bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Single-cell transcriptional regulations and accessible chromatin landscape of cell fate decisions in early heart development

Guangshuai Jia[1], Jens Preussner[1], Stefan Guenther, Xuejun Yuan, Michail Yekelchyk, Carsten Kuenne, Mario Looso, Yonggang Zhou and Thomas Braun*

Department of Cardiac Development and Remodeling, Max Planck Institute for Heart and Lung Research, 61231 Bad Nauheim, Germany

[1] Contributed equally

* Corresponding author: thomas.braun@mpi-bn.mpg.de

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SUMMARY

Formation and segregation of the cell lineages forming the vertebrate heart have been studied extensively by genetic cell tracing techniques and by analysis of single marker gene expression both in embryos and differentiating ES cells. However, the underlying gene regulatory networks driving cell fate transitions during early cardiogenesis is only partially understood, in part due to limited cell numbers and substantial cellular heterogeneity within the early embryo. Here, we comprehensively characterized cardiac progenitor cells (CPC) marked by Nkx2-5 and Isl1 expression from embryonic days E7.5 to E9.5 using single-cell RNA sequencing. By leveraging on cell-to-cell heterogeneity, we identified different previously unknown cardiac sub-populations. Reconstruction of the developmental trajectory revealed that Isl1$^+$ CPC represent a transitional cell population maintaining a prolonged multipotent state, whereas extended expression of Nkx-2.5 commits CPC to a unidirectional cardiomyocyte fate. Correlation-based analysis of cells in the unstable multipotent state uncovered underlying gene regulatory networks associated with differentiation. Furthermore, we show that CPC fate transitions are associated with distinct open chromatin states, which critically depend on Isl1 for accessibility of enhancers. In contrast, forced expression of Nkx2-5 eliminated multipotency of Isl1$^+$ cells and established a unidirectional cardiomyocyte fate. Our data provides a transcriptional map for early cardiogenic events at single-cell resolution and establishes a general model of transcriptional and epigenetic regulations during cardiac progenitor cell fate decisions.

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

INTRODUCTION

Cell fate mapping has revealed that cardiac progenitor cells (CPCs) form at mouse embryonic day (E) 7.5 from Mesp1^{+} cell leaving the primitive streak during gastrulation at E6.5. Although cells giving rise to myocardial cells can be defined by their position in the primitive streak at E6.5, a stable commitment is only achieved after gastrulation when endocardial, epicardial and myocardial progenitor lineages are established. In vitro differentiation of cardiac precursor cells derived from embryonic stem cells (ESC) or precardiac mesoderm indicates that Brachyury^{+}/Flk1^{+} cells give rise to Nkx2.5^{+}, Isl1^{+} and Nkx2.5^{+}/Isl1^{+} cells (reviewed by [1]). At this stage of mouse embryonic development (E7.5), Nkx2.5^{+}, Isl1^{+} and Nkx2.5^{+}/Isl1^{+} cells still retain a multilineage potential enabling them to generate cardiomyocytes, smooth muscle cells, endothelial cells and pericytes [2, 3]. Common CPCs labeled by expression of Mesp1 separate at early developmental stages into two distinct anatomical heart fields, the first (FHF) and second heart field (SHF), which contribute to the left ventricle and atria and the right ventricle, out flow tract and atria, respectively [4-6]. Aside from the different anatomical localization and contribution to different parts of the heart, the main difference between the FHF and SHF is the delayed differentiation of SHF cells, which unlike FHF cells do not differentiate immediately into myocardial cells but serve as a reservoir of multipotent CPCs during cardiogenesis (reviewed by [1]).

Isl1 is expressed in cardiac progenitor cells (CPCs) of the secondary heart field, although a broader, transient expression has been noted in the anterior intra-embryonic coelomic walls and proximal head mesenchyme encompassing both the FHF and the SHF [7, 8]. However, efficient nGFP labeling of CPCs is only achieved in the SHF making the Isl1$^{nGFP/+}$ knock-in reporter mouse line a reliable source for isolation of SHF cells [9, 10]. Moreover, inactivation of the *Isl1* gene leads to defects of SHF derivatives but does not compromise formation of the primary heart tube [11]. In contrast, Nkx-2.5 expression marks cells both of the FHF and SHF including the cardiac crescent and the pharyngeal mesoderm [1, 7, 12]. The expression of Nkx-2.5 in the pharyngeal mesoderm, the FHF and the primary heart tube suggests stage- or context-dependent cellular functions. Several lines of evidence indicate that Isl1 and Nkx-2.5 suppress

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

each other thereby allowing expansion of Isl1$^+$ CPCs and differentiation into Nkx-2.5$^+$ cardiomyocytes [7, 10].

Differentiated cells (e.g. cardiomyocytes) are assumed to acquire their identity in a successive step-wise manner from multipotent cells (e.g. CPCs) but the different intermediate states allowing transition from multipotent precursor cells to differentiated descendants still await further characterization. Furthermore, differentiation is usually not fully synchronized in a precursor cell population in vivo but is accompanied by significant cell-to-cell heterogeneity owing to spatial positioning of individual cells and temporal differences in the availability of inductive signals. Cellular fate transitions go along with gradual remodeling of underlying gene regulatory networks and dynamic chromatin landscape. Global analysis of transcriptional changes does not provide the resolution to precisely analyze such complexity and to identify specific cellular transition states. To obtain a full assessment of the developmental trajectories of precursor cells that is not only restricted to single markers but provides a comprehensive overview of involved genes, single-cell transcriptional analysis seems necessary. Recent advances in single-cell RNA sequencing (scRNA-seq) allows characterization of transcriptomes at the single cell level [13]. Analyses of large groups of closely related transcriptomes poses specific challenges for computational methods [14, 15] but have been successful to characterize mouse lung and early mesoderm development [16, 17]. Essentially, scRNA-seq data were used to project data points from the high-dimensional gene expression space into a low-dimensional latent space [17-20]. Inclusion of landmark genes [21] and microdissection of relevant structures of murine heart have further refined this approach [22, 23]. Furthermore, pseudotemporal ordering [24] has been adapted for scRNA-seq data [25, 26] allowing to reveal cellular decisions during progression of cells through developmental processes and to identify underlying gene regulatory networks (GRN) driving cell fate transitions [27]. Transcriptional changes are either preceded, followed, or accompanied by changes in chromatin organization [28]. ATAC-seq (an assay for transposase-accessible chromatin using sequencing) provides a robust means to identify chromatin closure and opening at enhancers and promoters

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

in a limited number of cells [29] but has not been applied yet to characterize chromatin accessibility and putative regulatory elements driving cardiogenesis.

Here, we used scRNA-seq to transcriptionally profile FACS-purified Nkx2.5$^+$ and Isl1$^+$ cells from E7.5, E8.5 and E9.5 mouse embryos and to monitor changes in transcription over closely linked time points for the capture of transiently unstable states. We decided to focus on native embryonic cells and not on ESC derivatives, since some in vitro results have to be viewed with a certain degree of caution despite some clear advantages of ESC-based approaches [30, 31]. By harnessing unsupervised bioinformatics analysis, we reconstructed the developmental trajectories of Nkx2.5$^+$ and Isl1$^+$ cells and identified a transition population in the second heart field, which became developmentally arrested after inactivation of Isl1. We show that forced expression of Nkx2.5 primes the cardiomyocyte fate and used ATAC-seq to characterize accompanying changes in the chromatin landscape. Our study provides a rich source for future studies aiming to dissect the functional role of newly identified genes in cardiac development and congenital heart disease.

RESULTS

Single cell transcriptomics reveals heterogeneity of cardiac progenitor cells

Isl1 is expressed in common cardiac progenitor cells (CPCs) in the secondary heart field, although a broader, transient expression has been noted in the anterior intra-embryonic coelomic walls and proximal head mesenchyme encompassing both the FHF and the SHF [7, 8]. However, efficient nGFP labeling of CPCs is only achieved in the SHF making the Isl1$^{nGFP/+}$ knock-in reporter mouse line a reliable source to isolate cells of the SHF [10] (Fig. 1a). In contrast, Nkx2-5 expression marks cells in both the FHF and SHF including the cardiac crescent and the pharyngeal mesoderm [1, 7, 12]. The expression of Nkx2-5 in the pharyngeal mesoderm, the FHF and the primary heart tube suggests stage- or context-dependent cellular functions. Several lines of evidence indicate that Isl1 and Nkx2-5 suppress each other thereby allowing expansion of Isl1$^+$ CPCs and differentiation into Nkx-2-5$^+$ cardiomyocytes [7, 10]. However, the interplay between Isl1 and Nkx2-5 at the singular cell level at different stages of heart

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

development is far from clear. To unravel the molecular composition of either Isl1$^+$ or Nkx2-5$^+$ CPCs, we isolated GFP$^+$ cells by FACS from Nkx2-5-emGFP transgenic and Isl1$^{nGFP/+}$ knock-in embryos (Fig. 1a) at E7.5, E8.5 and E9.5 and performed single-cell RNA sequencing using the Fludigm C1 workstation (Fig. 1b). At E8.5 and E9.5 we used dissected hearts instead of the whole embryo, to avoid contamination of non-cardiogenic cells that might be marked by Isl1 or Nkx2-5 expression. After removal of low quality cells (Supplementary Fig. 1a-g; Methods), we obtained 167 Nkx2-5$^+$ and 254 Isl1$^+$ cell transcriptomes which cover the whole early heart development stages (Fig. 1b).

We firstly sought to reveal the heterogeneous populations of Nkx2-5$^+$ and Isl1$^+$ CPCs that were sampled at successive developmental time-points described above. We analyzed the coefficient of variation and drop-out rates to defined heterogeneous genes as input for a neuronal network-based dimension reduction strategy (self-organizing map, SOM) [32], (Supplementary Fig. 2a & b; Methods). The resulting SOM maps were projected into 2D for visualization by t-distributed stochastic neighbor embedding (t-SNE). This strategy allowed us to identify three major subpopulations of Nkx2-5$^+$ and five subpopulations of Isl1$^+$ cells (Fig. 1c). The Nkx2-5$^+$ cluster 3 mainly comprised E7.5 cells, whereas cluster 1 contained cells from E8.5 and E9.5 implying an intermediate cell state. Cluster 2 predominantly contained cells from E9.5 (Fig. 1d). Stage-dependent clustering was less evident for the five Isl1$^+$ subpopulations, which might indicate that the specific cellular phenotypes of Isl1$^+$ subpopulations are maintained for longer time periods even in a changing developmental environment (Fig. 1d).

The existence of discrete clusters of CPCs within the developmental continuum suggests that subpopulation-specific genes might regulate proprietary cellular decisions. To identify such cluster-specific, differentially expressed genes (marker genes), we used MAST [33] and a gene ranking approach implemented in SC3 [34]. The top 269 and 216 genes that were differentially expressed in the Nkx2-5+ and the Isl1+ lineage, respectively, included several established cardiac regulatory genes such as *Hand1, Tbx3/4/5, Gata2/3, Smarcd3, Rbm24, Wnt5a, Bmp4, Notch1* and *Fgf3/15* (Fig. 1e; Supplementary Table 1, 2) [22, 23, 35-37]. Importantly, we detected numerous

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

differentially expressed genes that so far had not been linked to cardiogenesis probably due to expression in limited cell numbers, which renders them invisible to conventional bulk transcriptome analysis (Fig. 1e; Supplementary Table 1, 2). For example, Isl1$^+$ cluster 5 expressed several *Hox* genes including *Hoxa7/9/10, Hoxb6, Hoxc8* and *Hoxd8* which were co-expressed with cardiac transcription factor *Tbx3/4* and *Wnt5a* (Fig. 1e; Supplementary Fig. 3a) and might contribute to cardiac patterning in the second heart field between E7.5 and E8.5. Genes located at the 3'-ends of *Hox* clusters have not been implicated in early heart development before, although mutants of *Hoxa1, Hoxa3, Hoxb1* genes [38-42], located at 5'-ends of Hox clusters and were expressed more broadly across cardiac progenitor cell clusters (Supplementary Fig. 3b), and compound deletion of *Hox A/B* clusters caused cardiac defects [43]. In addition, we newly identified several transcription factors (Fig. 1e; Supplementary Table 1, 2), such as *Sox7/18, Sall3, Ldb3, Zbtb20, Zfp462/512b/711, Klf14*; G-proteins including *Arhgap1, Adgrf5, Arhgef15*; and the *de novo* DNA methyltransferase *Dnmt3b* in Nkx2-5$^+$ or Isl1$^+$ clusters.

Next, we assigned identities to each cluster based on the expression of key marker genes (Fig. 1f, g). Consistent with the gene ontology analysis of differentially expressed genes within each cluster, Nkx2-5$^+$ cluster 3 and Isl1$^+$ cluster 2, characterized by *cTnt* and *α-smooth muscle actin* expression appear to represent a myogenic fate whereas Isl1+ cluster 1, expressing *Cd31*, is characterized by endothelial cell features (Fig. 1f, g; Supplementary Fig. 4). Interestingly, expression of *Nkx-2.5* and *Isl1* varied among subpopulations within each lineage: (i) Nkx2-5 shows more pronounced expression in late stages (clusters 3 and 1, E8.5 and E9.5) (Fig. 1f, Fig. 2e); (ii) Isl1 expression decreases in the cells expressing differentiation markers (clusters 2 and 1) (Fig. 1g, Fig. 2f). This pattern suggests that Nkx-2.5 is associated with myogenic differentiation while Isl1 is linked to the maintenance of progenitor cell multipotency, which is consistent with previous models [10, 44, 45].

To test robustness of our approach and to analyze whether we have sequenced sufficient numbers of cells to unveil the entire heterogeneity of CPCS, we generated single-cell transcriptomes of additional 663 Nkx2-5$^+$ CPCs this time using WaferGen

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

iCell8 system (Fig. 1b). Since the sequencing depth of libraries generated with C1 and iCell8 systems (Supplementary Fig. 1a-e) differed, which masks the true biological heterogeneity, we corrected the resulting batch effects using the MNN method [46]. After successful merging and aligning data from both systems (Supplementary Fig. 5a, b), we performed the same analysis as described above. We identified 3 clusters of Nxk2-5$^+$ CPCs, essentially mirroring the C1 data (Supplementary Fig. 5c, d). The consistent recapitulation of subpopulations when using significantly more cells (663 vs 167 cells) suggests that even comparatively low numbers are sufficient to unravel the heterogeneity among Nxk2-5$^+$ CPCs. Furthermore, we compared genes that were found to be differentially expressed among subpopulations using the C1 platform with the merged data set. We detected a similar distribution of marker genes in cluster 1 and 3 but did not fully reproduce the marker gene pattern for cluster 2 (Supplementary Fig. 5e). We concluded that sequencing depth rather than cell numbers is the main limiting factor for discovery of novel genes in cardiac progenitor cells. Thus, we hereafter focused our analysis on the C1 data, which provided substantially deeper sequence coverage (Supplementary Fig. 1a-e).

Reconstruction of development trajectories reveals differential developmental potential of Isl1$^+$ and Nkx2-5$^+$ CPCs

Availability of scRNA-seq data allows ordering of cells by “pseudotime” based on cell-to-cell transcriptome similarity to reveal underlying developmental trajectories. We took advantage of diffusion maps to arrange cardiac progenitor cells according to their developmental pseudotime [26]. We mapped the cells collected at successive developmental stages along the pseudotime and reconstructed the developmental trajectories of Nkx2-5$^+$ and Isl1$^+$ CPCs (Fig. 2a, b). Interestingly, cells collected at the same embryonic stages aligned to broad pseudotime points, which suggested that CPCs are not synchronized at different embryonic stages but follow individual developmental traits. Next, we aligned the different Nkx2-5$^+$ and Isl1$^+$ cell clusters to developmental trajectories. Nkx2-5$^+$ CPCs showed one continuous trajectory suggesting a unipotent differentiation capacity (Fig. 2a). In contrast, the trajectory of Isl1$^+$ CPCs bifurcated into two distinct orientations, namely towards endothelial cells and

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

cardiomyocytes, respectively, suggesting the existence of a transition state, which separates multipotency of Isl1$^+$ CPCs from acquisition of distinct cellular identities (Fig. 2b).

Cells undergoing a critical fate decision (such as lineage bifurcation) have been postulated to pass transition states [27], which corresponds to a switch between different attractor states (i.e. stable states of the underlying gene regulatory network) [47]. To delineate such transition states, we calculated the critical transition index of Nkx2-5$^+$ and Isl1$^+$ cell clusters [abbreviated as $I_c(c)$] [48]. The $I_c(c)$ values of Nkx2-5$^+$ clusters showed similar numerical ranges essentially excluding existence of a transition state in the Nkx2-5$^+$ cell population, although cells from later stages (cluster 1) were characterized by a decrease of $I_c(c)$ indicating stable settlement into an attractor state (Fig. 2c, left), characterized by a cardiomyocyte-like expression (Fig. 1e). In contrast, computation of $I_C(c)$ values of Isl1$^+$ clusters revealed decreased values for cells at the bifurcation point (cluster 3 and 4), indicating appearance of a transition state. Cells that overcame this point (clusters 1 and 2) exhibited more coordinated changes of gene expression changes (Fig. 2c, right). Since the critical transition index only hints to the presence of a transition state but does not reveal its stability, we calculated pairwise cell-to-cell distances [49] (Supplementary Fig. 6a & b). High cell-to-cell distances in seemingly homogeneous populations indicate transcriptional noise, which occurs when several gene regulatory networks become active opening new opportunities for cellular decisions. As expected, cell-to-cell distances of Nkx2-5$^+$ CPCs did not change dramatically while cell-to-cell distances of Isl1$^+$ CPCs in cluster 3 and 4 increased substantially (Fig. 2d). Taken together our results show that Nkx2-5$^+$ CPCs follow a straight path towards their cardiomyocyte fate whereas Isl1$^+$ CPCs need to overcome a transition state with elevated noise levels.

To identify genes that are potentially required to determine and/or maintain correspondent cell states, we generated a list of 108 genes for Nkx2-5$^+$ cells and 130 genes for Isl1$^+$ cells positively correlated with pseudotime across clusters (Supplementary Fig. 7a & b) (Spearman rank correlation coefficient > 0.5;

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Table 3). Gene expression levels were plotted along the Nkx2-5$^+$ trajectory and along the two branching trajectories of the Isl1$^+$ lineage to visualize the establishment of stage-specific gene expression patterns (Supplementary Fig. 7c & d). Furthermore, we classified correlated genes expressed in multipotent progenitor cells at early stages into a "*priming*" category or as "*de novo*" genes when expressed at fate-restricted stages. Transcription factors (TFs) and their chromatin-modifying partners play pivotal roles in during heart development [36]. To understand the mechanism regulating cardiogenesis in CPCs, we focused on "*priming*" and "*de novo"* TFs (Fig. 2e & f). We note that along Nkx2-5$^+$ developmental trajectories, the expression of *de novo* DNA mthyltransferase *Dnmt3b, Gata2/3, Hand1* and *Msx1* were increased, serving as "*priming*" genes. By contrast, the expression of *Nkx2-5, Ankrd1, Asb2, Cdkn2d, Hopx, Mef2c, Myocd, Smyd1, Tgfb1i1 and Tbx20* were down-regulated during Nkx2-5$^+$ CPC differentiation (Fig. 2e). Moreover, the "*priming"* TFs for Isl1$^+$ CPCs include *Hox*-family genes (*Hoxa7, Hoxa9, Hoxb6, Hoxc8* and *Hoxd8*), cardiac TFs (*Gata2, Hand1,* and *Tbx3/4*), *Hhex, Msx1, Pitx1, Nkx1-2, Ets2, Sall4 and Snai1.* The "*de novo*" TFs for Isl1$^+$ CPCs are *Mef2c, Nkx2-5, Tgfb1i1, Asb2, Ankrd1, Myocd* and *Smyd1* (Fig. 2f). Overall, the distinct expression pattern of "*priming"* and "*de novo"* TFs in Nkx2-5$^+$ compared to Isl1$^+$ cells implies that different gene regulatory networks are active in either cell population.

Isl1 is indispensable to overcome a stable attractor state prior to bifurcation

The loss of Isl1 results in absence of outflow tract and right ventricle and early embryonic lethality [11], which prevents dissection of Isl1 dependent molecular processes in the SHF. To address the role of Isl1 in cell fate determination, we inactivated the *Isl1* gene by generating Isl1$^{nGFP/nGFP}$ embryos (serving as Isl1$^{-/-}$) and isolated Isl1-GFP$^+$ cells by FACS for scRNA-seq analysis at E9.5 (Fig. 3a). Projection of Isl1-KO single cells on the trajectory of the developing SHF revealed that Isl1-KO cells are stalled/trapped in the previously identified stable attractor state (Fig. 3b). We next asked whether the stalled/trapped state of Isl1-KO cells are resulted from cell proliferation defects. We scored each KO and WT cell for its likely G1/S and G2/M cell cycle phase (Supplementary Fig. 8a-c; Methods), and found cycling cells of Isl1-KO

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

were marginally, although not significantly, less comparing to wild type cells (χ^2 test: p= 0.062) (Fig. 3c). This hints that proliferation defects partially resulted in attractor state of Isl1-KO cells. To get biological function hints from deregulated genes of Isl1-KO cells, we performed gene ontology analysis and found regulation of cell differentiation genes (GO: 0045595) were deregulated upon inactivation of Isl1 (Fig. 3d). Consistent with that Isl1-KO cells failed to differentiate towards either endothelial or cardiomyocyte fates, GO term “endothelial cell migration” and “muscle organ development” were enriched in differentially expressed genes in WT cells comparing to Isl1-KO (Fig. 3d). Taken together, our results indicate the molecular mechanisms how Isl1 plays the crucial role for CPC fate bifurcation and transition.

Nkx-2.5 establishes a unidirectional fate for CPCs towards cardiomyocytes

Our pseudotime-based analysis of developmental trajectories revealed one continuous trajectory of Nkx2-5$^+$ CPCs suggesting that Nkx2-5$^+$ cells are exclusively committed to become cardiomyocytes. Of course, such a conclusion is not compatible with previous lineage tracing studies suggesting that Nkx2-5$^+$ cells contribute to other cell types such as smooth muscle cells [50]. We reasoned that Nkx2-5 expression is essential to maintain the ability of multipotent progenitor cells to differentiate into cardiomyocytes but that Nkx-2.5 expression is quickly terminated in cells, which acquire a stable smooth muscle cell fate thereby escaping Nkx2-5-emGFP based FACS-sorting. scRNA-Seq analysis indicated that Nkx2-5$^+$ cells at E8.5 express *α-SMA* as well as several other smooth muscle markers such as *Caldesmon*, *Tagln* and *Cnn1* but also cardiac markers such as *cTNT* (Fig. 1f, g; Supplementary Fig. 9), which clearly suggest the ability to differentiate into cardiomyocytes and smooth muscle cells. Cardiac priming at E8.5 due to continued expression of Nkx2-5 might overcome smooth muscle identity and induce a stable cardiomyocyte fate. To directly test this hypothesis, we pursued both loss of function and gain of function approaches. First, we re-analyzed the published scRNA-seq data of Nkx2-5 null embryonic hearts at E9.5 [22] and found significantly increased numbers of smooth muscle cells raising from 14.5% (138/949 cells) in wildtype to 31.2% (39/125 cells) in Nkx2-5 mutant hearts (Fig. 4a). Second, we specifically expressed Nkx2-5 fused EGFP in the Isl1$^+$ lineage using Isl1-Cre to initiate transcription from the

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Rosa26 locus (hereafter named Isl1$^+$/Nkx-2.5OE) [10]. Isolation of GFP$^+$ cells by FACS from E12.5 embryonic hearts and scRNA-seq (Fig. 4b) revealed that Isl1$^+$/Nkx-2.5OE cells align to the Nkx2-5$^+$ trajectories and the cardiomyocyte-like branch of the Isl1$^+$ trajectory (Fig. 4c, d). Importantly, Isl1$^+$/Nkx-2.5OE cells did not contain any endothelial cell- or smooth muscle cell- like populations, although by E12.5 Isl1$^+$ cells have given rise to multiple endothelial cells in wildtype conditions (Fig. 4d). Taken together our results indicate that Nkx2-5 is required and sufficient to induce cell fate bifurcation and to resolve the multipotent differentiation capacity of CPCs.

The landscape of chromatin accessibility in cardiac progenitor cells

To assess changes in genome-wide chromatin accessibility during early heart development, we performed ATAC-seq on 2,000–50,000 of FACS-sorted Nkx2-5$^+$ and Isl1$^+$ cardiac progenitor cells sampled at E7.5, E8.5 and E9.5 (Fig. 1b). For each condition and time-point, at least two biological replicates were sequenced to obtain an average of 36 million paired-end reads per sample (Supplementary Table 4) (Supplementary Fig. 10a). In addition, we obtained transcriptional profiles of each corresponding developmental stage by bulk RNA-seq using SMART-seq2 [51]. Only the biological replicates of high reproducibility were retained for further analysis (Supplementary Fig. 10a & b). Using edgeR [52] for sequential pairwise comparisons of the Nkx2-5$^+$ and Isl1$^+$ samples separately, we detected a total of 5,866 differential chromatin accessibility peaks (log2[fold changes]>2 or <-2, FDR <0.05) in Nkx2-5$^+$ cells and Isl1$^+$ cells at different developmental stages. Interestingly, many differentially accessible regions in in Nkx2-5$^+$ cells represented chromatin-opening events, although chromatin closure was fairly abundant. In contrast, chromatin-closing events represented the vast majority of chromatin accessibility changes in Isl1$^+$ cells (Fig. 5a) implying distinct chromatin states in either cell population. We observed robust closing chromatin peaks from E8.5 to E9.5 in the Nkx2-5 and the Isl1 lineage, suggesting that cell-fate restriction is associated with global loss of chromatin accessibility (Fig. 5a).

K-means clustering of the genome-wide distribution of differential peaks revealed 7 chromatin accessibility clusters for the three investigated developmental stages at

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

locations proximal and distal to transcriptional start sites (TSS) (Fig. 5b). Cluster 1 and 7 showed increased chromatin accessibility from E7.5 to E8.5 whereas clusters 2-6 represented closing chromatin, probably associated with differentiation of CPCs (Fig. 5b, Supplementary Fig. 11a). Between E8.5 and E9.5, we observed both a substantial loss of chromatin accessibility (cluster 6) and a gain of chromatin accessibility (cluster 7). Thus, we identified the combinatory patterns of distinct chromatin accessibility patterns that represent cell fate transition and terminal differentiation.

Next, we explored whether changes in proximal chromatin accessible peaks regulate gene expression and which genes were subject to regulation. In support of a functional role for transcription regulations, the proximal peaks are highly enriched at transcription start sites (TSS) (Fig. 5c). We selected the differential chromatin accessibility peaks located in the promoters and the changes of which is positively correlated with differential gene expression (DE genes) across all the conditions, and generated 805 of differential-peak/DE-gene pairs (Fig. 5d). Gene ontology (GO) analysis of this set of genes revealed enrichment for the terms cell signaling, regulation of cell signaling and cell communication. A minor group of genes was associated with heart-specific term such as heart process, heart contraction or cardiac muscle cell action potential (Fig. 5e). The genes involved in cell signaling comprised the *Wnt, BMP, FGF, Notch, TGF-β* and *Ras-MAPK* families suggesting that the dynamic changes in chromatin accessibility renders such cells responsive to environmental stimuli during development. For example, canonical Wnts like Wnt8a maintains proliferation of CPCs proliferation while non-canonical Wnts (e.g. Wnt11) promote differention [53, 54]. Correspondingly, the ATAC-seq analysis demonstrated that the promoter of Wnt8a is closing from E7.5 to E9.5 during Isl1^{+} CPCs differentiation while the promoter and gene body of *Wnt11* are more accessible from E7.5 to E9.5 in Nkx2-5^{+} CPCs (Fig. 5f, g).

ATAC-seq provides an excellent tool to identify transcription factor (TFs) binding regions that become accessible due to nucleosome eviction and/or chromatin remodeling [29] by interrogating enrichment of TF motifs. Such hotspots for transcription factor binding are potential enhancers. H3K4me1 and H3K27ac are the predominant histone

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

modifications found at enhancer elements [55]. Concordantly, we noted that H3K4me1 and active mark H3K27ac showed enriched distributions with distal ATAC-seq peaks but not the repressive mark H3K27me3 or random peaks (Fig. 5h).

After exclusion of transcription factors that are not expressed in corresponding development stages based on our RNA-seq results, we scanned accessible peaks in chromatin for clusters 4-7 (distal peaks) at E7.5-E9.5, using the motifs from a set of 364 transcription factors and the motif analysis package HOMER [56]. Gata family TFs were enriched in cluster 6 both in Isl1$^+$ and in Nkx2-5$^+$ cells at E7.5 and 8.5 but not at E9.5, suggesting they play roles in cell fate transition but not in differentiation (Fig. 5i). Interestingly, although Gata1/3/4 are expressed, they are not significantly enriched in clusters 4/5/7 at E7.5 and 8.5, suggesting the uneven genome distribution of developmental enhancers contributes to deployment of TF functions. In contrast to Gata-family TFs, Mef2a/b/c are enriched at E9.5 in cluster 7 in both Isl1$^+$ and in Nkx2-5$^+$ cells, suggesting Mef2 family TFs play role in CPC differentiation. Notably, we observed abrupt closure of sites containing motifs for several TFs including Zfx, Tcf12/21, Zfp809, POU3F1 and Nr2c2 from E7.5 to E8.5 in Isl1$^+$ cells (cluster 6) indicating that binding of such TFs has to be abrogated quickly once CPC fate is established and before CPC specification occurs. Consistent with the loss of chromatin accessibility in cluster 6 from E8.5 to E9.5 (Fig. 5b), most TFs found at E8.5 (cell transition state identified by our scRNA-seq as above) were completely absent at E9.5 in both Isl1$^+$ and in Nkx2-5$^+$ cells (Fig. 5i). Some TF motif patterns were conserved between Isl1$^+$ and Nkx-2.5$^+$ cells, such as the pattern for Gata1/3/4, but the majority was different, suggesting that the same TF binds in a cell-type specific manner at different developmental enhancers to direct the fate of CPCs. Taken together our newly established open chromatin atlas in combination with RNA-seq identified a set of transcription factors that seems to orchestrate early heart development through distinct developmental enhancers, which are differentially active in either Isl1$^+$ or Nkx2-5$^+$ cells.

Isl1 and Nkx2-5 shape chromatin accessibility during cardiogenesis

To analyze how the lack of Isl1 expression affects chromatin accessibility, we performed ATAC-seq on Isl1 mutant CPCs at E9.5. Inactivation of the *Isl1* gene resulted in

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

comparable numbers of opening and closing peaks (386 opening and 345 closing) (Fig. 6a). Since our scRNA-seq analysis at E9.5 indicated arrest of Isl1 mutant CPC development, essentially converting the transcriptional profile of E9.5 into an E8.5 state (Fig. 3), we compared the landscape of chromatin accessibility of both populations. Surprisingly, we found that E8.5 Isl1^{+} cells and E9.5 Isl1 mutant cells show significantly different open chromatin signatures (432 opening and 728 closing), although they share similar positions at the developmental pseudotime trajectory (Fig. 6a), which indicates that Isl1-dependent changes in chromatin accessibility occur ahead of transcriptional divergence.

Loss of Isl1 led to more robust opening than closing of chromatin regions as identified by *k*-means clustering (cluster 1 peaks depend on Isl1 for opening as they are closed upon Isl1 knock-out; cluster 2 peaks depend on Isl1 for closing) (Fig. 6b, Supplementary Fig. 11c), which suggests that Isl1 primarily maintain an open chromatin state. Annotation of opening and closing peaks by GREAT analysis [57] showed that 96.2% of the peaks (1,256) located in distal regions (> ±5 kb to the TSS sites) representing enhancer regions affecting cardiac progenitor cell decisions (Fig. 6c). In support of their enhancer occupies, the active mark H3K27ac and H3K4me1 for enhancers showed enriched distributions in both cluster 1 and 2 peaks comparing to repressive mark H3K27me3 or random peaks (Fig. 6d). Investigation of transcription factor motifs enriched in either opening or closing peaks using the motif analysis package HOMER [56] revealed that binding sites for GATA family factors and Tbx20 are enriched in both opening and closing peaks, suggesting an Isl1-independent mode of action. In contrast, binding sites for Mef2-family members were only enriched in Isl1-dependent opening peaks while binding sites for Forkhead box family TFs were enriched in Isl1-dependent closing peaks (Fig. 6e). We concluded that Isl1 acts together with Mef2 factors but prevents binding of Forkhead factors to guide cardiac progenitor cell fate decision. The enrichment of CTCF and CTCFL in open peaks implied that Isl1 alters the topology of the chromatin to achieve chromatin opening. Surprisingly, Isl1 binding motifs were not significantly enriched in either opening or closing peaks suggesting that Isl1 changes

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

chromatin accessibility not directly but by secondary effects, e.g. CTCF/CTCFL mediated-chromatin reorganization.

In contrast to the inactivation of Isl1, which led to a comparable number of opening and closing peaks, forced expression of Nkx2-5 resulted in more accessible chromatin, compared to either Nkx2-5$^+$ or Isl1$^+$ cardiac progenitor cells, and few closing peaks (opening peaks: 556 for Nkx2-5$^+$ and 1,526 for Isl1$^+$; closing peaks: 73 for Nkx2.5$^+$ and 86 for Isl1$^+$) (Supplementary Fig. 11c). Interestingly, forced expression of Nkx2-5 resulted in dramatic open chromatin state at E9.5 comparing to E12.5 and Nkx2-5$^+$ or Isl1$^+$ CPCs at E9.5 (Fig. 6f, g). We noted that such transient opening chromatin state was not stable and could not be sustained upon CPC differentiation, indicating a epigenetic plasticity (Fig. 6f). We speculated the opening chromatin state resulted from Nkx2-5 overexpression was reconfigured to closed state during cell differentiation. Similar to Isl1, most effects of forced Nkx2-5 expression were confined to distal regions (96.7% of 3,374 peaks) probably representing enhancers (Supplementary Fig. 6e). We performed motif analysis in these peaks and found most of the peaks corresponded to Sox family factor binding sites, although Nkx2-5 motif were not significantly enriched suggesting the altered chromatin accessibility upon Nkx2-5 overexpression is through secondary effects (Supplementary Fig. 11f).

Our analysis unveiled that Nkx2-5 and Isl1 mainly act at distal gene regions representing putative developmental enhancers to evoke dramatic effects on chromatin accessibility. Although more refined experiments are required to determine the impact of different TFs and individual enhances on cell fate decisions of CPC, our study provides a basic framework to understand the dynamic changes in Nkx2-5- and Isl1-driven chromatin accessibility that shape the early heart.

DISCUSSION

The long-lasting question in developmental biology is when and how the multipotent progenitor cells are restricted to final cell fates in a stepwise manner. Besides, it is still a challenge to capture the transition states in which time the fates of progenitor cells are

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

specified to their terminally differentiated descendants. In this study, we took an unprecedentedly detailed and closer glimpse into the transcription and epigenetic regulations of early heart development whereby we specifically focused on cardiac progenitor cells. Using single-cell RNA sequencing approach, we carried out deep sampling of more than 1,100 CPCs altogether at successive time points, revealed the heterogeneity of CPCs and reconstructed their developmental trajectories. Of note, we uncovered the bifurcation and transition state of Isl1$^+$ CPCs. Moreover, we performed both gain and loss of function study to demonstrate our conclusions. The analyses of big data pose significant challenges towards single-cell sequencing field [13]. Here, we applied a comprehensive analysis based on self-organizing maps, which mitigates the presence of dropouts by aggregating metagenes from highly similar individual genes in an unsupervised manner. Besides visualization of cellular transcriptomes, the lower dimensional representation of SOMs allowed us to employ advanced clustering strategies, like t-SNE and HDBSCAN, and uncover the presence of subpopulations irrespective of cell collection time-point.

Our scRNA-seq analysis provides a database of rich resource for novel gene discovery in the heart development study. For example, by comparing transcriptomes within cell subpopulations, we surprisingly found that homeobox genes *Hoxa7*, *Hoxa9*, *Hoxa10*, *Hoxb6*, *Hoxc8*, and *Hoxd8* are temporally expressed during early induction of the second heart field (Fig. 1). *Hox* genes are well known to establish anterior–posterior axis positioning during embryogenesis [58]. Interestingly, accumulating evidence has shown anterior *Hox* genes (*Hoxa1*, *Hoxb1*, and *Hoxa3*) are involved in cardiac development [39, 41, 42]. On the other hand, our analysis revealed posterior *Hox* genes are specifically expressed in cardiac progenitor cells as well. Taken together, we speculate that Hox family transcription factors contribute to patterning of heart. Gene disruption of some *Hox* genes identified by our analysis e.g. *Hoxb6* was not reported to show cardiac phenotypes, although there are other phenotypes such as hematopoietic and skeletal muscle defects [59]. We speculate that it is likely the heart development was not comprehensively studied in such models due to lacking their temporal expression profiles in early heart development. Another possibility is due to the

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

compensation effects, since as many as 39 murine *Hox* genes are of highly conserved and they can mutually compensate each other by overlapping protein expressions in a defined embryonic segment [60]. Consistently, deletion of either *HoxA* or *HoxB* gene cluster results in embryonic lethality but does not show cardiac defects, yet a compound deletion of *HoxA* and *HoxB* blocks cardiac morphogenesis before heart looping [43]. Signaling pathway retinoic acid regulates both heart development [36] and Hox expression [61]. Therefore, although there is no direct evidence, it implies that retinoic acid regulates Hox expression in cardiogenesis. *Hox* gene expression can be stimulated by histone H2B monoubiquitination in RNF20-dependent manner [62]. Interesting, *de novo* mutations of RNF20 have been identified in congenital heart disease (CHD) [63], which opens the possibility that *de novo* mutations of epigenetic-modifying enzyme RNF20 result in decreased level of H2B monoubiquitination which consequently influences *Hox* gene expression and eventually leads to abnormal heart development and CHD. Such candidate genes and hypothesis can be further tested in functional experiments in future.

Lineage tracing and clonal analysis shown Nkx2-5$^+$ cardiac progenitor cells contribute to cardiac endothelium and smooth muscle [50]. By reconstruction Nkx2-5$^+$ cell development, we found Nxk2-5 is not expressed outside of cardiomyocytes upon differentiation, which is consistent with previous studies that Nkx2-5 expresses principally in cardiomyocytes [23]. We reasoned that it is due to the differences of lineage tracing and reporter-based sorting approaches, since here we only focused on the cells that express Nkx2-5 at the sampling time point. In coincidence, our analysis revealed that once the progenitor cells differentiate to smooth muscle cells, Nkx2-5 expression is terminated rapidly. Furthermore, the forced expression of Nkx2-5 completely skewed bipotent fate of Isl1$^+$ cells to unipotent cardiomyocyte fate. Taken together, it suggests the Nkx2-5 guides cardiomyocyte differentiation as a master regulator at the hierarchical top of gene regulatory network. We believe the reassessment of the pivotal role of Nkx2-5 in cardiomyocyte benefits refinement of generation of stem-cell-derived cardiomyocyte which can be used to treat heart failure using cell-transplantation therapy. In addition, the cells in the cardiac crescent have started to express cardiomyocyte genes such as Myl7 [11, 64], suggesting they are

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

proceeding to cardiomyocyte at this stage, accompanying with increasing Nkx2-5 expression (Fig. 1f, 2e). In contrast to unidirectional cardiomyocyte fate maintained by Nkx2-5, our analysis showed that Isl1 essentially maintain the multipotency of progenitor cells, and once they differentiate to terminal cell fates, the expression of Isl1 decreases.

Although chromatin remodeling has been linked to heart development and BAF chromatin-remodeling complex is known to regulate heart development by permitting binding of cardiac transcription factors to cardiac genes [64, 65], the genome-wide accessible chromatin dynamics during early heart development was almost blank, while our ATAC-seq analysis in cardiac progenitor cells provides the first open chromatin atlas and new insight into understanding such important unsolved question. We characterized distinct chromatin accessibility modules at different developmental stages, suggesting unique open chromatin patterns are associated with cell fates and identities. It is worth noting that only ~1.5% of the mouse genome is surveyed by our ATAC-seq (data not shown), which theoretically covers most of the *cis*-regulatory elements. Thus, it enables our analysis as being highly informative and sensitive by merely focusing on active regulatory elements. Interestingly, we observed that the binding of CTCF and CTCFL in open peaks is altered upon Isl1 knock-out, which opens new avenues that chromatin dynamics during heart development is regulated by chromatin loops in addition to known Baf60c mediated remodeling. Since ATAC-seq peaks contain all *cis*-genomic elements such as enhancers, insulators as well as promoter, which leads to ambiguous correlations when we compare gene expression with chromatin open states. We therefore assigned and specifically selected the proximal peaks that positively correlate with gene expression, and revealed two gene categories which are regulated by open chromatin.

Of all the differential accessible peaks we identified, 55% are in distal regions, which enables us to analyze putative enhancer dynamic changes that are integration hubs of transcription factor occupancy [55]. By motif enrichment analysis approach, we identified several novel TFs such as POU3F1 and Nr2c2 that regulate heart development through putative enhancers. In addition, by loss/gain of function studies,

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

we demonstrate that two cardiac transcription factors Nkx2-5 and Isl1 execute their functions through synergetic TF networks on putative enhancers. Interestingly, although our scRNA-seq suggests *Hox* family genes are expressed in the early development stage, their binding motifs are identified in the late one. As the development and differentiation of cardiac progenitor cell occurs within a narrow window (E7.5-E9.5, 48 hours), it is likely that the Hox proteins can be sustained and play roles until CPC terminal differentiation. We compared the putative enhancers identified by our ATAC-seq with fetal heart enhancers identified based on distal P300 binding and H3K4me1 presence [66], and surprisingly found only limited chromatin regions overlap to each other (data not shown). We speculated that such differences are resulted from following reasons: (i) purified cardiac progenitor cells were used in this study, whereas fetal heart enhancer map was generated based on heterogonous heart cells; (ii) developmental enhancers undergo dramatic dynamics in different developmental stages. It is consistent with the latter speculation that we observed substantial different enrichment of many TF motifs at E8.5 and E9.5, including Gata4. Concordantly, ChIP-seq of GATA4 indicates that it binds to markedly different chromatin regions at different development stages [67]. Our ATAC-seq data in cardiac progenitor cells therefore conclude with a dynamic, development stage-specific transcription factor and enhancer interaction network that regulates cell fate decisions.

METHODS

Timed mating and sampling of single cells

All animal experiments were carried out following regulations of the Committee for Animal Rights Protection of the State of Hessen (Regierungspraesidium Darmstadt, Darmstadt, Germany). The transgenic mouse lines used in this study have been described previously [9, 10]. C57BL/6 mouse embryos were dissected at E7.5, E8.5, E9.5 or E12.5, and the embryonic hearts at E8.5-E12.5 were further dissected under the dissection microscope to remove other non-heart tissue. The samples were digested into single cells with 0.25% trypsin-EDTA after PBS wash, stained with DAPI for viability, and live GFP^{+} cells were purified by BD FACSAria II. To obtain *Isl1*$^{-/-}$ or Isl1^{+}/Nkx2-5OE cells, timed matings were set up between *Isl1*$^{nGFP/+}$ mice or *Isl1-Cre* and *Rosa26*$^{\text{Nxk2-5-}}$

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$^{IRES\text{-}GFP}$ mice, and genotyping PCR was performed using other non-heart tissue of the same embryos as previously described [10].

Single-cell RNA sequencing library preparation

Single-cell capture, lysis, reverse transcription, and pre-amplification were processed with C1 chips (#100-5763, 10-17 µm) in the C1 single-cell Auto Prep System (Fluidigm) or the ICELL8™ Single-Cell System (Wafergen) following the manufacturer's protocol. Libraries were sequenced on the Illumina NextSeq 500.

Single-cell RNA-seq data analysis

1. Raw data processing

Low quality bases were trimmed off the raw sequencing reads using Reaper with a minimum median quality of 53 in a window of 20 bases, omitting the first 50 bases of the read. Additionally, the -dust-suffix 20/AT option was used to trim remaining polyA or polyT stretches at the end of reads as well as stretches of B (a special Illumina Quality Score indicating non-thrustworthy bases) with the –bcq-late option.

The STAR alignment tool was used with default parameters to map trimmed reads to the mouse genome (version mm10) and transcriptome (--quantMode TranscriptomeSAM, together with the Gencode annotation in version vM10).

Mapping quality and statistics was assessed using Qualimap in rnaseq mode, setting the protocol to strand-specific-forward and using the same Gencode annotation. The Qualimap output was used later for single-cell filtering (see below).

RSEM was used with gene annotations from Gencode vM10 as well as a single-cell prior to assign reads to genes and extract gene-centered counts.

2. Cell quality and filtering

A SingleCellExpression-Set object (SCESet, R package scater) was created in R from all available metadata, cell quality data, gene annotations and the gene-centered count table. For each sequencing platform (Fluidigm C1 (C1) and Wafergen (WG)), an initial cell-quality map was generated with t-SNE (R package Rtsne), by grouping cells with similar quality metrics together (Supplementary Figure 1). The (per-cell) quality metrics

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

used as input were: number of features (genes) detected with at least 10 counts, the percentage of gene dropouts, the number of alignments, the number of alignments to exons, introns and intergenic regions, the number of secondary alignments, the expression of Rplp0 (also known as 36B4) as housekeeping gene, the percentage of read counts to mitochondrial genes, as well as the percentage of genes detected.
To define cells as low quality, we formulated and evaluated five criteria for each cell: The percentage of counts to mitochondrial genes is 1.5 median-absolute-deviations (MADs) above the median, the number of detected features is 2 MADs above or below the median, the percentage of gene dropouts is 2 MADs above the median, the Rplp0 expression is 2 MADs below the median and the percentage of genes is 1.5 MADs above or below the median. Cells failing more than one criterion were considered low quality and excluded from further analysis. See Supplementary Figure 1 for a graphical representation of the cell filtering.

3. Gene filtering and expression normalization
Similar to cell filtering, we defined two criteria for gene filtering: (1) A genes aggregated expression across all cells of a lineage (excluding cells from knockout and overexpression experiments) exceeds 2000 counts and (2) at least 10 cells from a lineage show a gene expression above 10 counts. A gene was filtered if it failed at least one criterion in both lineages. After filtering, count data of 12053 genes across 498 cells remained for further downstream analysis.

We normalized remaining count data by applying the sum factor method, as implemented in the R package scater, to cells from the two lineages separately.

4. Preprocessing of single-cell data from Li et. al [22].
We combined obtained count tables from wildtype cardiac single cells across time points E8.5, E9.5 and E10.5, as well as from Nkx2-5 knockout cells from E9.5 into a single SCESet object and filtered out cells that were marked as low-quality by the authors.

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

After filtering, count data from 11781 genes across 2358 cells was used to cluster cells using the quickCluster command from the R package scran and sum factor normalization was applied with deconvolution of size factors within obtained clusters.

5. Definition of heterogeneous genes

Sum factor normalized counts were used to define heterogeneous genes within lineages as well as within individual time points. Specifically, we calculated the coefficient of variation as well as the dropout-rate per gene and investigated their relationship to the mean expression of that gene. We next binned both (ordered) statistics into windows of size 200 and scaled values (zscore transformation) within windows. Genes for which one of the scaled statistics exceeded a 99-percentile within its window where called heterogeneous.

6. Clustering

We scaled normalized expression values of heterogeneous genes and used them as input to dimension reduction by self-organizing maps (SOMs) for each lineage, respectively. Briefly, SOMs or Kohonen Networks are a special case of neuronal networks, where no target vector containing class labels is necessary for training. Instead, a map is initialized randomly for each cell, consisting of fewer map tiles than input genes, effectively representing meta genes. During training, genes are subsequently placed onto map tiles with the most similar meta gene representation. Importantly, a gene ends up on the same map tile of all cell maps, therefore creating a lower dimensional representation of the cell's transcriptome using meta genes. After 2000 training epochs, cell maps were further projected into two dimensions by t-SNE (perplexity value of 15, 2000 epochs of convergence) and clustered with HDBSCAN using a minimum cluster size of 7 and min_samples 9 (Fig. 1 c, d and Supplementary Fig. 2 a, b).

7. Differential expression and lineage dynamics

We next assessed differentially expressed genes between cell clusters using MAST on sum factor normalized counts (log2 scale). Briefly, the MAST framework models gene

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

expression in a two-component generalized linear model, one component for the discrete expression rate of each gene across cells and the other component for the continuous expression level, given the gene is expressed. Additionally, we used a gene ranking approach (SC3) to define marker genes specific for each cluster (Supplementary Table 1, 2).

To define lineage dynamics, we used all protein coding genes that were marker genes for a cluster (AUROC > 0.8, FDR < 0.01) and differentially expressed in any cluster (lower bound of LFC > 2 or higher bound of LFC < -2, FDR < 0.01) as input to destiny (Figure 2 a, b).

8. Cell transition states and transcriptome noise

For the critical transition index ($I_C(c)$), we computed the absolute marker gene-to-gene and cell-to-cell correlations for each cluster and calculated the ratio of their means (Supplementary Figure 6 a, b). To reduce influence from differing cell numbers in clusters, we applied a bootstrapping procedure, randomly selecting 30 (20) cells from a given Nkx2-5 lineage cluster (Isl1 lineage cluster), repeating 1000 times.
Pairwise cell-to-cell distances were calculated as proposed by Mohammed and colleagues[49].

9. Gene correlation analysis

We next sought to define gene networks that play a role in lineage development. We reasoned that a genes expression will either increase or decrease with lineage progression and therefore calculated the (global) Spearman's Rank correlation of every genes expression to the diffusion pseudotime from destiny. Since it is also possible that a gene exhibits its expression dynamics only within discrete states (clusters), we also calculated the (local) Spearman's Rank correlation of gene expression to pseudotime within clusters. We defined a gene as correlated gene, if it shows a global correlation of at least 0.7 or a local correlation of at least 0.5 (Supplementary Table 3).

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We next used the lineage-specific correlated genes to identify gene networks. Genes within the same sub-network show a high correlation (measured as Pearson's Correlation), but a lower correlation between sub-networks (Supplementary Figure 5). To show the dynamics of correlated genes, we additionally smoothed their expression along pseudotime by calculating the mean expression in windows of 11 consecutive cells (Supplementary Fig. 7 c, d).

10. Integration with Wafergen data

To join data sets from the two different sequencing platforms, normalized expression values from heterogeneous genes were used as input into the mnnCorrect function from the R package scran. Briefly, mnnCorrect finds cells from different platforms that have mutually similar expression profiles. This is done by identification of pairs of cells that are mutual nearest neighbors, and can be interpreted as belonging to the same cell state. For each MNN pair, the method estimates a pair-specific correction vector. Those vectors are in turn averaged with nearby MNN pair vectors from the same hyperplane using a Gaussian-Kernel to obtain more stable cell-specific correction vectors. Importantly, this procedure also allows the correction of cells that are not part of any MNN pair, e.g. data set specific cells that were sampled only on one platform.

Corrected expression values were used for clustering and differential expression analysis analogous to steps 6 and 7 of this document (Supplementary Fig. 5).

11. Cell cycle scores of single cells

Cell cycle scores were calculated for each known cell cycle stage (G1/S, S, G2, G2/M, M/G1) using gene sets from Whitfield et. al. Specifically, a raw score was calculated as the average expression of genes in each set. To refine the score, we determined genes that correlated (rank correlation > 0.4) well with the raw score and calculated the cell cycle score using those genes. Cell cycle scores were z-score transformed (scaled) before plotting. A test of equal proportions was then conducted for cycling cells among $Isl1^{+}$ and $Isl1^{-/-}$ cells.

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ATAC-seq library preparation and sequencing

2, 000-20,000 of GFP^+ cardiac progenitor cells were FCAS-purified and used for ATAC-seq. The ATAC-seq libraries were prepared as previously described [29], and 2×50 paired-end sequencing was performed on Illumina NextSeq500 to achieve on average 36 million reads/sample.

ATAC-seq data analysis

1. Raw data processing

Raw ATAC-seq paired-end reads were trimmed and filtered for quality, and then aligned to the mouse genome GRCm38 (mm10) using STAR [68] with the following parameter: --outFilterMismatchNoverLmax 0.2 --outFilterScoreMinOverLread 0 --outFilterMatchNminOverLread 0 --outFilterMatchNmin 20 --alignIntronMin 2 --alignIntronMax 1 --outFilterMultimapNmax 1 --alignMatesGapMax 2000 --alignEndsProtrude 10 ConcordantPair. Reads that were unmapped, non-uniquely mapped, mapped to repetitive regions and chromosome M, and PCR duplicates were removed.

2. Normalization, peak calling and open chromatin atlas generation

For downstream analysis, the read counts were normalized to 1× depth (reads per genome coverage, RPGC) using “bamCoverage” function of deepTools2 [69]. Peak calling was performed using “callpeak” function of MACS2 [70] with following parameter: --nomodel --shift -100 --extsize 200 -q 0.05. Peaks in each sample were merged as union peaks for calculation of peak counts and quality control purpose. The normalized number of reads mapped to each peak of the union peaks in each sample was quantified using bigWigAverageOverBed (https://github.com/ENCODE-DCC/kentUtils). Peak counts of all samples were then merged to obtain a data matrix and normalize with edgeR [52]. To remove the not-reproducible replicates, we calculated pearson correlation using the log2 normalized counts, and removed the pairs whose pearson correlation is below 0.8, resulting in at least two replicates for each developmental stage. Differential accessible peaks were pairwise-compared sequentially across each

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

developmental stage using reproducible samples, and combined as a genome-wide atlas of accessible chromatin landscape in cardiac progenitor cells.

3. Genome-wide distribution of differential peaks

The normalized read counts in each developmental stage across replicates were merged, binned around all differential peak summit in 50 bp bins spanning ±1 kb region, clustered by *k*-means algorithm and visualized by heat map using deepTools2 [69].

4. Assignment of proximal and distal ATAC-seq peaks

The proximal and distal peaks are defined by the distance of differential ATAC-seq peaks towards annotated promoters (Gencode annotation): at least 2.5 kb away from promoters were selected as distal peaks, and the others were assigned as proximal peaks.

Bulk RNA-seq

5,000-20,000 of cardiac progenitor cells were sampled using the same protocol as described above for scRNA-seq. The bulk RNA-seq libraries were prepared using Smart-seq2 according to the manufacturer's protocol (#634889, Clontech), and sequenced on Illumina NextSeq500. The raw reads were processed using the same method for scRNA-seq as described above. The quantification and identification of differentially expressed genes were carried out using DEseq2 [71].

Analysis of ChIP-seq data

The ChIP-seq data of histone modification of mouse E10.5 embryonic heart was downloaded from ENCODE project (GEO: GSE86753, GSE86693, GSE86723, GSE86752). Metagene analysis was performed using HOMER [56] and random peaks of nearly same size were used as control.

Data and computation code availability

GEO Accession XXX

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

AUTHOR CONTRIBUTIONS

G.J. and T.B. designed and conceived the project. J.P. and G.J. analyzed the scRNA-seq data. G.J. analyzed the ATAC-seq data. G.J. performed embryo dissections and collections. G.J., S.G. and M.Y. performed scRNA-seq and ATAC-seq. X.Y. provided transgenic mouse lines. S.G., C.K., M.L. and Y.Z. contributed to data processing, discussions and advice. G.J., J.P. and T.B. contributed with interpretation of data and writing the manuscript. All authors have contributed with helpful discussions of the manuscript.

Acknowledgements

We thank A. Atzberger for FACS assistance and H. Qi for embryo dissection assistance. This work was supported by XXX.

REFERENCES

1. Meilhac, S.M., et al., *Cardiac cell lineages that form the heart.* Cold Spring Harb Perspect Med, 2014. **4**(9): p. a013888.
2. Moretti, A., et al., *Multipotent embryonic isl1+ progenitor cells lead to cardiac, smooth muscle, and endothelial cell diversification.* Cell, 2006. **127**(6): p. 1151-65.
3. Wu, S.M., et al., *Developmental origin of a bipotential myocardial and smooth muscle cell precursor in the mammalian heart.* Cell, 2006. **127**(6): p. 1137-50.
4. Harvey, R.P., *Patterning the vertebrate heart.* Nat Rev Genet, 2002. **3**(7): p. 544-56.
5. Meilhac, S.M., et al., *The clonal origin of myocardial cells in different regions of the embryonic mouse heart.* Dev Cell, 2004. **6**(5): p. 685-98.
6. Buckingham, M., S. Meilhac, and S. Zaffran, *Building the mammalian heart from two sources of myocardial cells.* Nat Rev Genet, 2005. **6**(11): p. 826-35.
7. Prall, O.W., et al., *An Nkx2-5/Bmp2/Smad1 negative feedback loop controls heart progenitor specification and proliferation.* Cell, 2007. **128**(5): p. 947-59.
8. Laugwitz, K.L., et al., *Islet1 cardiovascular progenitors: a single source for heart lineages?* Development, 2008. **135**(2): p. 193-205.
9. Hsiao, E.C., et al., *Marking embryonic stem cells with a 2A self-cleaving peptide: a NKX2-5 emerald GFP BAC reporter.* PLoS One, 2008. **3**(7): p. e2532.
10. Yuan, X., et al., *Disruption of spatiotemporal hypoxic signaling causes congenital heart disease in mice.* J Clin Invest, 2017. **127**(6): p. 2235-2248.
11. Cai, C.L., et al., *Isl1 identifies a cardiac progenitor population that proliferates prior to differentiation and contributes a majority of cells to the heart.* Dev Cell, 2003. **5**(6): p. 877-89.
12. Zhang, L., et al., *Mesodermal Nkx2.5 is necessary and sufficient for early second heart field development.* Dev Biol, 2014. **390**(1): p. 68-79.

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

13. Shapiro, E., T. Biezuner, and S. Linnarsson, *Single-cell sequencing-based technologies will revolutionize whole-organism science.* Nat Rev Genet, 2013. **14**(9): p. 618-30.
14. Brennecke, P., et al., *Accounting for technical noise in single-cell RNA-seq experiments.* Nat Methods, 2013. **10**(11): p. 1093-5.
15. Tallulah S. Andrews, M.H., *Modelling dropouts for feature selection in scRNASeq experiments.* bioRxiv, 2017. **doi:** https://doi.org/10.1101/065094.
16. Treutlein, B., et al., *Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq.* Nature, 2014. **509**(7500): p. 371-5.
17. Scialdone, A., et al., *Resolving early mesoderm diversification through single-cell expression profiling.* Nature, 2016. **535**(7611): p. 289-293.
18. Pierson, E. and C. Yau, *ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis.* Genome Biol, 2015. **16**: p. 241.
19. Segerstolpe, A., et al., *Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes.* Cell Metab, 2016. **24**(4): p. 593-607.
20. Fan, J., et al., *Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis.* Nat Methods, 2016. **13**(3): p. 241-4.
21. Satija, R., et al., *Spatial reconstruction of single-cell gene expression data.* Nat Biotechnol, 2015. **33**(5): p. 495-502.
22. Li, G., et al., *Transcriptomic Profiling Maps Anatomically Patterned Subpopulations among Single Embryonic Cardiac Cells.* Dev Cell, 2016. **39**(4): p. 491-507.
23. DeLaughter, D.M., et al., *Single-Cell Resolution of Temporal Gene Expression during Heart Development.* Dev Cell, 2016. **39**(4): p. 480-490.
24. Magwene, P.M., P. Lizardi, and J. Kim, *Reconstructing the temporal ordering of biological samples using microarray data.* Bioinformatics, 2003. **19**(7): p. 842-50.
25. Trapnell, C., et al., *The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells.* Nat Biotechnol, 2014. **32**(4): p. 381-386.
26. Haghverdi, L., et al., *Diffusion pseudotime robustly reconstructs lineage branching.* Nat Methods, 2016. **13**(10): p. 845-8.
27. Bargaje, R., et al., *Cell population structure prior to bifurcation predicts efficiency of directed differentiation in human induced pluripotent cells.* Proc Natl Acad Sci U S A, 2017. **114**(9): p. 2271-2276.
28. Thurman, R.E., et al., *The accessible chromatin landscape of the human genome.* Nature, 2012. **489**(7414): p. 75-82.
29. Buenrostro, J.D., et al., *Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position.* Nat Methods, 2013. **10**(12): p. 1213-8.
30. Martin-Puig, S., Z. Wang, and K.R. Chien, *Lives of a heart cell: tracing the origins of cardiac progenitors.* Cell Stem Cell, 2008. **2**(4): p. 320-31.
31. Van Vliet, P., et al., *Early cardiac development: a view from stem cells to embryos.* Cardiovasc Res, 2012. **96**(3): p. 352-62.
32. Kim, D.H., et al., *Single-cell transcriptome analysis reveals dynamic changes in lncRNA expression during reprogramming.* Cell Stem Cell, 2015. **16**(1): p. 88-101.

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

33. Finak, G., et al., *MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data.* Genome Biol, 2015. **16**: p. 278.
34. Kiselev, V.Y., et al., *SC3: consensus clustering of single-cell RNA-seq data.* Nat Methods, 2017. **14**(5): p. 483-486.
35. Masino, A.M., et al., *Transcriptional regulation of cardiac progenitor cell populations.* Circ Res, 2004. **95**(4): p. 389-97.
36. Bruneau, B.G., *Signaling and transcriptional networks in heart development and regeneration.* Cold Spring Harb Perspect Biol, 2013. **5**(3): p. a008292.
37. Yang, J., et al., *RBM24 is a major regulator of muscle-specific alternative splicing.* Dev Cell, 2014. **31**(1): p. 87-99.
38. Bertrand, N., et al., *Hox genes define distinct progenitor sub-domains within the second heart field.* Dev Biol, 2011. **353**(2): p. 266-74.
39. Chisaka, O. and Y. Kameda, *Hoxa3 regulates the proliferation and differentiation of the third pharyngeal arch mesenchyme in mice.* Cell Tissue Res, 2005. **320**(1): p. 77-89.
40. Kameda, Y., et al., *Disruption of the Hoxa3 homeobox gene results in anomalies of the carotid artery system and the arterial baroreceptors.* Cell Tissue Res, 2003. **311**(3): p. 343-52.
41. Makki, N. and M.R. Capecchi, *Cardiovascular defects in a mouse model of HOXA1 syndrome.* Hum Mol Genet, 2012. **21**(1): p. 26-31.
42. Roux, M., et al., *Hoxb1 regulates proliferation and differentiation of second heart field progenitors in pharyngeal mesoderm and genetically interacts with Hoxa1 during cardiac outflow tract development.* Dev Biol, 2015. **406**(2): p. 247-58.
43. Soshnikova, N., et al., *Duplications of hox gene clusters and the emergence of vertebrates.* Dev Biol, 2013. **378**(2): p. 194-9.
44. Dorn, T., et al., *Direct nkx2-5 transcriptional repression of isl1 controls cardiomyocyte subtype identity.* Stem Cells, 2015. **33**(4): p. 1113-29.
45. Watanabe, Y., et al., *Fibroblast growth factor 10 gene regulation in the second heart field by Tbx1, Nkx2-5, and Islet1 reveals a genetic switch for down-regulation in the myocardium.* Proc Natl Acad Sci U S A, 2012. **109**(45): p. 18273-80.
46. Laleh Haghverdi, A.T.L.L., Michael D. Morgan, John C. Marioni, *Correcting batch effects in single-cell RNA sequencing data by matching mutual nearest neighbours.* bioRxiv, 2017. **doi:** https://doi.org/10.1101/165118.
47. Huang, S., *The molecular and mathematical basis of Waddington's epigenetic landscape: A framework for post-Darwinian biology?* BioEssays, 2012. **34**(2): p. 149-157.
48. Mojtahedi, M., et al., *Cell Fate Decision as High-Dimensional Critical State Transition.* PLoS Biol, 2016. **14**(12): p. e2000640.
49. Mohammed, H., et al., *Single-Cell Landscape of Transcriptional Heterogeneity and Cell Fate Decisions during Mouse Early Gastrulation.* Cell Reports, 2017. **20**(5): p. 1215-1228.
50. Ma, Q., B. Zhou, and W.T. Pu, *Reassessment of Isl1 and Nkx2-5 cardiac fate maps using a Gata4-based reporter of Cre activity.* Dev Biol, 2008. **323**(1): p. 98-104.
51. Picelli, S., et al., *Full-length RNA-seq from single cells using Smart-seq2.* Nat Protoc, 2014. **9**(1): p. 171-81.

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

52. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.
53. Cohen, E.D., et al., *Wnt/beta-catenin signaling promotes expansion of Isl-1-positive cardiac progenitor cells through regulation of FGF signaling.* J Clin Invest, 2007. **117**(7): p. 1794-804.
54. Lalit, P.A., et al., *Lineage Reprogramming of Fibroblasts into Proliferative Induced Cardiac Progenitor Cells by Defined Factors.* Cell Stem Cell, 2016. **18**(3): p. 354-67.
55. Buecker, C. and J. Wysocka, *Enhancers as information integration hubs in development: lessons from genomics.* Trends Genet, 2012. **28**(6): p. 276-84.
56. Heinz, S., et al., *Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.* Mol Cell, 2010. **38**(4): p. 576-89.
57. McLean, C.Y., et al., *GREAT improves functional interpretation of cis-regulatory regions.* Nat Biotechnol, 2010. **28**(5): p. 495-501.
58. Deschamps, J. and J. van Nes, *Developmental regulation of the Hox genes during axial morphogenesis in the mouse.* Development, 2005. **132**(13): p. 2931-42.
59. Kappen, C., *Disruption of the homeobox gene Hoxb-6 in mice results in increased numbers of early erythrocyte progenitors.* Am J Hematol, 2000. **65**(2): p. 111-8.
60. Duboule, D. and P. Dolle, *The structural and functional organization of the murine HOX gene family resembles that of Drosophila homeotic genes.* EMBO J, 1989. **8**(5): p. 1497-505.
61. Alexander, T., C. Nolte, and R. Krumlauf, *Hox genes and segmentation of the hindbrain and axial skeleton.* Annu Rev Cell Dev Biol, 2009. **25**: p. 431-56.
62. Zhu, B., et al., *Monoubiquitination of human histone H2B: the factors involved and their roles in HOX gene regulation.* Mol Cell, 2005. **20**(4): p. 601-11.
63. Zaidi, S., et al., *De novo mutations in histone-modifying genes in congenital heart disease.* Nature, 2013. **498**(7453): p. 220-3.
64. Takeuchi, J.K. and B.G. Bruneau, *Directed transdifferentiation of mouse mesoderm to heart tissue by defined factors.* Nature, 2009. **459**(7247): p. 708-11.
65. Bruneau, B.G., *Chromatin remodeling in heart development.* Curr Opin Genet Dev, 2010. **20**(5): p. 505-11.
66. Shen, Y., et al., *A map of the cis-regulatory sequences in the mouse genome.* Nature, 2012. **488**(7409): p. 116-20.
67. He, A., et al., *Dynamic GATA4 enhancers shape the chromatin landscape central to heart development and disease.* Nat Commun, 2014. **5**: p. 4907.
68. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
69. Ramirez, F., et al., *deepTools2: a next generation web server for deep-sequencing data analysis.* Nucleic Acids Res, 2016. **44**(W1): p. W160-5.
70. Zhang, Y., et al., *Model-based analysis of ChIP-Seq (MACS).* Genome Biol, 2008. **9**(9): p. R137.
71. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure Legend

Figure 1. Single-cell RNA-seq of cardiac progenitor cells and identification of heterogeneous subpopulations and maker genes. (**a**) Schematic representation of Nkx2-5-emGFP transgenic reporter and Isl1$^{nGFP/+}$ allele (top). Expression patterns of Nkx2-5-emGFP and Isl1-nGFP at E8.5 in mouse embryonic hearts. (bottom). (**b**) Sampling time points for scRNA-seq, bulk RNA-seq, ATAC-seq. The table shows cell numbers of scRNA-seq. QC: quality control. (**c**) t-SNE visualization of Nkx2-5$^+$ and Isl1$^+$ cardiac progenitor cells reveal subpopulations. Colors denote corresponding clusters, and (**d**) development stages. (**e**) Hierarchical clustering of expression heatmap showing differentially expressed marker genes (AUROC > 0.8, FDR < 0.01; and lower bound of LogFC > 2 or higher bound of LogFC < -2, FDR < 0.01) across different clusters in Nkx2-5 (top) and Isl1 (bottom). (**f**) Expression of individual genes in Nkx2-5$^+$ and Isl1$^+$ (**g**) cells. The color represents expression level of cells that are plotted in t-SNE as (**c**). (Scale bars: 300 μm).

Figure 2. Reconstruction of developmental trajectories of cardiac progenitor cells and computation of transition states. (**a**) t-SNE plots show diffusion pseudotime of Nkx2-5$^+$ and (**b**) Isl1$^+$ CPCs. The clusters and development stages of individual cells are colored as the label. (**c**) Boxplots represent the distribution of $I_C(C)$ values from all marker genes of each cluster of Nkx2-5$^+$ (left) and Isl1$^+$ (right) cells. P-value for comparison between clusters (Kolmogorov–Smirnov test and Wilcoxon rank sum test). (**d**) Boxplots showing pairwise cell-to-cell distances across each cluster of Nkx2-5$^+$ (left) and Isl1$^+$ (right) cells. (**e** and **f**) Expression levels different transcription factors and key genes along pseudotime in Nkx2-5$^+$ (**e**) and Isl1$^+$ (**f**) cells.

Figure 3. scRNA-seq of cardiac progenitor cells upon Isl1 knock-out. (**a**) Schematic illustration showing generation of Isl1 KO embryos and scRNA-seq. (**b**) Predicted diffusion pseudotime of Isl1 KO CPCs is plotted to t-SNE plot of Isl1$^+$ cells. (**c**) The proportion of cycling and non-cycling cells of Isl1 KO and Isl1$^+$ wild type CPCs. (**d**)

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Heatmap showing expression of deregulated genes compared among Isl1 KO, Isl1$^+$ E8.5 and E9.5 (cluster 1, 2 and 5) wild type cells.

Figure 4. scRNA-seq of Isl1$^+$/Nkx2-5OE cells. (**a**) Plot of re-analyzed published data showing the proportion of smooth muscle cell in unbiased-sampled embryonic heart at E9.5 in wild type and Nkx2-5 KO. The smooth muscle cells are "scored" by evaluating the expression of *Nkx2-5* (LogTPM < 1, null expression), while the expression of the other genes *Tagln, Cnn1, Acta2, Cald1, Mylk, Hexim1* and *Smtnl2* are moderate to high (LogTPM >2) for at least 5 of these 7 genes. χ^2 test: p < 2.37e-6. (**b**) Schematic illustration of forced expression of Nkx2-5 in Isl1$^+$ cells and scRNA-seq. (**c**) t-SNE plots showing the prediction of diffusion pseudotime of Isl1$^+$/Nkx2-5OE cells in Nkx2-5$^+$ and Isl1$^+$ (**d**) CPCs.

Figure 5. Accessible chromatin dynamics and gene expression of cardiac progenitor cells. (**a**) Number of differential chromatin accessibility peaks in each transition state (log2(FC)>2, false discovery rate [FDR]<0.05). (**b**) Genome-wide distribution of differential open chromatin peaks grouped first by distance to the promoters and then by *K*-means. Each row represents one of 5,866 selected peaks in sequential comparisons (log2[FC]>2, FDR<0.05). (**c**) Metagene plot of proximal ATAC-seq peaks. (**d**) Plot showing accessible chromatin levels and gene expression levels of 805 selected differential-peak/DE-gene pairs. (**e**) Bubble chart showing enrichment of gene ontology terms of selected genes. (**f**) ATAC-seq signal profiles across Wnt8a and Wnt11 loci in Isl1$^+$ and Nkx2-5$^+$ CPCs, respectively. (**g**) Boxplots showing gene expression of Wnt8a and Wnt11 in Isl1$^+$ and Nkx2-5$^+$ CPCs. (**h**) Plot showing ChIP-seq intensity of histone modifications at E10.5 in embryonic heart across distal ATAC-seq peaks. (**i**) Heatmap showing enrichment of transcription factor motifs in each accessible chromatin cluster of Isl1$^+$ and Nkx2-5$^+$ CPCs across E7.5 to E9.5.

Figure 6. Putative enhancer dynamics in Isl1 KO and Isl1$^+$/Nkx2-5OE cardiac progenitor cells. (**a**) Number of differential chromatin accessibility peaks in Isl KO cells comparing to wild type Isl1$^+$ cells at E8.5 and E9.5 (log2[FC]>2, FDR<0.05). (**b**)

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Heatmap showing genome-wide distribution of differential open chromatin peaks grouped by *K*-means in Isl1 KO and Isl1 WT CPCs. Each row represents one of differential peaks in pairwise comparison (log2[FC]>2, FDR<0.05). (**c**) Bar plot showing the distance of differential peaks to their nearest promoters. (**d**) Plot showing ChIP-seq intensity of histone modifications at E10.5 in embryonic heart across cluster 1 (left) and 2 (right) ATAC-seq peaks. (**e**) Bubble chart showing the enrichment of transcription factor motifs in each cluster of differential peaks. The color and area of circles represent the *p*-value. (**f**) Venn diagram showing overlap of genome coverage of open chromatin peaks between E9.5 and E12.5 Isl1^{+}/Nkx2-5OE cell. (**g**) Heatmap showing genome-wide distribution of differential open chromatin peaks grouped by *K*-means in Isl1^{+}/Nkx2-5OE at E9.5 and E12.5 comparing with Nkx2-5^{+} and Isl1^{+}.

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Figure Legend

Supplementary Figure 1. Cell quality filtering procedure. **(a-e)** Distributions of relevant quality criteria for single cells from the C1 platform (left panel) and Wafergen platform (right panel). Shown are (in order) the number of detected features (genes, log2 scale) per cell, gene dropout rates per cell, percentage of counts to mitochondrial features, percentage of reads aligned to genes, Rplp0 counts (expression, log10 scale). A t-SNE projection based on quality criteria aggregates cells with similar number of detected features in close proximity **(f)** and can be used to identify low quality cells (marked in red) **(g)**.

Supplementary Figure 2. Single cell SOM mapping and clustering. (**a**) Exemplification of SOM mapping and clustering for Nkx2-5$^+$ data. For each cell, an individual SOM (right panel) is created, effectively summarizing the cells transcriptome in reduced dimensionality. Reduced individual cell representations are then clustered and visualized using t-SNE (Figure 1c,d). Aggregated signals from individual SOMs in clusters resemble key characteristics from individual SOMs (left panel). (**b**) Key characteristics of aggregated and individual SOMs for Isl1$^+$ data.

Supplementary Figure 3. Expression of posterior (**a**) and anterior (**b**) *Hox* genes in Isl1$^+$ CPCs.

Supplementary Figure 4. Gene Ontology enrichment analyses of heterogenours clusters of Isl1$^+$ (**a**) and Nkx2-5$^+$ (**b**) CPCs. Gene enrichment was performed using the R package clusterProfiler.

Supplementary Figure 5. Single cell alignment and batch correction between two sequencing platforms for Nkx2.5$^+$ data. Principle component analysis (PCA) of **(a)** unaligned single cells shows a batch effect between sequencing platforms (right), which is diminished after **(b)** alignment of single cells using MnnCorrect. **(c)** t-SNE/HDBSCAN clustering of aligned single cells from two sequencing platforms identifies three clusters

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

that aggregate by **(d)** capture time point, not sequencing platform. The clustering resembles the clustering using data from the C1 platform alone (Figure 1c and 1d, left). **(e)** Expression of the top 30 differentially expressed genes between clusters from Figure 1c, shown for cells from both sequencing platforms. Cells in cluster 2 that have been processed on the Wafergen system fail to show a consistent expression signature for cluster 2, as compared to cells that have been processed on the C1 system. **(f)** Those cells inevitably influence the differential expression analysis using cells from both sequencing platforms, as there are no differentially expressed genes for cluster 2 from (**c**).

Supplementary Figure 6. Gene-to-gene and cell-to-cell correlations reveal different characteristics of heart development. **(a** and **b)** Pearson coefficients of pairwise gene-gene and cell-cell correlations for clusters from Nkx2.5$^+$ (**a**) and Isl1$^+$ (**b**) data.

Supplementary Figure 7. Genes correlated with developmental pseudotime. **(a** and **b)** Gene-gene correlation matrices for genes playing a role in Nkx2.5$^+$ (**a**) and Isl1$^+$ (**b**) developmental trajectories. **(c** and **d)** Heatmap showing the expression of genes positively correlated with pseudotime (Spearman rank correlation coefficient > 0.5) in Nkx2-5$^+$ cells (**c**) and Isl1$^+$ cells (**d**). Two bifurcation trajectories of Isl1$^+$ cells are shown separately (**d**).

Supplementary Figure 8. Determination of cycling Isl1$^+$ and Isl1$^{-/-}$ cells. (**a**) Cells ordered by the maximum of their G1/S and G2/M cell cycle score. Cells with a maximum score lower than -0.1 where considered non-cycling and marked in red. The cutoff was determined empirically. (**b**) Scatterplot showing G1/S cell cycle score vs. G2/M cell cycle score. Cells identified as cycling in (**a**) accumulate in the first quadrant, indicating their cell cycle status. (**c**) Gene-to-gene correlation of G1/S and G2/M phase genes shows distinct cell cycle signatures for cycling cells identified in (**a**).

Supplementary Figure 9. Expression of individual smooth muscle cell markers in Nkx2-5$^+$ CPCs.

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Figure 10. Reproducibility of biological replicates of ATAC-seq and bulk RNA-seq. Scatter plots show Spearman correlation coefficient of ATAC-seq samples (**a**) and bulk RNA-seq samples (**b**).

Supplementary Figure 11. (**a, b**) Boxplots showing the accessible chromatin level of each clusters. P-value for comparison between clusters are based on Kolmogorov–Smirnov test and Wilcoxon rank sum test. (**c**) Volcano plots showing differential chromatin accessible peaks of Isl1$^+$/Nkx2-5OE comparing with Nkx2-5$^+$ (left) and Isl1$^+$ (right). Each dot represents one differential peak and the colored ones were filtered as log2[FC]>2 or < -2, FDR<0.05. (**d**) Boxplots showing the accessible chromatin level of Isl1$^+$/Nkx2-5OE CPCs at E9.5 and 12.5. (**e**) Bar plot showing the distance of Isl1$^+$/Nkx2-5OE differential peaks to their nearest promoters. (**f**) Enrichment of known transcription factor motifs in Isl1$^+$/Nkx2-5OE differential peaks. The height of logo represents frequency of each nucleobase in the cognate motif.

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Tables

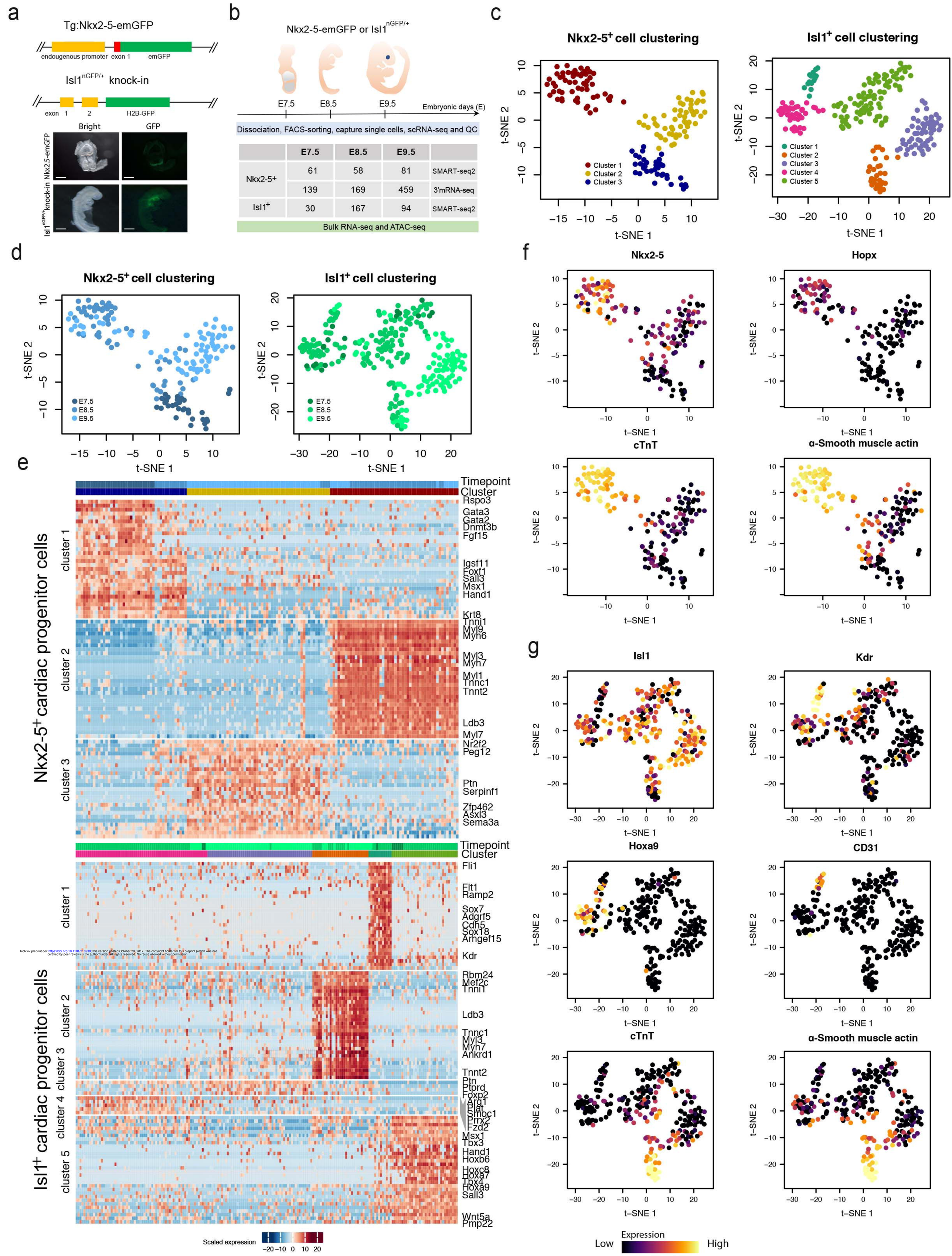
Supplementary Table 1. Top genes that were differentially expressed across each cluster in the Nkx2-5$^+$ CPCs.

Supplementary Table 2. Top genes that were differentially expressed across each cluster in the Isl1$^+$ CPCs.

Supplementary Table 3. List of genes for Nkx2-5$^+$ and Isl1$^+$ cells that are positively correlated with developmental pseudotime.

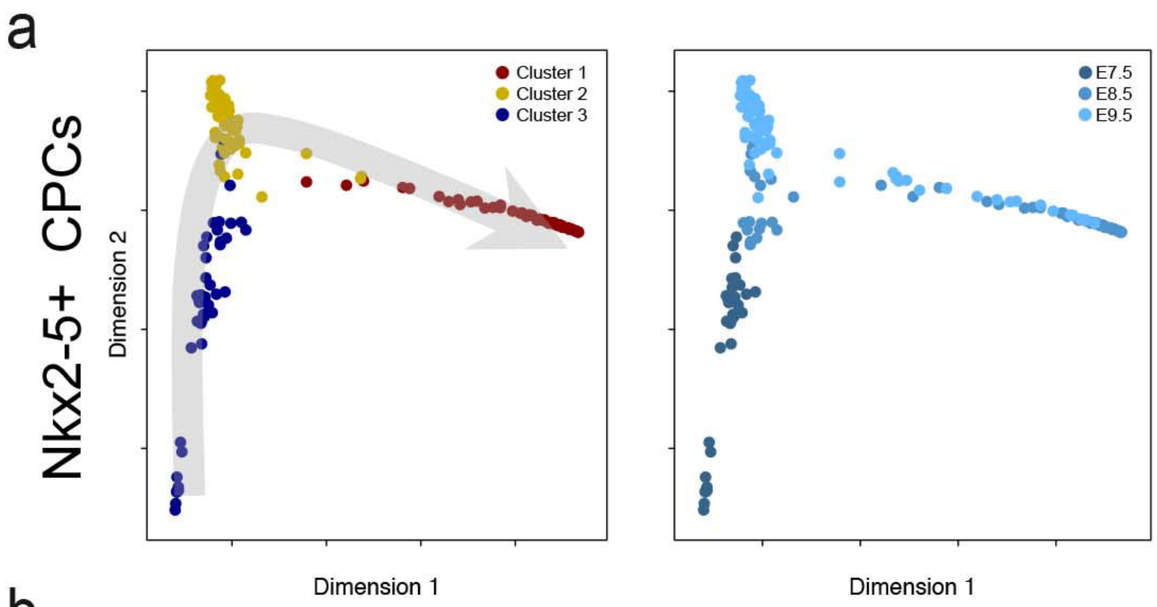
Supplementary Table 4. Number of raw and processed reads of ATAC-seq.

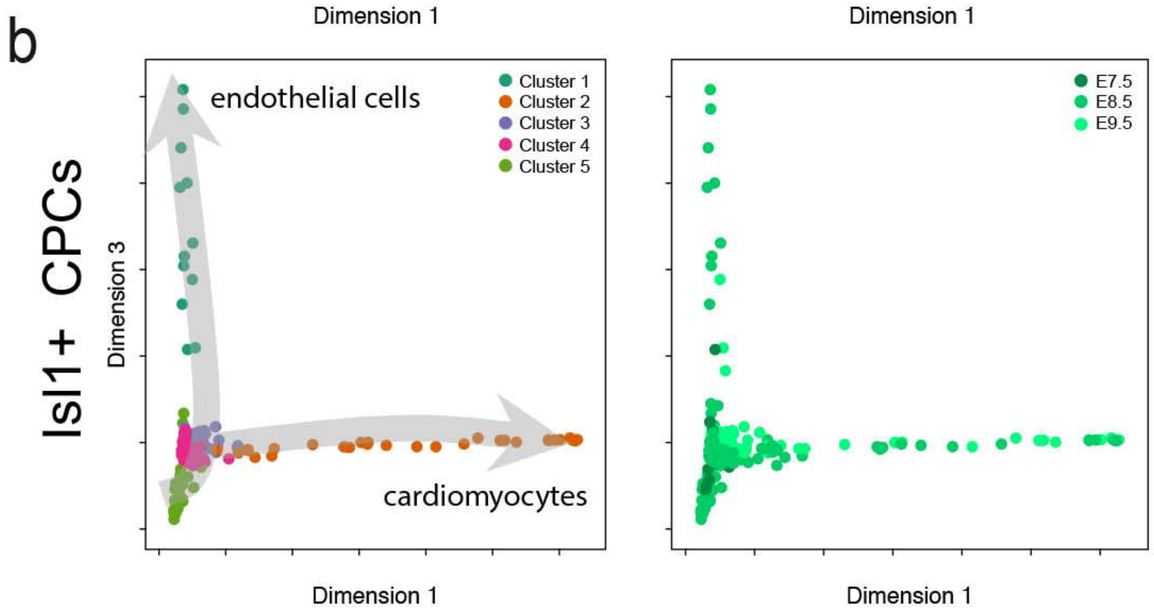
Jia, Preussner, et al Figure 1

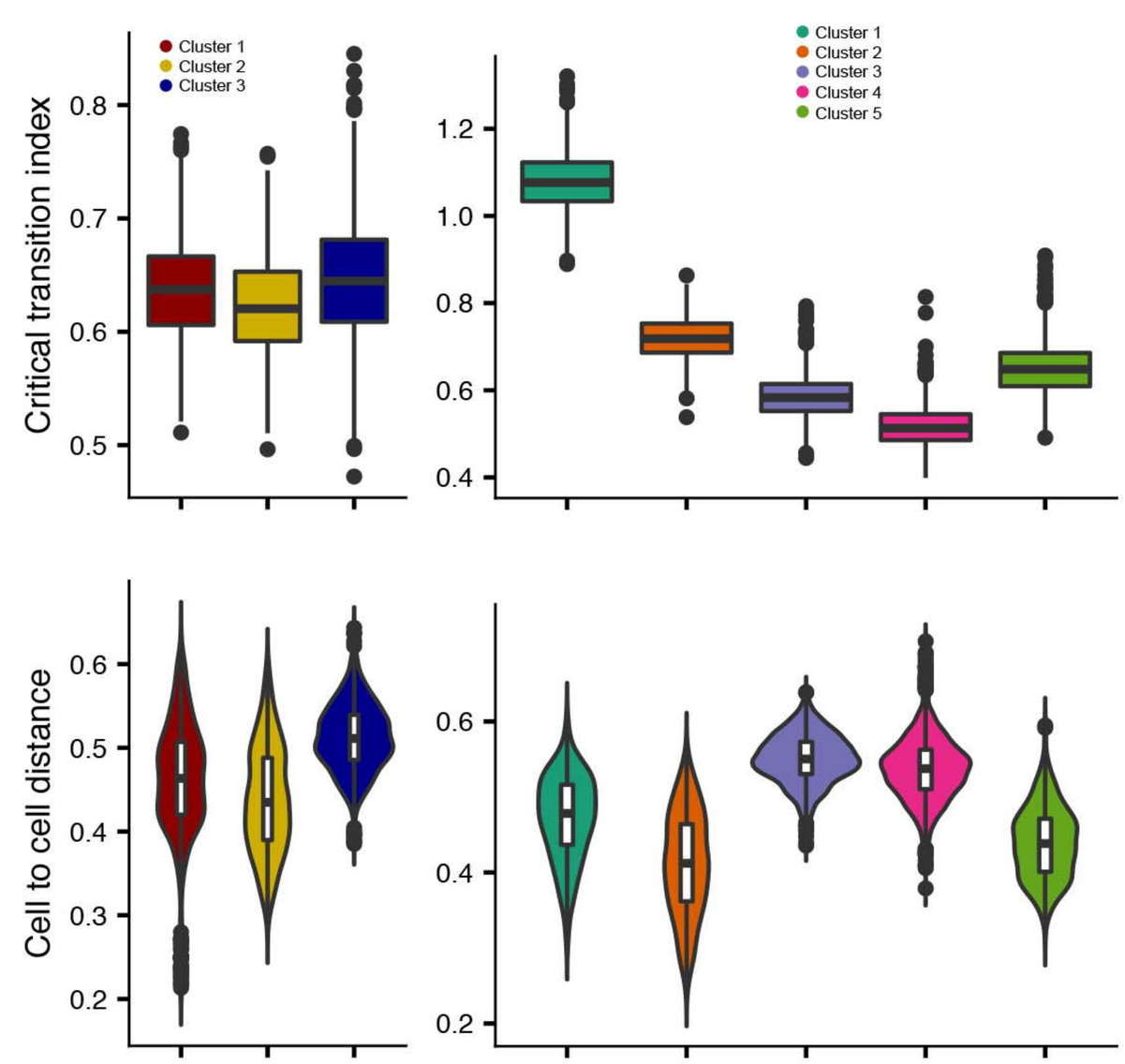


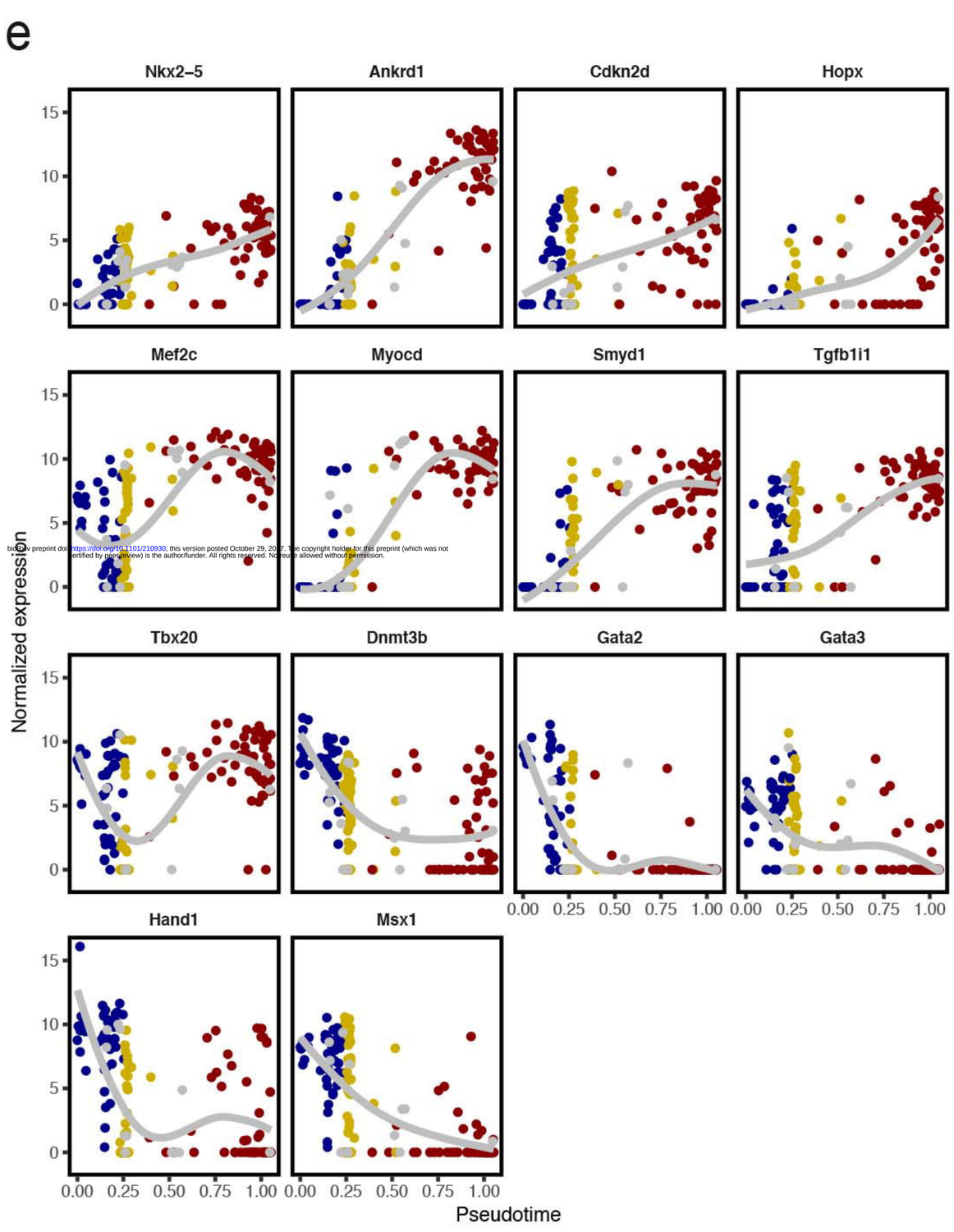
bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

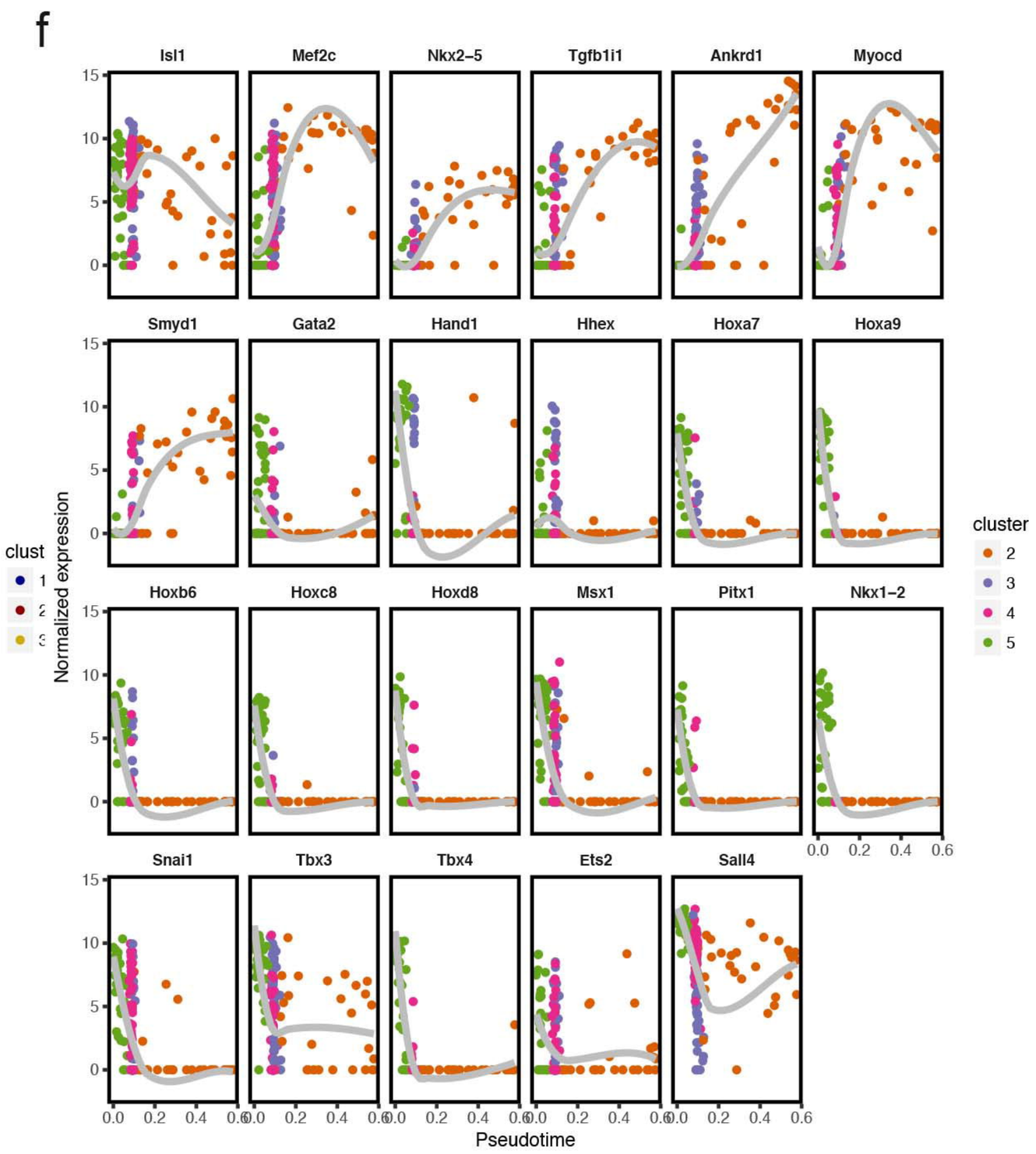
Jia, Preussner, et al Figure 2







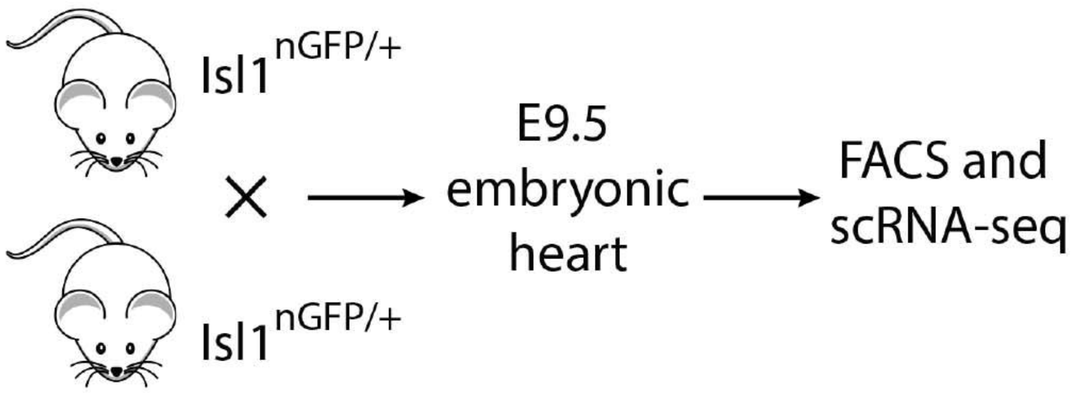




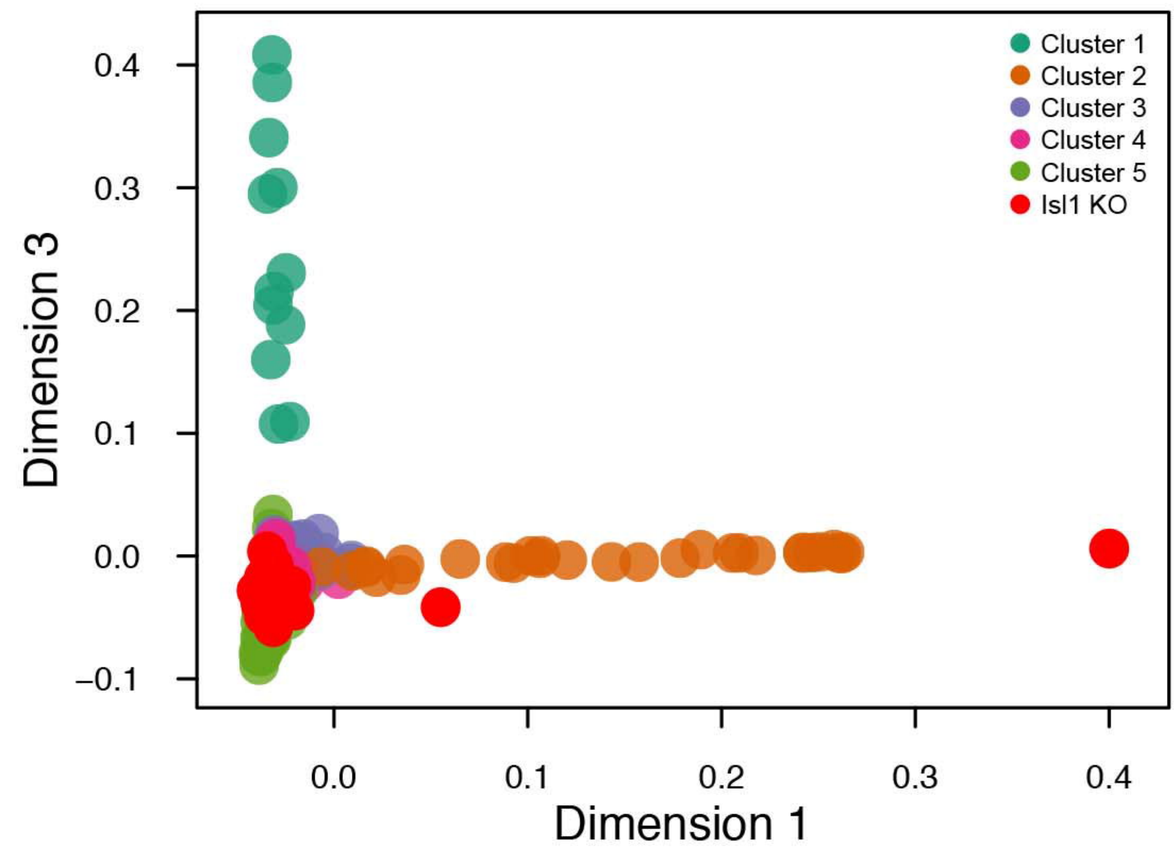
biorxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Jia, Preussner, et al Figure 3

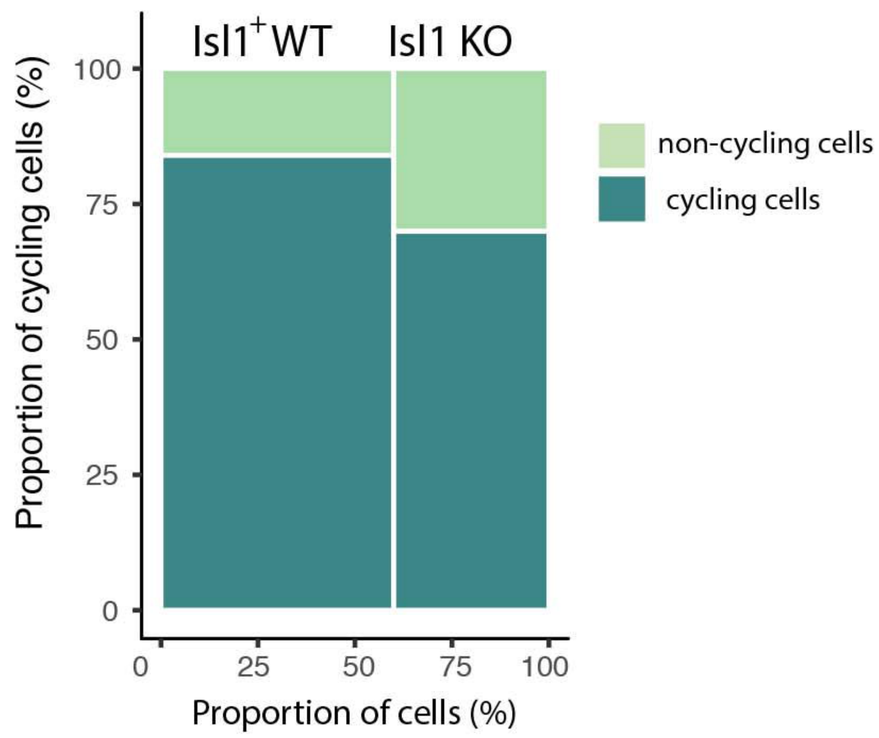
a



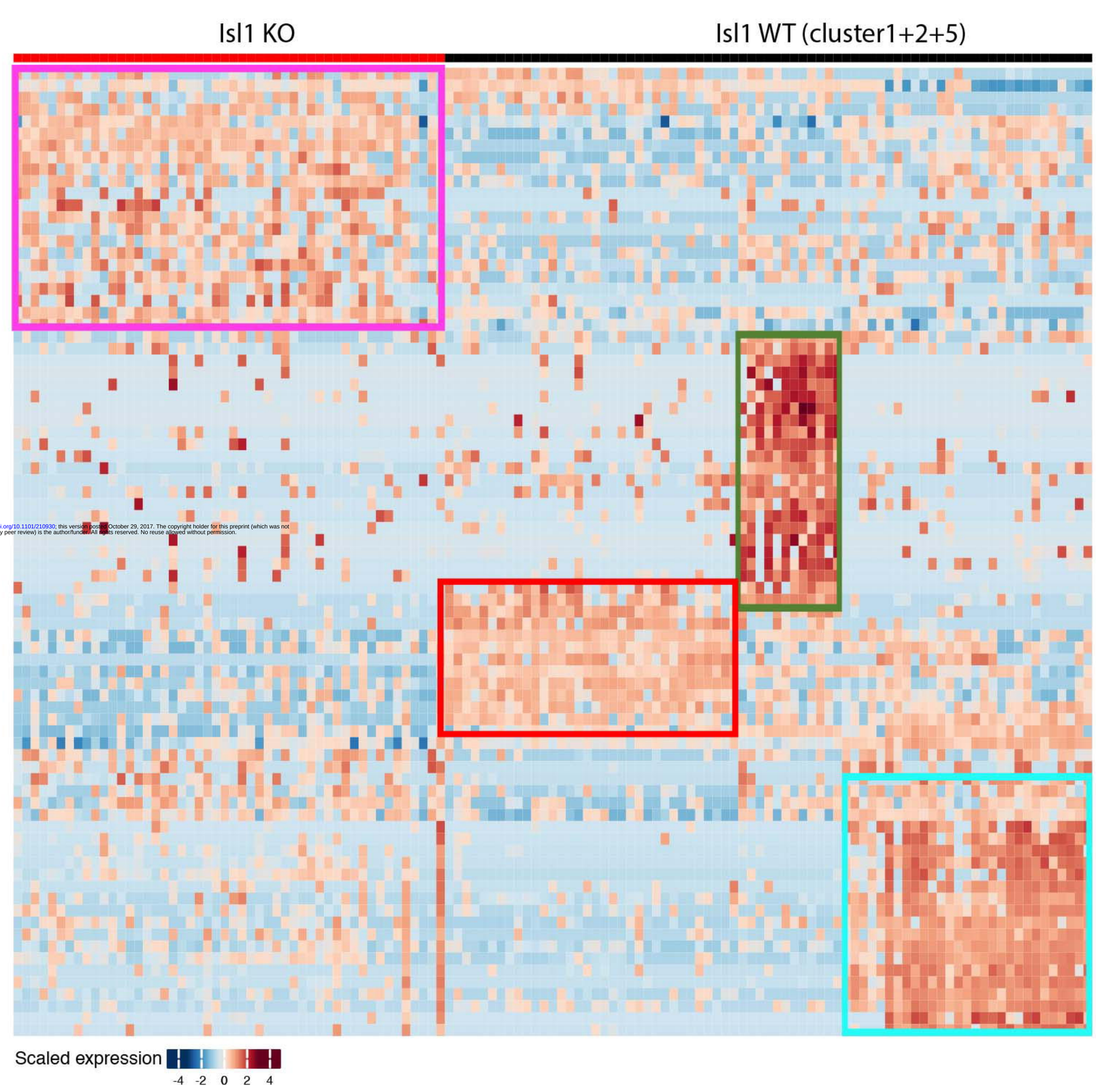
b



c



d



GO term:
regulation of cell differentiation (GO:0045595) (*p*: 9.30e-07)
Genes:
Gja1 Tgif1 Pdcd4 Nrep Ptn Fzd7 Crabp2 Adnp Nr2f1 Sema6a Npnt Rbp1 Sox11

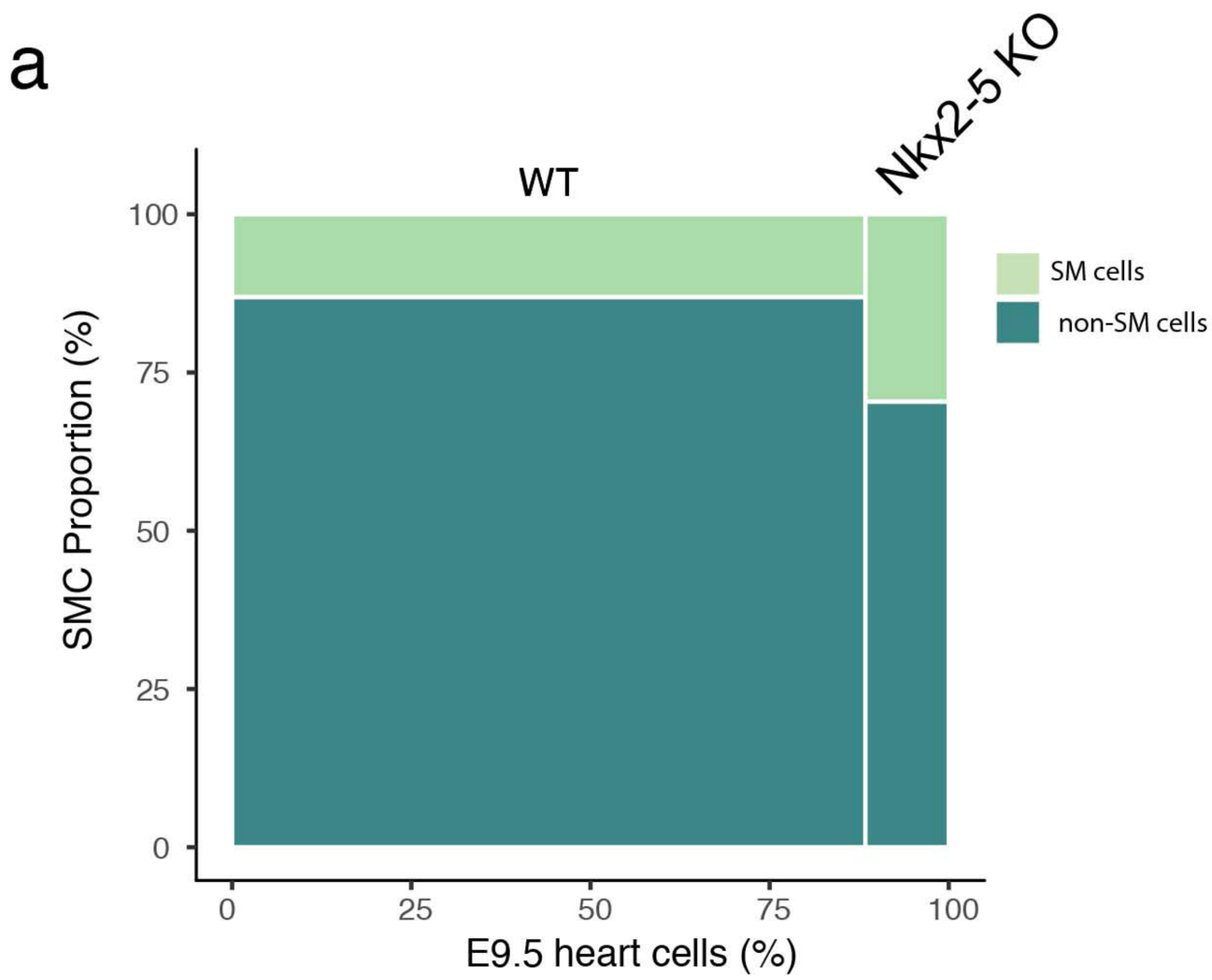
GO term:
endothelial cell migration (GO:0043542) (*p*: 2.08e-05)
Genes:
Flt4 Sox18 Plxnd1 Pecam1 Tek Kdr

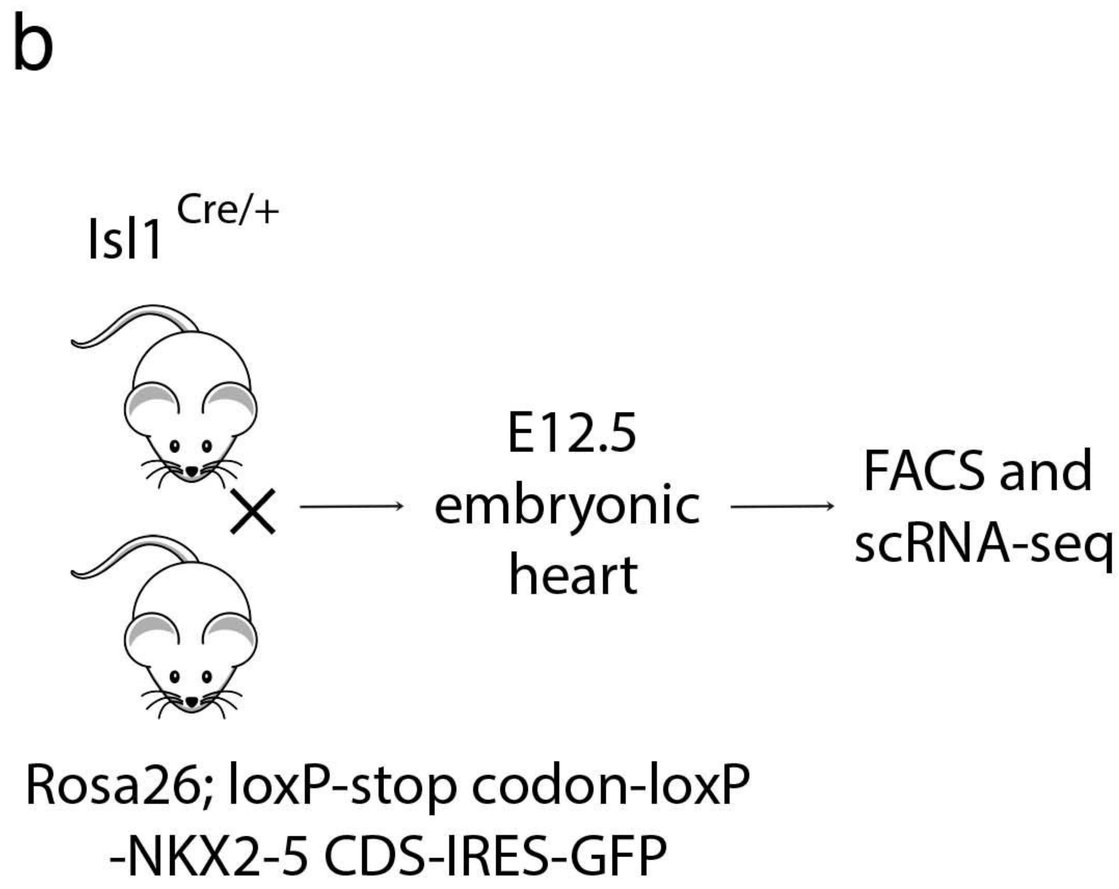
Genes:
Etv2 Kdr Sall3 Msx1 Lin28a Krt8 Kifc2 Tmem88 Spin2c Amot Rbpms2

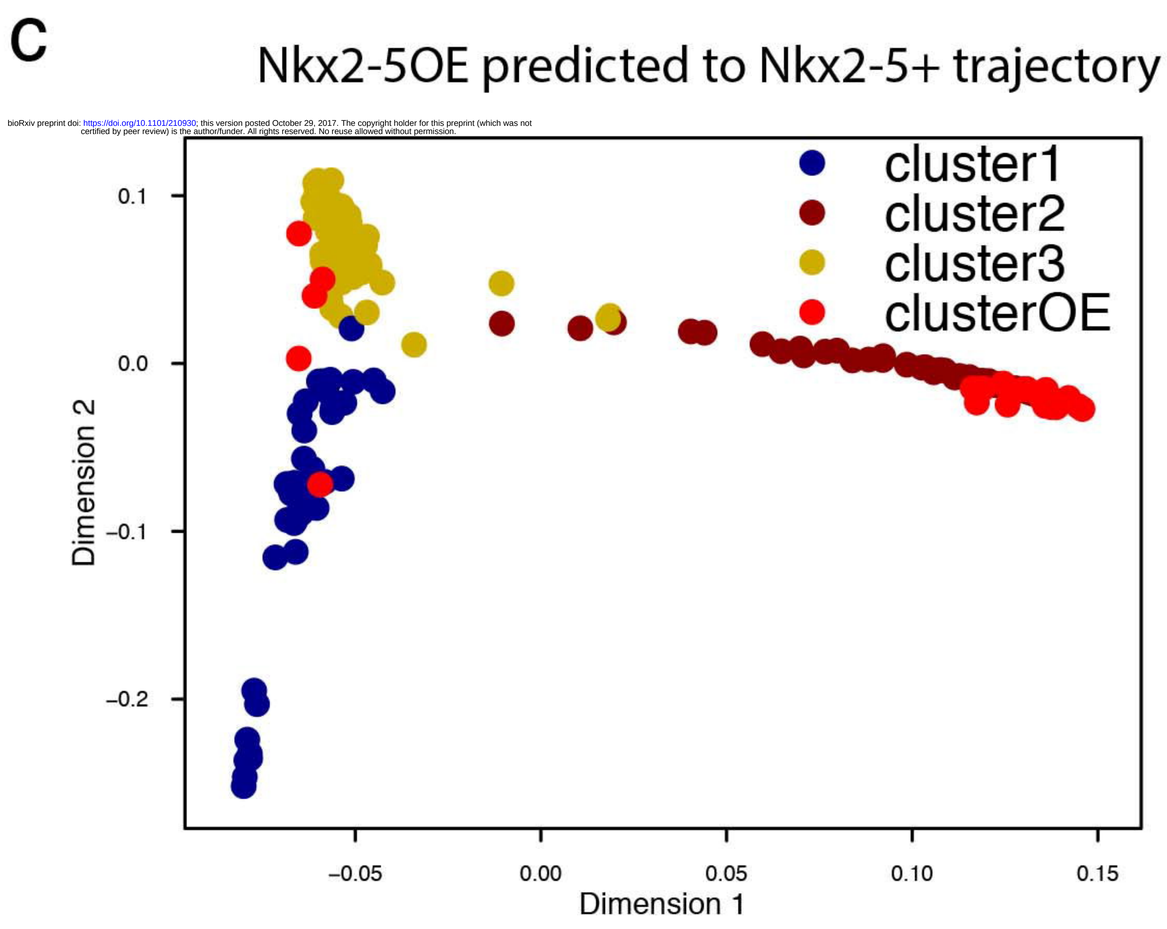
GO term:
muscle organ development (GO:0007517) (*p*: 2.03e-14)
Genes:
Tpm1 Plagl1 Smyd1 Myom1 Myh6 Mybpc3 Ankrd1 Cryab Mef2c Actc1 Myocd Unc45b Csrp3

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

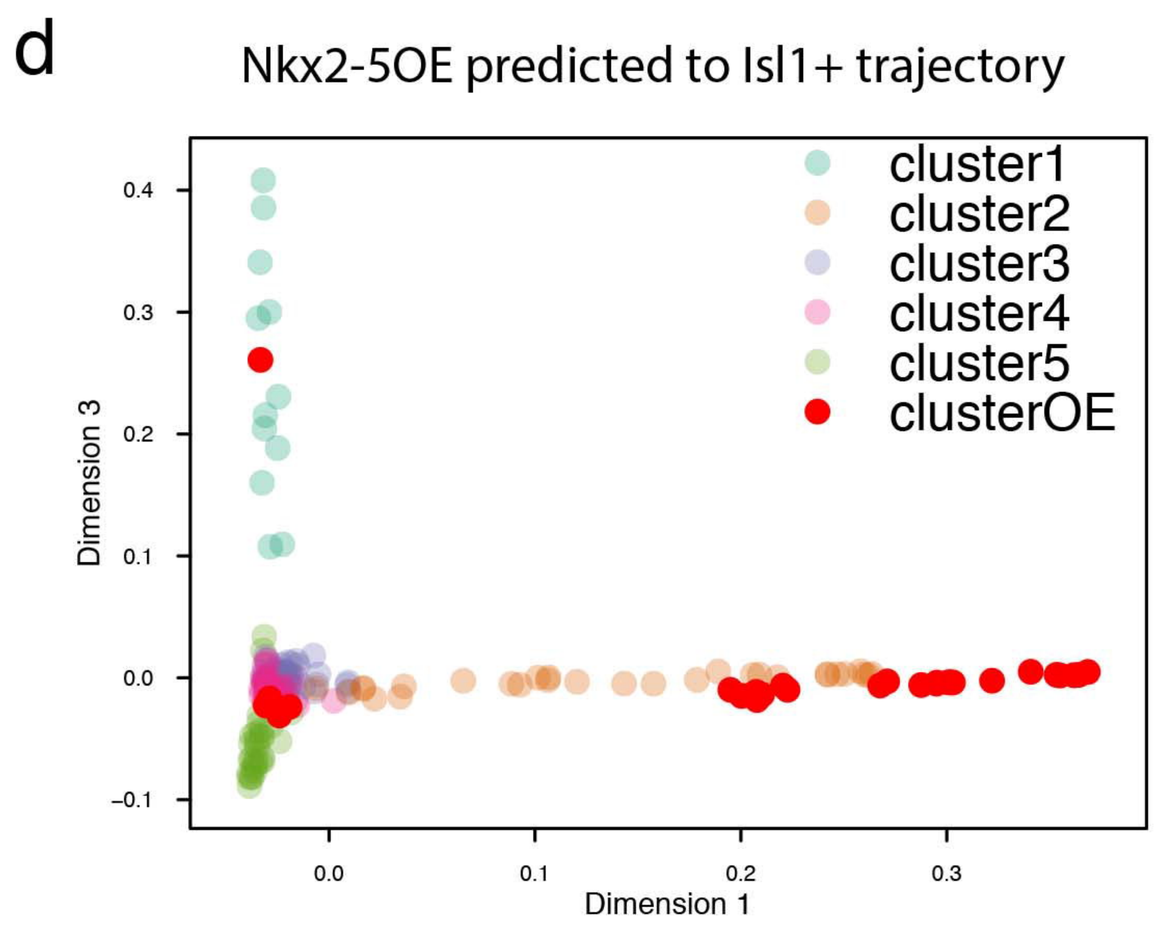
Jia, Preussner, et al Figure 4



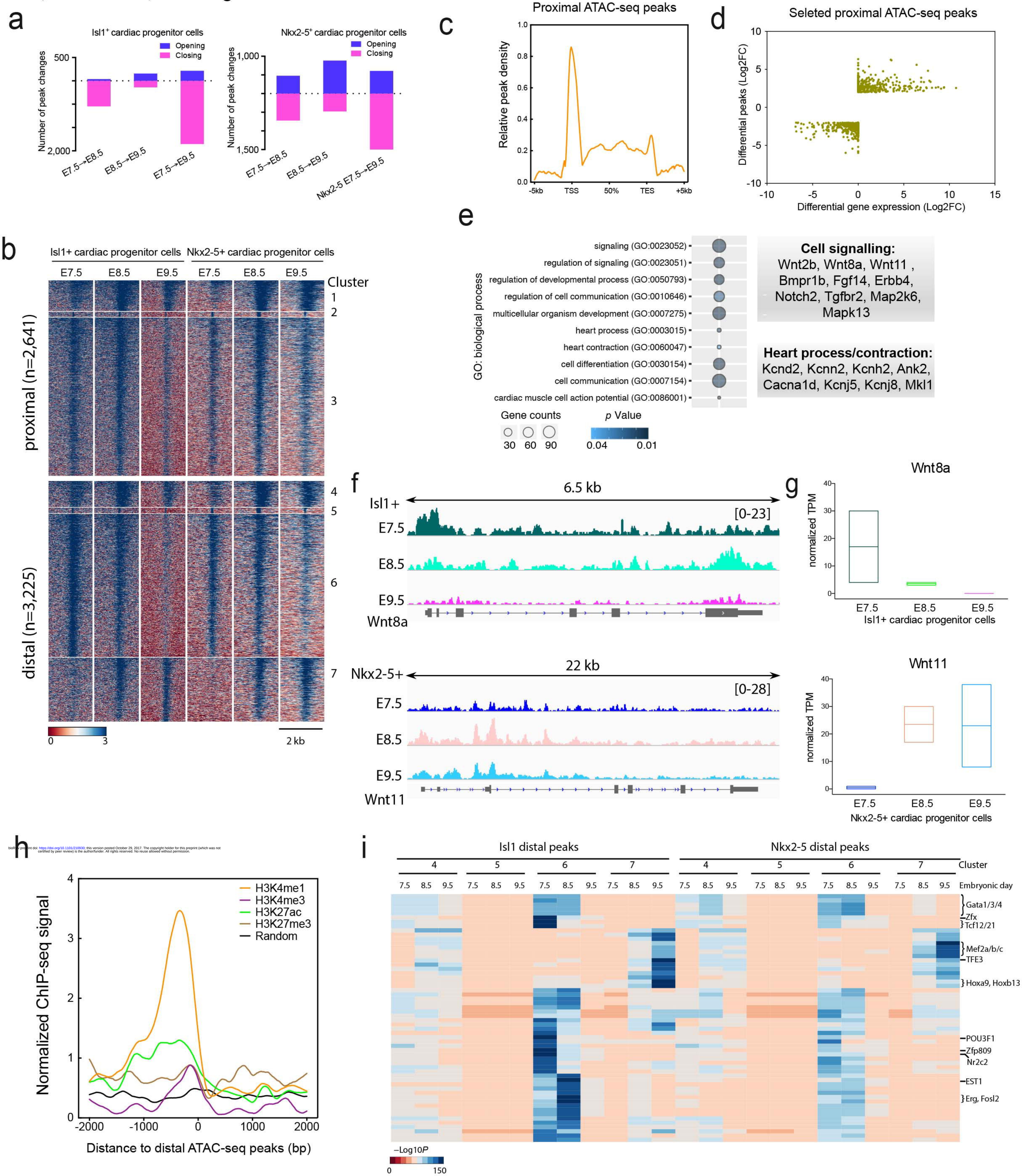




bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

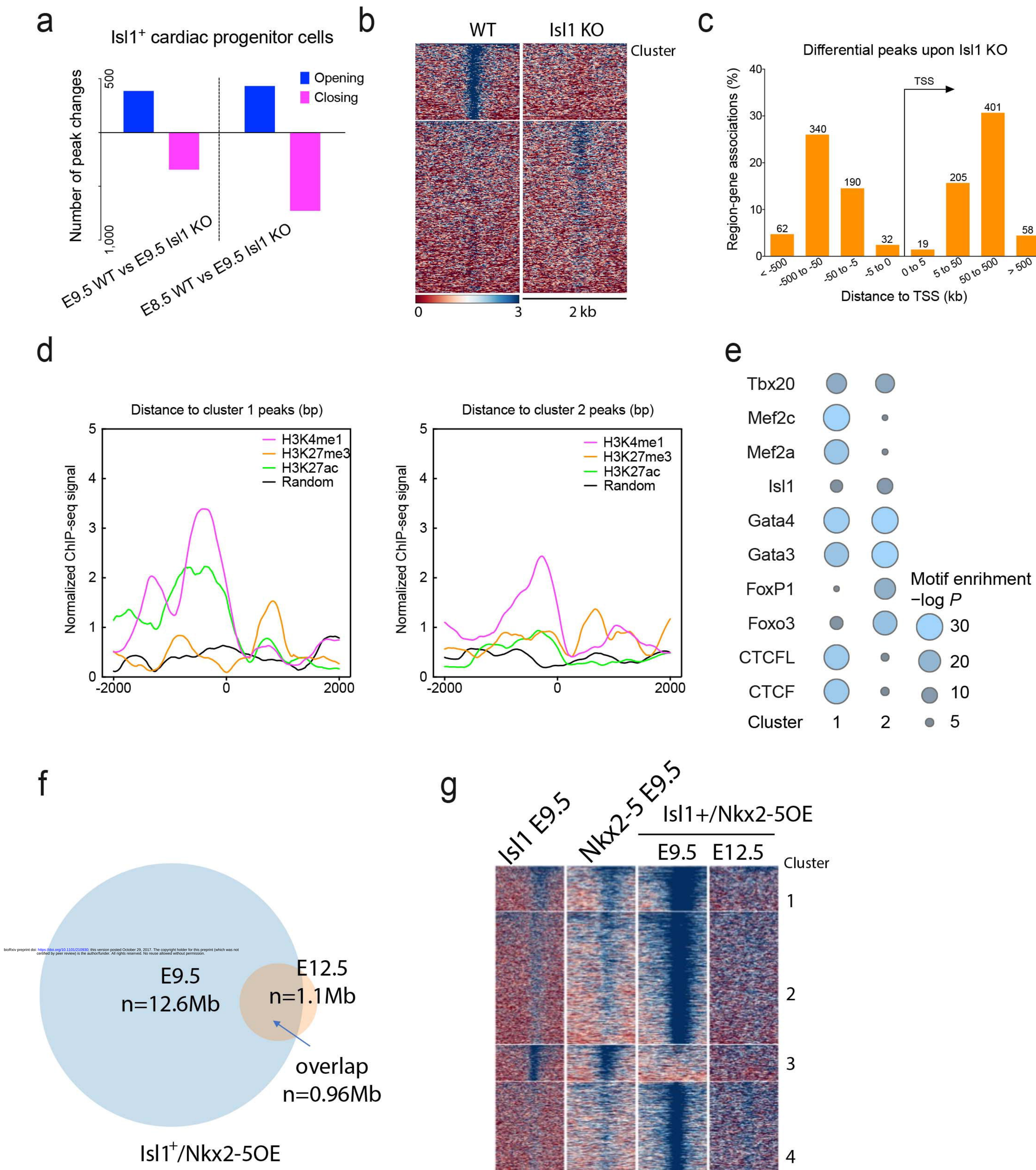


Jia, Preussner, et al Figure 5

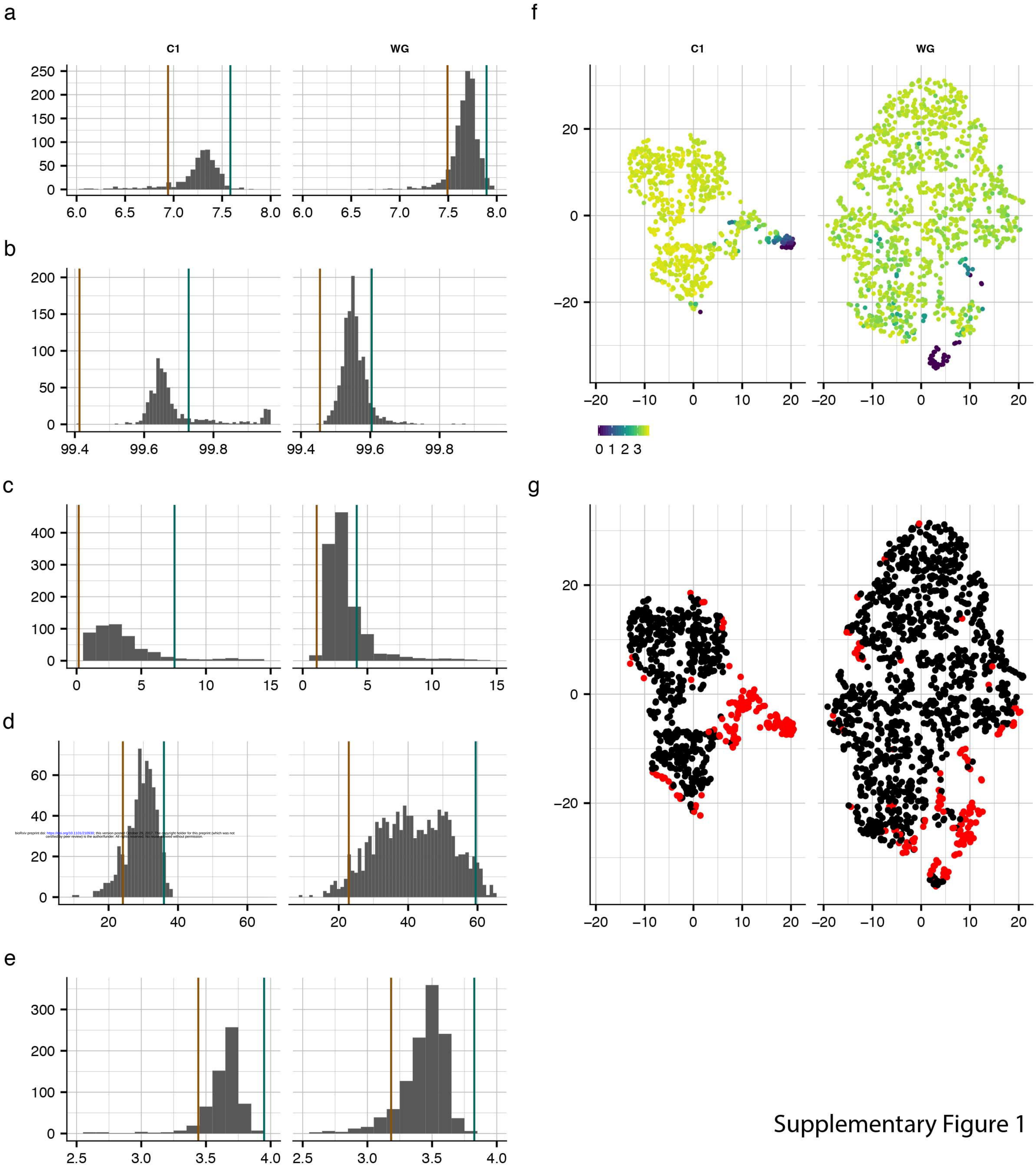


bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Jia, Preussner, et al Figure 6

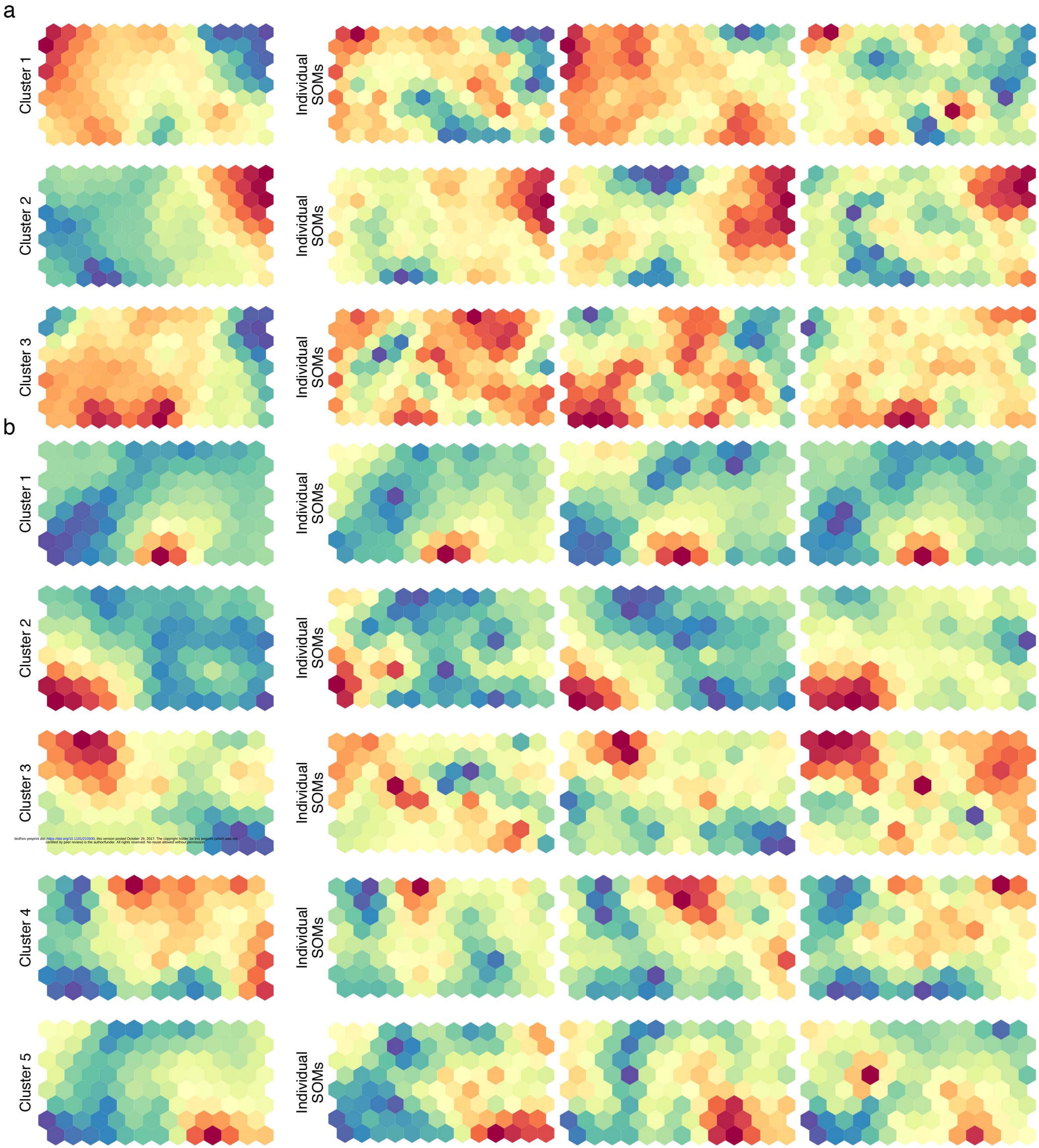


bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Figure 1

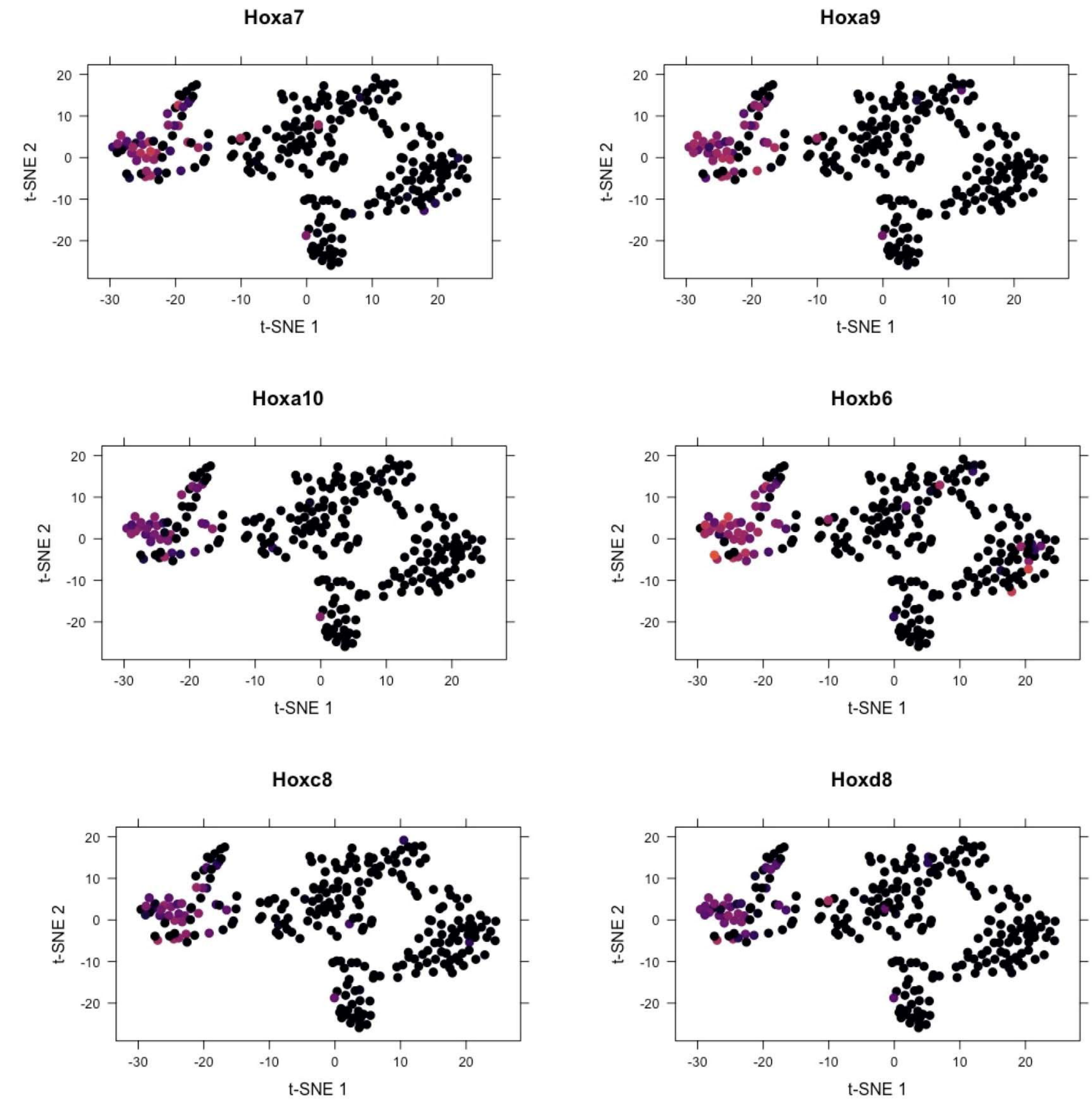


bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

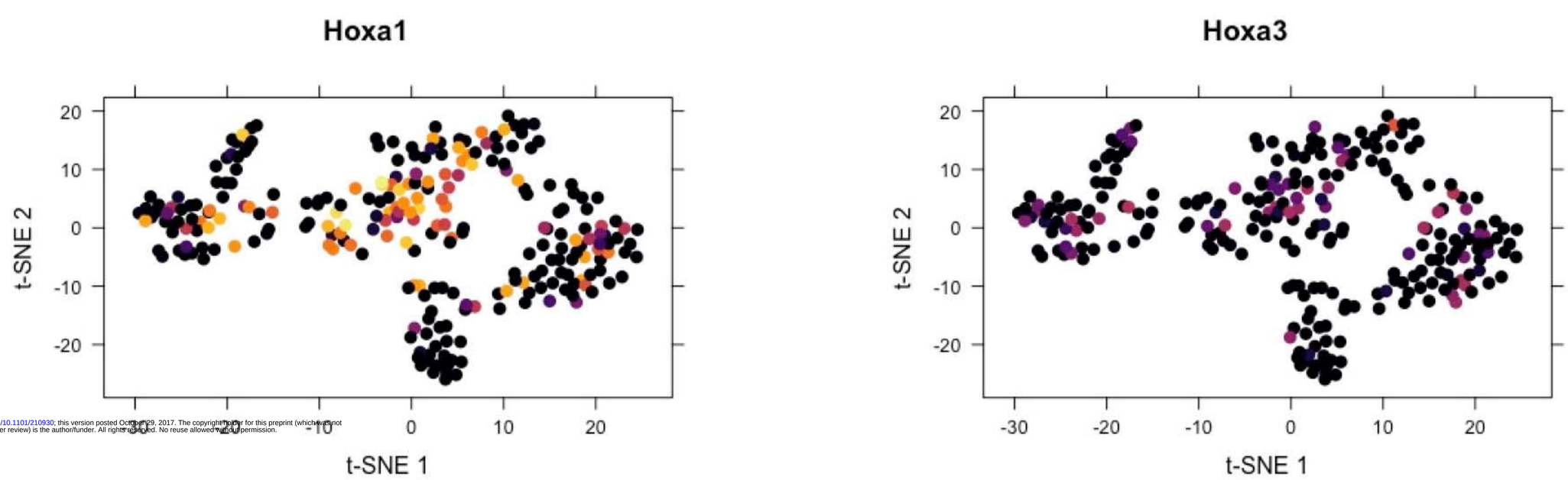
Supplementary Figure 2

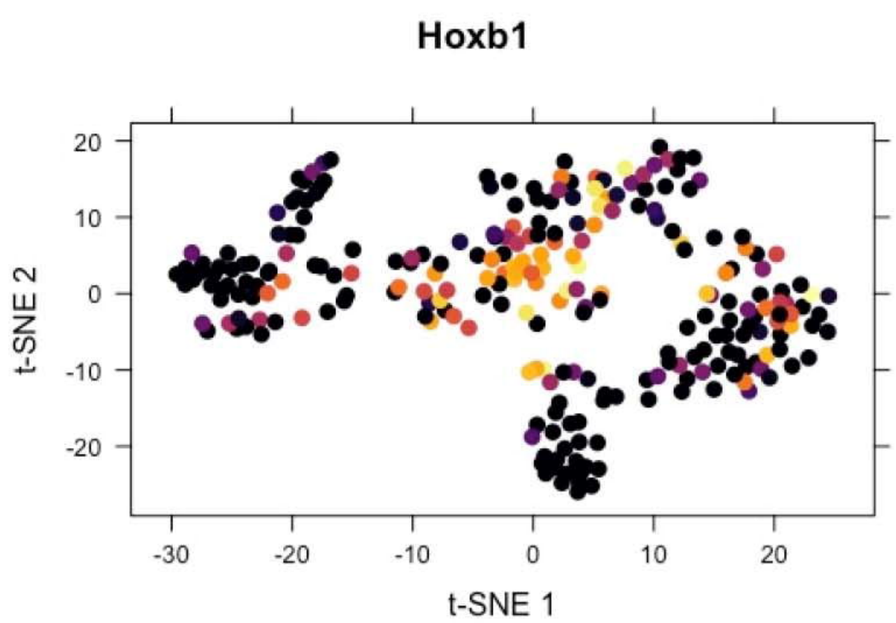
bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

a



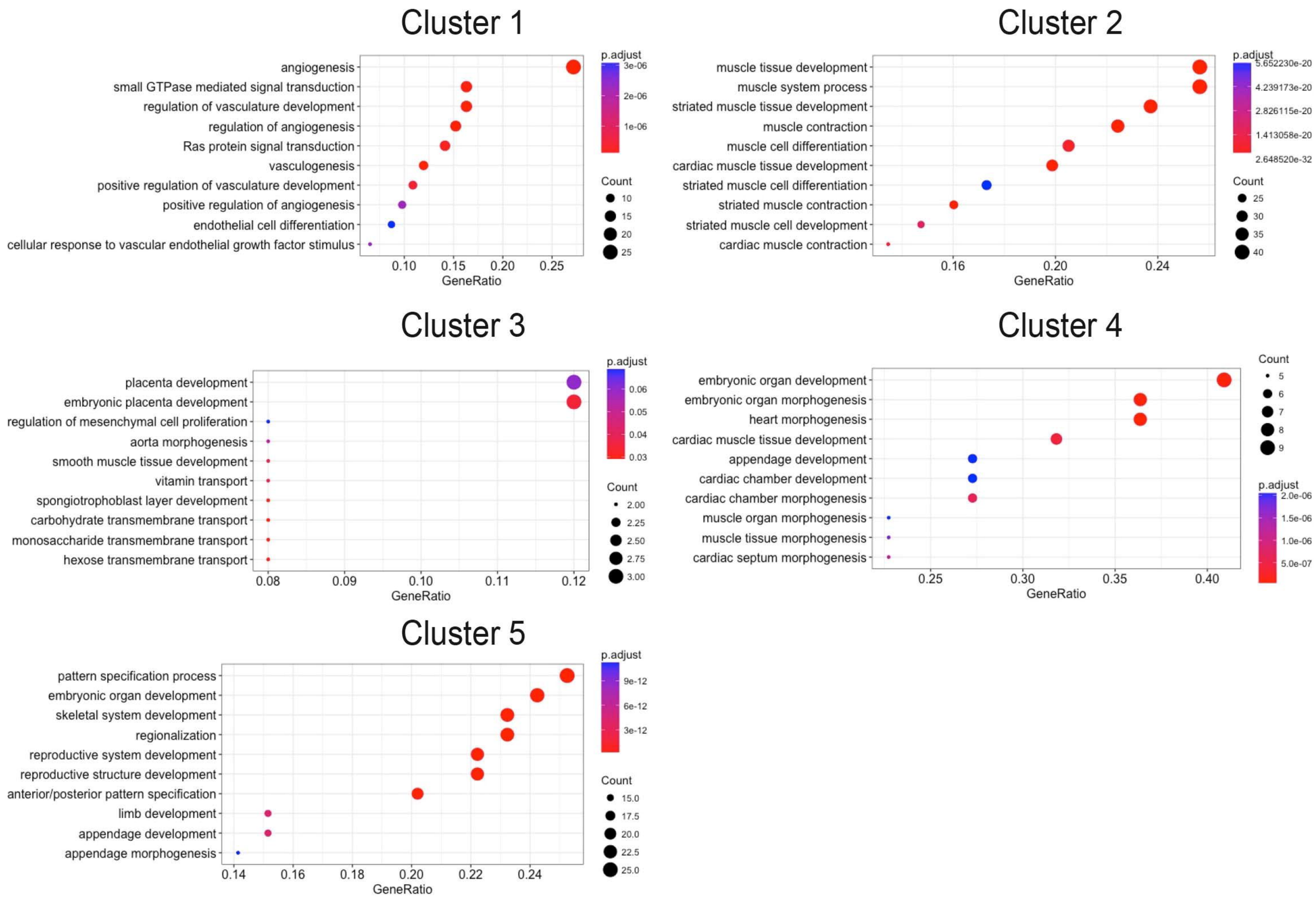
b



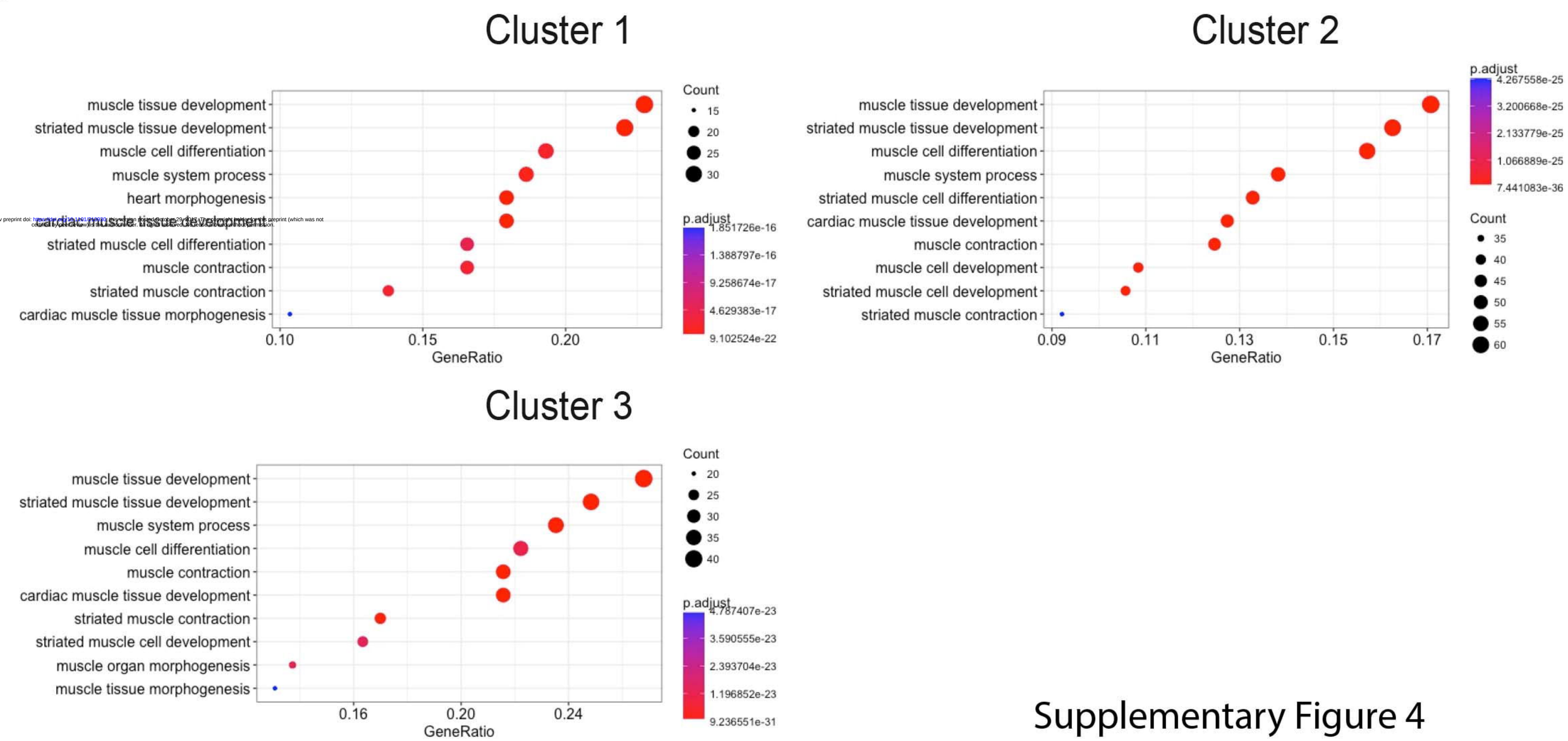


Supplementary Figure 3

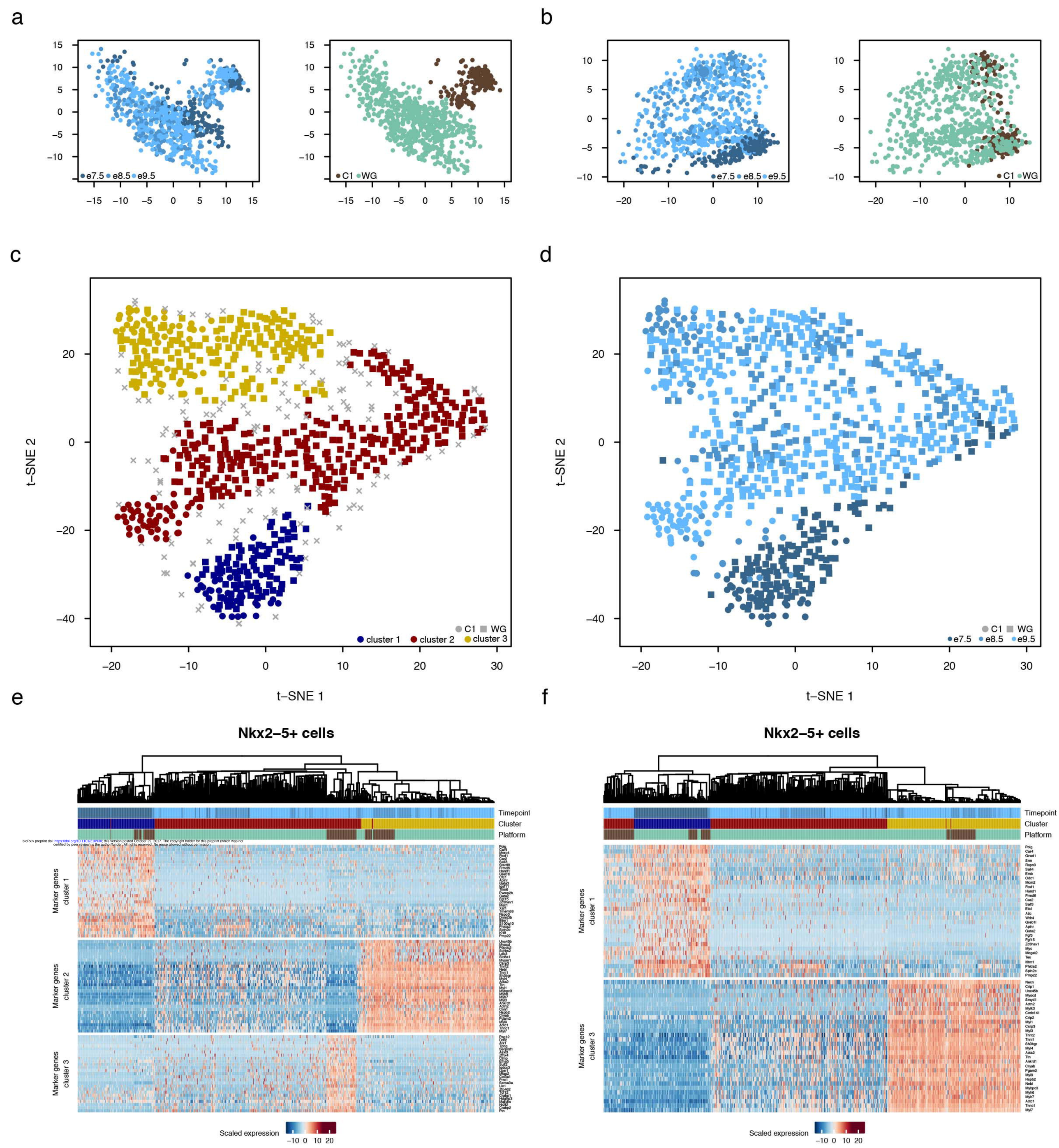
a



b

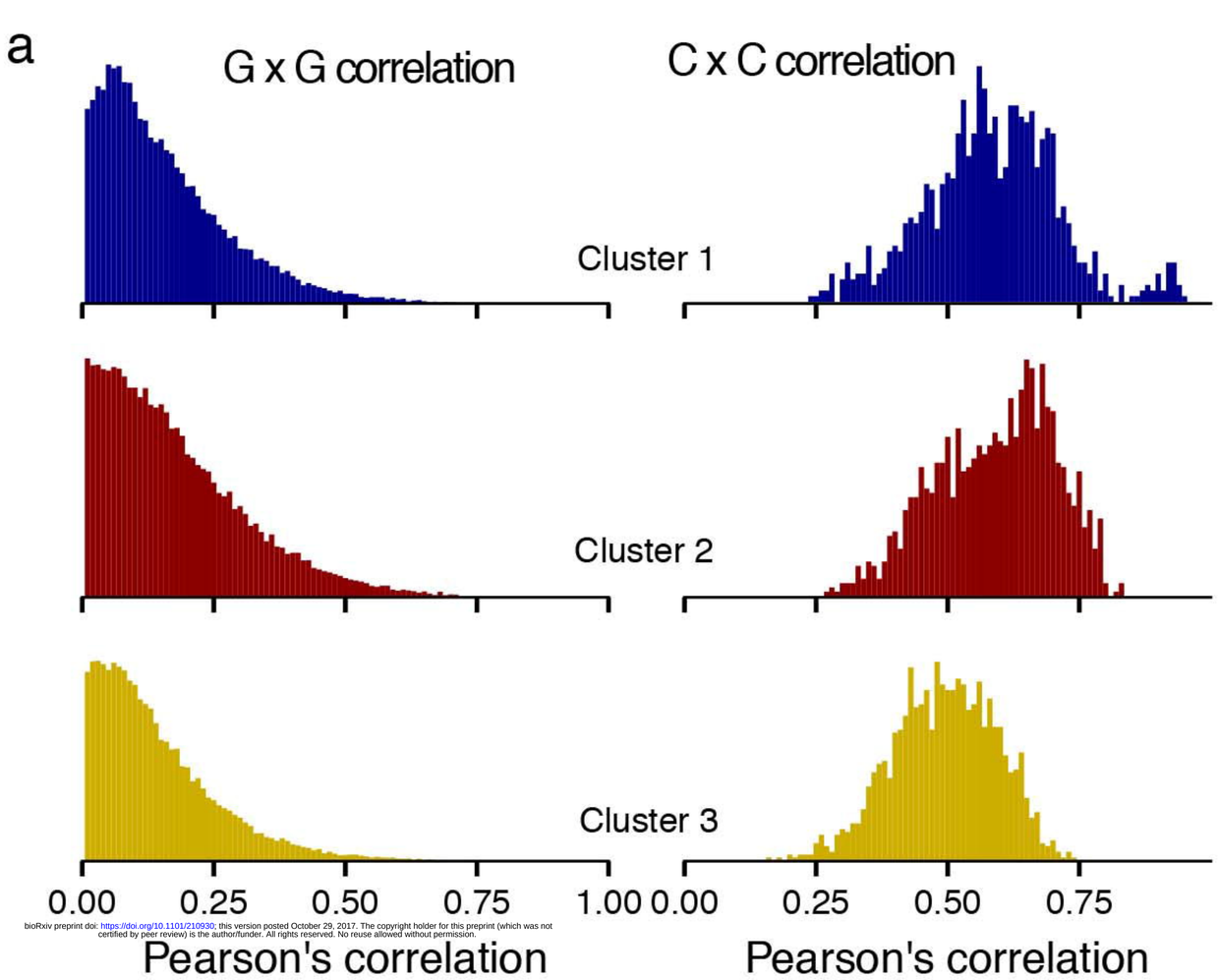


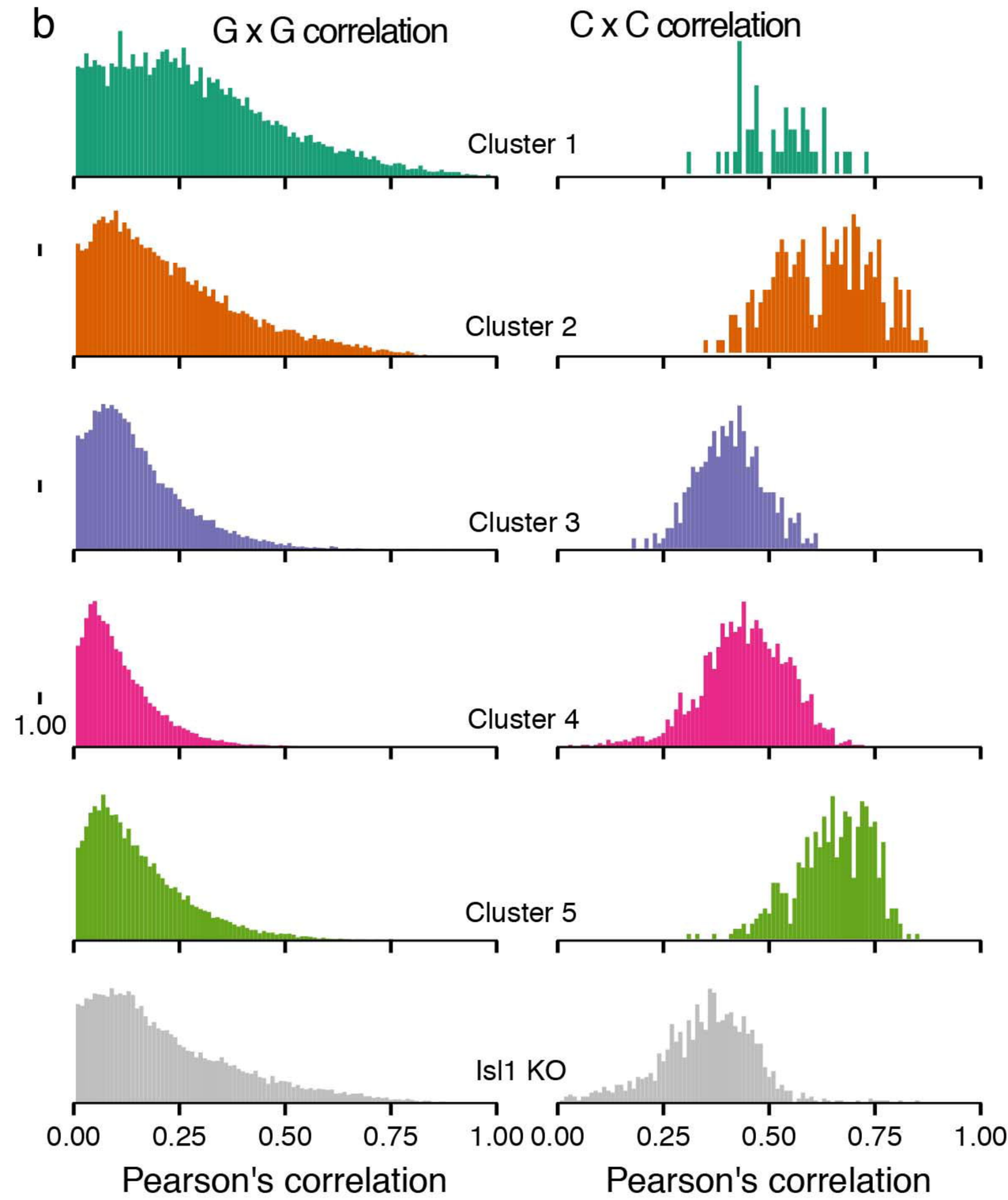
Supplementary Figure 4



bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

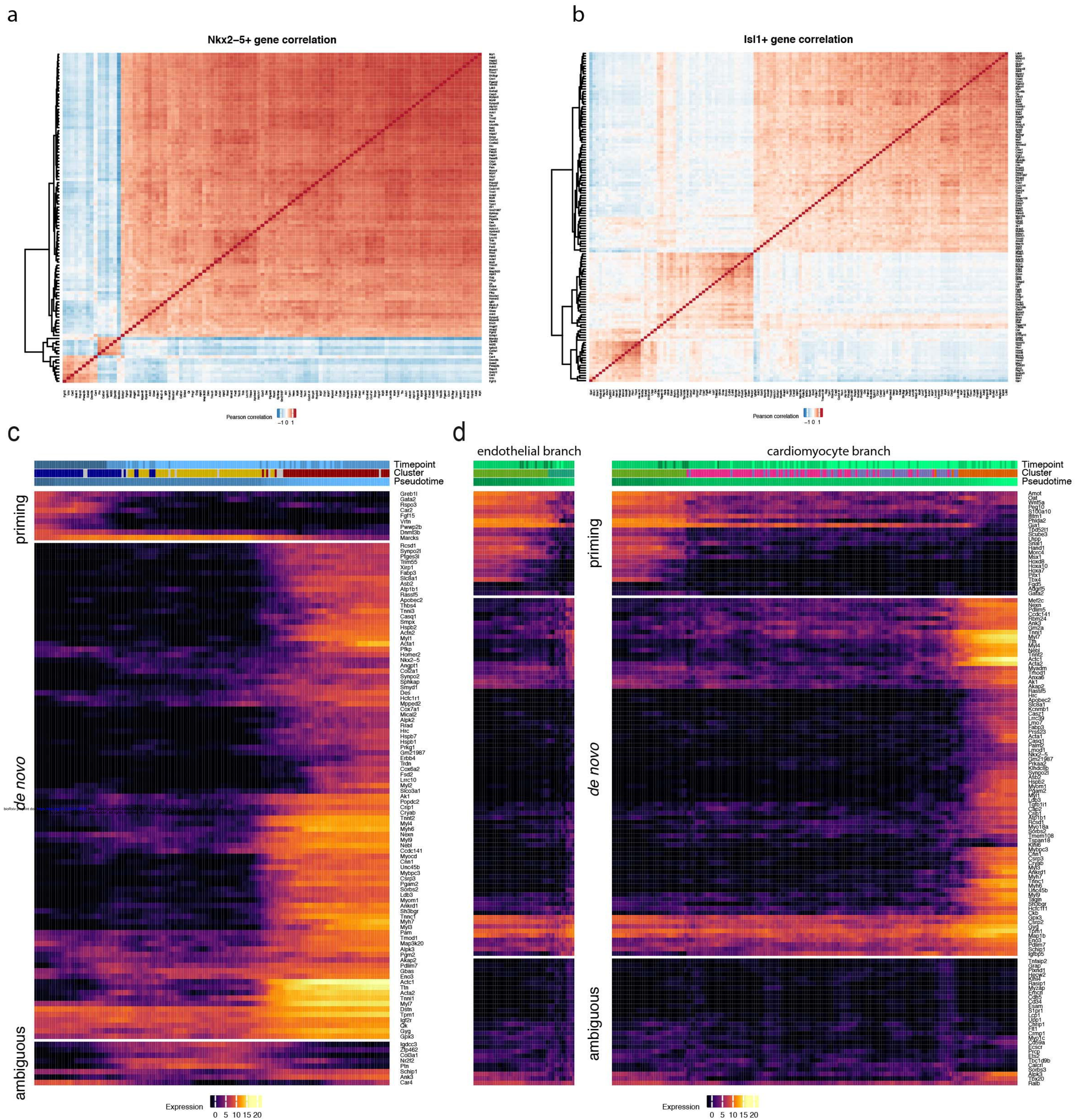
Supplementary Figure 5



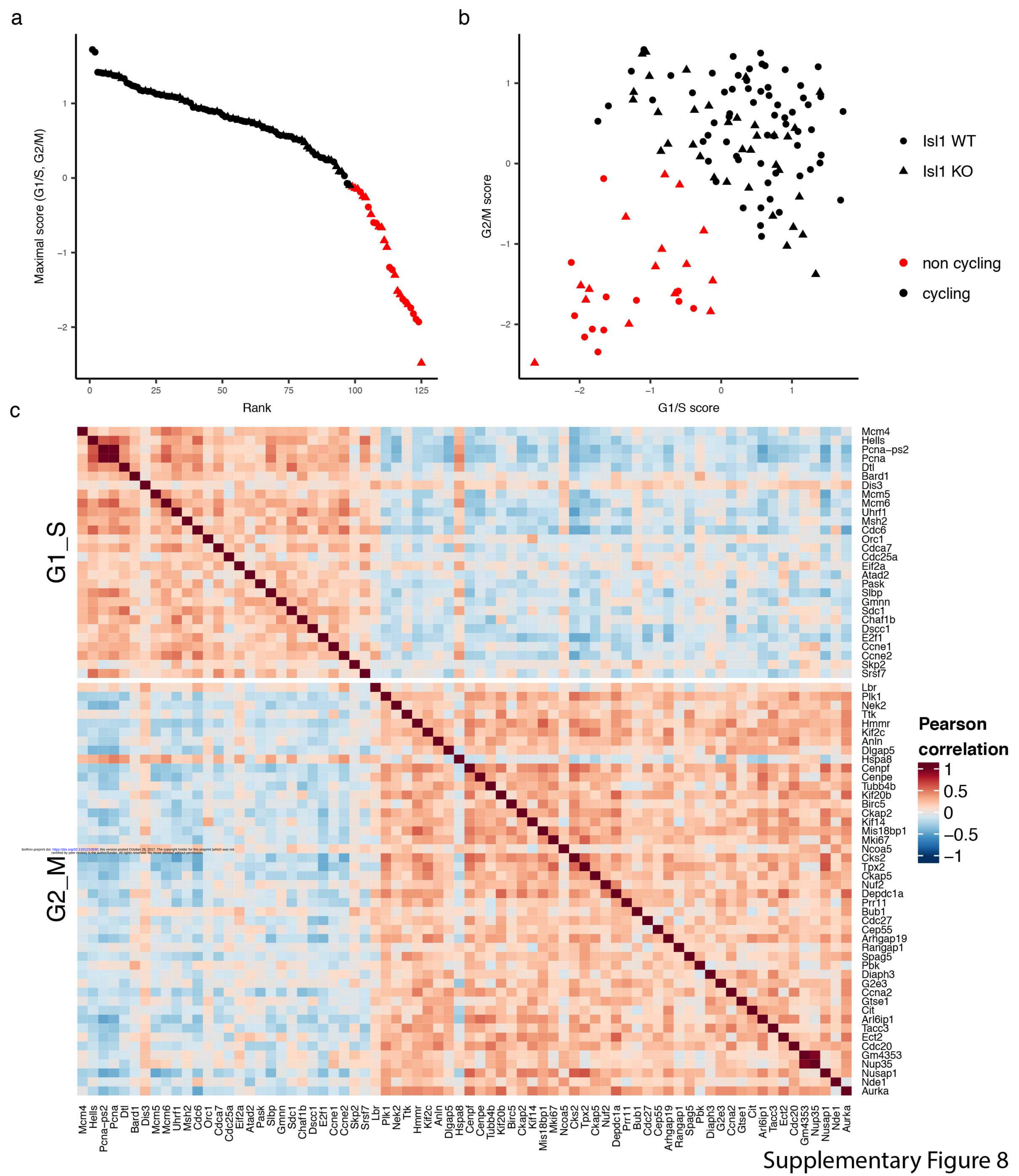


bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Figure 6

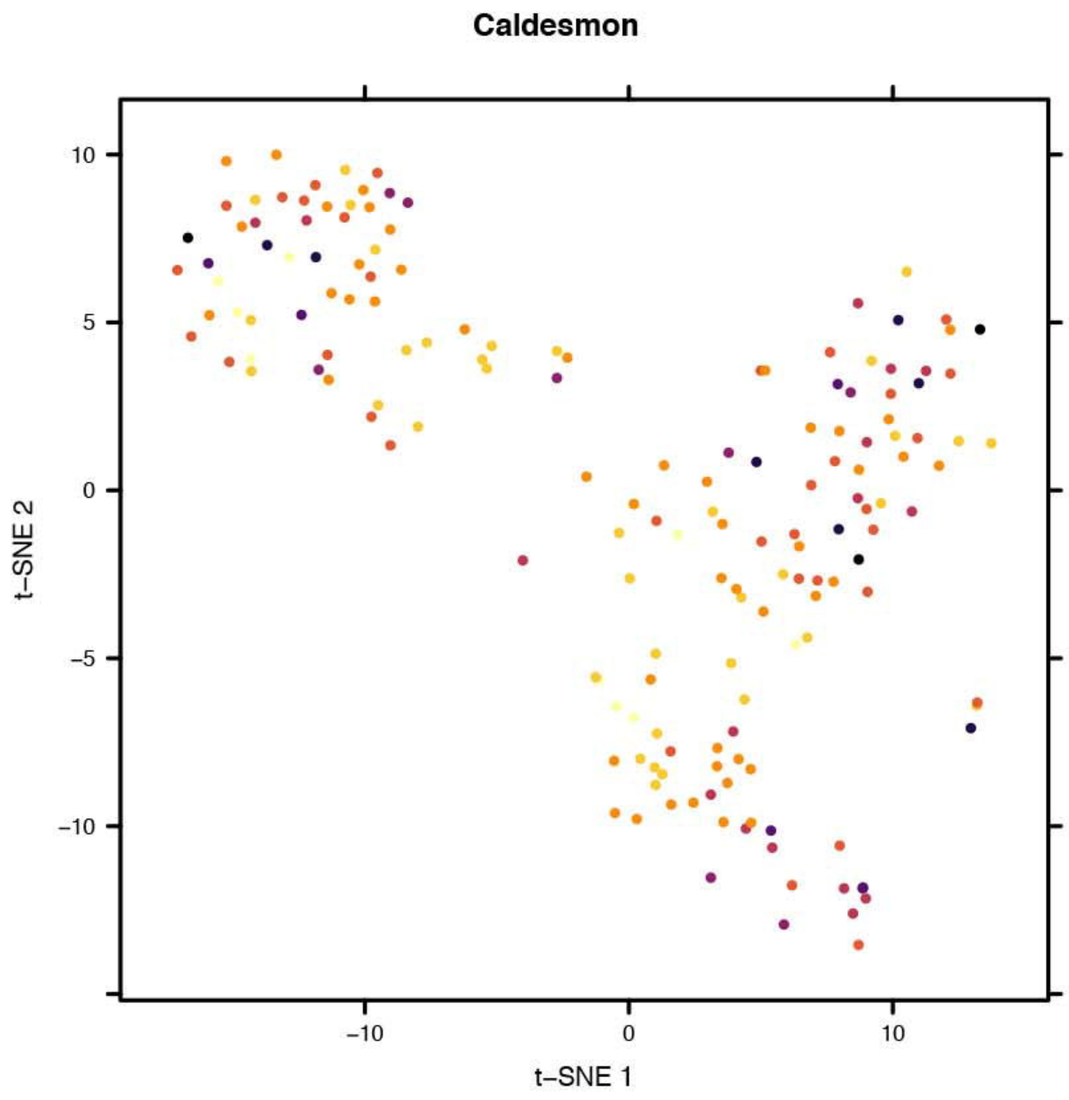


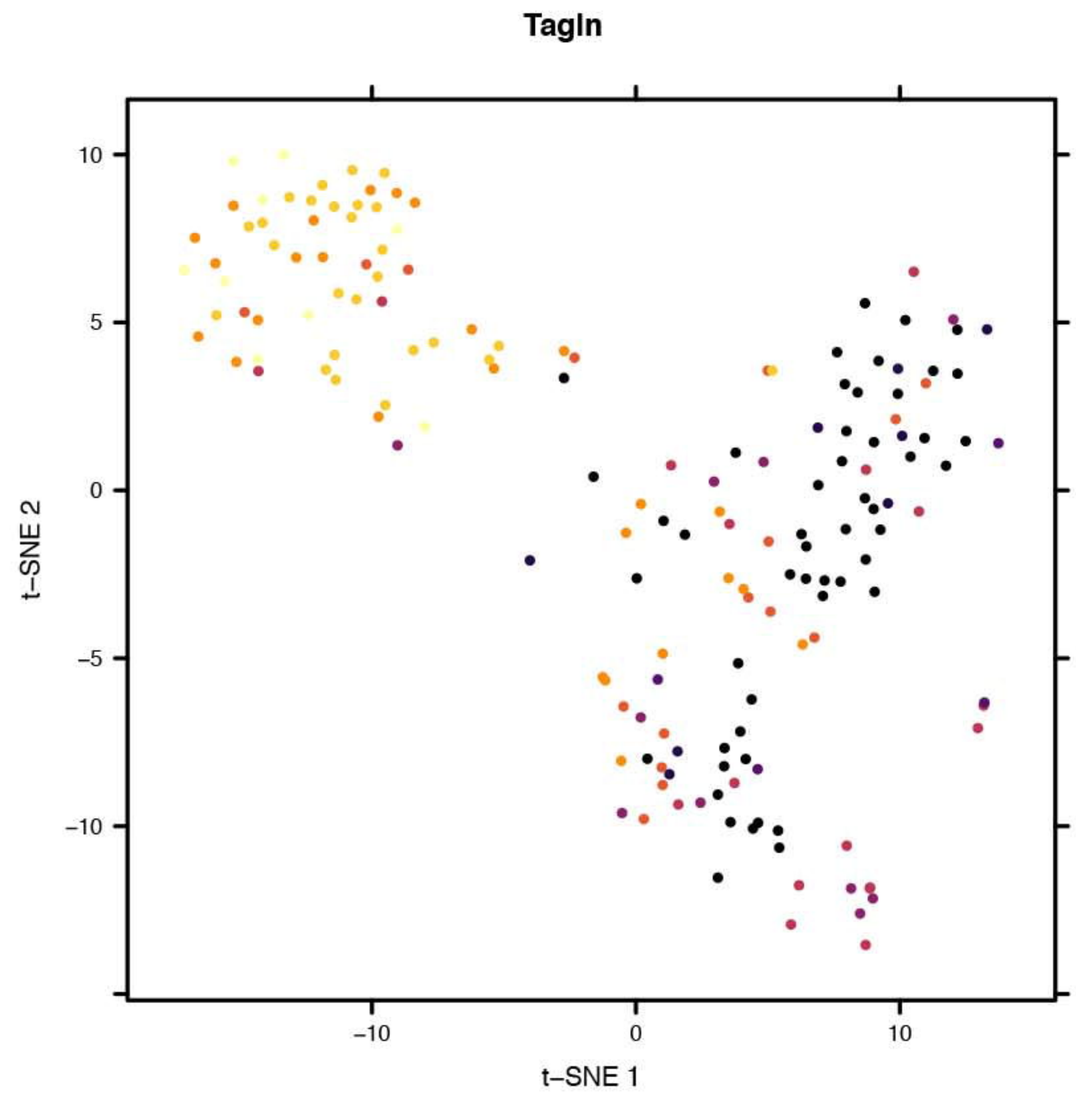
Supplementary Figure 7

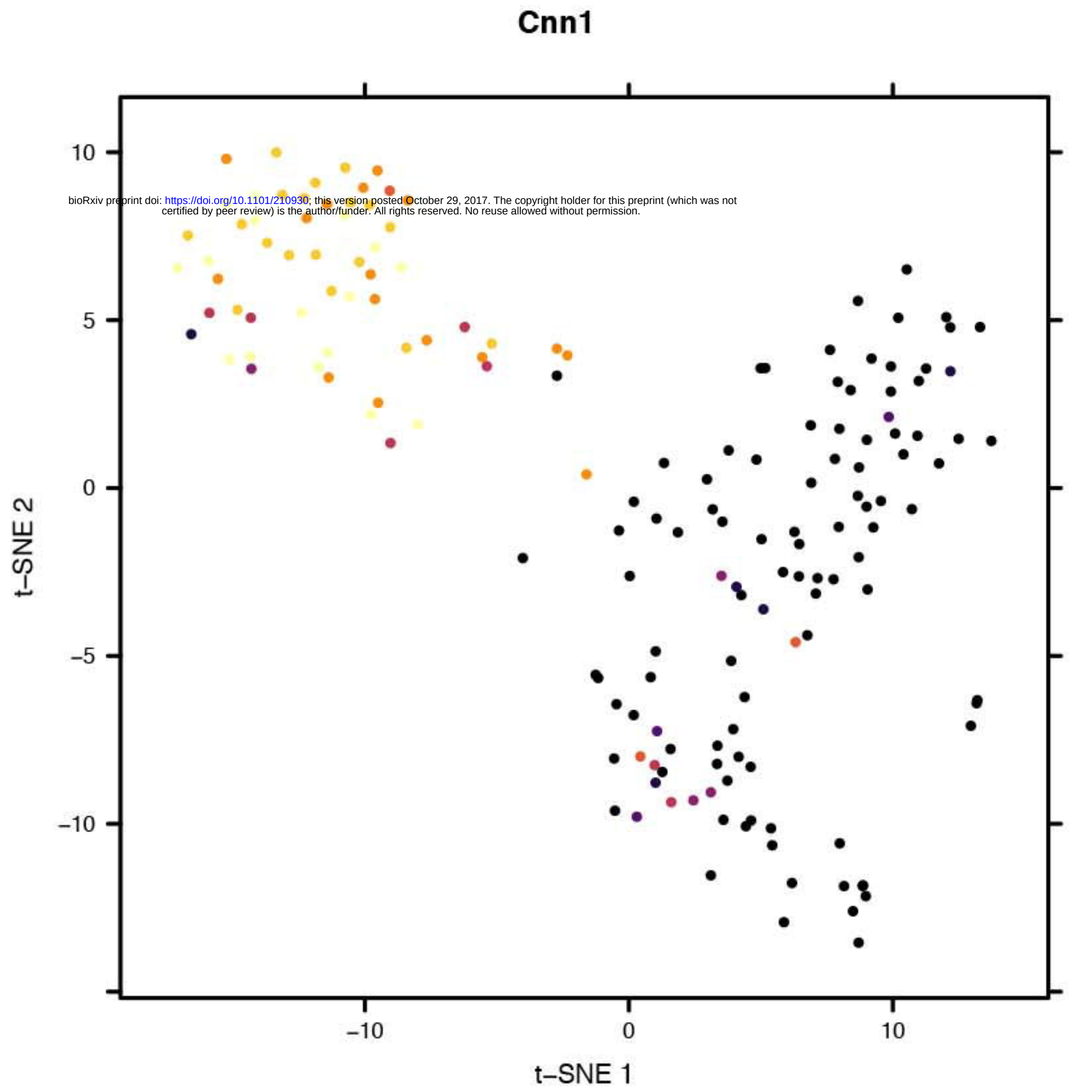


bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Figure 8



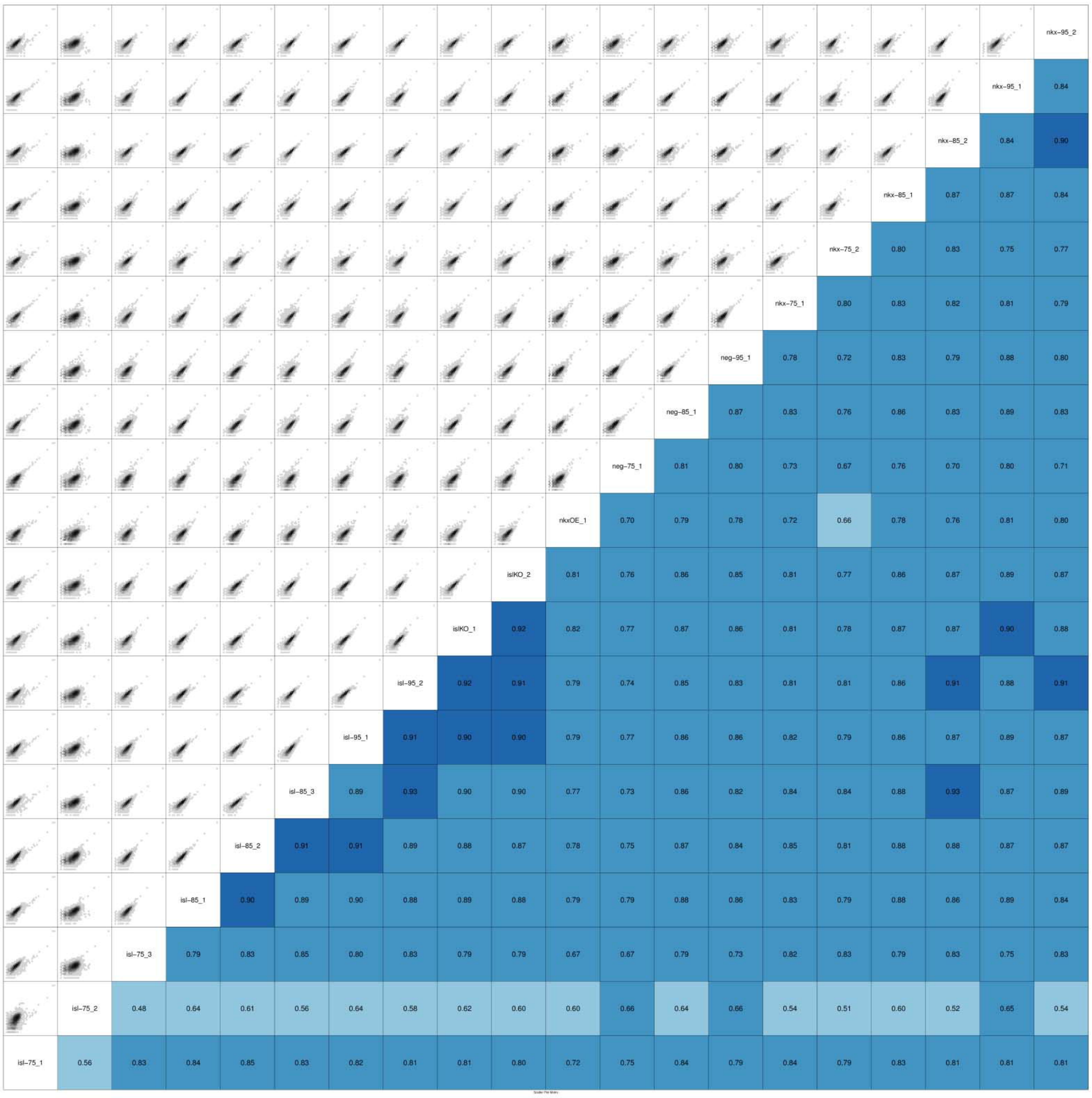




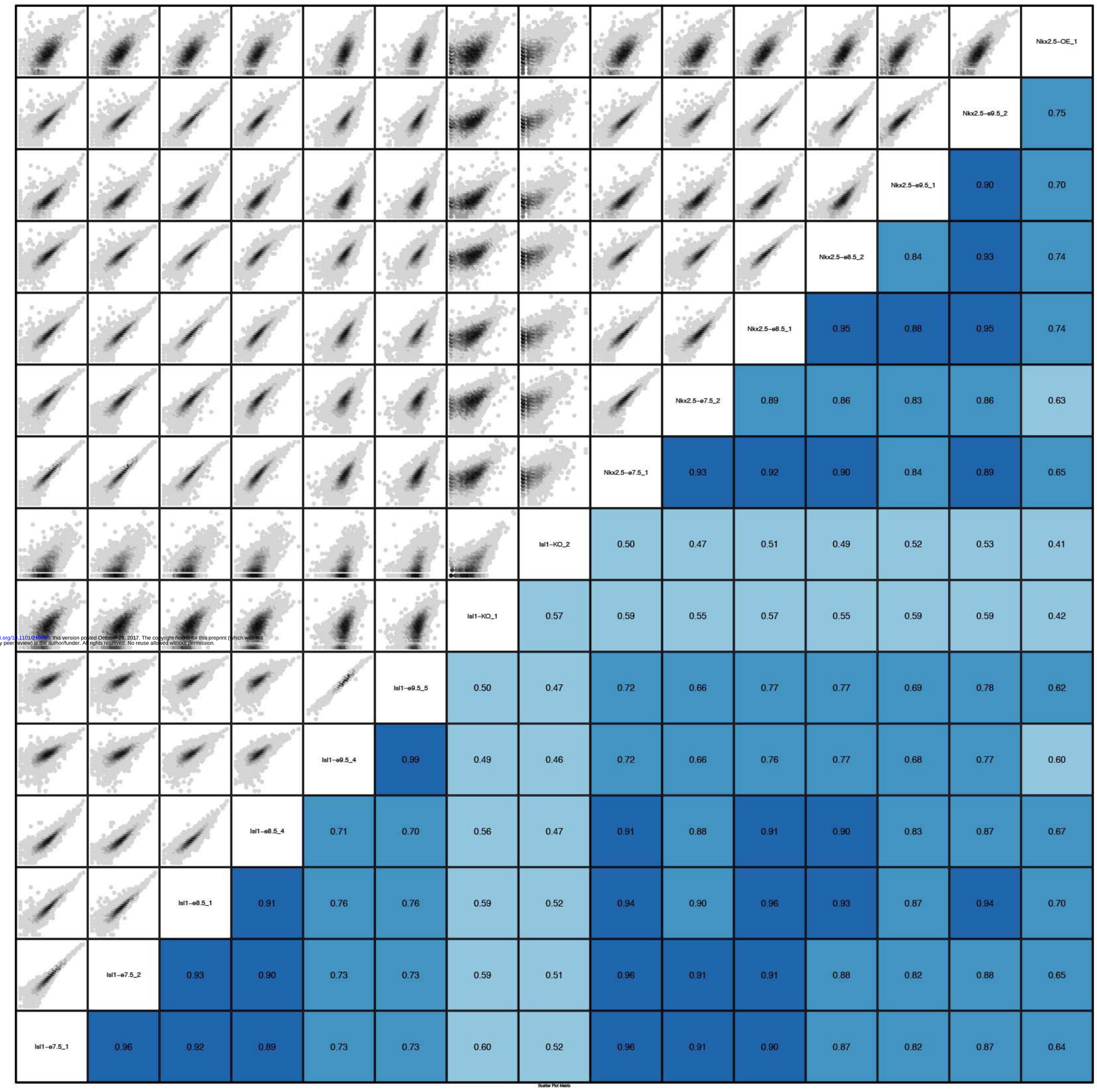
bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Figure 9

a

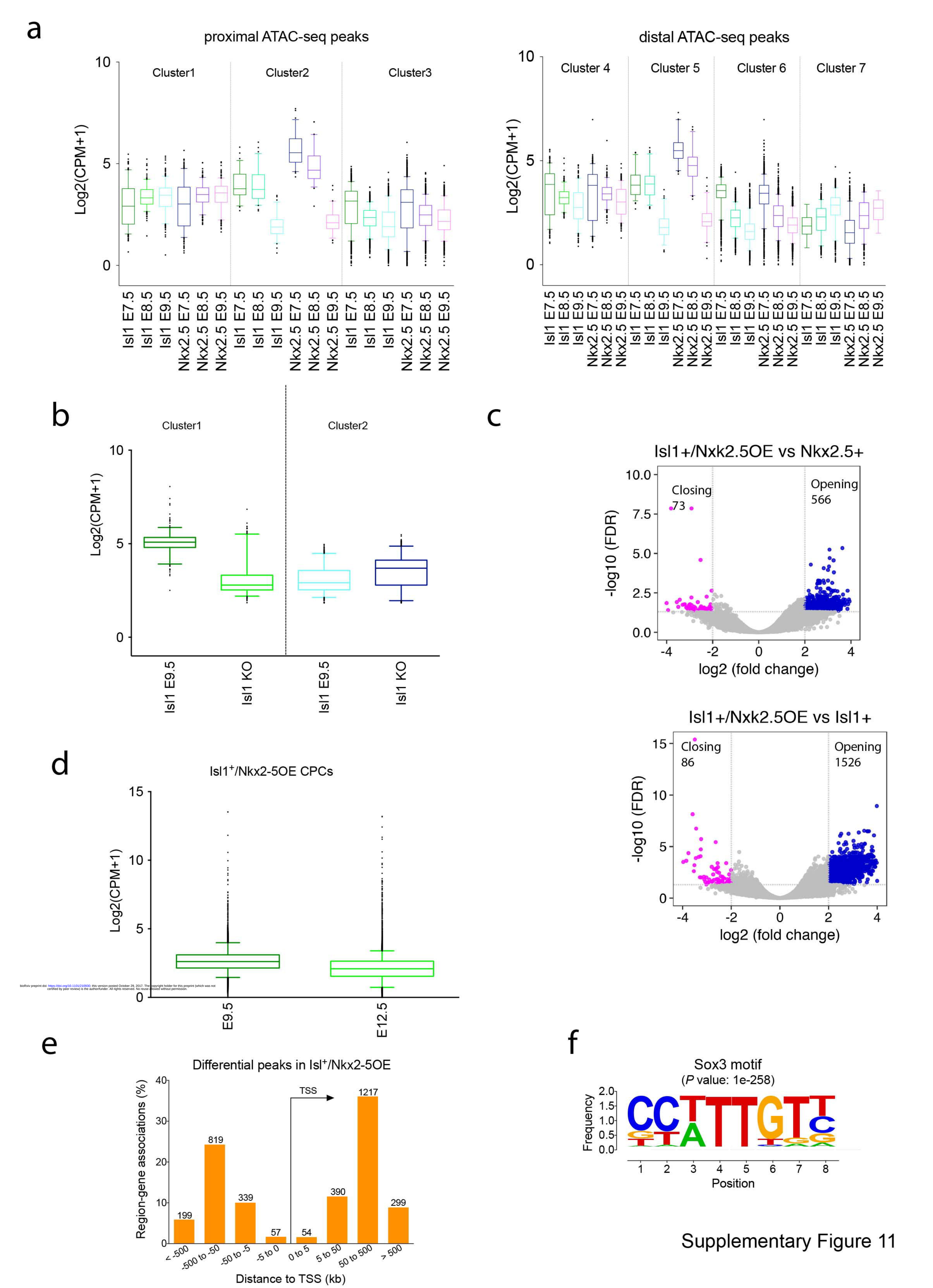


b



bioRxiv preprint doi: https://doi.org/10.1101/240869, this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Figure 10



Supplementary Figure 11

bioRxiv preprint doi: https://doi.org/10.1101/210930; this version posted October 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.