Continuous single cell transcriptome dynamics from pluripotency to hemangiogenic lineage

Haiyong Zhao[1] and Kyunghee Choi[1, 2, *]

[1]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO, USA
[2]Graduate School of Biotechnology, Kyung Hee University, Yong In, South Korea
*Corresponding author, kchoi@wustl.edu

Running title: Single cell RNA-seq analysis of hemangiogenic lineage development

Key words: single cell RNA-seq, ETS transcription factor, Etv2/Er71, hemangiogenesis, smooth muscle cell, lineage specification

Recent single cell RNA-sequencing (scRNA-seq) studies of early stages of embryos or human embryonic stem (ES) cell-derived embryoid bodies (EBs) provided unprecedented information on the spatiotemporal heterogeneity of cells in embryogenesis [1-8]. Nonetheless, these snapshots offer insufficient information on dynamic developmental processes due to inadvertently missing intermediate states and unavoidable batch effects. Blood and endothelial cells arise from hemangiogenic progenitors that are specified from FLK1-expressing mesoderm by the transcription factor ETV2 [9, 10]. To delineate the entire transcriptome dynamics from pluripotency to hemangiogenic lineage, we performed scRNA-seq of mouse ES cells in asynchronous EB differentiation. We found the exit from naïve pluripotency and hemangiogenic program activation were the two major transitions in the trajectory, while the intermediate gastrulation stages were gradually specified by 'relay'-like highly overlapping transcription factor modules. Unexpectedly, we found smooth muscle lineage might be the 'default' fate of FLK1 mesoderm, from which ETV2 initiates the hemangiogenic commitment. We also identified cell adhesion signaling required for ETV2-mediated activation of the hemangiogenic program. This continuous transcriptome map will benefit both basic and applied studies of mesoderm and its derivatives.

Embryoid bodies (EBs), in vitro differentiated ES cells, contain undifferentiated pluripotent cells, differentiated cells of specific lineages, and the intermediates[11] (Fig. 1A). The co-existence of the cells representing the full spectrum of differentiation in EBs avoids batch effects and allows a unique opportunity to delineate a continuous differentiation process[12]. By taking advantage of the asynchronous nature of differentiation of mouse ES cells and the culture conditions that favor hemangiogenic lineage output, we performed scRNA-seq of day 4 EB cells to elucidate a continuous developmental process of hemangiogenic lineage (Fig. 1A, 1B). After filtering out low quality cells, cells were clustered into 18 populations (Fig. 1C). Among them, 936 genes were differentially upregulated (supplementary Table 1). t-Stochastic Neighbor Embedding (t-SNE) was used to visualize the populations (Fig. 1D). Most clusters did not form a distinct, well-segregated population, underpinning continuous developmental processes. Notably, 6 major distinct branches/populations were observed at the edges of the trunk, to which different fates were assigned according to expression of related marker genes (Fig. 1D). High expression of *Sox2*, *Pou5f1/Oct4*, *Nanog* and *Zfp42/Rex1* highlights the maintenance of undifferentiated naïve pluripotent cells (Fig. 1E). Co-expression of the pluripotency markers *Sox2*, *Nanog* and *Pou5f1* with *Prdm1* and *Prdm14* indicated that primordial germ cells were generated in this differentiation system (Fig. 1E, F). A separated endoderm population was observed, as indicated by the expression of *Foxa2* and *Sox17* (Fig. 1G). *Hand1* and *Tbx20* were expressed by both cardiac and smooth muscle progenitors (Fig. 1H, I). However, differential upregulation of *Isl1* and *Foxf1* could distinguish these two lineages (Fig. 1H, I). As expected, we observed hemangiogenic lineage commitment, as marked by high expression of *Flk1*/*Kdr* and *Etv2* and the following upregulation of the ETV2 target genes *Tal1* and *Gata2* (Fig. 1J). Importantly, smooth muscle and hemangiogenic lineages directly bifurcated from a common *Flk1/Kdr*-expressing (*Flk1*$^+$) population (Fig. 1I, J). The immediate upstream population of *Flk1*$^+$ mesoderm expressed *Pdgfra* and *Mesp1* (Fig. 1K). This population also expressed low levels of *Tbx6*, implying that they are to some extent 'primed' for paraxial mesoderm, which generates tissues like skeletal muscles and cartilage, and might represent an ancestral state for both FLK1 lateral plate mesoderm and paraxial mesoderm. The cardiac lineage branched out largely from the *Pdgfrα^+Flk1$^-$* early mesoderm population, and partially from *Flk1$^+$* mesoderm (Fig. 1H-K), consistent with the previous lineage tracing report that cardiomyocytes partially derive from FLK1$^+$ mesoderm[12]. Cells upstream of the early mesoderm down regulated the naïve pluripotency marker *Zfp42*, and up regulated the primitive streak marker *T/Brachyury*, indicating the initiation of gastrulation (Fig. 1K). These results suggested that our populations covered a continuous process from pluripotency to *Flk1$^+$* mesoderm and hemangiogenic lineage.

To examine the dynamics of the loss of pluripotency and the bifurcation of mesoderm and endoderm fates, we picked out clusters 4, 5, 8, 9, 11 and 17 (Fig. 1D) and re-clustered the cells based on 163 most differentially expressed genes among these clusters (Fig. 2A, 2B). A narrow “path” from *Utf1*-expressing pluripotent population leads to the *Sox17*-expressing endoderm (red dash line-surrounded region, Fig. 2A), next to the sprouting *Mesp1$^+$* early mesoderm. We ordered cells in this path into a pseudo-time development line and examined the kinetics of 163 most differentially expressed genes (Fig. 2B). It is suggested that pluripotency is maintained in two states: the ‘naïve’ and the ‘primed’[13]. The former reserves better pluripotency and highly expresses *Zfp42/Rex1*; while the latter loses *Zfp42* expression and maintains a balanced state from diversified differentiation cues, marking an early stage of gastrulation/mesendoderm[13]. At the beginning of the endoderm differentiation trajectory, we observed an exit from naïve pluripotency (Fig. 2C, blue dash line). Unexpectedly, expression of a small group of genes (represented by *Dppa3*, in brown box in Fig. 2B, and in Fig. 2C) already displayed strong heterogeneity in the *Zfp42* -expressing population. Gradual loss of the expression of these genes preceded the exit from naïve pluripotency. This may suggest an extra layer of guarding mechanisms of pluripotency beyond the naïve circuit (as represented by the genes in the yellow box in Fig. 2C, which have relatively constant expression in the naïve stage).

Primitive streak markers such as *T* and *Fgf5* were up regulated in cells exiting from the naïve pluripotent state and entering the mesendoderm stage (Fig. 2A, 2B). Shortly after, the endoderm and mesoderm lineages bifurcated. *Fgf5* expression became more enriched in endoderm, while *T* expression was exclusively enriched in mesoderm (Fig. 2A). *Gata6* and *Gsc* were expressed in both mesoderm and endoderm, while more extensively expressed in the latter. Most cells in the endoderm branch were separated from the trunk, with very few *Foxa2*-high expressing cells in the intermediate state. This might reflect that the intermediate state undergoing endoderm specification is relatively unstable. After *Foxa2* expression was initiated (depicted by the red-dash line in Fig. 2D), almost all the important early endoderm transcription factors, including *Gata6*, *Sox17*, *Foxa2*, *Foxq1* and *Lhx1* became highly expressed in the whole endoderm branch (Fig. 2B, 2D). This implies an activation of intertwined strong positive feedback modules of transcription factors, and may explain why the intermediate state in endoderm specification is unstable. Meanwhile, *Dkk1* and *Frzb*, two WNT signaling antagonist genes, and *Cer1*, encoding a BMP4 antagonist, were extensively up regulated throughout the endoderm population (Fig. 2B, 2D), implicating that active exclusion of WNT and BMP4 signaling is important for endoderm establishment. Consistently, a mesendoderm gene module, represented by *T*, and primed *Mesp1* expression, were terminated in the endoderm population (Fig. 2B, 2D). The cell surface markers *Cd24a*, *Epcam* and *Cldn7* were up regulated early in differentiation, while becoming more extensively expressed in the committed endoderm (Fig. 2B, 2D).

Next, we retrieved a continuous process of hemangiogenic lineage development by manually choosing the populations on the shortest path from the naïve pluripotent state to the hemangiogenic branch and re-constituted a fine developmental route (Fig. 3A). A group of *Flk1*-expressing cells achieved the highest level of *Etv2* expression, followed by dramatic up regulation of the ETV2 target genes, *Tal1*(*Scl*) and *Gata2*, confirming that the threshold-level of *Etv2* expression is necessary for hemangiogenic lineage specification[9] (Fig. 3A). We ordered the cells into a 1-D pseudo-time line and assigned the populations to specific stages according to the dynamics of marker gene expression (Fig. 3B). 68 most variable transcription regulation-related factors were clustered based on their expression along the pseudo-time line (Fig. 3C). They aggregated into a few obvious modules covering different developmental stages (yellow dash line boxed regions in the heatmap). Notably, the naïve pluripotency module (box II) and the hemangiogenic module (box III) were relatively independent, while each of the intermediate gastrulation modules were shared by adjacent stages, displaying 'relay'-like patterns (Fig. 3C). This stepwise combinatorial usage of the regulatory circuits fits the gastrulation's role in specifying diverse lineages. Consistently, the intermediate populations were uncommitted, largely interchangeable[14], and similar to each other in transcriptome (Fig. 3D). Therefore, exit from the naïve pluripotency and activation of the hemangiogenic transcription program are the two most critical transitions in hemangiogenic lineage development, with the intermediate states constituting a multistep specification process.

Due to intracellular noises and microenvironment influences, gene expression within individual cells keeps fluctuating. scRNA-seq analysis offers an opportunity to utilize this luxuriant gene expression heterogeneity to assess a regulatory network. We previously reported that *Etv2* expression above a threshold-level is essential to activate hemangiogenic genes[9]. We confirmed this at a single cell endogenous mRNA level (Fig. 4A). Importantly, we noticed that upregulation of ETV2 target genes soon became extremely steep after the *Etv2* threshold expression level was achieved. We explored potential genes required for *Etv2* to trigger this sensitive switch by examining potential upstream regulators of *Tal1* and *Lmo2*, two ETV2 direct target genes that are critical for establishing the hemangiogenic lineage[15, 16]. First, we assumed that the possibility of gene A-expressing cells while simultaneously expressing gene B is correlated with the dependence of gene A expression on gene B (Fig. 4B, see methods for more details). From the population where ETV2 starts to activate hemangiogenic genes (population 12 in Fig. 3A), we assessed the possibility of a given gene being required for *Tal1* or *Lmo2*'s expression (Fig. 4C, 4D). *Etv2* and *Kdr/Flk1* lie at the top of the list of genes required for *Tal1* or *Lmo2* expression (Fig. 4C, 4D, supplemental Table 2), consistent with the finding that VEGF-FLK1 signaling ensures *Etv2* threshold expression to activate hemangiogenic genes[9]. The comparison of the top 100 genes required for *Tal1* expression and the top 100 required for *Lmo2* expression generated 66 in common, of which 22 were annotated to cell adhesion-related signaling (Fig. 4E, highlighted in supplemental Table 2). Most of these 22 genes were already highly expressed before entering the hemangiogenic stage (Fig. 4F). To test if local cell adhesion signaling is required for *Etv2*-mediated hemangiogenic gene activation, we utilized doxycycline (DOX)-inducible *Etv2*-expressing ES cell line[17]. Since extracellular matrix/cell adhesion signaling converges on the SRC kinase[18], we added SRC inhibitor PP2 to day 3 EBs, while simultaneously inducing exogenous ETV2 protein expression. After 24 hours, as expected, DOX-induced *Etv2* expression greatly enhanced the FLK1$^+$PDGFRα^- hemangiogenic population[17]. Remarkably, Inhibition of SRC using PP2 was sufficient to reduce *Etv2* overexpression-induced hemangiogenic lineage skewing (Fig. 4G). Given that cell adhesion is necessary for gastrulation as well[19], it was possible that PP2 affected earlier stages rather than directly affecting ETV2's function. To minimize this possibility, we ectopically induced *Etv2* expression in undifferentiated ES cells, where endogenous *Etv2* level is ignorable and exogenous *Etv2* can induce *Flk1* expression[9]. It has been established that reciprocal activation between *Etv2* and FLK1 signaling is an important mechanism for hemangiogenic fate determination[9]. Exogenous *Etv2* expression induced more than 3% of FLK1-positive cells in undifferentiated ES cells, while this ratio was significantly reduced by the PP2 treatment (Fig. 4H). These results suggest that cell adhesion signaling regulates ETV2-mediated hemangiogenic fate specification.

Based on the unexpected observation of bifurcations of smooth muscle and hemangiogenic lineages from FLK1 mesoderm, we analyzed the cell populations corresponding to the developmental route of the smooth muscle lineage (Fig. 5A), and visualized expression of the 427 most varied genes along the pseudo-time of differentiation (Fig. 5B). Similar to the transcription factors in Fig. 3C, most of the variable genes aggregated as modules. The modules corresponding to mesendoderm, early mesoderm or *Flk1*-expressing mesoderm stages displayed largely overlapping ‘relay’-like patterns (summarized in the bottom plots of Fig. 5B). Three groups of genes were most extensively expressed in the smooth muscle branch (Fig. 5B, gene clusters a, b and c in the yellow dash line boxed region). Unexpectedly, two groups of them (b and c) were already prominently up regulated in *Flk1*$^{+}$ mesoderm, with cluster c being up regulated even earlier from the mesendoderm stage. The transcription factors *Hand1*, *Tbx20*, and *Foxf1* were among the gene cluster b (Fig. 5B, 5C). Of these, *Hand1* and *Foxf1* have been reported to be important for smooth muscle development[20, 21]. These two clusters of genes were not further up regulated in the hemangiogenic branch. In contrast, only a small cluster of genes was exclusively up regulated after cells entered the smooth muscle branch (gene cluster a). These results suggest that smooth muscle lineage could be the default fate of FLK1 mesoderm.

We zoomed into the populations 0, 3, 10, 12 and 13 in Fig. 1C, corresponding to the bifurcation of hemangiogenic and smooth muscle fates (Fig. 5D). BMP4 can enhance smooth muscle development [22]. *Bmp4* was dramatically up regulated in the smooth muscle branch (from black to orange), but not in population 12 (green), where *Etv2* starts to specify hemangiogenic fate (Fig. 5D). *Hand1* and *Foxf1* were dramatically down regulated in population 12 compared to the upstream population 3, implying an active repression of them by *Etv2* (Fig. 5E). Consistently, by analyzing Wareing and colleagues' work[23] we found that re-expression of *Etv2* in *Etv2*-knockout EBs inhibited expression of *Hand1* and *Foxf1* (Fig. 5F). These results suggest that *Etv2* represses the default smooth muscle program when specifying the hemangiogenic lineage from FLK1 mesoderm.

The lineage relationships among cardiac, smooth muscle, skeletal muscle, and hemangiogenic tissues concerning FLK1 mesoderm remain elusive. Our single cell transcriptome profiling of EB cells representing a continuous differentiation trajectory revealed that the cardiac branch partially overlaps with *Flk1*$^+$ mesoderm, while largely segregates out from *Pdgfr*α^+*Flk1*$^-$ early mesoderm. An unexpected finding is that the smooth muscle and hemangiogenic lineages have the closest developmental relationship and directly bifurcate from a common group of *Flk1*-expressing cells. Importantly, most *Flk1*$^+$ mesoderm stage markers were further up regulated in the smooth muscle branch, suggesting that smooth muscle might be a default fate of the *Flk1*-expressing mesoderm. We propose that *Etv2* drives the branching of the hemangiogenic lineage from the smooth muscle fate. Expression of cell adhesion signaling seems required for ETV2 target gene activation for hemangiogenic lineage specification. Consistently, inhibition of SRC, a kinase important for cell adhesion signaling, was sufficient to repress hemangiogenic lineage skewing induced by *Etv2* overexpression. Meanwhile, critical smooth muscle transcription factors, *Hand1* and

Foxf1, become actively down regulated. Notably, neither *Hand1* nor *Foxf1* are direct ETV2 target genes[16]. It will be critical in the future to explore how this repression is achieved.

Hemangiogenic lineage development can be clearly annotated by a series of transcription factor modules, therefore providing strong molecular support for current definition of early embryo developmental stages (Fig. 3C). Cells in the gastrulation stages after exiting pluripotency and before entering hemangiogenic state are overall similar to each other in transcriptome and the underlying transcription factor modules show a ‘relay’-like highly overlapping pattern. Consistently, FLK1$^-$PDGFRα^+, FLK1$^+$PDGFRα^+ and FLK1$^-$PDGFRα^- populations in early differentiation of ES cells are reversible/interchangeable[14]. These results suggested that exit from naïve pluripotency and activation of the hemangiogenic program are the two rate-limiting steps in hemangiogenic lineage development, while the intermediate gastrulating stages are plastic and adapt for stepwise specification to multiple lineage fates by combinational usage of limited regulatory circuits (Fig. 5G).

In summary, our work described a continuous process of mesoderm development leading to hemangiogenic lineage emergence and the underlying molecular networks, thereby providing comprehensive information on early embryo and blood/endothelium development. These are fundamental for the study of basic cell fate determination and gene network structures, also for designing more effective strategies to generate hematopoietic and endothelial cells for regenerative medicine.

Methods

Mouse ES cell culture and differentiation

Mouse ES cells were maintained and for EB differentiation in serum as previously reported[9]. Briefly, ES cells were maintained on mouse embryo fibroblast (MEF) feeder cell layers in Dulbecco-modified Eagle medium containing 15% fetal bovine serum, 100 units/mL LIF, 1× MEM Non-Essential Amino Acids Solution (Gibco), 1× Glutamax^{TM} Supplement (Gibco), and 4.5×10^{-4} M 1-Thioglycerol (MTG, Sigma). For feeder-free culture, ES cells were maintained in gelatin-coated dish in Iscove's modified Dulbecco medium (IMDM) with the same supplements as used for maintenance on feeder. For EB differentiation, ES cells were first transferred into feeder-free condition for 2 days, then single-cell suspensions were prepared, and 8,000 cells were added per mL to a differentiation medium of IMDM containing 15% differentiation-screened fetal calf serum, 1× Glutamax, 50 μ g/mL ascorbic acid, and 4.5×10^{-4} M MTG on a bacteriological Petri dish. The SRC inhibitor PP2 (Sigma) and Doxycycline (Sigma) treatment were performed as indicated in the text.

Single cell RNA-seq and data analysis

Day 4 EBs were dissociated with Accutase solution (Sigma). Single cell suspension at 300 cells/μL were subjected to the Chromium 10X Genomics library construction and HiSeq2500 sequencing (The Genome Technology Access Center, Washington University in St. Louis). The sequenced reads were mapped to the GRCm38 assembly using Cell Range 2.0.1 (10x Genomics). The output was imported into Seurat 1.4 [24], and genes expressed in at least 3 cells were kept for analysis. Cells with more than 5% mitochondria reads or less than 2000 unique genes detected were filtered out. 1848 cells were kept for further analysis, which had 4373 genes/27272 unique mRNAs detected per cell in average. The remaining cells were clustered into 18 populations, based on the expression of 427 most variable genes. Cells in selected populations were imported into Monocle 2[25] for re-clustering, PCA plotting and pseudo-time ordering. For hierarchical clustering and heatmap plotting of selected genes or cells, the R package 'pheatmap' was used. All scRNA-seq data analyses were finished in R environment.

To identify genes required for ETV2-target gene activation, we assumed that if gene B is required for gene A expression, then in a cell where gene A is expressed gene B should also be expressed. In a population, more gene A expressing-cells have gene B expression, it is more likely that gene B is required for gene A expression. In fact close gene-gene correlations can root from three possibilities, take *Tal1* as an example: 1) gene x lies upstream of *Tal1* and is required for its expression, 2) gene x is TAL1's target, 3) both gene x and *Tal1* are ETV2's target, thus respond to ETV2 in a similar way. In the population we chose where*Tal1* just starts to be expressed, it is not likely that TAL1activates its own target genes yet, therefore reduces possibility 2. We can refer to a gene's expression pattern in the differentiation trajectory to test possibility 3. In fact, most of the genes tightly following *Tal1* and *Lmo2* were already extensively activated before ETV2 starts to function.

Microarray data analysis

Wareing and colleagues' microarray data[23] was re-analyzed using the Expression Console software (Affymetrix).

Gene functional annotation

Gene funcitonal annotation was performed using Metascape (http://metascape.org/gp/index.html)[26].

Flow cytometry

Single cells were incubated with α-mouse FLK1 (BioLegend) and α-mouse PDGFRα (BioLegend). Data were acquired on LSR-Fortessa flow cytometer (BDbiosciences) and analyzed using the FlowJo (Treestar) software.

Statistical analysis

The results of flow cytometry were analyzed using Students' t test. $P < 0.05$ was considered significant.

Acknowledgements

We thank members of the Choi lab for helpful discussion and support. The work was supported by NIH grant HL55337.

Figure legend

Figure 1. Single cell RNA-seq analysis of mouse embryoid bodies.

(A) Experimental scheme of scRNA-seq. In a sub-optimal differentiation condition, residual pluripotent cells reside in the center of EBs, while differentiated cells are prone to localize close to the peripheral parts. Coexistence of pluripotent cells and differentiated cells imply a continuous spectrum of cell states in differentiation process, which can be detected by scRNA-seq. (B) Distribution of genes and transcripts detected in individual cells. (C) Representative marker gene expression in each cluster. Only at most 10 representative marker genes are displayed for each cluster. (D) t-SNE projection of all cells. Cluster numbers is corresponding to that in (C), with different color for each cluster. Potential lineage branches were annotated according to specific marker gene expression patterns. (E-K) Expression of the indicated marker genes for specific cell states/lineage fates in the t-SNE.

Figure 2. scRNA-seq captures exit from naïve pluripotency and bifurcation of endoderm and mesoderm.

(A) Re-clustering of populations in exit from naïve pluripotency and bifurcation of endoderm and mesoderm. Clusters 4, 5, 8, 9, 11 and 17 in Fig. 1C were picked to re-cluster based on Principle Component Analysis. The first two components are shown. The clusters are in the same colors as they are in Fig. 1C. “naïve”, naïve pluripotency. (B) Cells in the endoderm differentiation route (red dash line-surrounded region) were ordered into pseudo-time line. Arrows indicate the differentiation direction. Heatmap of 163 differentially expressed genes are shown. Related genes were labelled on the right. Brown box encircles genes heterogeneously expressed in the naïve stage; yellow box marks the constantly expressed naïve stage-specific genes. (C) Barplot of expression of selected marker genes along the endoderm pseudotime line. Shaded region covers cells maintaining high *Dppa3* expression in the naïve population. Blue dash line marks the exit point from naïve pluripotency. (D) Barplot of expression of marker genes along the pseudotime line of endoderm differentiation. Red dash line marks the point, where endoderm fate is committed.

Figure 3. A continuous development process of hemangiogenic lineage from pluripotent cells.

(A) Re-clustering of populations in the route of hemangiogenic lineage development. Clusters 3, 4, 8, 9, 10, 12 and 17 in Fig. 1C were picked to re-cluster. The first two components of PCA are shown for re-clustered populations. The clusters are in the same colors as they are in Fig. 1C. Arrows indicate the differentiation direction. Expression of the representative marker genes in the re-clustered populations is shown. (B) The re-clustered cells are ordered into a pseudo-time line of differentiation. Expression dynamics of representative marker genes along the pseudo-time line is shown. Gene expression data was smoothened. The expression curves and the gene names are in same colors. At the bottom, color bars indicate the cells' original identities as in Fig. 1C. Cell states were annotated: "naïve, naïve pluripotent state; "mesendo", mesendoderm; "early meso", early mesoderm; "FLK1 meso", FLK1 mesoderm; "hemangio", hemangiogenic lineage. (C) Dynamics of transcription-related factors from the 427 most variable genes are shown along the pseudo-time line of hemangiogenic lineage development. The genes are hierarchically clustered based on Pearson's correlation. Boxes in a yellow dash-line indicate the obvious stage-specific gene modules. (D) Cell-cell Pearson's correlations along the pseudo-time line based on expression of 427 most variable genes. The brown dash boxes mark the three relatively isolated naïve pluripotent population, 'gastrulation' stage, and hemangiogenic lineage.

Figure 4. *Etv2* triggers a sensitive switch initiating hemangiogenic program, depending on cell adhesion signaling.

(A) *Etv2* target genes sensitively respond to *Etv2* dosage above the threshold. The input data was from smoothed and re-scaled expression of related genes along the pseudo-time line of hemangiogenic lineage development. Datasets only before *Etv2* achieves its highest level were displayed, because after that *Etv2* starts to be down regulated. (B) Scheme for identification of potential regulators of *Etv2*-mediated hemangiogenic fate specification. We assumed that if gene B is required for gene A expression, then in a cell where gene A is expressed gene B should also be expressed. In a population, more gene A expressing-cells have gene B expression, it is more likely that gene B is required for gene A expression. (C) All genes' relationship with *Tal1*. x-axis is the possibility of *Tal1* expression requires a given gene x in group 12 in Fig. 1C, where *Etv2* achieves threshold expression and starts to activate hemangiongenic genes. y-axis is the possibility of *Tal1* in all cells in return is required for a specific gene's expression. *Etv2* and *Kdr/Flk1*, two known upstream regulators of *Tal1*, are labeled in the plot. (D) Comparison of genes required for *Tal1* expression to that required for *Lmo2* or *Eomes*. *Eomes* is a FLK1 mesoderm-expressing transcription factor but not reported regulating hemangiogenic genes. For the top 100 genes that most tightly required for *Tal1* expression in the cluster 12 (dots in shadow regions), the Pearson's correlation coefficient (*r*) between x-axis and y-axis are shown respectively. Genes of interest are shown in color. House keeping genes *Actb* (beta-actin) and *Gapdh* are shown as control. (E) Top 100 genes that are most required for *Tal1* expression and the top 100 most required for *Lmo2* expression have 66 overlap. These 66 genes were analyzed for functional enrichment. (F) Expression of 22 cell adhesion-related genes along the pseudo-time line of hemangiogenic lineage development. The expression values were smoothened for visualization. (G) A2lox ES cells with doxycycline (DOX)-inducible exogenous *Etv2* expression were differentiated in serum medium for 3 days, EBs were then treated with or without the SRC kinase inhibitor PP2 (5μM) for another 24 hours in the presence of 2μg/mL of DOX. EBs without PP2 or DOX treatment were used as control. EBs were collected and analyzed using flow cytometry for surface marker examination. (H) A2lox ES cells with DOX-inducible *Etv2* were transferred to feeder-free condition and cultured in the presence of 2μg/mL of DOX, simultaneously treating with or without PP2 (5μM) for 24 hours. Cells were then analyzed for FLK1 expression using flow cytometry. The statistics summary is shown on the right. ***, P<0.001.

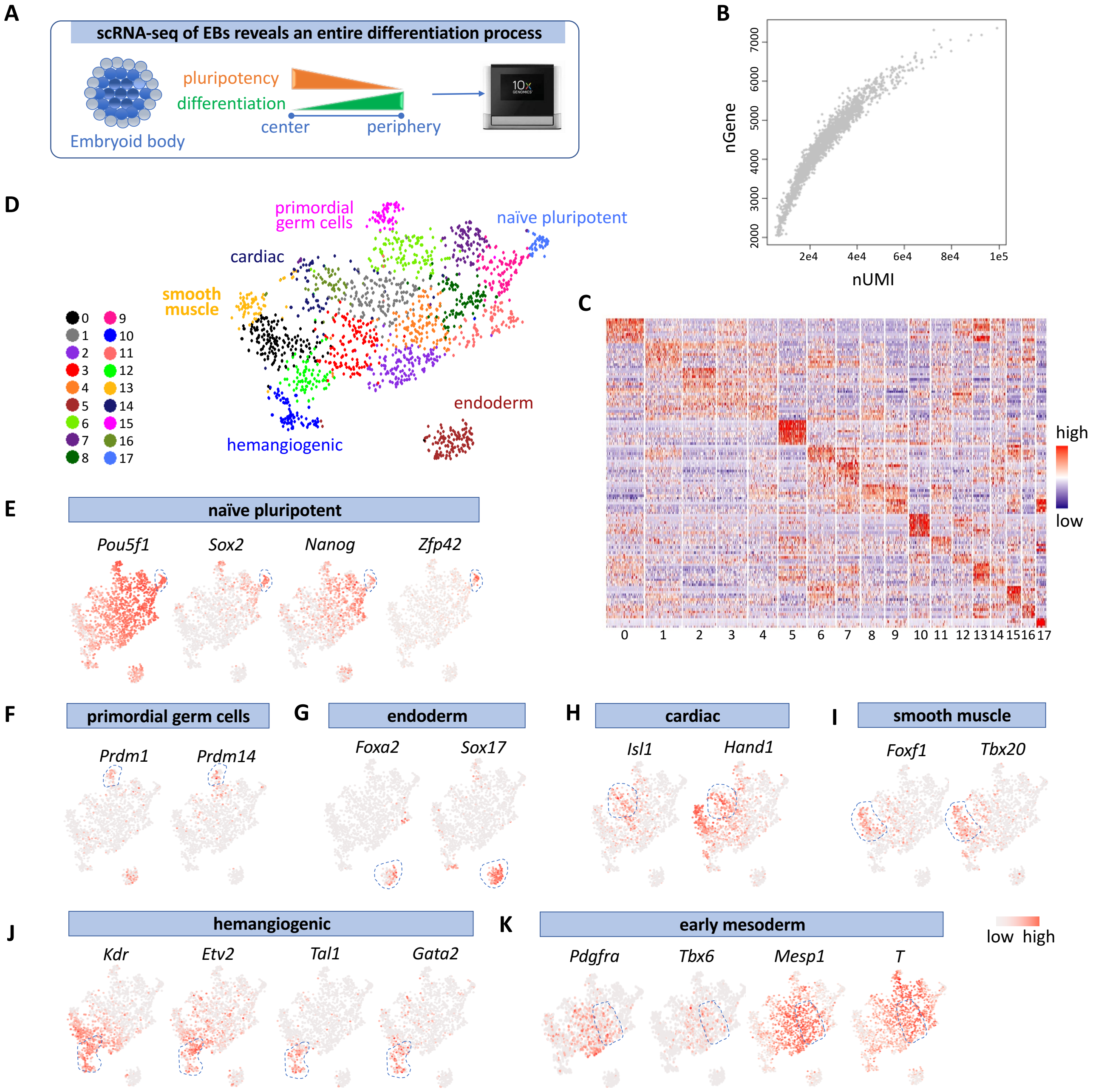
Figure 5. *Etv2*-mediated branching of hemangiogenic lineage from smooth muscle-prone FLK1 mesoderm. (A) Reclustering of populations in the route of smooth muscle development. Clusters 0, 3, 4, 8, 9, 13 and 17 in Fig. 1C were picked, and the first components of PCA are shown. (B) Pseudo-time ordering of the reclustered cells. The comparison of the most variable 427 genes' dynamics along the smooth muscle or hemangiogenic routes is shown. The genes were clustered based on Pearson's correlation. The cells corresponding to columns were ordered along the pseudo-time line. Color bars indicate cells' original identities. Plots at the bottom summarize the gene modules in the heatmap. Labels at the bottom annotate gene groups. "a", "b", and "c", shown in a yellow dash-box, mark three groups of genes that are mostly enriched in the smooth muscle branch. (C) Expression dynamics of indicated genes along the pseudo-time lines of hemangiogenic and smooth muscle lineage development is shown. The expression values were smoothened. The colors of curves and gene names are consistent. (D) Gene expression at the fate fork. Clusters 0, 3, 10, 12 and 13, corresponding to the smooth muscle/hemangiogenic lineage bifurcation fork in Fig. 1C were picked and reclustered. Cells are colored according to the expression levels of indicated genes. (E) Violin plots comparing indicated genes' expression in different populations. **, P<0.01; ***, P<0.001. (F) Important smooth muscle transcription factors respond to Etv2 overexpression. Day 2.5 *Etv2*-knockout EBs were induced for exogenous *Etv2* expression using DOX. Two replicates of samples corresponding to DOX

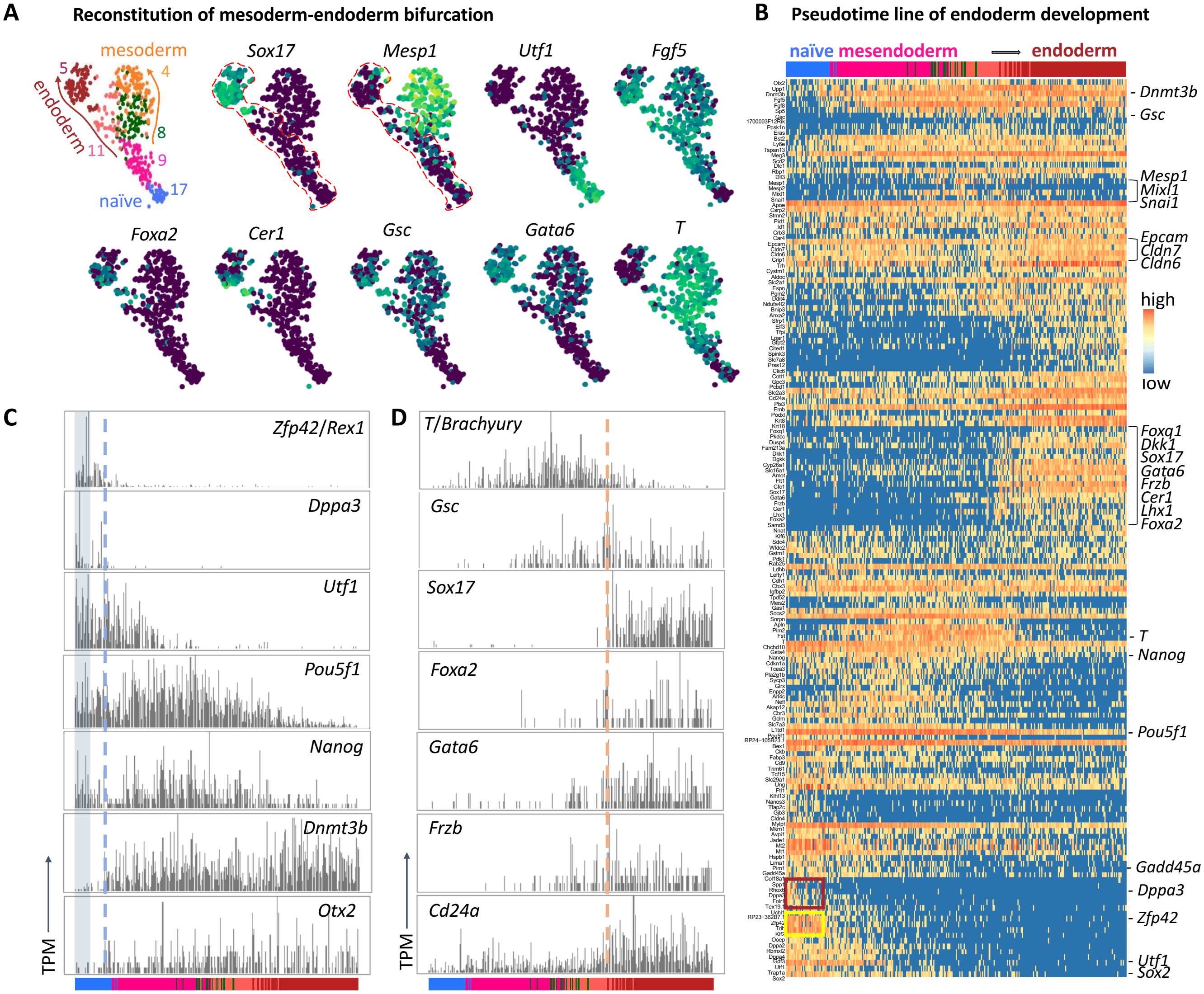
induction for 0 hour, 12 hours or 24 hours were collected for microarray analysis. (G) Model for the hemangiogenic lineage development. Note that exit from naïve pluripotency and hemangiogenic fate determination are the two rate-limiting steps in the process (the two valves). The intermediate states are largely reversible and plastic, adapting multiple fate specifications.

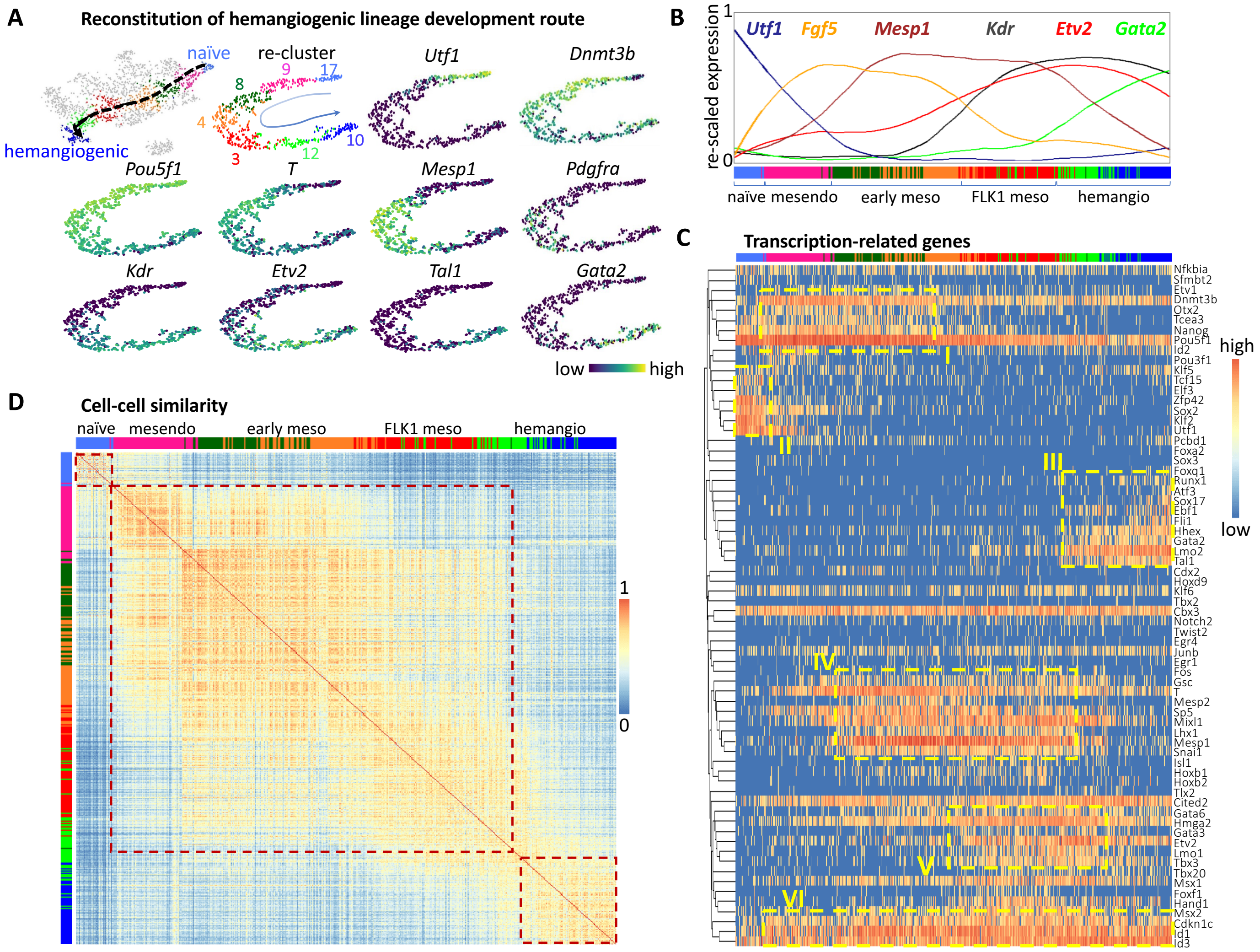
References

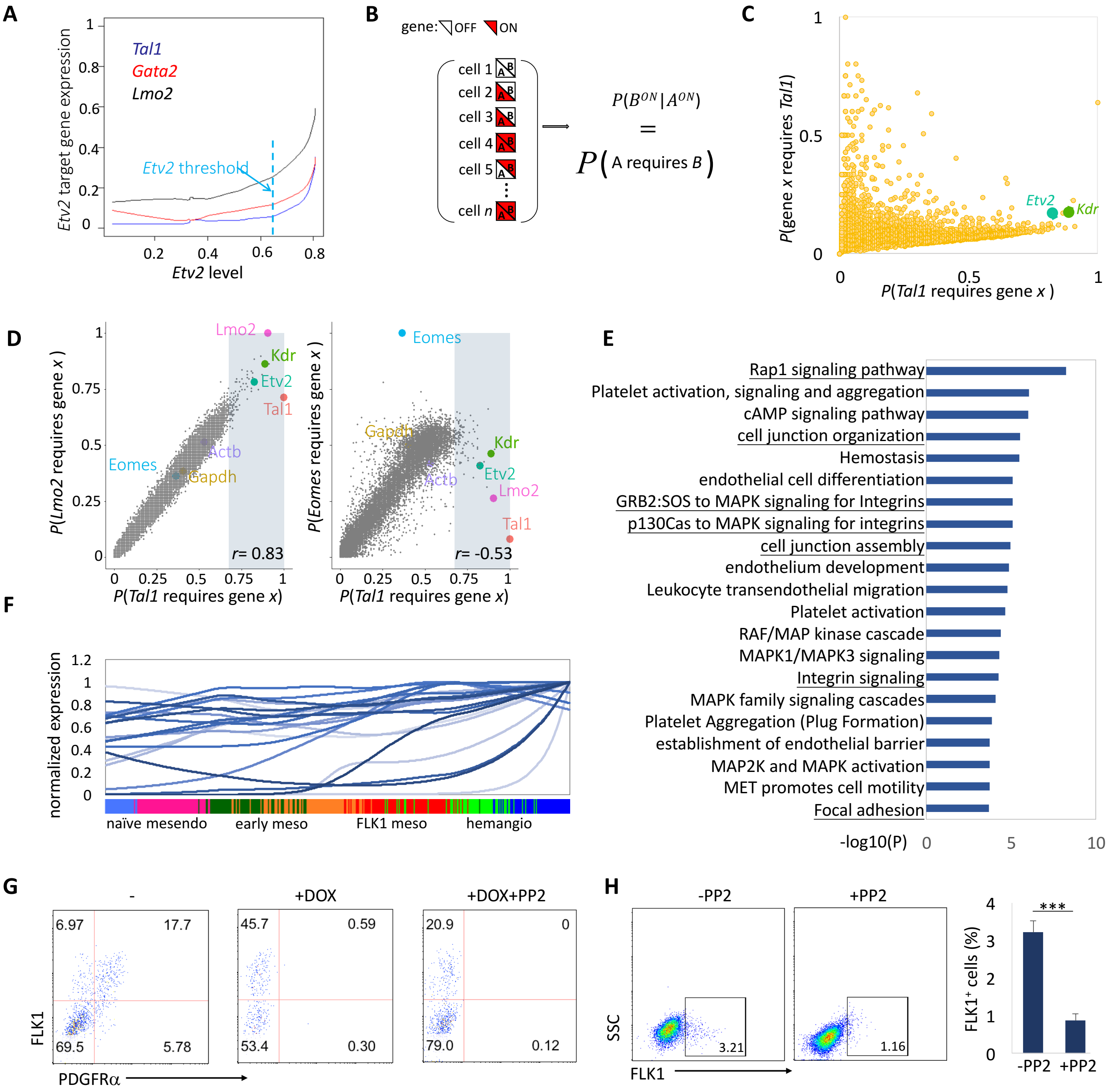
1. Scialdone, A. *et al.* Resolving early mesoderm diversification through single-cell expression profiling. *Nature* **535**, 289-293 (2016).
2. Mohammed, H. *et al.* Single-Cell Landscape of Transcriptional Heterogeneity and Cell Fate Decisions during Mouse Early Gastrulation. *Cell reports* **20**, 1215-1228 (2017).
3. Farrell, J.A. *et al.* Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* (2018).
4. Han, X. *et al.* Mapping human pluripotent stem cell differentiation pathways using high throughput single-cell RNA-sequencing. *Genome biology* **19**, 47 (2018).
5. Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**, 1091-1107 e1017 (2018).
6. Ibarra-Soria, X. *et al.* Defining murine organogenesis at single-cell resolution reveals a role for the leukotriene pathway in regulating blood progenitor formation. *Nature cell biology* **20**, 127-134 (2018).
7. Lescroart, F. *et al.* Defining the earliest step of cardiovascular lineage segregation by single-cell RNA-seq. *Science* **359**, 1177-1181 (2018).
8. Chu, L.F. *et al.* Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome biology* **17**, 173 (2016).
9. Zhao, H. & Choi, K. A CRISPR screen identifies genes controlling Etv2 threshold expression in murine hemangiogenic fate commitment. *Nature communications* **8**, 541 (2017).
10. Lugus, J.J., Park, C., Ma, Y.D. & Choi, K. Both primitive and definitive blood cells are derived from Flk-1+ mesoderm. *Blood* **113**, 563-566 (2009).
11. Park, K.S. *et al.* Transcription elongation factor Tcea3 regulates the pluripotent differentiation potential of mouse embryonic stem cells via the Lefty1-Nodal-Smad2 pathway. *Stem cells* **31**, 282-292 (2013).
12. Motoike, T., Markham, D.W., Rossant, J. & Sato, T.N. Evidence for novel fate of Flk1+ progenitor: contribution to muscle lineage. *Genesis* **35**, 153-159 (2003).
13. Kalkan, T. & Smith, A. Mapping the route from naive pluripotency to lineage specification. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **369** (2014).
14. Sakurai, H. *et al.* In vitro modeling of paraxial and lateral mesoderm differentiation reveals early reversibility. *Stem cells* **24**, 575-586 (2006).
15. Porcher, C. *et al.* The T cell leukemia oncoprotein SCL/tal-1 is essential for development of all hematopoietic lineages. *Cell* **86**, 47-57 (1996).
16. Liu, F. *et al.* Induction of hematopoietic and endothelial cell program orchestrated by ETS transcription factor ER71/ETV2. *EMBO reports* **16**, 654-669 (2015).
17. Liu, F. *et al.* ER71 specifies Flk-1+ hemangiogenic mesoderm by inhibiting cardiac mesoderm and Wnt signaling. *Blood* **119**, 3295-3305 (2012).
18. Cabodi, S., del Pilar Camacho-Leal, M., Di Stefano, P. & Defilippi, P. Integrin signalling adaptors: not only figurants in the cancer story. *Nature reviews. Cancer* **10**, 858-870 (2010).
19. Adams, J.C. & Watt, F.M. Regulation of development and differentiation by the extracellular matrix. *Development* **117**, 1183-1198 (1993).
20. Mahlapuu, M., Ormestad, M., Enerback, S. & Carlsson, P. The forkhead transcription factor Foxf1 is required for differentiation of extra-embryonic and lateral plate mesoderm. *Development* **128**, 155-166 (2001).
21. Morikawa, Y. & Cserjesi, P. Extra-embryonic vasculature development is regulated by the transcription factor HAND1. *Development* **131**, 2195-2204 (2004).

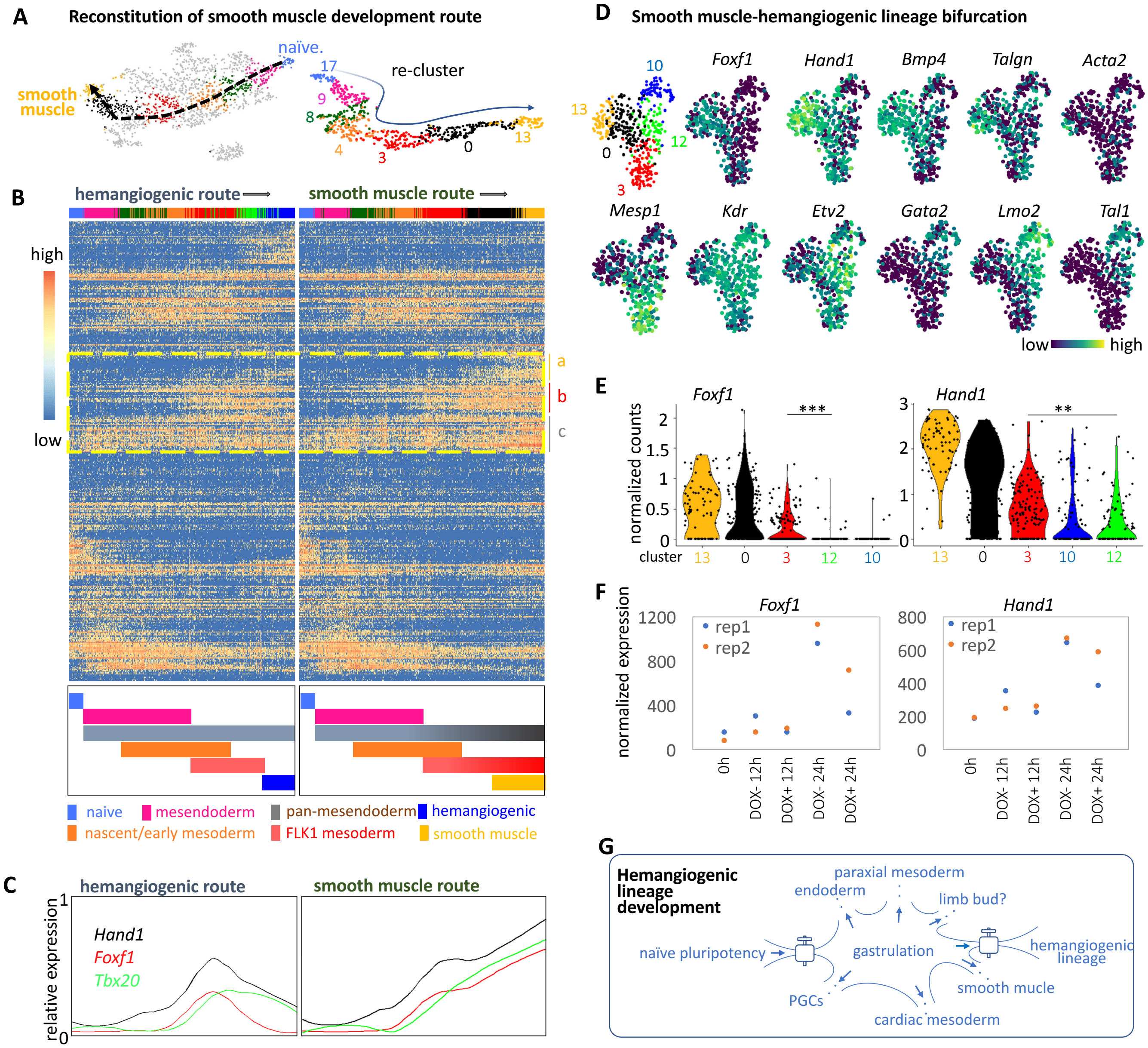
22. Cheung, C., Bernardo, A.S., Trotter, M.W., Pedersen, R.A. & Sinha, S. Generation of human vascular smooth muscle subtypes provides insight into embryological origin-dependent disease susceptibility. *Nature biotechnology* **30**, 165-173 (2012).
23. Wareing, S. *et al.* The Flk1-Cre-mediated deletion of ETV2 defines its narrow temporal requirement during embryonic hematopoietic development. *Stem cells* **30**, 1521-1531 (2012).
24. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology* **36**, 411-420 (2018).
25. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology* **32**, 381-386 (2014).
26. Tripathi, S. *et al.* Meta- and Orthogonal Integration of Influenza "OMICs" Data Defines a Role for UBR4 in Virus Budding. *Cell host & microbe* **18**, 723-735 (2015).


A
scRNA-seq of EBs reveals an entire differentiation process
pluripotency
differentiation
center
periphery
Embryoid body
10x GENOMICS
B
nGene
nUMI
2000
3000
4000
5000
6000
7000
2e4
4e4
6e4
8e4
1e5
C
high
low
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
D
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
primordial germ cells
naïve pluripotent
cardiac
smooth muscle
endoderm
hemangiogenic
E
naïve pluripotent
Pou5f1
Sox2
Nanog
Zfp42
F
primordial germ cells
Prdm1
Prdm14
G
endoderm
Foxa2
Sox17
H
cardiac
Isl1
Hand1
I
smooth muscle
Foxf1
Tbx20
J
hemangiogenic
Kdr
Etv2
Tal1
Gata2
K
early mesoderm
Pdgfra
Tbx6
Mesp1
T
low high


A
Reconstitution of mesoderm-endoderm bifurcation
mesoderm
endoderm
naïve
5
4
8
11
9
17
Sox17
Mesp1
Utf1
Fgf5
Foxa2
Cer1
Gsc
Gata6
T
B
Pseudotime line of endoderm development
naïve mesendoderm
endoderm
Otx2
Upp1
Dnmt3b
Fgf5
Fgf8
Sp5
Gsc
1700003F12Rik
Pcsk1n
Eras
Bst2
Ly6e
Tspan13
Meg3
Scd2
Dlc1
Rbp1
Dll3
Mesp1
Mesp2
Mixl1
Snai1
Apoe
Csrp2
Stmn2
Pid1
Id1
Crb3
Car4
Epcam
Cldn7
Cldn6
Crip1
Trh
Cystm1
Aldoc
Slc2a1
Espn
Pgm2
Ddit4
Ndufa4l2
Bnip3
Anxa2
Sfrp1
Elf3
Tfpi
Lpar1
Gfpt2
Cited1
Spink3
Slc7a8
Prss12
Clic6
Cotl1
Gpc3
Pcbd1
Slc2a3
Cd24a
Pls3
Emb
Podxl
Krt8
Krt18
Foxq1
Pkdcc
Dusp4
Fam213a
Dkk1
Dgkk
Cyp26a1
Slc16a1
Amot
Flt1
Cfc1
Sox17
Gata6
Frzb
Cer1
Lhx1
Foxa2
Samd3
Nnat
Klf6
Sdc4
Wfdc2
Gstm1
Pdk1
Rab25
Ldhb
Lefty1
Cdh1
Cbx3
Igfbp2
Tpd52
Meis2
Gas1
Socs2
Snrpn
Apln
Pim2
Fst
T
Chchd10
Gsta4
Nanog
Cdkn1a
Tcea3
Pla2g1b
Sycp3
Glrx
Enpp2
Arl4c
Nefl
Akap12
Cbr3
Gclm
Slc7a3
L1td1
Pou5f1
RP24−105B23.1
Bex1
Ckb
Fabp3
Cd9
Trim61
Tcf15
Slc29a1
Ung
Ftl1
Klhl13
Nanos3
Tfap2c
Gjb3
Cldn4
Mylpf
Mkrn1
Avpi1
Jade1
Mt2
Mt1
Hspb1
Lima1
Pim1
Gadd45a
Col18a1
Spp1
Rhox6
Dppa3
Folr1
Tex19.1
Uchl1
RP23−362B7.1
Zfp42
Tdh
Klf2
Ooep
Dppa2
Rbmxl2
Dppa4
Gdf3
Utf1
Trap1a
Sox2
- Dnmt3b
- Gsc
Mesp1
Mixl1
Snai1
Epcam
Cldn7
Cldn6
high
low
Foxq1
Dkk1
Sox17
Gata6
Frzb
Cer1
Lhx1
Foxa2
- T
- Nanog
- Pou5f1
- Gadd45a
- Dppa3
- Zfp42
- Utf1
- Sox2
C
Zfp42/Rex1
Dppa3
Utf1
Pou5f1
Nanog
Dnmt3b
Otx2
TPM
D
T/Brachyury
Gsc
Sox17
Foxa2
Gata6
Frzb
Cd24a
TPM


A
Reconstitution of hemangiogenic lineage development route
naïve
hemangiogenic
re-cluster
8
9
17
4
3
12
10
Utf1
Dnmt3b
Pou5f1
T
Mesp1
Pdgfra
Kdr
Etv2
Tal1
Gata2
low
high
B
re-scaled expression
1
0
Utf1
Fgf5
Mesp1
Kdr
Etv2
Gata2
naïve
mesendo
early meso
FLK1 meso
hemangio
C
Transcription-related genes
Nfkbia
Sfmbt2
Etv1
Dnmt3b
Otx2
Tcea3
Nanog
Pou5f1
Id2
Pou3f1
Klf5
Tcf15
Elf3
Zfp42
Sox2
Klf2
Utf1
Pcbd1
Foxa2
Sox3
Foxq1
Runx1
Atf3
Sox17
Ebf1
Fli1
Hhex
Gata2
Lmo2
Tal1
Cdx2
Hoxd9
Klf6
Tbx2
Cbx3
Notch2
Twist2
Egr4
Junb
Egr1
Fos
Gsc
T
Mesp2
Sp5
Mixl1
Lhx1
Mesp1
Snai1
Isl1
Hoxb1
Hoxb2
Tlx2
Cited2
Gata6
Hmga2
Gata3
Etv2
Lmo1
Tbx3
Tbx20
Msx1
Foxf1
Hand1
Msx2
Cdkn1c
Id1
Id3
I
II
III
IV
V
VI
high
low
D
Cell-cell similarity
naïve
mesendo
early meso
FLK1 meso
hemangio
1
0


A
Tal1
Gata2
Lmo2
Etv2 threshold
Etv2 target gene expression
Etv2 level
0
0.2
0.4
0.6
0.8
1
B
gene: OFF ON
cell 1
cell 2
cell 3
cell 4
cell 5
cell n
A
B
$P(B^{ON}|A^{ON})$
=
$P($ A requires B $)$
C
P(gene x requires Tal1)
P(Tal1 requires gene x)
Etv2
Kdr
0
0.5
1
D
P(Lmo2 requires gene x)
P(Eomes requires gene x)
P(Tal1 requires gene x)
Lmo2
Kdr
Etv2
Tal1
Actb
Eomes
Gapdh
r= 0.83
r= -0.53
0
0.25
0.5
0.75
1
E
Rap1 signaling pathway
Platelet activation, signaling and aggregation
cAMP signaling pathway
cell junction organization
Hemostasis
endothelial cell differentiation
GRB2:SOS to MAPK signaling for Integrins
p130Cas to MAPK signaling for integrins
cell junction assembly
endothelium development
Leukocyte transendothelial migration
Platelet activation
RAF/MAP kinase cascade
MAPK1/MAPK3 signaling
Integrin signaling
MAPK family signaling cascades
Platelet Aggregation (Plug Formation)
establishment of endothelial barrier
MAP2K and MAPK activation
MET promotes cell motility
Focal adhesion
-log10(P)
0
5
10
F
normalized expression
0
0.2
0.4
0.6
0.8
1
1.2
naïve mesendo
early meso
FLK1 meso
hemangio
G
-
+DOX
+DOX+PP2
6.97
17.7
69.5
5.78
45.7
0.59
53.4
0.30
20.9
0
79.0
0.12
FLK1
PDGFRα
H
-PP2
+PP2
SSC
FLK1
3.21
1.16
FLK1⁺ cells (%)
***
0
1
2
3
4
-PP2
+PP2


A
Reconstitution of smooth muscle development route
smooth muscle
naïve.
17
9
8
4
3
0
13
re-cluster
B
hemangiogenic route
smooth muscle route
high
low
a
b
c
naive
mesendoderm
pan-mesendoderm
hemangiogenic
nascent/early mesoderm
FLK1 mesoderm
smooth muscle
C
hemangiogenic route
smooth muscle route
relative expression
1
0
Hand1
Foxf1
Tbx20
D
Smooth muscle-hemangiogenic lineage bifurcation
10
13
12
0
3
Foxf1
Hand1
Bmp4
Talgn
Acta2
Mesp1
Kdr
Etv2
Gata2
Lmo2
Tal1
low
high
E
Foxf1
Hand1
normalized counts
2
1.5
1
0.5
0
***
3
2
1
0
**
cluster
13
0
3
12
10
13
0
3
10
12
F
normalized expression
Foxf1
1200
800
400
0
rep1
rep2
Hand1
800
600
400
200
0
rep1
rep2
0h
DOX- 12h
DOX+ 12h
DOX- 24h
DOX+ 24h
0h
DOX- 12h
DOX+ 12h
DOX- 24h
DOX+ 24h
G
Hemangiogenic lineage development
paraxial mesoderm
endoderm
limb bud?
naïve pluripotency
gastrulation
hemangiogenic lineage
smooth mucle
PGCs
cardiac mesoderm