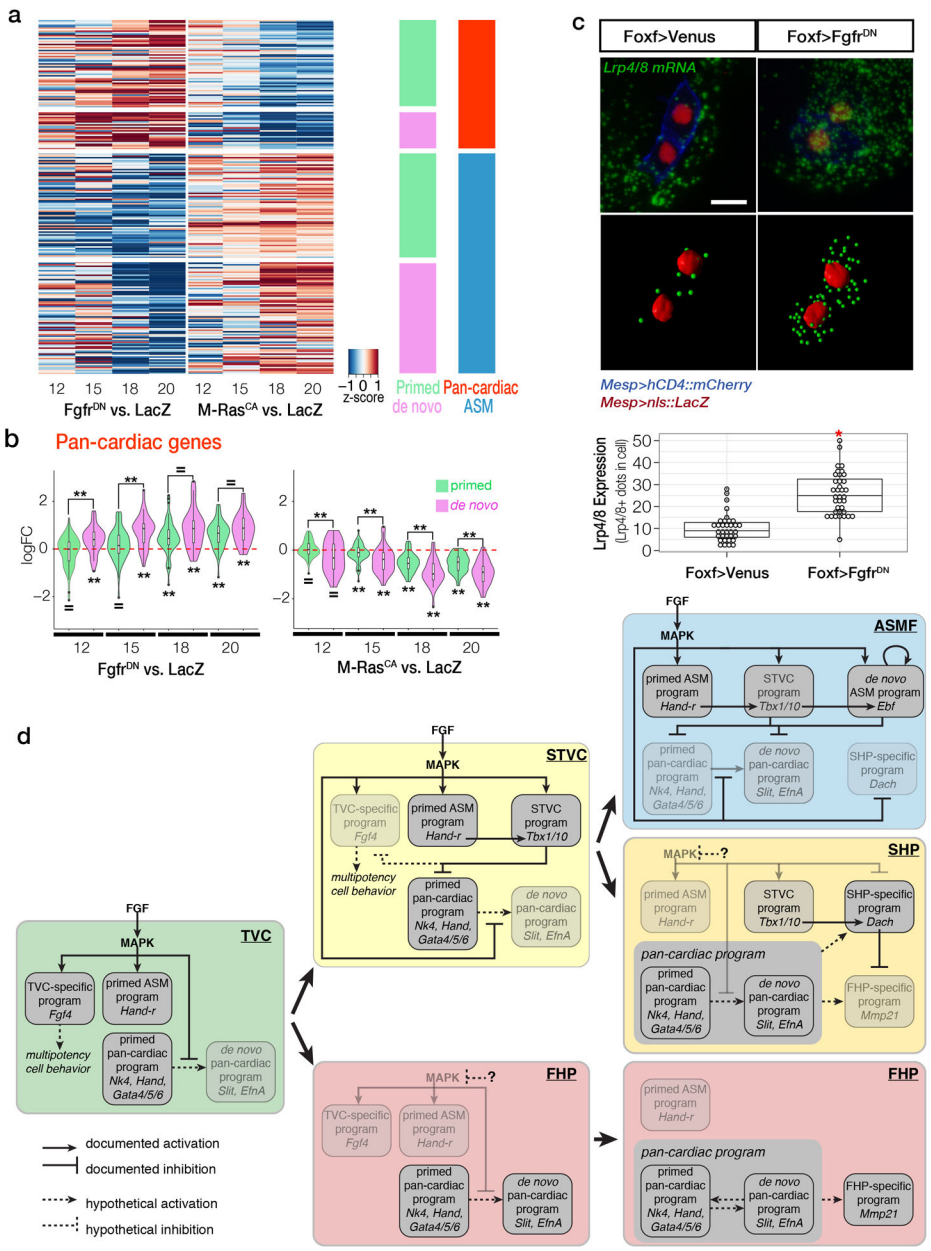

Figure 5l. FGF-MAPK signaling regulates the pan-cardiac program for heart fate specification.

(**a**) Differential expression of primed and *de novo*-expressed ASM and pan-cardiac genes in indicated conditions vs. LacZ. (**b**) Violin plots represent the distributions of $\log_2$ fold changes in indicated conditions and time points relative to LacZ controls, and parsed by primed or *de novo*-expressed pan-cardiac genes. The white bars indicate the interquartile range. The black whiskers extended from the bars represent the upper (max) and lower (min) adjacent values in the data. The black lines in the middle of the bars show the median values. Sample size: Primed Pan-cardiac genes, n= 58; *de novo* Pan-cardiac genes, n= 26. Summary statistics: results of two-tailed t-test for significant difference from 0 are indicated below violin plots, results for KS tests for significant differences between “Primed” and “*de novo*”

gene sets in each condition are indicated above violin plots. “=”, no difference, “**”, P-value <0.01. (n = 2 biological replicates; Supplementary Table 6). (**c**) FGF-MAPK inhibition induces precocious *Lrp4/8* expression in multipotent cardiopharyngeal progenitors (TVCs). *Lrp4/8* mRNAs (green) visualized by FISH and processed by Imaris (green dots). Anti-beta-galactosidase antibody (red) marks TVC nuclei expressing *Mesp>nls::LacZ*. Mesp>hCD4::mCherry, revealed by anti-mCherry antibody (blue), marks cell membranes. Anterior to the left. Scale bar, 10 μm. Box plots represent the distributions of numbers of *Lrp4/8*+ dots per cell in indicated conditions. Bars indicate the median value. * $p=4.46e^{-11}$ (One-tailed student’s t-test, n = 2 biological replicates). (**d**) Summary model showing the maintenance and progressive restriction of FGF-MAPK signaling in the multipotent progenitors (TVCs, trunk ventral cells, and STVCs, second trunk ventral cells) and atrial siphon muscle founder cells (ASMFs). Inhibition of MAPK activity permit the deployment of de novo-expressed pan-cardiac genes in both cardiac lineages (FHP, first heart precursors, and SHP, second heart precursors). FHPs specifically activate genes like Mmp21, and later produce most Mhc2+ cardiomyocytes, whereas SHPs descend from Tbx1/10+ multipotent progenitors, and activate Dach, which contributes to inhibiting the FHP-specific program. See discussion and Razy-Krajka et al.[24] for details.)

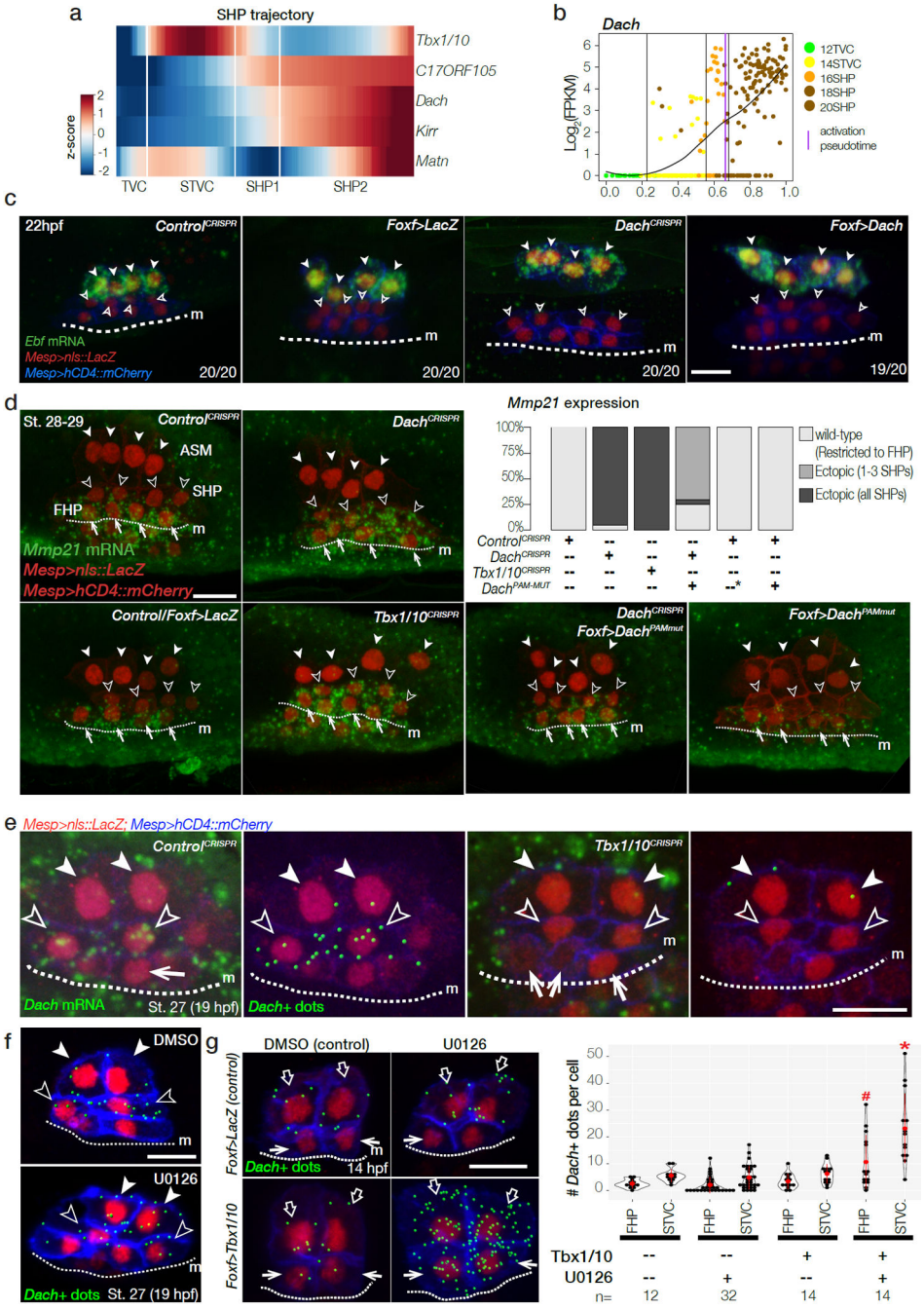

Figure 6l. A second-heart-lineage specific *Tbx1/10-Dach* pathway.

(**a**) Smoothed gene expression along the SHP trajectory for SHP-specific genes. White vertical lines: transitions between regulatory states. (**b**) *Dach* expression pattern along the SHP trajectory. Purple dashed line: predicted induction time. Black dashed lines: transitions between regulatory states. (**c**) Perturbations of Dach function does not alter Ebf expression in ASMPs. Numbers: observed/total. (**d**) Dach and Tbx1/10 are required to restrict *Mmp2*1 activation to the FHP. Barplots: proportions of larvae showing indicated phenotypes in each experimental condition. Wt, wild-type. Foxf>Dach^{PAMmut}: TVC-specific Foxf enhancer

driving expression of CRISPR-resistant Dach cDNA. (**e**) Tbx1/10 function is required for *Dach* expression in SHP. Panels 2 and 4: segmented *Dach+* dots superimposed on cell patterns. **(c-e)** Confocal stacks acquired for 10 larvae in each condition in biological duplicates. None of the 20 $Tbx1/10^{CRISPR}$ larvae showed *Dach* expression in SHPs. Solid arrowheads: ASMPs, open arrowheads: SHPs, arrows: FHPs. (**f**) FGF-MAPK signaling negatively regulates *Dach* expression in Tbx1/10+ ASMFs. Representative confocal stacks showing segmented Dach+ dots in 18.5 hpf (St. 27) larvae. Blocking MEK activity causes ectopic *Dach* expression in the ASMFs (solid arrowheads), in addition to its endogenous expression in the SHPs (open arrowheads) (n=10/10 for each condition). (**g**) Combined Tbx1/10 over-expression and MAPK inhibition induced precocious Dach expression in 14 hpf B7.5 lineage cells. Open arrows: STVCs, arrows: FHPs; dotted line: midline (m). Violin plots represent the distribution of counts Dach+ dots per cell. Black dots: cells with identify and experimental perturbation indicated below. Red dots: mean values. The thin red line: the upper (max) and lower (min) adjacent values in the data. One-tailed student's t-test indicates the precocious Dach expression in both FHPs and STVCs in combined Tbx1/10 over-expression and MAPK inhibition condition. * p=9.972e-05; # p=0.005275. **(c-g)** mRNAs (green) visualized by FISH or segmented (green dots). Mesp>nls::LacZ, revealed by anti-beta-galactosidase antibody (red), marks nuclei. Mesp>hCD4::mCherry, revealed by anti-mCherry antibody (blue), marks cell membranes. Dotted line: midline (m). Anterior to the left. Scale bar, 10 μm.

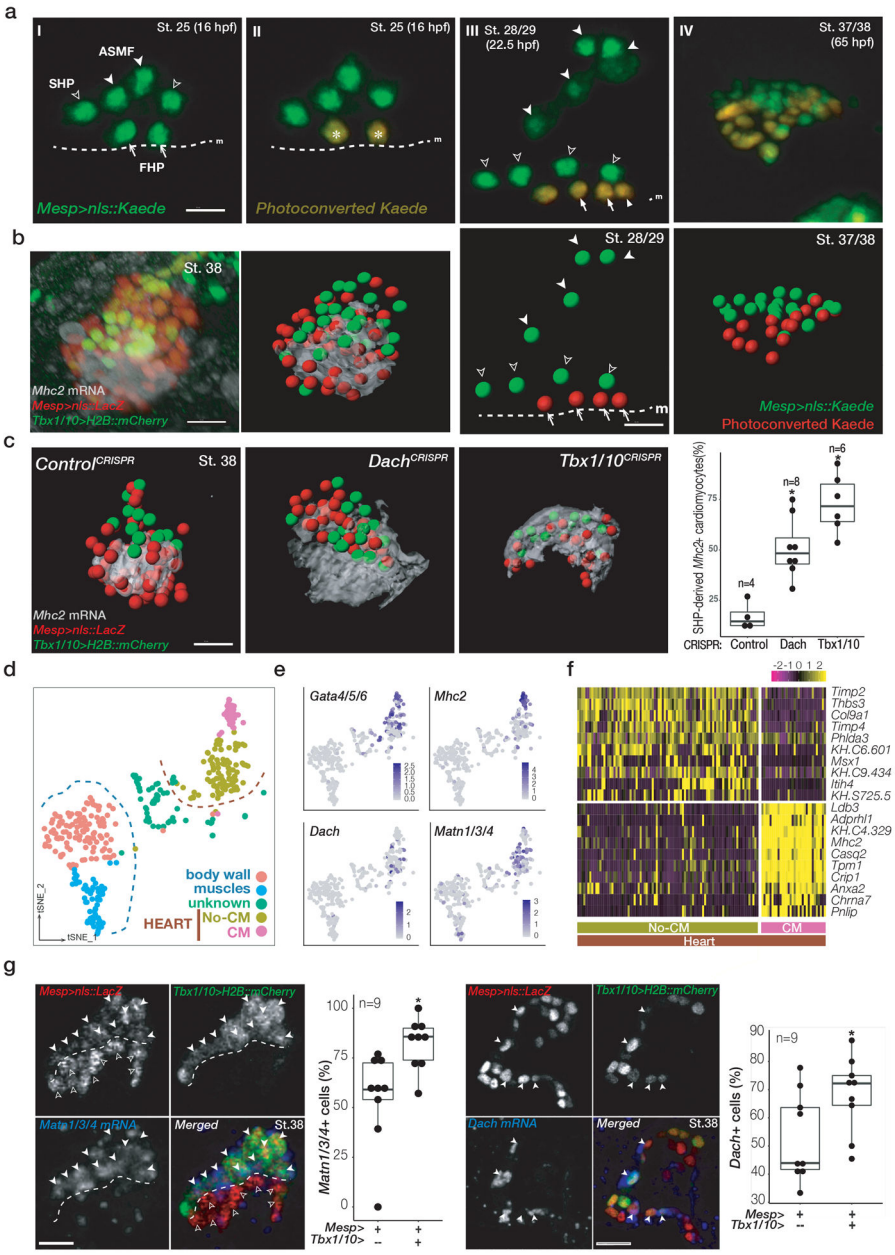

Figure 7| Characteristics and origins of the intracardiac cell diversity in beating hearts.

(**a**) Lineage tracing by photoconversion of nuclear Kaede. Mesp>nls::Kaede::nls (green) marks nuclei of live B7.5 lineage cells. Kaede photoconverted from green (I) to red (II) specifically in the FHPs of a 16hpf larva, which is shown at successive time points (III and IV). Segmented nuclei are shown below. Open arrowheads, ASMPs; solid arrowheads, SHPs; arrows, FHPs, Dotted line: midline. arrowheads, SHPs; arrows FHPs; Anterior to the left. Scale bar, 10 μm. Experiment performed in biological duplicates. (**b**) Confocal data (left) and segmented image (right) showing *Mhc2* expression (grey) primarily excluded from SHP-derived cells in wild-type juvenile heart (St. 38). *Mesp>nls::LacZ* marks nuclei of FHP and SHP-derived cells (red), *Tbx1/10>H2B::mCherry* marks only SHP-derived cells (green).

Scale bar, 10μm. (**c**) Dach and Tbx1/10 antagonize the production of *Mhc2*+ cardiomyocytes from the second heart lineage. Rendered segmented signals are shown. Grey: *Mhc2* mRNA. Green: Tbx1/10>H2B::mCherry, revealed by an anti-mCherry antibody, marks SHP-derived cells; Red: Mesp>nls::LacZ, revealed by an anti-beta-galactosidase antibody, marks all B7.5 lineage cells. Scale bar, 10μm. Boxplots: proportions of *Mhc2*+ cells among the *Tbx1/10>H2B::mCherry*+ SHP-derived cells in juvenile hearts. Bars in the box indicate the median value. * $p=1.41e^{-4}$ and * $p=2.80e^{-5}$ for $Dach^{CRISPR}$ and $Tbx1/10^{CRISPR}$, respectively (One-tailed student's t-test). N numbers represent the embryos analyiszed for the experimental perturbation as indicated. (**d**) t-SNE plots of scRNA-seq data acquired in n=386 FACS-purified cardiopharyngeal lineage cells from juveniles at St. 38. (**e**) Feature plots: expression of indicated markers in clusters shown in (d). (**f**) Top predicted differentially expressed genes across the st. 38 juvenile heart. (**g**) *Matn1/3/4* and *Dach* are enriched in the *Tbx1/10>H2B::mCherry*+ SHP-derived cells (green). Scale bar, 5μm. Images are XY cross section of juvenile hearts. Boxplots: proportions of *Matn1/3/4*+ or *Dach*+ cells among indicated cell populations. Both *Matn1/3/4* ($p = 4.459e-11$) and *Dach* ($p = 0.006795$) are significantly enriched in SHP-derived cells. N numbers represent juveniles used to quatifiy the gene expression among indicated cell populations. Bars in the box: median value in each condition. * $p<0.05$ (One-tailed student's t-test).

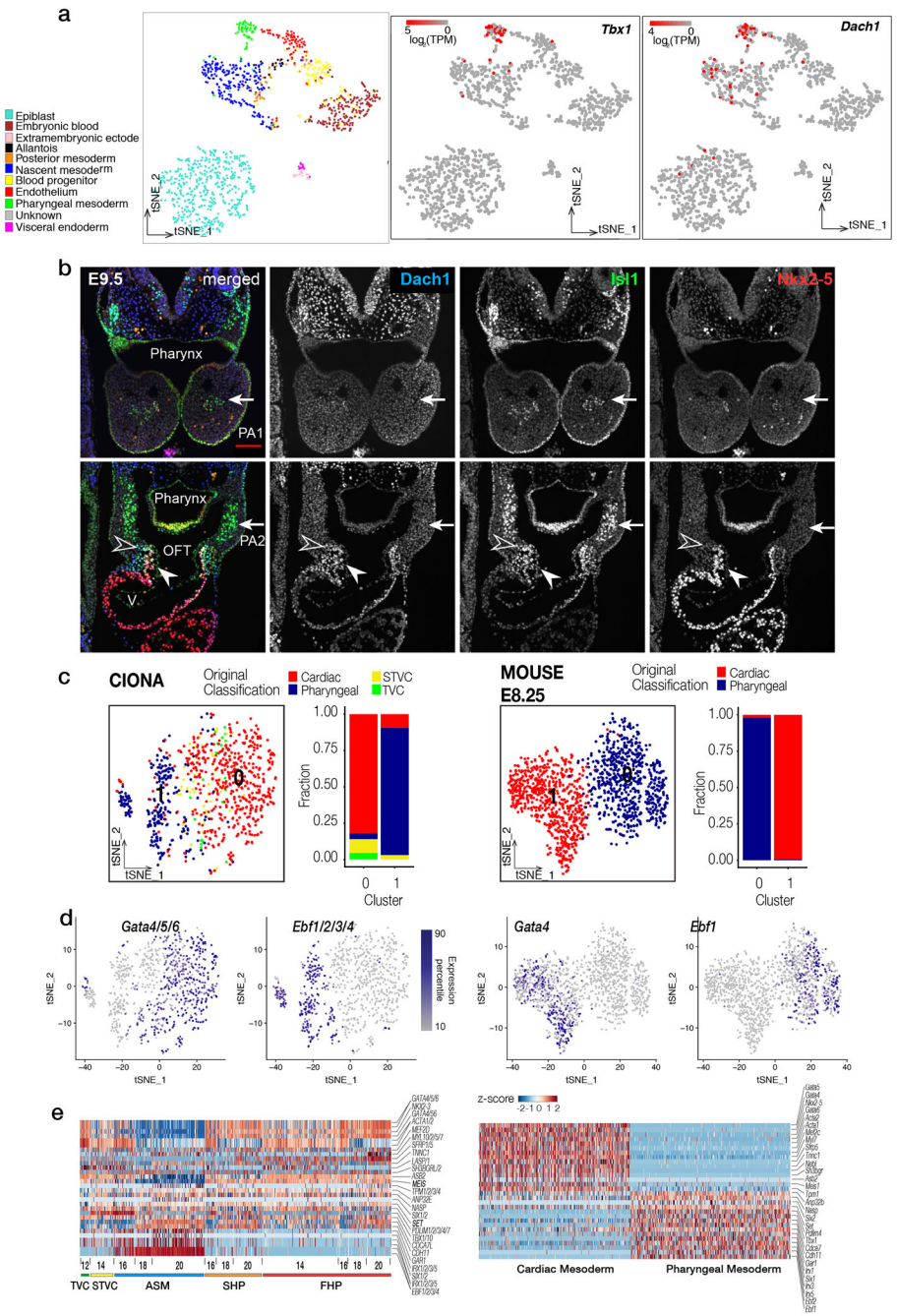

Figure 8|. Conserved cardiopharyngeal programs in chordates.

(**a**) t-SNE plots of mouse scRNA-seq data[41] (n=1,205 cells) showing *Tbx1* and *Dach1* expression patterns. Cluster identities are as determined in the original publication, with the pharyngeal mesoderm shown in green. (**b**) Expression patterns of Dach1, Islet1 and Nkx2.5 proteins in E9.5 mouse embryos. Representative image of 4 analysed embryos. Arrows: Islet1+ head muscle progenitor cells in the mesodermal core of the first (PA1, top) and second (PA2, bottom) pharyngeal arches, showing absence of Dach1 and Nkx2.5 expression. Open arrowhead: Dach1+, Islet1+, Nkx2.5− second heart field cells in the dorsal pericardial

wall, solid arrowhead: Triple Nkx2.5+, Dach1+, Islet1+ second heart field-derived cells in the outflow tract (OFT). Note the Nkx2.5+, Dach1−, Islet1− cells in the ventricle (V). Scale bar, 100 μm. (**c**) Aligned structure of Ciona and Mouse E8.25 cardiopharyngeal cells (n=2291 cells). tSNE plots showing the clustering of Ciona and Mouse E8.25 cardiopharyngeal cells, respectively, using conserved markers determined by canonical correlation (CC). Barplots indicate that original cell identities, defined in each species independently, as recovered in the clustering using conserved markers. (**d**) tSNE plots of Ciona and Mouse scRNA-seq data as described in (c), with the expression patterns of Ebf1 and Gata4. (**e**) Single cell expression profiles for the top 30 conserved markers in each species, separately.