A single-cell transcriptome atlas of human early embryogenesis

Yichi Xu[1*], Tengjiao Zhang[2*], Qin Zhou[3], Mengzhu Hu[4], Yao Qi[4], Yifang Xue[3], Lihui Wang[3], Zhirong Bao[1†], Weiyang Shi[5†]

1. Developmental Biology Program, Sloan Kettering Institute, New York, NY 10065, USA.
2. Institute for Regenerative Medicine, Shanghai East Hospital, School of Life Sciences and Technology, Tongji University, Shanghai 200123, China; Developmental Biology Program, Sloan Kettering Institute, New York, NY 10065, USA.
3. Traditional Chinese Medicine Hospital of Kunshan, Suzhou, Jiangsu 215300, China
4. Ministry of Education Key Laboratory of Marine Genetics and Breeding, College of Marine Life Sciences, Ocean University of China, Qingdao 266003, China
5. Ministry of Education Key Laboratory of Marine Genetics and Breeding, College of Marine Life Sciences, Ocean University of China, Qingdao 266003, China; Institute for Regenerative Medicine, Shanghai East Hospital, Tongji University, Shanghai 200123, China

* equal contribution.

† Correspondence: WS (wshi@ouc.edu.cn); ZB (baoz@mskcc.org)

Abstract

The early window of human embryogenesis is largely a black box for developmental biologists. Here we probed the cellular diversity of 4- to 6-week human embryos when essentially all organs are just laid out. Based on over 100,000 single-cell transcriptomes, we generated a comprehensive atlas of 333 cell types that belong to 18 developmental systems, and identified hundreds of cell type specific markers as well as dynamic gene changes. Combined with data of other vertebrates, the rich information shed light on spatial patterning of axes, systemic temporal regulation of developmental progression and potential human-specific regulation. Our study provides a compendium of early progenitor cells of human organs, which can serve as the root of lineage analysis in organogenesis.

Introduction

Human embryogenesis finishes gastrulation by 2.5 week and at 4 week major embryonic organ and tissue types start to differentiate[1]. The transition from gastrulation to organogenesis is marked by sharp increase in cellular diversity generated from early progenitors. It is also at this stage that most developmental defects start to arise which could lead to miscarriage or birth defects[1]. Studies on these processes in vertebrates are mostly carried out in model systems such as the mouse and zebrafish[2–5] but to which degree they are conserved in human embryo is unknown due to technical difficulties and ethical limitations. While single-cell data have been examined for human embryogenesis at later stages, either systematically[6,7] or in an organ-specific manner[8–17], the critical time window of great expansion of cellular diversity remains to be explored. Here, we examine this developmental window by studying human embryos at Carnegie stages (CS) 12~16 (4- to 6-week) and provide a comprehensive single-cell transcriptome atlas of early human embryo.

Comprehensive map of human embryonic developmental systems and cell types

To systematically define the developmental landscape of human organogenesis, we obtained five aborted human embryos from CS12 to 16 (Fig. 1A, Extended Data Fig. 1), and used the 10x Genomics Chromium platform to obtain scRNA expression profiles. To retain the spatial origin for cell types, we dissected the embryos into major parts including head, trunk, viscera, and limb (Extended Data Fig. 1A). In total, we obtained 103,264 cells from 15 embryonic dissection parts. We captured an average of 8,836 transcripts and 2,474 genes per cell (Extended Data Fig. 1B). Based on sex-specific gene expression, we were able to determine the sex of each embryo (2 males and 3 females, Extended Data Fig. 1C).

To identify cell types, we conducted semi-supervised clustering, where knowledge of human and mammalian lineage differentiation was used to optimize *de novo* clustering to best recapitulate the known lineage hierarchy. We curated a total of 179 publications that define the developmental systems (major lineages, organs and tissues) and the known cell types within each system at human CS12 to 16 or the corresponding mouse stages (E9.5 to E11.5), with a total of 227 diagnostic markers. To first resolve the developmental systems and then individual cell types within each, we applied iterative clustering (Extended Data Fig. 2-5) and identified a total of 333 cell types/clusters in 18 developmental systems (Fig. 1B, Extended Data Fig. 2A, Supplementary Table 1). Notably, the mesoderm displays the most tissue-type and transcriptional diversity at this stage. Eight mesodermal systems are present, including two paraxial mesoderm (head mesoderm and somite), the intermediate mesoderm, three lateral plate mesoderm (LPM; limb, somatic LPM and splanchnic LPM), as well as blood and endothelium. Technically, we found that clustering based on transcription factors (TFs) better resolves the developmental systems than using all highly variable genes (HVGs) (Extended Data Fig. 2, Methods), possibly because of convergent expression of genes such as epithelial and extracellular matrix pathways across lineages.

Among the 333 cell types (Fig. 1B), we were able to assign 217 to known cell type identity and another 95 less specifically to a tissue identity (e.g., telencephalon) (Fig. 2A, 2B, Supplementary Table 1). That is, 94% of the 333 cell types have literature support. The identified cell types include small anatomical structures such as known signaling centers and sensory placodes (see below), as well as migrating cell types that would have been difficult to identify without the whole-embryo approach, such as the neural crest derived second heart field (SHF)[18] marked by *ISL1*. Finally, 21 clusters could not be assigned to known cell or tissue types. Notably, 14 of these were from the head mesoderm, reflecting our lack of understanding of this system compared to the others. Thus, our analysis provides a deeply annotated atlas of cell types of human early organogenesis.

In addition to the literature and marker support, we further assessed the quality of our clusters/cell types through a series of quantitative tests. First, because we dissected embryos into different parts, each cell preserves coarse spatial information that can be used to test the quality of cell type identification. We found that more than 99% of all cells came from the expected dissection parts of the annotated cell type (Fig. 2C, Supplementary Table 1). For example, dorsal telencephalon and lateral motor columns were only from head and trunk, respectively, whereas embryo-wide cell types such as endothelium and erythroid cells were from all dissection parts (Fig. 2D). Cell types from dissection boundaries show multiple origins, such as pharyngeal arches. Second, we examined stage distribution of cells in each cell type. Fifteen cell types (5%) may have bias on sampling or clustering because the middle stage CS13 is the only underrepresented stage in these cell types (Extended Data Fig. 6). Third, we calculated pairwise correlation between cell types based on their gene expression (Extended Data Fig. 7) and found that cell types show tiered similarity within and between the 18 developmental systems that is consistent with the known biological relationship.

Finally, the cell types are not biased by batch, sequencing depth, or cell cycle states of cells (Extended Data Fig. 8). These results suggest that computational artifacts are rare in the defined cell types.

To characterize cell types from the perspective of gene expression, we identified 3,390 differentially expressed genes (DEGs) across all cell types (Methods), including 213 (94%) of the canonical markers we collected. The number of DEGs for each cell type ranged from less than 30 in epidermis to more than 200 in neurons (Fig. 2E). To identify cell type specific markers, we ranked DEGs by z-score for each cell type. ~80% of the canonical markers belong to the top 20 most specific DEGs in each cell type (Supplementary Table 1). Many of the top DEGs by z-score have not been identified as cell specific markers in prior studies, e.g., *FOXL2* and *MMP23B* for SHF.

The comprehensive map of cell types and gene expression allows in-depth analysis of human early embryonic development both globally and in individual cell types. For example, we examined the global pattern for five major signaling pathways, namely Fgf, TGF-β, Hedgehog, Notch and Wnt, in terms of their ligands, receptors, and major antagonists[19] (Extended Data Fig. 9). Notably, the nervous system, limb and head mesoderm show the most prevalent expression of pathway components, indicating the most complex scheme of induction and patterning[20,21]. More specifically, we interrogated 9 known signaling centers identified in our data, including the ZPA (zone of polarizing activity) and AER (apical ectodermal ridge) in limb patterning[21], and 7 in the neural tube and brain[22], which consists of less than 1.5% of the cells sampled (Fig. 2F). Our data captured the known signaling molecules in these 9 signaling centers, such as *SHH* in the floor plate[22], *SHH* and *BMP4* in ZPA[21] and *FGF8/9* in AER[21], demonstrating the fidelity of our data. Moreover, we identified 36 additional signaling molecules from these centers[23], ranging from 2 to 13 per signaling center. This result suggests that the patterning function of these centers may be more complex than what is known.

To allow easy access of the expression data, we constructed a web-based database (Human early organogenesis atlas, HEOA, https://heoa.shinyapps.io/base/) for visualization, extraction of cell type specific expression profiles as well as comparison of gene expression differences between cell types.

Spatial patterning of axes from single cell data

Bilaterian embryos consist of three body axes. Axis patterning is a prominent and ubiquitous process at this stage of development as each organ elaborates the three axes via diverse mechanisms. Studies have provided classic paradigms of spatial patterning albeit in model organisms. While scRNA-seq do not contain explicit spatial information, we found that they can be useful in examining axis patterning. Here we examine axis formation in limb and neural tube development in human.

Morphologically, the limb buds start to appear at CS12. During this time, the proximal-distal (PD), anterior-posterior (AP) and dorsal-ventral (DV) axes are established through interactions of signaling gradients along different directions[24]. *In situ* patterns of gene expression have been instrumental in revealing spatial domains of differentiation as cell types are elaborated under the axes. We annotated cell types from our scRNA-seq data with *in situ* results in mouse[25–33] to locate cell types in the forelimb mesenchyme at each stage. Altogether, we defined 6, 10, 14 spatial domains ('domain', a cell type with identified spatial location) along the PD and AP axes for CS12, CS13 and CS15~16 forelimb, respectively (Fig. 3A, Extended Data Fig. 10-12, Methods). The spatial organization of cell types was largely

recapitulated by UMAP, where domains are ordered along the PD and AP axes (Fig. 3B). As *post hoc* validation, the expression of HOXA genes[34], which were not used in annotating domains, showed largely consistent pattern across spatial domains with very few exceptions (Fig. 3C). Notably, at CS15~16, we identified a domain (m) marked by *PTHLH* in the zeugopod territory, which has not been characterized by *in situ* and has its own DEGs (Fig. 3D, Extended Data Fig. 12). Moreover, we found that four domains (a, c, m and n) are at different degrees of chondrogenesis[35,36] (Fig. 3A, D), emphasizing the spatial and temporal heterogeneity during chondrogenesis in limb bud development. Our results provide a fine map of the limb bud with, to our knowledge, the highest resolution as well as a digital *in situ* database for thousands of genes in limb development (see HEOA database).

The neural tube is patterned along the AP axis by HOX genes[37] and the DV axis by opposing signaling gradients to form regionalized cell types[38]. We examined these patterns in human in contrast to those in mouse.

Based on the expression of HOX genes, we reconstructed the AP axis for the neural progenitors of the hindbrain and the spinal cord (Fig. 4A). HOX expression agreed with the spatial origin of cells in terms of the known HOX codes[37]. In addition, cell types along the DV axis were dispersed along the pseudo AP axis, demonstrating that our cell type identification was not affected by the AP gradient of gene expression. We then identified genes whose expression level display an AP gradient. Besides the HOX genes, we identified 16 genes (Fig. 4B). Among these, the *CYP26C1* gene is known to be involved in establishing the retinoic acid gradient that is responsible for activating some of the HOX genes along the AP axis[39]. Notably, these genes include five long non-coding RNAs (lncRNA) situated in the HOX gene clusters, namely *HOTAIRM1*, *RP11-357H14.17*, *RP11-834C11.4*, *RP11-834C11.6* and *FLJ12825*. Other than *HOTAIRM1*, these lncRNAs are unique to the human genome[40], indicating human-specific regulatory networks in neural tube AP patterning. The spatial pattern of *HOTAIRM1* expression in the neural tube has not been reported. We found that its expression is not correlated with the adjacent HOX genes (*HOXA1* and *HOXA2*) as by the general rule of lncRNA expression in the HOX gene cluster[41] and the cases of the other four lncRNAs with AP gradient (Extended Data Fig. 13A). Instead, it appears to be correlated to the more distant *HOXA4*. Thus, our analysis expands the known AP gradient gene expression in vertebrate neural tube patterning and sheds light on human-specific regulation.

Vertebrate neuronal diversity is most evident along the DV axis[20]. We identified all the 13 neural progenitor and 13 neuronal cell types in the spinal cord which can be recognized by the known marker expression in mouse[42] (Fig. 4C). Given the conservation of neuronal cell types, we next asked to what extent homologous cell types have conserved gene expression between mouse and human. For this, we compared our data with a recent scRNA-seq dataset of mouse neural tube[42]. Neurons display virtually the same expression patterns of canonical markers for each type (Extended Data Fig. 13B). In contrast, progenitors show more difference (Extended Data Fig. 13B). Most of the differences are quantitative, *i.e.*, markers display the same DV spatial range and relative expression levels but at different absolute levels of expression, such as *MSX1* and *OLIG3*. A recent scRNA-seq study of human neural tube identified two major TFs that are specific to human, namely *PAX7* and *NKX6*-2, which are confirmed in our dataset (Extended Data Fig. 13C)[43]. More importantly, we found that *MSX2* expression is different between human and mouse in dp2-4 (Fig. 4D). The expression of *MSX2* in dp2-4 is supported by the recent human dataset[43] (Fig. 4E). This difference is not caused by erroneous clustering in human or mouse datasets but reflects true cross-species divergence (Extended Data Fig. 13D). We speculate that *MSX2* likely compensates the function of its paralog *MSX3* in

human neural tube: *MSX3* is expressed in mouse dp2-4 but the gene is lost in the human genome (Fig. 4D). Further experimental studies on transcriptional regulation may uncover the mechanism of active compensation, which is poorly understood[44].

In summary, spatial information in scRNA-seq can shed light on potential new cell types, heterogeneity of lineage differentiation, as well as divergences of developmental regulation cross species.

Systemic temporal regulation of vertebrate embryogenesis

Vertebrate embryogenesis progresses through stages with characteristic developmental events, but it is not clear whether there exist conserved regulatory mechanisms that may control stage transitions. Here we examine systemic gene expression changes in our data in comparison to other vertebrate models, and implicated the potential role of lin28a, a conserved RNA binding protein known to regulate developmental stage transition in *C. elegans*[45,46], in the regulation of vertebrate embryonic stage transition.

While examining temporal gene changes in the early human embryo, we observed a sharp decrease of *LIN28A* expression (>2 fold) from CS12/13 to CS15~16 embryo in 97% cell types (Fig. 5A). Similarly, at the corresponding stages in mouse embryos (E9.5/10.5 vs. E11.5), *Lin28a* showed >2 fold decrease in 90% of reported cell types[2] (Fig. 5B, 5C). Combining published bulk and single-cell RNA-seq in other vertebrate embryos in a broader time window[2,4,5,47–50], we observed a conserved pattern of *LIN28A* expression: *LIN28A* is dramatically up-regulated at the beginning of gastrulation, peaking at early organogenesis, and diminishing at hatching in zebrafish/frog or E11.5/CS15~16 in mouse/human (Fig. 5D). Previous studies of LIN28A protein in mouse[51] and in frog[52] showed protein dynamics of LIN28A is consistent with mRNA dynamics. Thus, LIN28A displays a conserved and systemic dynamics in vertebrate embryogenesis.

Furthermore, we found that gene expression through vertebrate embryogenesis displays distinct temporal boundaries that corresponds to landmarks in development, which we name punctuated developmental stages. Four such stages were revealed in each species by hierarchical clustering based on pairwise correlations of TF expression between timepoints (Fig. 5E, Extended Data Fig. 14A, Supplementary Table 2). In zebrafish and frog, stage 2 corresponds to gastrulation, stage 3 to organogenesis, and hatching at the boundary between stage 3 and stage 4[53,54]. In addition, these 4 stages in zebrafish and frog show strong pairwise correlations of expression of homologous TFs (Extended Data Fig. 14B), which is consistent with the temporal alignment based on two sparsely sampled time series of zebrafish and frog embryos[55]. Similarly, 3 stage boundaries were observed in mouse (Fig. 5E). Interestingly, the beginning of the 4th stage in mouse, E11.5, corresponds to hatching in zebrafish and frog. Strikingly, a previous study based on tissue-level transcriptomes suggested that E11.5 in mouse corresponds to birth in opossum[56]. These results suggest that vertebrate embryogenesis consists of conserved and punctuated stages and raise an intriguing possibility that over the course of evolution of mammals some of the post embryonic development was shifted in uterus.

LIN28A dynamics is correlated with these punctuated stages, in that the up- and down-regulation of *LIN28A* coincide with the transition from stage 1 to 2 and 3 to 4, respectively (Fig. 5F). Several lines of evidence suggest that LIN28A may regulate development during stage transition. First, three of the four major processes known to be regulated by LIN28A[57] stood out among systemically changing genes (genes that show consistent changes across cell types) from stage 1 to stage 2 and from stage 3 to stage 4 (Fig. 5G, Extended Data Fig. 14C, Supplementary Table 3, Methods), namely cell cycle, mRNA splicing, and

translation (Supplementary Table 4). The three pathways have 17, 36, and 10 genes, respectively, that are positively correlated to *LIN28A* expression in both zebrafish and frog (Extended Data Fig. 14D, Supplementary Table 4). Second, LIN28A may regulate systemically changing genes through direct binding of mRNAs, instead of the well-studied *let-7*-dependent pathway. The systemically down-regulated genes from stage 3 to stage 4 in human are enriched for LIN28A binding[58] (Fig. 5H, hypergeometric test $p = 10^{-24}$), while up-regulated genes and ubiquitously expressed genes with no significant changes in the time window are not. We found that none of these categories is enriched for the known *let-7* binding motif in human[59] (Fig. 5H). Third, growth phenotypes of *Lin28a*(-) mouse embryos suggest that *Lin28a* is necessary and sufficient to promote growth no later than E9.5[60], which corresponds to our stages 3 (organogenesis). Perturbation of *Lin28a* in mouse tailbud[61] and lung[62] led to heterochronic phenotypes, where loss of function caused precocious development and prolonged expression caused retarded development. Taken together, our results suggest punctuated stage transitions in vertebrate embryogenesis and regulation by a conserved mechanism via LIN28A.

Integration with later stage human data

Due to practical limitations of specimen availability, our dataset provides the earliest timepoint in terms of human organogenesis when major organs are being laid out. It practically provides the root for inferring the developmental trajectories of different organs. In this regard, we conducted systematic data integration with scRNA-seq from later stages of human fetus (10~19 week old)[7], which profiled fetal organs covering diverse systems but with some absent, such as limb and head mesoderm. Joint embedding after batch correction shows a global alignment of datasets between early and late stages in which developmental systems are aligned with distinct temporal trajectories (Fig. 6A, Extended Data Fig. 15A-B, Methods).

An effective alignment between datasets should reflect the lineage relationship between cell types. To evaluate this, we assigned one or two best matched cell types in our dataset for each of 73 cell types at the later stages based on the distances calculated by Slingshot[63] (Fig. 6B, Supplementary Table 5, Methods). These mapping generally respect the unfolding of cell types in embryos from three aspects. First, all cell types in our dataset from the structures that were not sampled at the later stages were not matched to any cell type at the later stage, e.g., limb, head mesoderm, craniofacial mesoderm, and spinal cord. Second, 90% cell types of later stage are linked to the correct developmental system at the early stage (Supplementary Table 5). Third, on the cell type level, although most cell types do not have known lineage to be referenced, there are many unambiguous cases standing out. For example, endoderm-originated hepatoblasts and mesoderm-originated stellate cells from liver samples at the later stages were matched to hepatocyte and stellate cells in our dataset, respectively. Blood and endothelium at the later stages were identified from multiple organs, contributing to 5 mutual best matches from the same lineage of early cell types, which are erythrocytes to erythroblasts, megakaryocyte to megakaryocyte, lymphocyte to lymphoid cells, macrophage to myeloid cells, and vascular endothelium to vascular endothelium. However, a few exceptions exist. For example, in eye development, only 2 out of 6 downstream cell types of retinal progenitor cells at the later stages were matched to retinal progenitor cells in our dataset. This analysis shows that one could establish eligible but imperfect mapping of cell types between early organogenesis and fetal stage.

Data from intermediate developmental stages would likely improve trajectories. In this regard, we further examined the trajectory of eye development, where organ-specific data is available at the intermediate

stages. We collected scRNA-seq data in human eye spanning from 5- to 13-week[9] to fill the stage gap and integrate cell types in eye from three datasets (Fig. 6C, Extended Data Fig. 15C-D). Vertebrate eye forms from the optic vesicle that subsequently develops into a two-layer structure, i.e., retinal pigmented epithelium (RPE) and retinal progenitor cells (RPC). The multipotent RPC in turn gives rise to various types of retinal neurons, such as retinal ganglion cells (RGC)[64] (Extended Data Fig. 15C). Using Chu-Liu/Edmonds' algorithm[65] to incorporate temporal information on transcriptome similarity, we reconstructed the trajectories of cell types in eye, which recapitulated the known lineage (Fig. 6D, Methods). Pseudo-time analysis on the RPE and RGC branches confirmed the known expression patterns of key regulators (Fig. 6E). *RAX* is initially expressed in the entire vesicle and then restricted to RPC[66]. The expression of *PMEL* and *MITF* in RPE is important for the differentiation of pigment cells[67,68]. As shown in Fig. 6E, data from consecutive stages convey contiguous trajectories. Furthermore, we identified key genes and TFs that are specifically upregulated in each cell type, which predicts the key regulators that underly the specification of each lineage from the optic vesicle as well as the temporal evolution within each lineage (Supplementary Table 5, Methods). More broadly, with improved temporal resolution in data collection, one could systematically infer organ-specific trajectories rooted by our dataset.

Discussion

In this study, we provide the single-cell transcriptional landscape of 4- to 6-week human embryos, which is the critical period of early organogenesis when major embryonic tissue and diverse cell types appear[1]. As demonstrated in the case studies, our study provides an opportunity to further our understanding of the developmental schemes of vertebrate embryogenesis and the molecular mechanisms that drive the specification of organs and the founding cell types. In developmental systems where our current understanding lacks depth, such as the head mesoderm, the cell types defined in our study are particularly valuable. Subsequent single-cell characterization of these systems at slightly later stages will bring in the lineage context to further understand the significance of these cell types. In the meantime, the systematic nature of the dataset also provides the opportunity to gain global views that are not obvious from studying individual developmental processes or organs, such as temporal gene regulation on whole-embryo level by LIN28A that is conserved in vertebrates.

Powerful as it may be, scRNA-seq relies on cell dissociation during which the spatial information is lost. However, in some cases the spatial information may be inferred from gene expression profiles[69]. In our studies, we successfully reconstructed the spatial axes and domains of the developing neural tube and limb bud. We reason that such success reflects the developmental mechanisms of axes specification that generate a gradient of cell differentiation states at the initial stage of axes formation, such as molecular gradients and the coupling of time and space through a differentiation front in limb bud development. Since axis specification is prominent at this stage of embryogenesis and occurs in many organs and developmental systems, systematic examination of gene expression gradients will not only enrich our understanding of the molecular mechanisms of axis specification and pattern formation across the embryo, but also help focus future efforts of spatial transcriptomics[13].

With the accumulation of whole embryo datasets in vertebrates, cross species analysis becomes increasingly powerful. In one such analysis, we showed that vertebrate embryogenesis consists of 4 conserved developmental stages with sharp temporal boundaries, which implies an unexpected concept that vertebrate embryogenesis may require systemic regulation of stage transition. LIN28A appears to regulate

stage-specific growth through direct binding of target mRNAs. Further studies are needed to examine potential heterochronic phenotypes of differentiation across organs and tissues. Besides systemic insights, detailed comparison of individual organs and cell types can reveal potential human-specific regulation, such as the human-mouse differences in the AP and DV patterning of the neural tube.

Accumulation of single-cell studies and data integration are key toward consortium efforts such as the human cell atlas[70,71]. In terms of human development, our data covers a critical period. As demonstrated in the developing retina, by defining the founding cell population and early progenitor cell types of organs, our data provide the root of organ-specific differentiation trajectories for most of the organs. Meanwhile, our attempt for systematic integration with fetal organ data from 10 to 26 weeks suggest that weeks 6 to 10 is a critical gap to be filled in terms of single-cell omics to achieve informative trajectories of lineage differentiation systematically. In contrast, clinical specimens for stages prior to 4 weeks are extremely rare[72], and our best hope in the foreseeable future to study the earliest stages may be the *in vitro* models that are being perfected such as the organoids and gastruloids[73]. In this regard, our dataset provides the *in vivo* benchmark to evaluate how well different models mimic development: the end product of such a model should match the *in vivo* cell types. Similarly, our dataset provides valuable guidance for iPS and ESC-based stem cell engineering for regenerative medicine.

References

1. Schoenwolf, G. C. & Larsen, W. J. (William J. *Larsen's human embryology*. (Churchill Livingstone/Elsevier, 2009).
2. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* 1 (2019) doi:10.1038/s41586-019-0969-x.
3. Farrell, J. A. *et al.* Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science (80-.).* eaar3131 (2018) doi:10.1126/SCIENCE.AAR3131.
4. Wagner, D. E. *et al.* Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science (80-.).* eaar4362 (2018) doi:10.1126/SCIENCE.AAR4362.
5. Pijuan-Sala, B. *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* 1 (2019) doi:10.1038/s41586-019-0933-9.
6. Han, X. *et al.* Construction of a human cell landscape at single-cell level. *Nature* **581**, 303–309 (2020).
7. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science (80-.).* **370**, eaba7721 (2020).
8. Polioudakis, D. *et al.* A Single-Cell Transcriptomic Atlas of Human Neocortical Development during Mid-gestation. *Neuron* **0**, (2019).
9. Hu, Y. *et al.* Dissecting the transcriptome landscape of the human fetal neural retina and retinal pigment epithelium by single-cell RNA-seq analysis. *PLOS Biol.* **17**, e3000365 (2019).
10. Li, L. *et al.* Single-Cell RNA-Seq Analysis Maps Development of Human Germline Cells and Gonadal Niche Interactions. *Cell Stem Cell* **20**, 891–892 (2017).
11. Gao, S. *et al.* Tracing the temporal-spatial transcriptome landscapes of the human fetal digestive tract using single-cell RNA-sequencing. *Nat. Cell Biol.* **20**, 721–734 (2018).
12. Popescu, D.-M. *et al.* Decoding human fetal liver haematopoiesis. *Nature* 1–7 (2019) doi:10.1038/s41586-019-1652-y.
13. Asp, M. *et al.* A Spatiotemporal Organ-Wide Gene Expression and Cell Atlas of the Developing Human Heart. *Cell* **179**, 1647-1660.e19 (2019).
14. Nowakowski, T. J. *et al.* Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science (80-.).* **358**, 1318–1323 (2017).
15. Zhong, S. *et al.* A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature*

(2018) doi:10.1038/nature25980.

16. Young, M. D. *et al.* Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science (80-.).* **361**, 594–599 (2018).

17. Li, M. *et al.* Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science (80-.).* **362**, eaat7615 (2018).

18. Dyer, L. A. & Kirby, M. L. The role of secondary heart field in cardiac development. *Dev Biol* **336**, 137–144 (2009).

19. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* vol. 28 27–30 (2000).

20. Wilson, L. & Maden, M. The mechanisms of dorsoventral patterning in the vertebrate neural tube. *Dev Biol* **282**, 1–13 (2005).

21. Capdevila, J. & Izpisua Belmonte, J. C. Patterning mechanisms controlling vertebrate limb development. *Annu Rev Cell Dev Biol* **17**, 87–132 (2001).

22. Echevarria, D., Vieira, C., Gimeno, L. & Martinez, S. Neuroepithelial secondary organizers and cell fate specification in the developing brain. *Brain Res Brain Res Rev* **43**, 179–191 (2003).

23. Ramilowski, J. A. *et al.* A draft network of ligand–receptor-mediated multicellular signalling in human. *Nat. Commun.* **6**, 7866 (2015).

24. Benazet, J. D. & Zeller, R. Vertebrate limb development: moving from classical morphogen gradients to an integrated 4-dimensional patterning system. *Cold Spring Harb Perspect Biol* **1**, a001339 (2009).

25. Yokoyama, S. *et al.* A Systems Approach Reveals that the Myogenesis Genome Network Is Regulated by the Transcriptional Repressor RP58. *Dev. Cell* **17**, 836–848 (2009).

26. Capellini, T. D. *et al.* Pbx1/Pbx2 requirement for distal limb patterning is mediated by the hierarchical control of Hox gene spatial distribution and Shh expression. *Development* **133**, 2263–2273 (2006).

27. Coudert, A. E. *et al.* Expression and regulation of the Msx1 natural antisense transcript during development. *Nucleic Acids Res.* **33**, 5208–18 (2005).

28. B, X. *et al.* Hox5 interacts with Plzf to restrict Shh expression in the developing forelimb. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 19438–19443 (2013).

29. Capellini, T. D. *et al.* Scapula development is governed by genetic interactions of Pbx1 with its family members and with Emx2 via their cooperative control of Alx1. *Development* **137**, 2559–2569 (2010).

30. Kuijper, S. *et al.* Function and regulation of Alx4 in limb development: complex genetic interactions with Gli3 and Shh. *Dev Biol* **285**, 533–544 (2005).

31. Cooper, K. L. *et al.* Patterning and post-patterning modes of evolutionary digit loss in mammals. *Nature* **511**, 41–45 (2014).

32. Firulli, B. A. *et al.* Altered Twist1 and Hand2 dimerization is associated with Saethre-Chotzen syndrome and limb abnormalities. *Nat. Genet.* **37**, 373–381 (2005).

33. Akiyama, H. *et al.* Osteo-chondroprogenitor cells are derived from Sox9 expressing precursors. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 14665–70 (2005).

34. Zakany, J. & Duboule, D. The role of Hox genes during vertebrate limb development. *Curr. Opin. Genet. Dev.* **17**, 359–366 (2007).

35. Lefebvre, V. & Smits, P. Transcriptional control of chondrocyte fate and differentiation. *Birth Defects Res. Part C Embryo Today Rev.* **75**, 200–212 (2005).

36. Kobayashi, T. *et al.* Indian hedgehog stimulates periarticular chondrocyte differentiation to regulate growth plate length independently of PTHrP. *J. Clin. Invest.* **115**, 1734–42 (2005).

37. Philippidou, P. & Dasen, J. S. Hox genes: choreographers in neural development, architects of circuit organization. *Neuron* **80**, 12–34 (2013).

38. Le Dreau, G. & Marti, E. Dorsal-ventral patterning of the neural tube: a tale of three signals. *Dev Neurobiol* **72**, 1471–1481 (2012).

39. Uehara, M. *et al.* CYP26A1 and CYP26C1 cooperatively regulate anterior-posterior patterning of the developing brain and the production of migratory cranial neural crest cells in the mouse. *Dev Biol* **302**, 399–411 (2007).

40. Sarropoulos, I., Marin, R., Cardoso-Moreira, M. & Kaessmann, H. Developmental dynamics of lncRNAs across mammalian organs and species. *Nature* **571**, 510–514 (2019).

41. De Kumar, B. & Krumlauf, R. HOXs and lincRNAs: Two sides of the same coin. *Sci Adv* **2**, e1501402 (2016).

42. Delile, J. *et al.* Single cell transcriptomics reveals spatial and temporal dynamics of gene expression in the developing mouse spinal cord. *Development* **146**, (2019).

43. Rayon, T., Maizels, R., Barrington, C. & Briscoe, J. Single-cell transcriptome profiling of the human developing spinal cord reveals a conserved genetic programme with human-specific features. *Development* **148**, (2021).

44. Diss, G., Ascencio, D., DeLuna, A. & Landry, C. R. Molecular mechanisms of paralogous compensation and the robustness of cellular networks. *J. Exp. Zool. Part B Mol. Dev. Evol.* **322**, 488–499 (2014).

45. Moss, E. G. Heterochronic Genes and the Nature of Developmental Time. *Curr. Biol.* **17**, R425–R434 (2007).

46. Ambros, V. & Horvitz, H. Heterochronic mutants of the nematode Caenorhabditis elegans. *Science (80-.).* **226**, 409–416 (1984).

47. Boroviak, T. *et al.* Lineage-Specific Profiling Delineates the Emergence and Progression of Naive Pluripotency in Mammalian Embryogenesis. *Dev. Cell* **35**, 366–382 (2015).

48. Tan, M. H. *et al.* RNA sequencing reveals a diverse and dynamic repertoire of the Xenopus tropicalis transcriptome over development. *Genome Res.* **23**, 201–16 (2013).

49. Briggs, J. A. *et al.* The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science (80-.).* eaar5780 (2018) doi:10.1126/SCIENCE.AAR5780.

50. White, R. J. *et al.* A high-resolution mRNA expression time course of embryonic development in zebrafish. *Elife* **6**, (2017).

51. Yang, D.-H. & Moss, E. G. Temporally regulated expression of Lin-28 in diverse tissues of the developing mouse. *Gene Expr. Patterns* **3**, 719–726 (2003).

52. Faas, L. *et al.* Lin28 proteins are required for germ layer specification in Xenopus. *Development* **140**, 976–86 (2013).

53. Nieuwkoop, P. D. & Faber, J. *Normal Table of Xenopus laevis (Daudin)*. (Oxford, UK: Taylor and Francis, 1994).

54. Carroll, E. J. & Hedrick, J. L. Hatching in the toad Xenopus laevis: Morphological events and evidence for a hatching enzyme. *Dev. Biol.* **38**, 1–13 (1974).

55. Irie, N. & Kuratani, S. Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. *Nat. Commun.* **2**, 248 (2011).

56. Cardoso-Moreira, M. *et al.* Gene expression across mammalian organ development. *Nature* 1 (2019) doi:10.1038/s41586-019-1338-5.

57. Shyh-Chang, N. & Daley, G. Q. Lin28: Primal Regulator of Growth and Metabolism in Stem Cells. *Cell Stem Cell* **12**, 395–406 (2013).

58. Wilbert, M. L. *et al.* LIN28 Binds Messenger RNAs at GGAGA Motifs and Regulates Splicing Factor Abundance. *Mol. Cell* **48**, 195–206 (2012).

59. Agarwal, V., Bell, G. W., Nam, J.-W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* **4**, (2015).

60. Shinoda, G. *et al.* Fetal deficiency of lin28 programs life-long aberrations in growth and glucose metabolism. *Stem Cells* **31**, 1563–73 (2013).

61. Robinton, D. A. *et al.* The Lin28/let-7 Pathway Regulates the Mammalian Caudal Body Axis Elongation Program. *Dev. Cell* **48**, 396-405.e3 (2019).

62. Komarovsky Gulman, N., Armon, L., Shalit, T. & Urbach, A. Heterochronic regulation of lung development via the Lin28-Let-7 pathway. *FASEB J.* **33**, 12008–12018 (2019).

63. Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).

64. Marquardt, T. & Gruss, P. Generating neuronal diversity in the retina: one for nearly all. *Trends Neurosci.* **25**, 32–38 (2002).

65. CHU Y. & LIU, T. On the shortest arborescence of a directed graph. *Sci. Sin.* **14**, 1396–1400 (1965).

66. Fuhrmann, S. Eye morphogenesis and patterning of the optic vesicle. *Curr. Top. Dev. Biol.* **93**, 61–84 (2010).

67. Theos, A. C., Truschel, S. T., Raposo, G. & Marks, M. S. The Silver locus product Pmel17/gp100/Silv/ME20: controversial in name and in function. *Pigment Cell Res* **18**, 322–336 (2005).

68. Ma, X. *et al.* The transcription factor MITF in RPE function and dysfunction. *Prog Retin Eye Res* **73**, 100766 (2019).

69. Nowotschin, S. *et al.* The emergent landscape of the mouse gut endoderm at single-cell resolution. *Nature* (2019) doi:10.1038/s41586-019-1127-1.

70. Regev, A. *et al.* The Human Cell Atlas. *Elife* **6**, (2017).

71. Regev, A. *et al. The Human Cell Atlas White Paper*. (2018).

72. Tyser, R. C. V. *et al.* Single-cell transcriptomic characterization of a gastrulating human embryo. *Nature* 1–5 (2021) doi:10.1038/s41586-021-04158-y.

73. Moris, N. *et al.* An in vitro model of early anteroposterior organization during human development. *Nature* 1–6 (2020) doi:10.1038/s41586-020-2383-9.

74. Kaufman & Matthew, H. *Atlas Of Mouse Development*. (Springer Berlin Heidelberg, 2007).

75. SL, W., R, L. & AM, K. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst.* **8**, 281-291.e9 (2019).

76. Zhang, T. *et al.* A single-cell analysis of the molecular lineage of chordate embryogenesis. *Sci. Adv.* **6**, eabc4773 (2020).

77. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).

78. Santos, A., Wernersson, R. & Jensen, L. J. Cyclebase 3.0: A multi-organism database on cell-cycle regulation and phenotypes. *Nucleic Acids Res.* **43**, D1140–D1144 (2015).

79. Giotti, B. *et al.* Assembly of a parts list of the human mitotic cell cycle machinery. *J. Mol. Cell Biol.* **11**, 703–718 (2019).

80. Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* **174**, 999-1014.e22 (2018).

81. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).

82. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–50 (2005).

83. Packer, J. S. *et al.* A lineage-resolved molecular atlas of C. elegans embryogenesis at single cell resolution. doi:10.1101/565549.

84. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).

85. Irie, N. & Kuratani, S. Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. *Nat. Commun.* **2**, 248 (2011).

Acknowledgements. We thank Drs. Rachna Sheth, Chaolin Zhang, Danwei Huangfu, Lorenz Studer, and Gufa Lin for discussions. We thank Leo Xiao for website design. This work is supported by Funding Project of National Key Research and Development Program of China (2018YFD0900604), Natural Science Foundation of China (41676119 and 41476120), and Fundamental Research Funds for the Central Universities from the Ocean University of China to W.S.. Z.B. is supported by funding from Memorial Sloan Kettering Cancer Center. T.Z. is a visiting student at Memorial Sloan Kettering Cancer Center sponsored by the China Scholarship Council.

Author Contributions. W.S. and Z.B. conceived the project. T.Z., M.H., Y.Q., L.W., Q.Z., and Y.X. collected the embryos and generated the atlas dataset. Y.X. and T.Z. performed data processing, clustering, and all the bioinformatics analysis. Y.X. created the atlas website. W.S., Z.B., Y.X., and T.Z. wrote the

manuscript.

Competing Interests. The authors declare no competing interests.

A

Embryo	Emb.01	Emb.02	Emb.03	Emb.04	Emb.05
Stage	CS12	CS13	CS15-16	CS15-16	CS15-16
Sex	Male	Female	Female	Male	Female
# of cells	18190	18997	44281	10122	11674

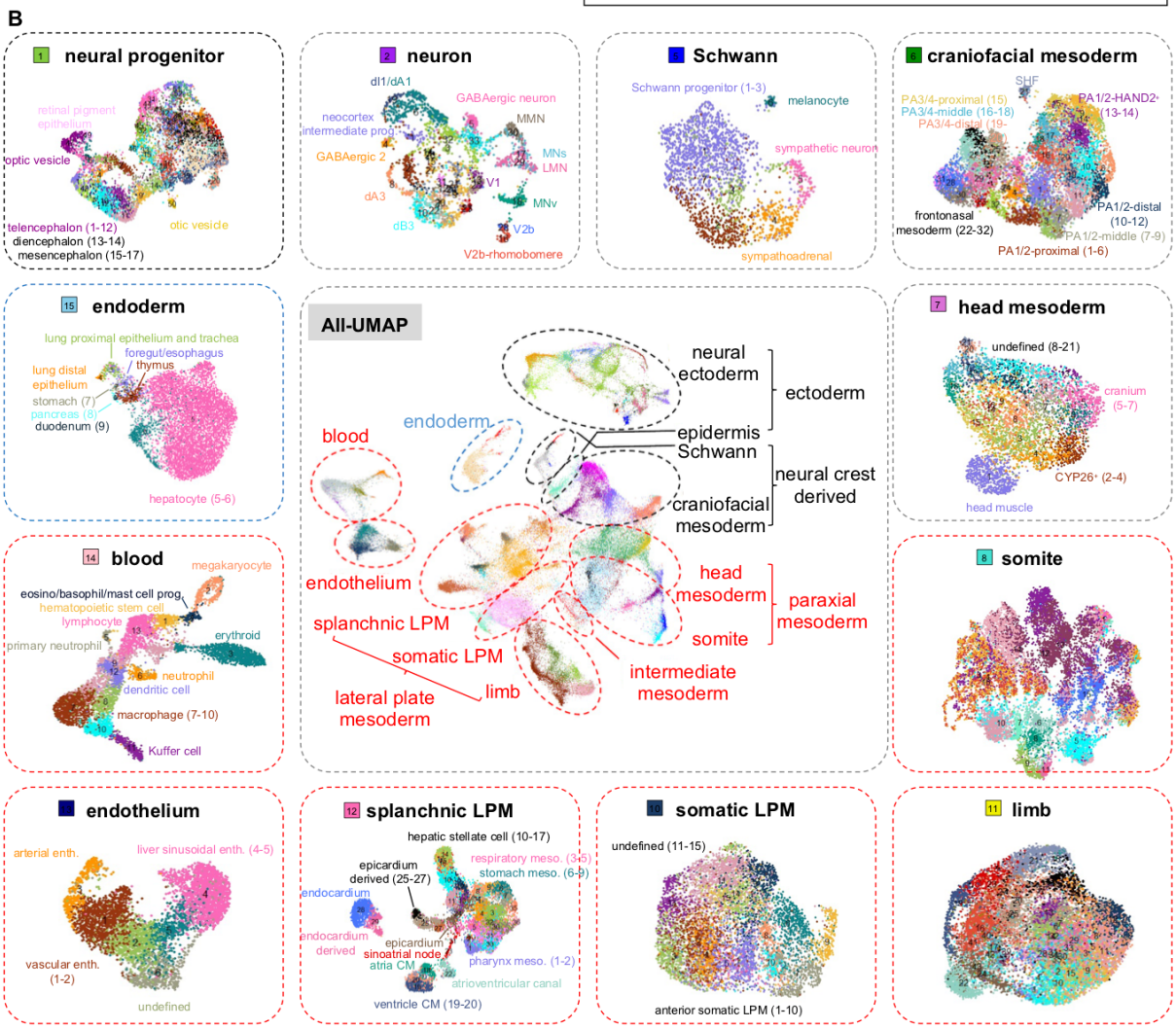


Fig. 1. Single-cell transcriptome atlas of early human embryo. (**A**), Carnegie stage, sex, and number of cells of sampled embryos. (**B**), UMAP visualization of all profiled cells (center panel) with major developmental systems demarcated by dash circles. UMAP visualization of cell type classification for 12 developmental systems shown in outer panels (for the rest, see HEOA database). IM, intermediate mesoderm; LPM, lateral plate mesoderm; PGC, primordial germ cell.

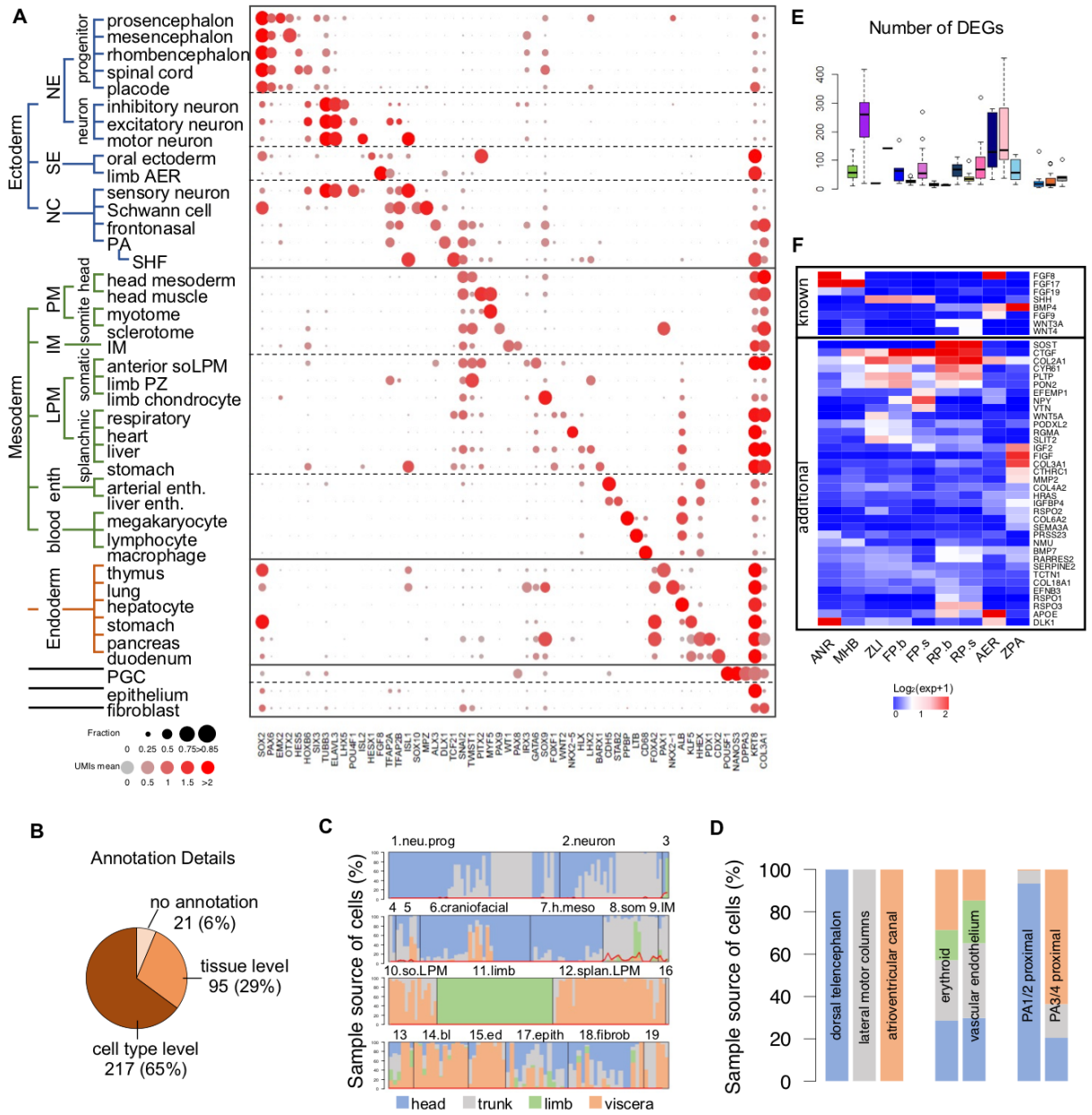


Fig. 2. Characterization of cell types and DEGs. (**A**), Expression of known markers for 41 representative cell/tissue types. Dot size represents the fraction of expressing cells and dot color shows average expression level. NE, neural ectoderm; SE, surface ectoderm; NC, neural crest; PM, paraxial mesoderm; IM, intermediate mesoderm; LPM, lateral plate mesoderm; enth, endothelium; AER, Apical ectodermal ridge; PA, pharyngeal arch; SHF, second heart field; PZ, progress zone; PGC, primordial germ cell. (**B**), Number and percentage of cell types annotated at cell type level, tissue level or without annotation. (**C**), The origin of dissection parts for each cell by cell type. The red lines show the ratio of cells from unexpected to expected origin. (**D**), The origin of dissection part for each cell in selected cell types. (**E**), Number of DEGs per cell type for each developmental system (see Fig. 1 for convention). (**F**), Known and previously unidentified signaling molecule expression in 9 signaling centers. ANR, anterior neural ridge; MHB, mid-hindbrain boundary; ZLI, zona limitans intrathalamica; FP, floor plate (spinal, brain); RP, roof plate (spinal, brain); AER, apical ectodermal ridge; ZPA, zone of polarizing activity.

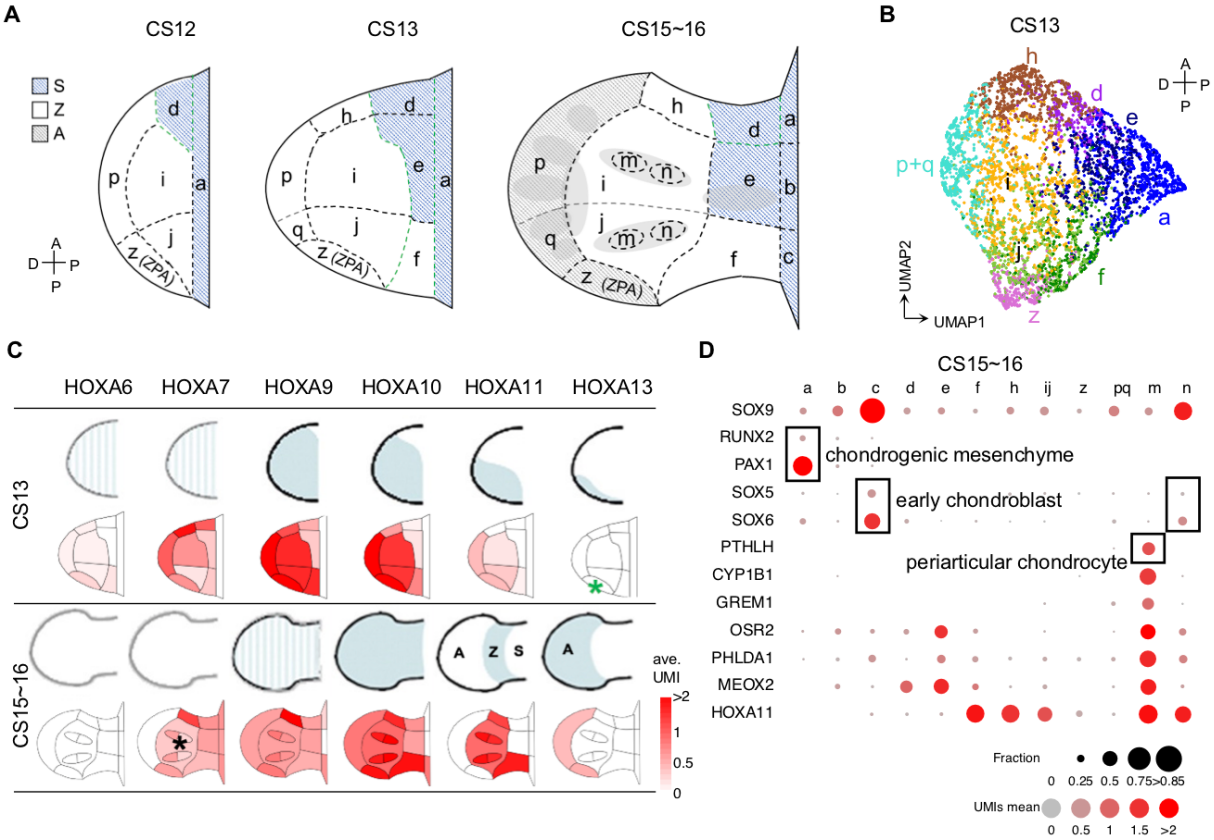


Fig. 3. Spatial domains in limb bud. (**A**), Diagram depicting domains in forelimb mesenchyme defined in this study at each developmental stage. Domains from stylopod (S), zeugopod (Z) and Autopod (A) are indicated by different colors. D-P, distal-proximal; A-P, anterior-posterior. (**B**), UMAP visualization of cell types in forelimb at CS13. Cells are colored by domains, as labeled in panel A. (**C**), Comparison of expression pattern of HOXA genes between *in situ* and scRNA-seq. The expression in scRNA-seq was summarized as mean UMIs in each domain. Black asterisk, only detected in scRNA-seq. Green asterisk, only detected by *in situ*. (**D**), The expression of markers in chondrocyte differentiation (box) and selected DEGs of domain m at CS15~16.

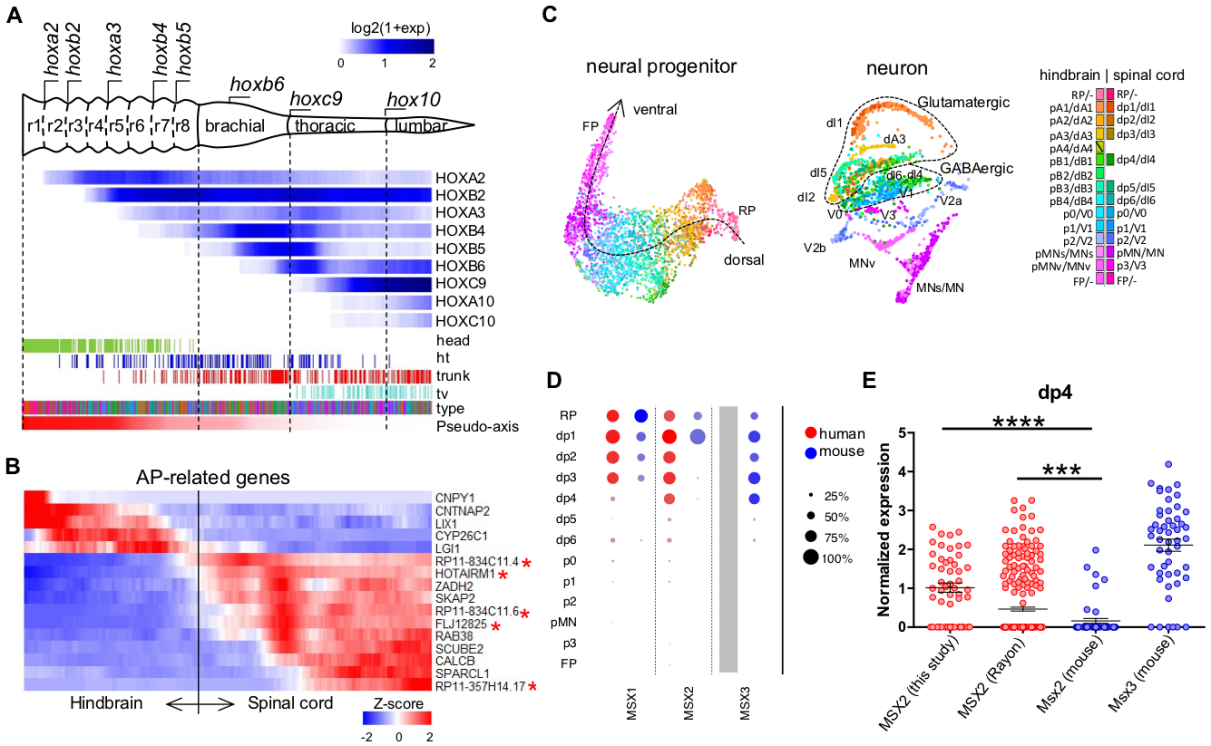
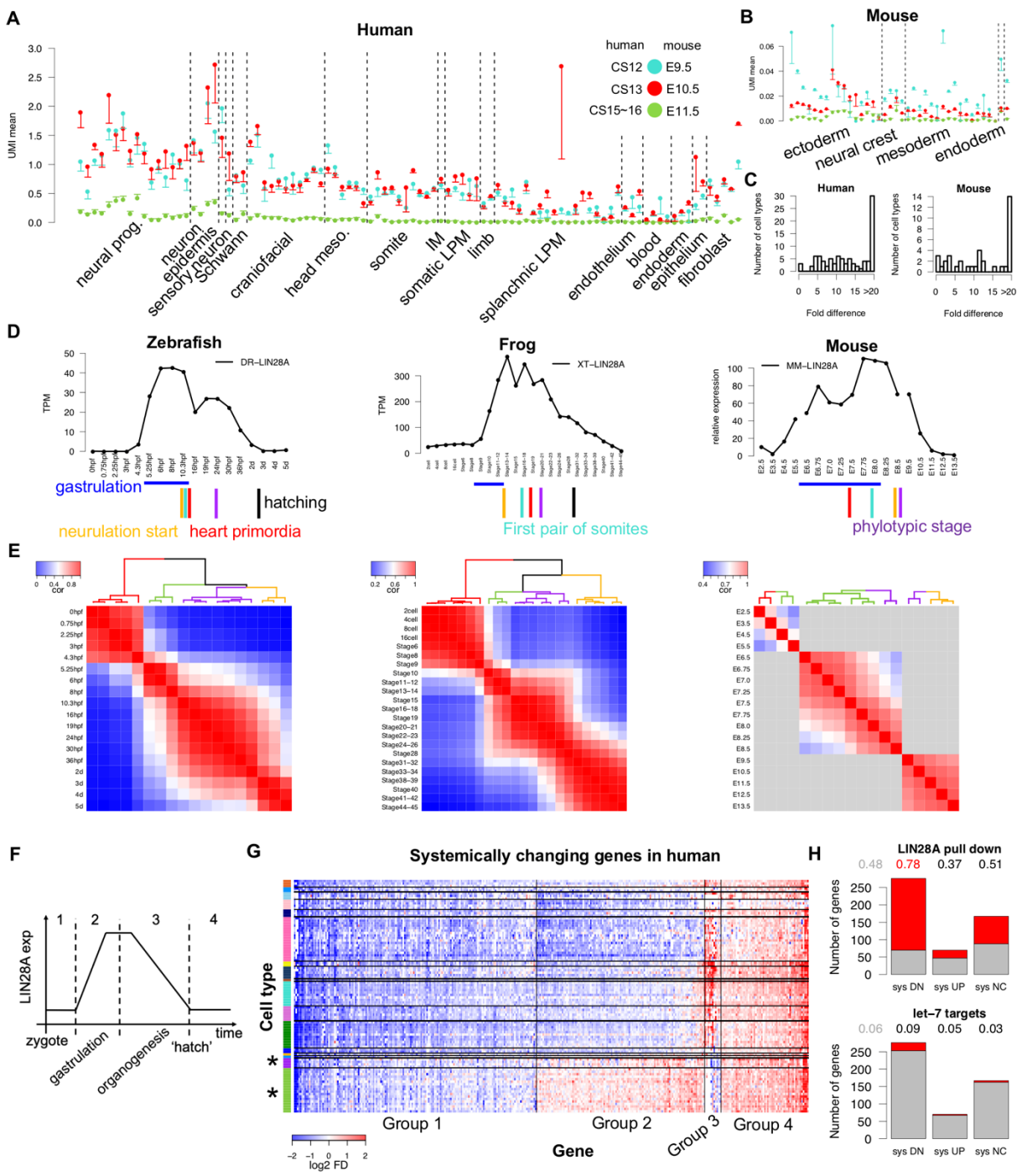


Fig. 4. Spatial patterning of neural tube. (**A**), Schematic diagram modified from Philippidou *et al*.[37] (upper panel) illustrating HOX gene expression along the vertebrate anterior-posterior axis. Heatmap (lower panel) displaying the expression of HOX genes along the pseudo-axis constructed with neural progenitor cells from hindbrain and spinal cord. Origin of dissection parts, cell types (see panel C for color code), and pseudo-axis value of each cell were shown below. (**B**), Expression patterns of previously unidentified AP-related genes. Asterisks denote lncRNAs. (**C**), UMAP visualization of neural progenitors (left) and neurons (right) from hindbrain and spinal cord. Cells are colored by cell types. (**D**), Comparison of expression of MSX paralogs in neural progenitor between human and mouse. Grey box indicates no *MSX3* in the human genome. **(E**), The expression of *MSX2* in dp4 in human (our dataset and Rayon *et al.*[43]) and mouse, as well as the expression of *MSX3* in dp4 in mouse. Each dot denotes a cell. ****, p value < 10^{-7}; ***, p value < 0.001 (Mann-Whitney U test).



521

522 **Fig. 5. Systemic regulation in vertebrate embryogenesis.** (**A**), *LIN28A* expression (normalized UMIs +/-
523 SE) in human embryos at CS12, CS13, and CS15~16 by cell type. (**B**), *LIN28A* expression (normalized
524 UMIs +/- SE) in mouse embryos at E9.5, E10.5, and E11.5 by cell type. (**C**), The distribution of fold
525 difference of *LIN28A* in each cell type between early and late stages in human (left) and mouse (right). (**D**),
526 *LIN28A* expression in zebrafish, frog, and mouse embryos by time point. For mouse, lines between time
527 points from different datasets were not drawn. Vertical lines below show the timing of major
528 developmental events. (**E**), Heatmaps of Pearson's correlation between time points in each species. Colors
529 of dendrogram show clusters of time points. For mouse, pairwise correlation and clustering were only
530 performed within each of the three mouse datasets (grey, inter-dataset entries). (**F**), A schematic diagram

showing the dynamics of *LIN28A* across four developmental stages. (**G**), Heatmap showing fold differences of 346 systemically changing genes in each cell type in human from stage 3 to stage 4. Four patterns of genes were separated by vertical lines. * indicates neural ectoderm where change direction is reversed in groups 2 and 3. (**H**), Enrichment analysis of LIN28A pull-down targets and *let-7* targets in systemically down-regulated ('sys DN', groups 1 and 2 in panel G), systemically up-regulated ('sys UP', groups 3 and 4 in panel G), and unchanged genes ('sys NC') in human from stage 3 to stage 4. Red portion of bars represent targets in each group. Numbers above indicate the percentage of targets in whole genome (grey) and each group (red, hypergeometric test p value < 0.001).

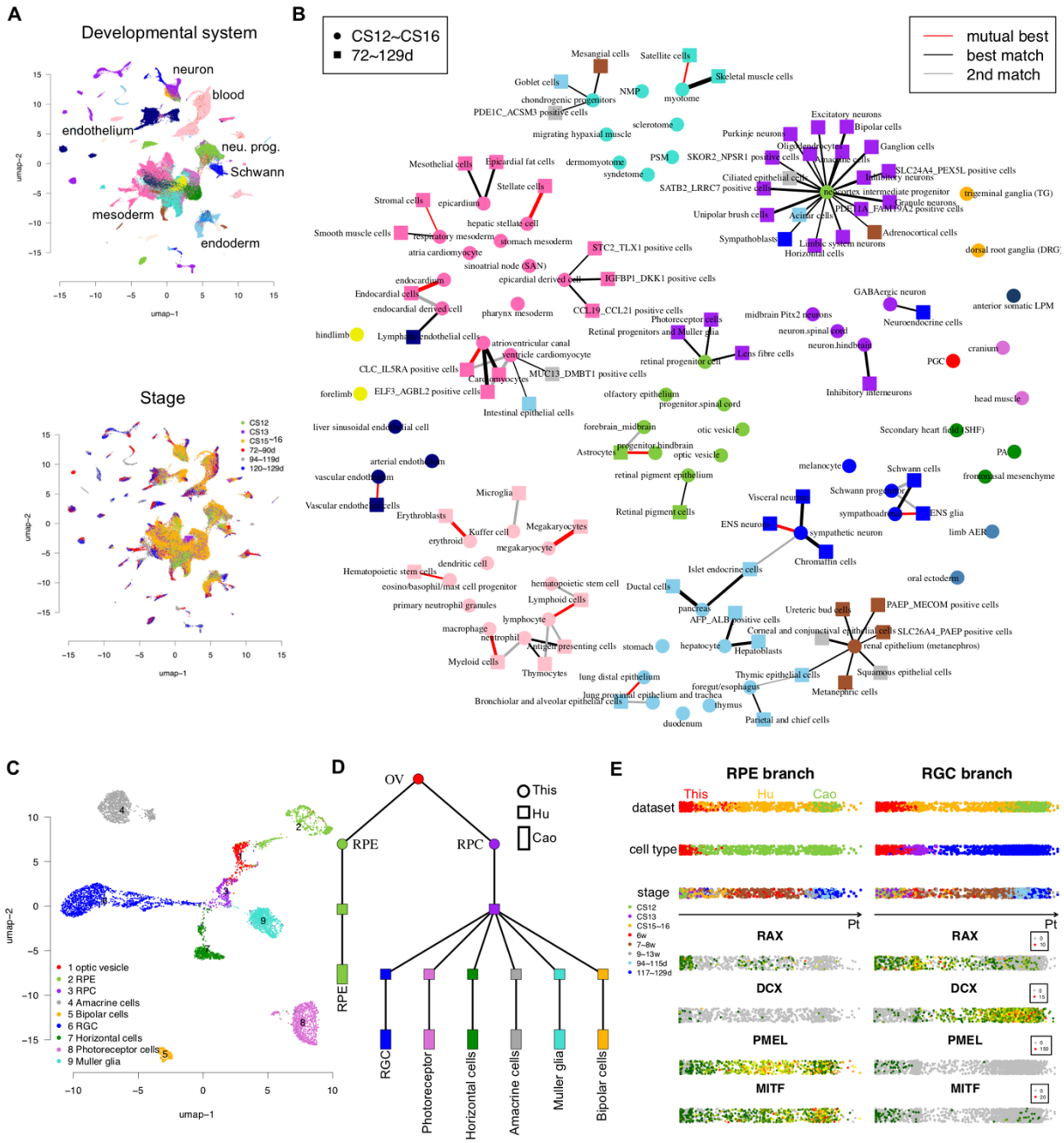


Fig. 6. Systemic data integration. (**A**), Umap of visualization of our dataset and Cao's dataset[7] colored by developmental system (upper panel) and developmental stage (lower panel). Small isolated clusters in Umap are formed mostly by highly differentiated cell types, such as bipolar cells and ductal cells. (**B**), The mutual best match, and 2nd-best match in our dataset for each cell type in Cao's dataset by Slingshot (also see Supplementary Table 5). Cell types are colored by developmental system (see Fig. 1 for conventions). The thickness of lines is inversely proportional to z-score across distances between a cell type in Cao's dataset and all cell types in our dataset (**C**), Umap visualization of cell types of the developing eye in three studies colored by cell type. (**D**), Inferred cell trajectory in eye development by Slingshot and temporal information. Nodes are colored by cell type as in panel C. (**E**), Pseudo-time ordering of cells in RPE branch and RGC branch by Slingshot. The color codes from top to bottom are dataset, cell type (see panel C), and stage. The heatmaps below show normalized UMIs of 4 known marker genes along pseudo-time.

Methods

Ethics statement

This study was approved by the Ethics Committee of Tongji University School of Medicine and Life Sciences (No. 2019tjdx280) and the Medical Ethics Committee of Traditional Chinese Medicine Hospital of Kunshan (No. 2019-17). The collection and use of human embryos are compliant to current ISSCR guidelines. Informed consent was agreed and signed by each patient only after each patient was provided with all the necessary information about the study. Embryos were collected only after voluntary informed consent was obtained from each patient undergoing legal pregnancy termination. All procedures were performed in strict accordance with the 'Management of Human Genetic Resources' formulated by the Ministry of Science and Technology of China (No. 717, effective July 1, 2019).

Embryo collection and sequencing

Aborted embryos from donations were washed with PBS and immediately transported to the laboratory in cold Hibernate-E medium (Gibco). Embryos with good quality were staged based on anatomical characters. We dissected each embryo on ice into four main parts, *i.e.*, head, trunk, viscera, and limb. The dissection parts of embryo were digested with 0.05mg/ml Liberase TM (Roche) in Hibernate-E medium at 37°C for 10 minutes and stopped with FBS. Single cells were subsequently washed twice with cold PBS, filtered through a 35 μm cell strainer and loaded onto 10x Genomics Chromium system using Single Cell 3' Reagent Kits v2 at a concentration of 16,000 single cells per sample. Libraries were sequenced on Illumina NovaSeq system in PE150 mode.

Cell type identification

1. Collection of ontology and diagnostic markers

The quality of annotation of cell types is the key for a valuable single-cell database. Towards a high-quality annotation, we need precise and comprehensive ontology and diagnostic markers. Two principals were considered during the literature search. First, the ontology should follow a hierarchical structure from developmental systems to cell types to capture the known lineage hierarchy. Second, the developmental stage of the diagnostic markers should match the time window we study.

Because studies on human embryos from CS12~CS16 are scarce, we expanded our search to studies in mouse embryos at the corresponding stages (E9.5~E11.5)[74]. In total, 159 references for developmental systems and cell types were collected, which describe their organization on lineage/anatomy (ontology), as well as 49 diagnostic markers for developmental systems and 215 markers for cell types (Supplementary Table 1). Diagnostic markers were defined as genes with specificity expression on global (developmental systems) or local (cell types) scale at the time window we studied.

2. Knowledge-based semi-supervised clustering

2.1 Overview

Embryos in our study are at the beginning of organogenesis with a dramatic increase of cell types. To handle such a complex system, we designed an iterative clustering approach to first resolve the developmental systems (organs, tissues and major sublineages) and then cluster cell types within each system. Some developmental systems do not have well-documented ontology information to be referenced,

such as splanchnic LPM. To cover known cell types and allow for searching unknown cell types, we built a semi-supervised approach that applied knowledge to guide and optimize the iterative clustering. This approach combines the advantage of unsupervised clustering[2] and knowledge-guided identification[42]. Details of the approach follow.

2.2 Preprocessing of 10x Genomics data

2.2.1 The determination of sex of embryos

The expression of *XIST* and *RSP4Y1* were used to determine the sex of embryos (Extended Data Fig. 1C). Embryos with high *XIST* expression and no *RSP4Y1* expression were defined as female, and *vice versa*, as male.

2.2.2 Quality control on cells

Raw reads were demultiplexed, filtered and mapped to human reference genome (GRCh38) using Cell Ranger 3.0.2 with default settings. The filtered feature-barcode matrices generated by Cell Ranger were used for downstream analysis. To exclude low-quality cells and doublets, cells with fewer than 1,000 or over 60,000 UMIs were discarded. The doublet ratios were estimated by scrublet pipeline with 'expected_doublet_rate'=0.07[75] (Supplementary Table 1C). The expression matrix was normalized into UMIs per 10,000 UMIs.

2.2.3 Isolation of erythroid cells

We observed that erythroid cells are the major source of lysis contamination. The number of erythroid cells is disproportionately large. To prevent potential influence of erythroid cells on downstream analysis, we identified and isolated them before clustering using canonical markers. To do that, the total UMIs of hemoglobin genes (*HBA1*, *HBA2*, *HBE1*, *HBG1*, *HBG2*, *HBZ*) were plotted against total UMIs in each cell (Extended Data Fig. 1D). The threshold of erythroid cells was set as the upper boundary of 99% confidence interval of linear regression on cells with low UMIs of hemoglobin genes (Extended Data Fig. 1D, dash green line). The cells with total UMIs of hemoglobin genes above the threshold were considered as erythroid cells (Extended Data Fig. 1D, red dots) and excluded from clustering. A total of 1,000 randomly selected red blood cells were included for the downstream analysis after clustering.

2.3 Identification of developmental systems

Cells from 5 embryos were pooled before clustering. Aggregate information from pooled samples increased power of clustering especially for cell types with low differentiated status[76]. As shown in the section of benchmark, the temporal difference within a cell type does not have a profound effect on the identification of cell types (except limb mesenchyme, see section of *Special cases* below).

The top level of clustering is aimed to identify super-clusters at the level of developmental systems. To avoid batch effect between samples, the detection of HVGs were performed by sample based on Poisson distribution[76] and then merged from each sample. Fifteen super-clusters were annotated with canonical markers, out of total 18 super-clusters identified by TFs in HVGs with R package Seurat 3.0[77] (Extended Data Fig. 2A, 'FindClusters', r=0.5, PCs=20). Stripped nucleus removal was performed as described in Pijuan-Sala *et al.*[5]. Three super-clusters with considerably lower mitochondrial gene expression compared to the 15 annotated super-clusters (total UMIs of 13 mitochondrial protein-coding genes 51±64 vs.

638 195±144, mean ± SD) were excluded from downstream analysis. In summary, a total of 103,264 high-
639 quality cells were included in the downstream analysis.
640
641 Only TFs in HVGs were used in the top level of clustering because we found TFs in HVGs are better in
642 preserving lineage organization than all HVGs (Extended Data Fig. 2). The distribution of expression of
643 canonical markers across clusters show that markers of intermediate mesoderm (*PAX1*, *PAX9*) and
644 endoderm (*FOXA2*) significantly lose their specificity in the clustering by HVGs compared to that by TFs
645 (Kolmogorov-Smirnov test $p < 0.05$). The global mapping between two sets of clusters shows that clusters
646 3 and 6-10 by HVGs do not have clear match to clusters by TFs (Extended Data Fig. 2C). Cluster 3 by
647 HVGs is composed of cells from many clusters by TFs, including almost all the *FOXA2*+ endoderm
648 (Extended Data Fig. 2D). The non-TF HVGs shared in cluster 3 are likely the common feature of
649 epithelium, causing the mixture of epithelial cells from different germ layers in cluster 3. Furthermore,
650 clusters 6-10 by HVGs are strongly biased to embryonic stage (Extended Data Fig. 2E), which leads to the
651 misclassification on mesoderm, including intermediate mesoderm.
652
653 *2.4 Identification of cell types*
654 The low level of clustering is aimed to identify clusters at the level of cell type, which is performed within
655 each of the 15 super-clusters.
656
657 *2.4.1 HVGs*
658 HVGs were calculated by sample based on Poisson distribution[76] and then merged from each sample.
659 Ubiquitously expressed HVGs in each sample (>70% cells) were excluded. Also, hemoglobin genes, cell
660 cycle genes[78,79], sex-specific genes (*XIST*, *RPS4Y1*, *RPS4Y2*), and batch-effect genes[13,80] were excluded
661 from HVGs during clustering.
662
663 *2.4.2 Parameter optimalization*
664 *2.4.2.1 Rounds of clustering*
665 The rounds of clustering were controlled so that known cell types were well-resolved according to
666 diagnostic markers and clustering is not too fragile. For relatively complex systems, i.e., brain, neural
667 progenitor, neuron, craniofacial, head mesoderm, LPM, and endoderm, we performed two rounds of
668 clustering. For other systems, i.e., Schwann, somite, intermediate mesoderm, limb, epidermis, endothelium,
669 blood, and heart, one round of clustering was performed. We considered whether the second round of
670 clustering is needed according to the support of divisions of the second round by diagnostic markers
671 (Extended Data Fig. 3A-B). For example, in brain, 9 clusters were derived from first round of clustering,
672 each of which was further divided into 1~4 clusters in the second round of clustering. All the divisions in
673 the second round were supported by 2 or more markers. In comparison, in endothelium, only 1 out of 8
674 clusters from the first round were supported to be further divided. Thus, two rounds of clustering were
675 chosen for brain and one round of clustering were chosen for endothelium.
676
677 *2.4.2.2 Parameters in Louvain clustering*
678 The resolution 'r' in 'FindClusters' (Seurat) were determined by systems with more prior knowledge and
679 then applied to systems with less prior knowledge. The 'r' was initially trained in spinal neuron, which is
680 composed of 12 well-studied cell types. A list of 'r' value spanning from 0.1 to 1 with interval of 0.2 were
681 examined to see which one gives the closest number of clusters to expected number (Extended Data Fig.

682 3C). Under the condition of 10, 20, and 30 PCs, the number of clusters resulted from r=0.5 is closest to the
683 expected number 12. The clusters start to be convergent when r>0.5 (Extended Data Fig. 3D, adjusted Rand
684 index, ARI). As a *post hoc* examination, known cell types in blood were resolved under r=0.5
685 (Supplementary Note 1). Thus, r=0.5 was propagated to the clustering of all other super-clusters. The
686 number of PCs were determined by ranking of PCs based on the percentage of variance explained by each
687 one (elbow point, PCs=20).
688
689 The robustness of clustering was tested by comparing the result of the chosen r or PCs to results from a
690 range of values in each super-cluster (Extended Data Fig. 3E), which is measured by ARI. PC = 20 was
691 compared to a series of PCs (15~25, increment 1, excluding 20) when r was fixed to 0.5. Resolution 'r' =
692 0.5 was compared to a series of 'r' (0.3~0.7, increment 0.1, excluding 0.5) when PCs was fixed to 20.
693
694 *2.4.3 Special cases: untangling axes in the identification of cell types*
695 During clustering, we found different developmental systems may have different profound factors (or axes)
696 entangling upon the axis of cell types. Axis formation is very common among organs at this stage. These
697 axes are reflected in the single-cell data and sometimes orthogonal to the axis of cell types. They will not
698 be sequentially decomposed by clustering, which may cause mistakes on cell type identification in blind
699 clustering. Thus, untangling axes properly could improve the correctness of cell type identification. This
700 section will introduce two cases we handled, which can be implication for other studies with similar
701 characteristics.
702
703 *2.4.3.1 Temporal axis in limb bud*
704 Temporal axis is profound in the mesenchyme of limb bud. We found removal of temporal axis improved
705 the cell type identification at CS12. The specificity of marker expression across clusters was lower in
706 clustering of limb cells from all stages than that in clustering of limb cells from CS12 only (Extended Data
707 Fig. 4, Kolmogorov-Smirnov test). Limb bud is fast growing in the time window of our samples. The limb
708 cells harvested at CS13 or CS15~16 are 10 folds more than cells at CS12. It is likely that signals from the
709 relatively low number of cells at CS12 was overshadowed by other stages in pooled clustering. Thus,
710 clustering in the super-cluster limb was performed by stage.
711
712 *2.4.3.2 HOX genes*
713 Cell types running along the AP axis adapt HOX code, which is supposed to make cell type dissimilar by
714 AP position. We excluded HOX genes in HVGs for developmental systems neural tube, Schwann,
715 craniofacial, somites, IM, and LPM. In this way, distinguished segmentation on AP axis (e.g., neural cell
716 types in rhombomere and spinal cord) can still be identified and clustering is not over fragmented.
717
718 *3. Annotation*
719 *3.1 Annotation of cell types*
720 The biological identity of each cluster was annotated with 2~5 diagnostic genes. Further detail can be found
721 in Supplementary Note 1 "annotation_vignette.pdf" using blood as an example. For all other systems, the
722 detail can be found at https://heoa.shinyapps.io/code/. In total, we obtained 333 cell types.
723
724 *3.2 Post-processing*

To organize resulted cell types in reflecting lineage history, we regrouped cell types into 18 developmental systems (Extended Data Fig. 5). Cell types in super-cluster brain at this stage are all neural progenitors, and thus were re-grouped with the neural progenitor super-cluster. Sensory neurons are different from other neurons in spinal cord and were separated as a new system "sensory neuron". Somatic LPM and splanchnic LPM were separated as two systems. Cell types in super-cluster heart were merged into splanchnic LPM. PGC (*POU5F1*, *DPPA3*) is separated as a new system "PGC". Cell types with epithelium feature (*KRT18/19*, *CLDN4*) were grouped as a new system "epithelium" and cell types with fibroblast feature (*COL1A2*, *COL3A1*) were grouped as "fibroblast". Nine cell types were moved to other systems according to their identities (Extended Data Fig. 5, red arrows), which are likely to be misclassified at the top level clustering. Six cell types without clear belonging of developmental system were grouped as "miscellaneous".

4. Identification of DEGs

The traditional way of identifying DEGs by comparing one cell type versus all other cell types may not be appropriate here considering large number of cell types and most DEGs are not specific to only one cell type. We considered a gene as a DEG if it is not expressed in many cell types and is highly expressed in at least one cell type. First, genes with mean UMIs < 0.3 in more than 20% of cell types were selected. Second, the DEGs for a given cell type were defined as genes with high expression in the given cell type (detected in >=40% of cells, mean UMIs >=0.5, and z-score of mean UMIs >= 7 with non-expressing types as background). A *post hoc* test by Mann-Whitney U test showed that adjusted p values of identified DEGs are less than 0.005. Genes associated with batch effect (see above) were excluded from DEGs.

Signaling pathway analysis

To systematically analyze signaling gene expression, we collected ligand, receptor and antagonist of the Fgf, TGF-β, Sonic Hedgehog, Notch and Wnt pathways provided by KEGG[19]. To identify novel signaling molecules expressed in signaling centers, we examined the expression of all signaling ligands provided by Ramilowski *et al.*[23]. Ligands with robust expression (detected in >=40% of cells and mean UMIs >=0.5) in at least one signaling center are showed in Fig. 2F.

Spatial domains in limb development

In situ results on limb staged from E9.5 to E11.5 mouse were collected from literatures and the EMBRYS database[25–33] (Extended Data Fig. 11-12). The spatial location of cell types of forelimb was determined by comparing expression of DEGs in scRNA-seq and expressed regions from *in situ* results. Limb domains at three stages were summarized as schematic diagrams. Please see Supplementary Note 1 "annotation_vignette.pdf" for the detail.

Neural tube analysis

Neural progenitors from the hindbrain and the spinal cord were used for pseudo-spatial analysis. All expressed HOX genes and 5 housekeeping genes (*RPL5*, *ELAVL1*, *ATP2C1*, *ARPC2*, *ARPC1A*, to prevent omitting cells with zero HOX gene expression) were used to reconstruct the pseudo-axis using Monocle[81]. Known expression patterns of HOX genes were used to evaluate whether the anterior-posterior pattern could be resolved. We then defined a group of AP-related genes by calculating the correlation between gene expression and cell ordering. Genes with notable positive (3.5×SD) or negative (-4.5×SD) correlation were considered as AP-related genes. For patterning along the dorsal-ventral axis, cells were visualized by

UMAP with HVGs excluding HOX genes. The same method was applied on neuron cells from the hindbrain and the spinal cord.

To compare gene expression in neural type between human and mouse, we collected single-cell RNA-seq data in neural tube in mouse[42] for comparison and data in neural tube in human[43] for cross-validation. Raw UMI matrix were normalized into UMIs per 10,000. The canonical markers of neuronal cell types of these two species were collected from Delile *et al.*[42].

LIN28A and embryogenesis in vertebrates

1. RNA-seq data and preprocessing

RNA-seq in other vertebrates were downloaded from published data (Supplementary Table 2). Bulk RNA-seq were normalized into transcripts per million reads (TPM). Single-cell RNA-seq were normalized into UMI per 10,000 UMIs.

2. Defining embryonic stages upon timepoints

For each species of zebrafish, frog, and mouse, the pairwise Pearson's correlation between samples were calculated based on the expression of TFs (Gene Ontology term GO:0003700). Clustering of time points was based on pairwise correlation between samples (R package dendextend, k=4). Considering the time series of mouse data are from three sources using different techniques (E2.5~E5.5 using bulk RNA-seq, E6.5~E8.5 using scRNA-seq, and E9.5~E13.5 using sci-RNA-seq for nuclear sequencing), correlation calculation and clustering were performed within each experiment. Single-cell RNA-seq at E6.5~E8.5 and E9.5~E13.5 were treated as pseudo bulk-seq for correlation analysis by averaging all cells. Similarity between zebrafish and frog samples were measured by pairwise correlation based on the expression of homologous TFs. The orthology relationship was downloaded from Ensembl (Ensembl Genes 100). Dynamic time warping was performed to align zebrafish and frog samples based on their correlations (R package dtw).

3. Detection of systemically changing genes

For systemically changing genes from stage 1 to stage 2, we compared transcriptome of zebrafish (6~8 hpf vs. 3~4.3 hpf), frog (stages 13~14 vs. stages 6~8), and mouse (E4.5~5.5 vs. E2.5~3.5). No human transcriptome data is available at this stage. From stage 3 to stage 4, we compared transcriptome of zebrafish (3~4 d vs. 10.3~16 hpf), frog (stages 31~34 vs. stages 15~18), mouse (E11.5 vs. E9.5~10.5) and human (CS15~16 vs. CS12~13). For zebrafish and frog, as well as mouse from stage 1 to stage 2, where bulk-seq is available, systemically changing genes were defined as those differentially expressed between stages (> 2 fold) in bulk RNA-seq and are systemically expressed (> 50% cell types) in scRNA-seq at the corresponding stages. For human from stage 3 to stage 4, systemically changing genes were defined as differentially expressed between stages across cell types (> 2 fold in at least 1 cell type and > 1.2 fold in at least 50% cell types). Genes with inconsistent change across cell types (opposite direction of changes in > 30% cell types) were removed. For mouse from stage 3 to stage 4, since the mouse scRNA-seq data in this time window is based on nuclear sequencing and the low detection level makes the identification less reliable, we filtered the identified human systemically temporal genes for those whose mouse ortholog have consistent direction of changes and are systemically expressed in mouse scRNA-seq.

For each species, cell types were as defined in the scRNA-seq publication. To avoid distortion by cell types with low number of cells, we only considered cell types with more than 40 cells at each of the relevant

developmental stage in zebrafish, frog, and mouse. For human, similar cell types were merged. Similar cell types for a given one were defined as Pearson's correlation greater than 0.7 based on their transcriptome. Each cell type was merged with its most similar cell type by a greedy approach until it met the thresholds on cell number, which is more than 40 cells at early and late stages.

4. Pathway enrichment analysis

Pathway enrichment analysis for systemically changing genes was performed against gene sets of Reactome on MSigDB[82]. P values were calculated by hypergeometric test and corrected by the Benjamini & Hochberg method. The enriched pathways were defined as pathways with adjusted p value < 10^{-4} in at least 2 species. Pathways that are both enriched from stage 1 to stage 2 and from stage 3 to stage 4 were highlighted (Supplementary Table 4).

Data integration

Single-cell RNA-seq data in human embryos between 72 days and 129 days in estimated postconceptual age were downloaded from Cao *et al.*[7] (raw UMI matrix of 377,456 cells from https://descartes.brotmanbaty.org/). To compensate the low detected UMIs per cell (median 863) in sci-RNA-seq3, we pooled 5 cells per type per embryo into 1 meta-cell, which resulted median UMIs 2871 per meta-cell (Extended Data Fig. 15A). Two Seurat objects were created with raw UMI matrix of our data and Cao's data, respectively. Reciprocal PCA in Seurat package was applied to integrate two datasets by top 2,000 HVGs. The joint PCA (30 PCs) is used as the input of dimension reduction by Umap.

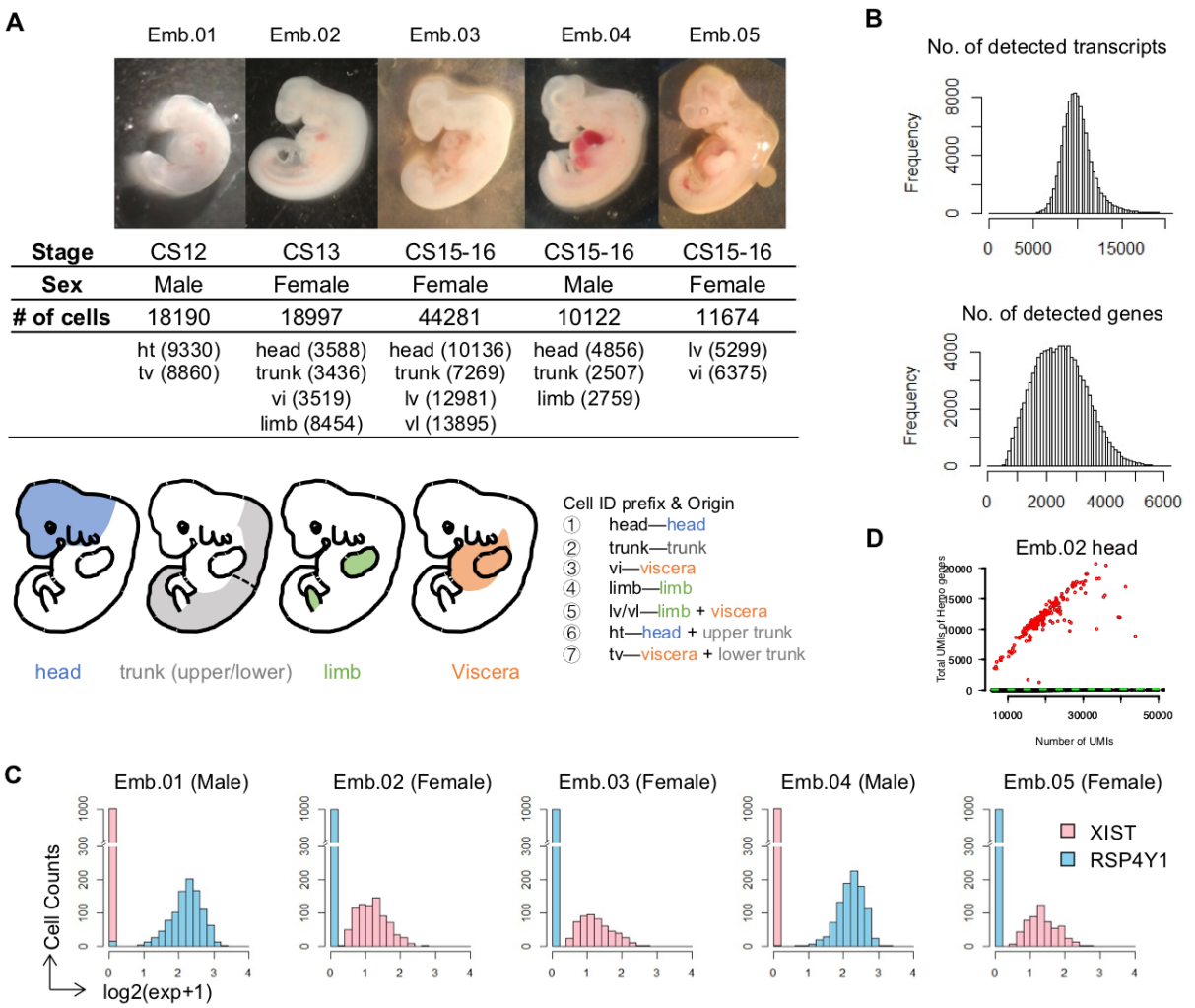
To calculate the similarity between our cell types and cell types at the later stages (Cao's dataset), we first cleaned cell types in both datasets. 73 cell types in Cao's dataset were included except 4 cell types from trophoblasts. To balance cell types between two datasets, we regrouped cell types in our data into 72 cell types (*e.g.*, all types of frontonasal mesenchyme were merged as one cell type because it was not sampled in Cao's dataset), and exclude cell types from epithelium, fibroblast, and miscellaneous (Supplementary Table 1). Slingshot[63] was used to calculate the pairwise distance between cell types from two datasets in the joint PCA space from Seurat. We considered three types of linkages based on Slingshot distance, mutual best match, best match, and second-best match. First, mutual best matches of cell types between two datasets were taken. Second, for each cell type in Cao's dataset, the best matched cell type in our dataset was defined as the cell type with minimal distance. For 6 terminal cell types in our dataset (Supplementary Table 1), only 1 best match from Cao's dataset was allowed. Other cell types in Cao's dataset best matched to terminal cell types in our dataset were relinked to next best match. Third, for each cell type in Cao's dataset, the second-best matched cell type in our dataset was defined if the z-score of distance of the cell type with second minimal distance is less than -2. For a cell type in Cao's dataset with mutual best match, any cell type in our dataset with distance great than 3 fold of the minimal distance was excluded from second-best match. The three types of linkages between two datasets were visualized by R package igraph (layout 'layout_nicely').

Single-cell RNA-seq data in eye of human embryos (Hu's dataset)[9] were collected and cell types of eye in three datasets (our data, Hu's data, and Cao's data) were integrated. Three Seurat objects were created with raw UMI matrix of cell types of eye in three datasets, respectively (meta-cells defined above were used for Cao's dataset). Reciprocal PCA in Seurat package was applied to integrate three datasets by top 2,000 HVGs. The joint PCA (30 PCs) is used as the input of dimension reduction by Umap. In an ideal trajectory,

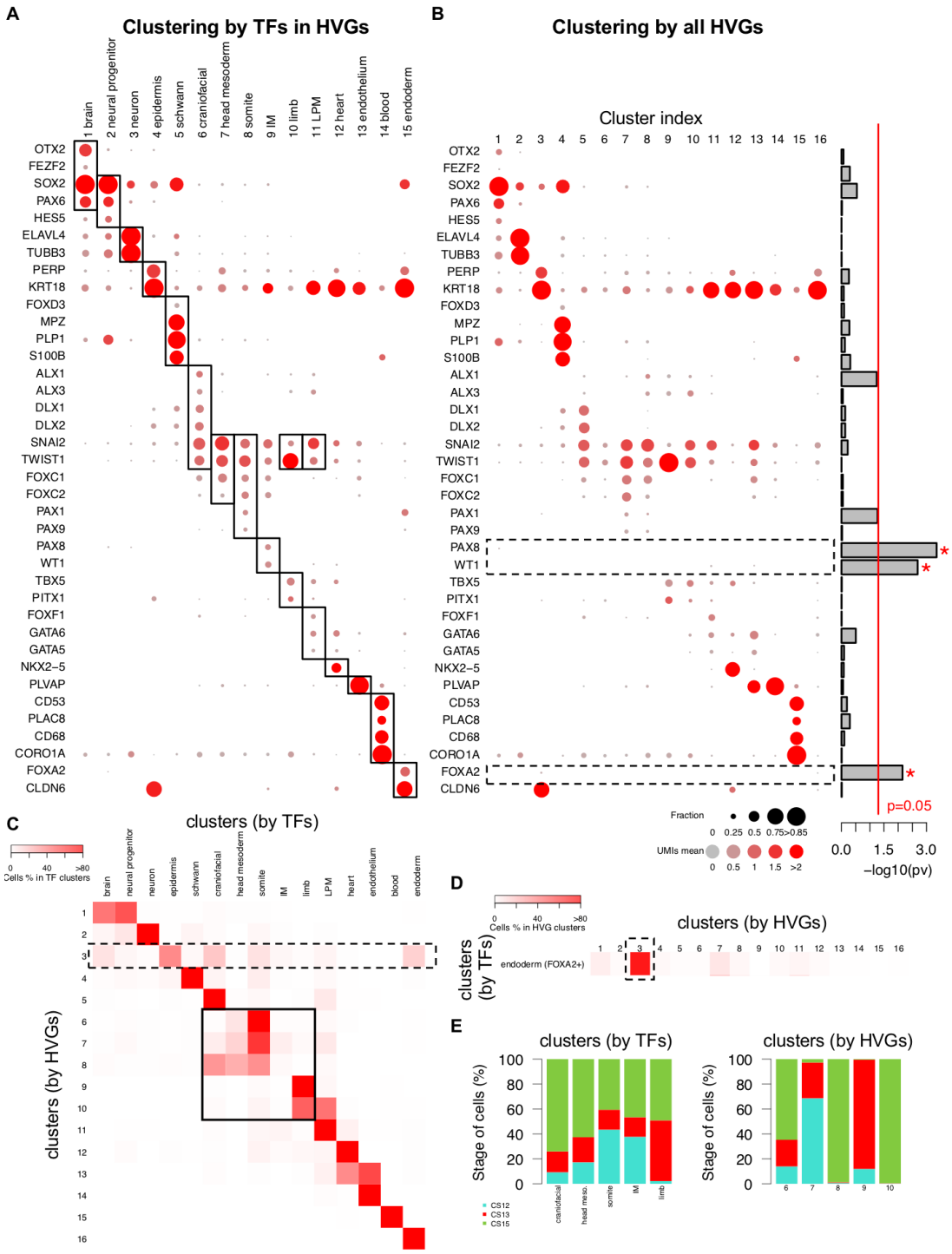
cells proceed from earlier to later time point. To incorporate temporal information as a constraint, we modeled trajectory reconstruction as minimum span tree problem in a directed graph. First, pairwise distance between cell types were calculated by Slingshot as the weight of edge. Second, cell types were assigned to 4 groups according to their stages (Supplementary Table 5). Third, Chu-Liu/Edmonds' algorithm[65] was performed to find minimum spanning tree only allow links from early group to late group (R package optrees, 'msArborEdmonds'). Pseudotime of cells in RPE and RGC lineage were analyzed in Slingshot ('slingPseudotime'), respectively. To identify key genes potentially involved in the differentiation of each cell type, we normalized UMI matrix into UMIs per 10,000 in each dataset. The key genes of a cell type were defined as genes that expressed in this cell type (mean UMI >= 0.5 & fraction >= 0.4), upregulated relative to its pseudo-mother in trajectory (> 2 fold), and not upregulated in any of its pseudo-sister (Supplementary Table 5).

Data and code availability

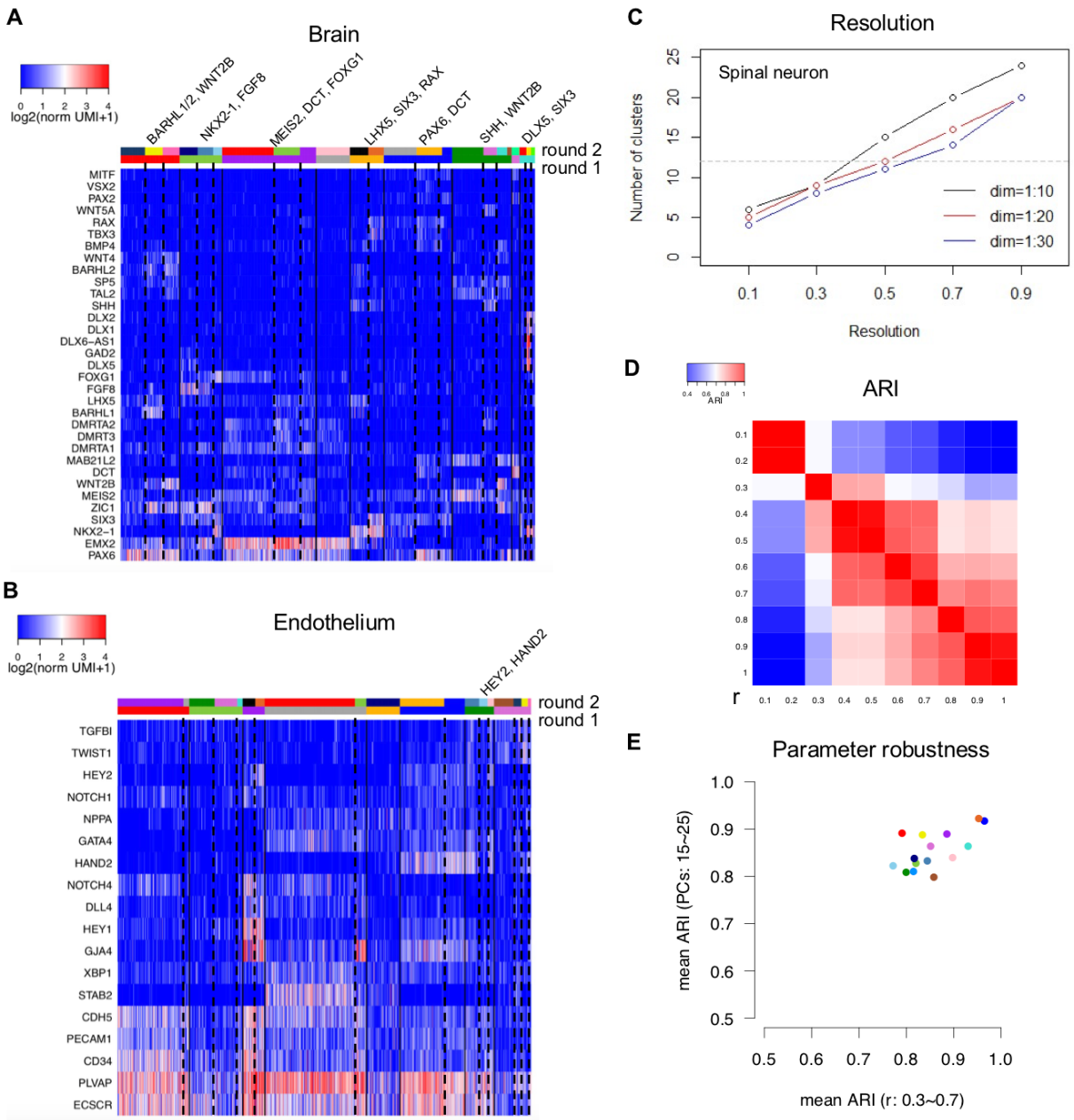
The raw and processed data generated in this study can be downloaded from the NCBI Gene Expression Omnibus (GSE157329). An online depository for cell types and gene expression [developed with R package VisCello[83]] is available at https://heoa.shinyapps.io/base/. R Scripts for analysis and figures are available at https://heoa.shinyapps.io/code/.



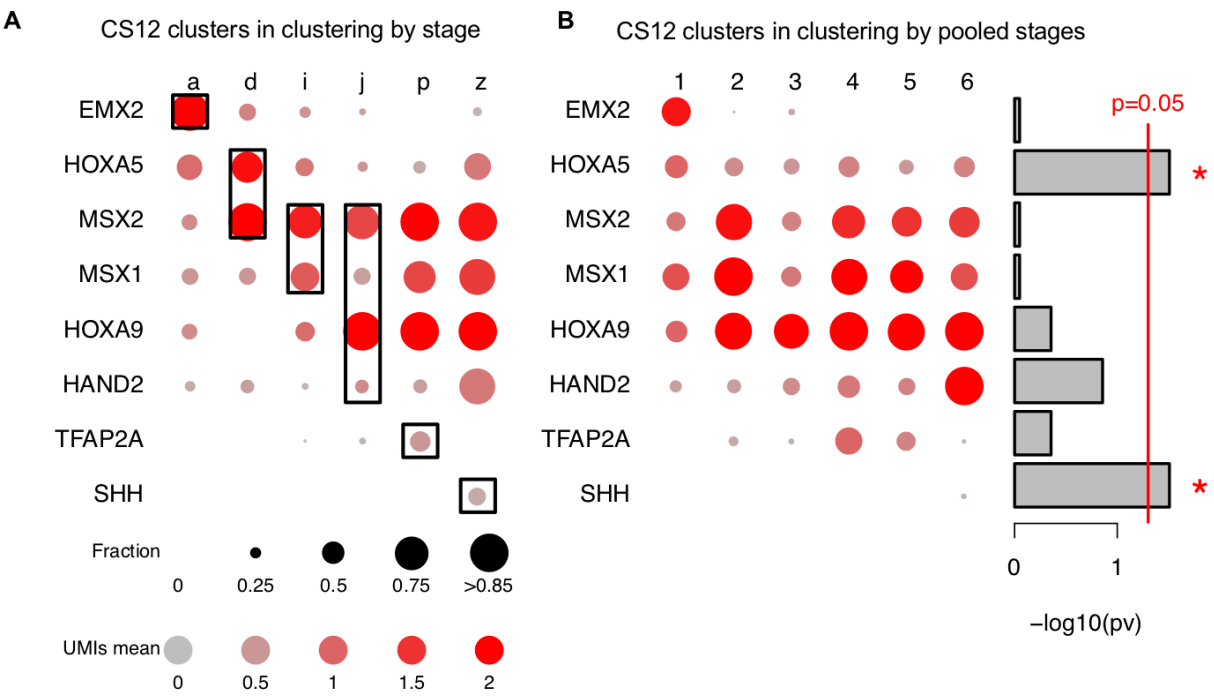
Extended Data Fig. 1. Embryo images and data quality. (**A**), Images of embryos and the number of cells retained for data analysis. Schematic diagrams (lower panel) showing the spatial territory of embryo dissection parts. (**B**), Histograms showing the distribution of transcript numbers (upper) and gene numbers (lower) per cell. (**C**), Distribution of *XIST* (female) versus *RSP4Y1* (male) expression from randomly sampled 1000 cells per embryo. (**D**), UMIs of hemoglobin genes (*HBA1*, *HBA2*, *HBE1*, *HBG1*, *HBG2*, *HBZ*) against total UMIs. Each dot denotes a cell. Cells with high linear slope (above the green line) was identified as red blood cells (red) (see Methods).



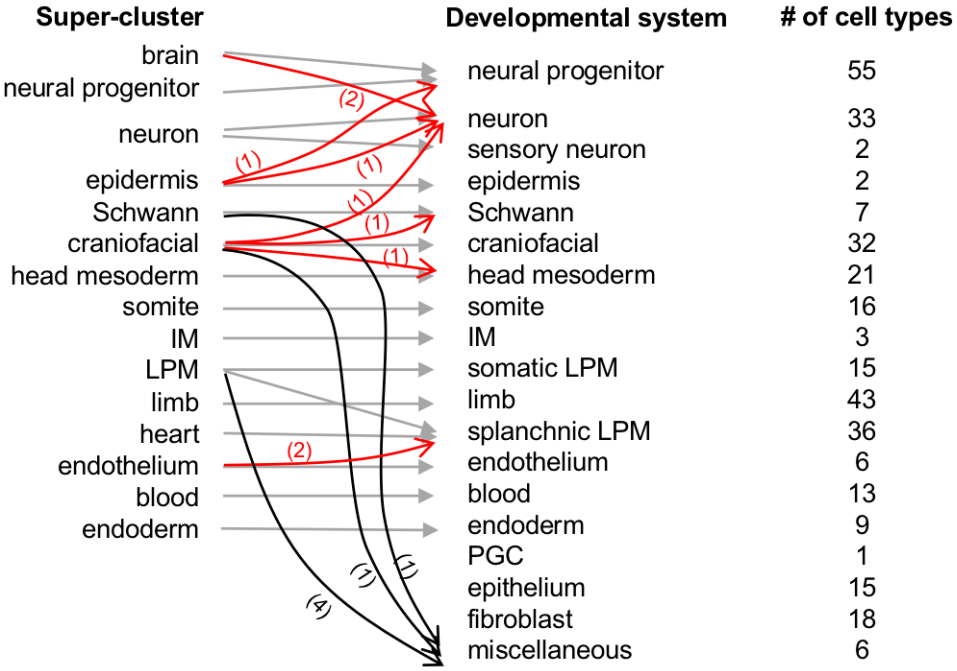
Extended Data Fig. 2. Clustering with TFs in HVGs and with all HVGs. (**A**), The expression pattern of marker genes in super-clusters of clustering with TFs in HVGs. The annotations were showed on the column. (**B**), The expression pattern of marker genes in super-clusters of clustering with all HVGs. Right panel, p values of Kolmogorov-Smirnov test of mean UMIs of marker genes across clusters between two ways of clustering. Red line, p=0.05; box, genes with $p < 0.05$. (**C**), The percentages of cells from clusters by TFs (column) for each cluster by HVGs (row). Boxes indicate cluster 3 and 6-10 by HVGs that do not have clear match to clusters by TFs. (**D**), The percentages of cells from clusters by HVGs (column) for *FOXA2*+ cells of endoderm cluster by TFs (row). (**E**), The stage distribution of cells in mesodermal clusters by TFs (left) and in cluster 6-10 by HVGs (right).



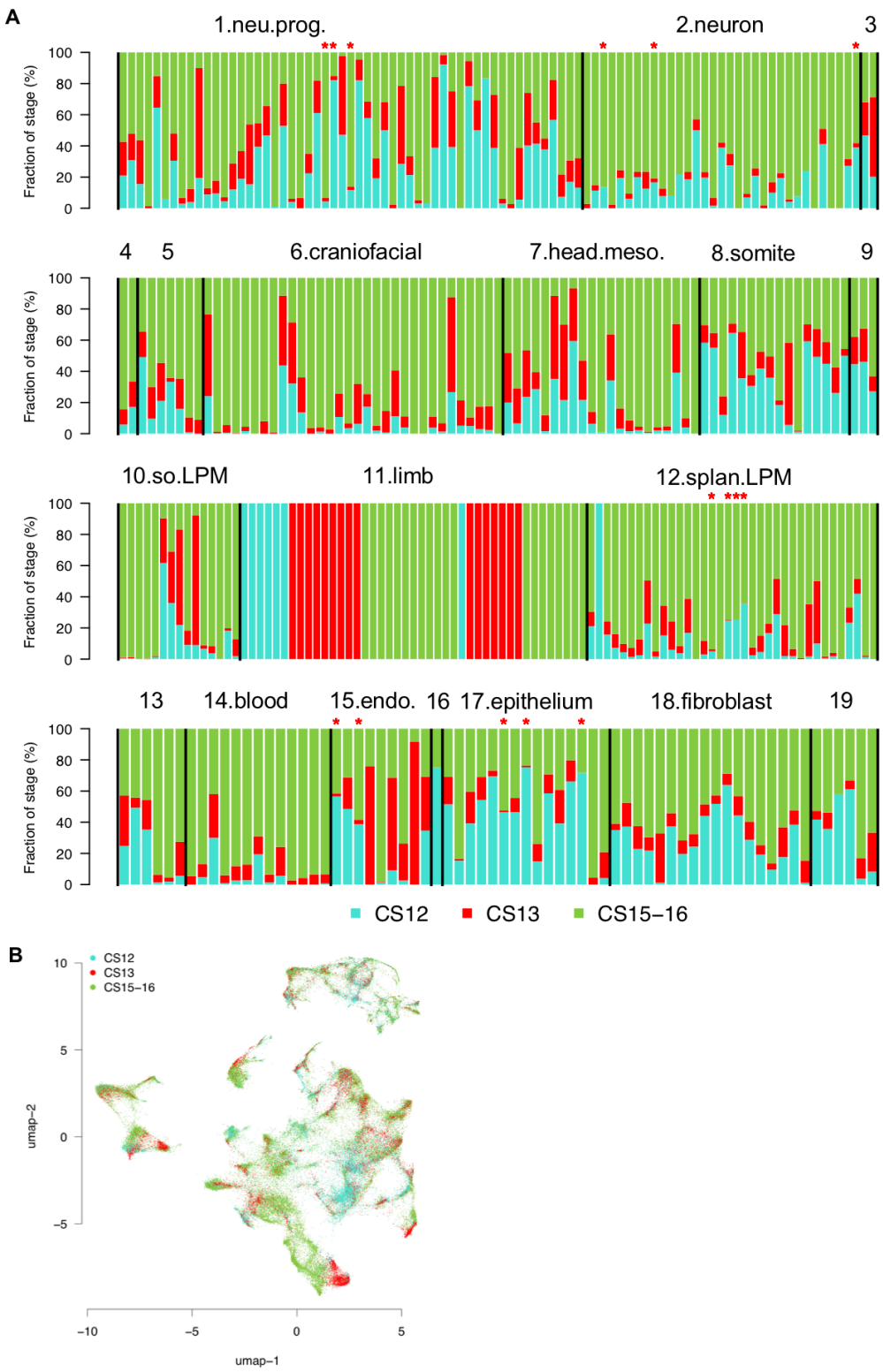
Extended Data Fig. 3. The choice of parameters in iterative clustering. (**A**), The expression of markers relative to boundaries of first round of clustering (solid lines) and second round of clustering (dash lines) in brain. Markers that support the division of second round of clustering were indicated on the top. (**B**), The expression of markers relative to boundaries of one round of clustering and second round of clustering in endothelium. Convention follows panel a. (**C**), The number of clusters resulted from a series of resolution 'r' and PCs in the clustering of spinal neuron. (**D**), The pairwise ARI between clusters resulted from different resolution 'r' in spinal neuron. (**E**), X-axis, the mean ARI of clusters between chosen 'r' (0.5) and a series of 'r' (0.3~0.7, increment 0.1, excluding 0.5) when PCs was fixed to 20. Y-axis, the mean ARI of clusters between chosen PCs (20) and a series of PCs (15~25, increment 1, excluding 20) when r was fixed to 0.5. Each dot denotes a super-cluster.



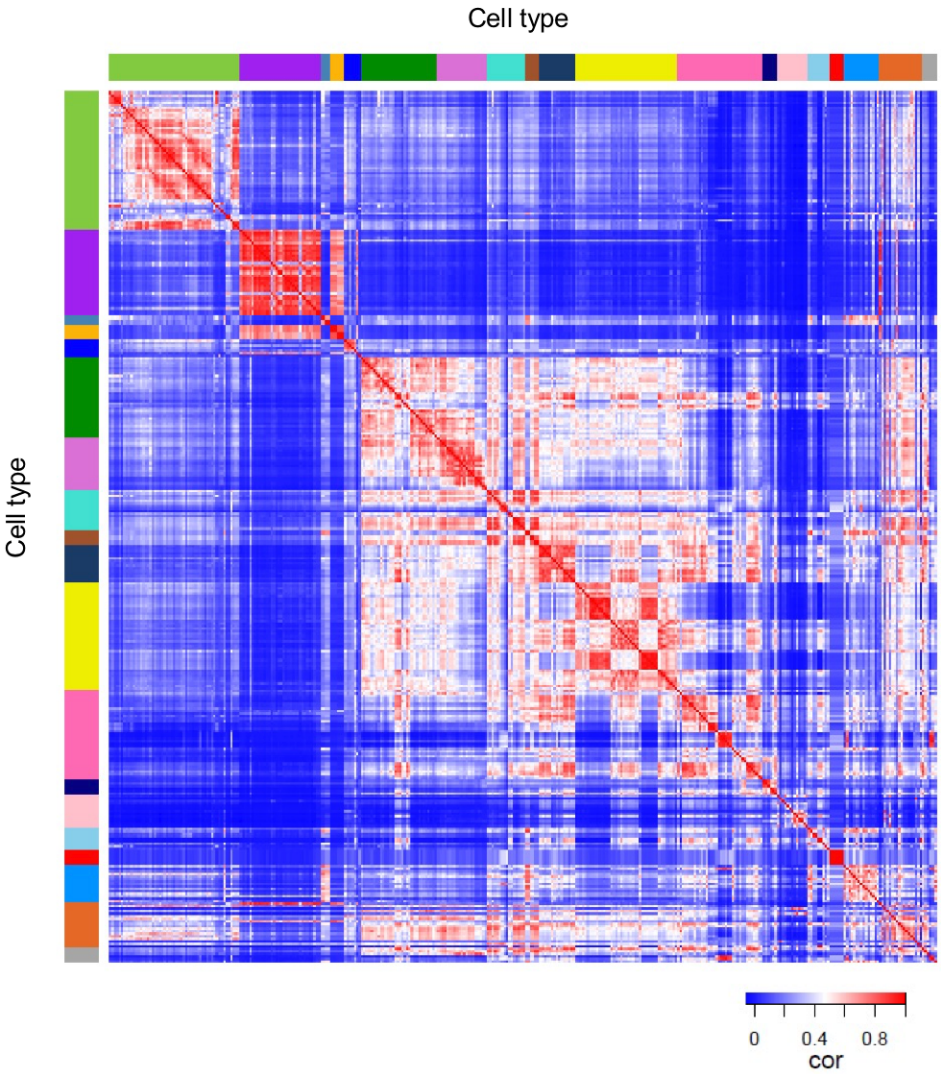
Extended Data Fig. 4. Temporal axis in limb bud. (**A**), The expression pattern of marker genes in clusters resulted from clustering only CS12 cells in limb bud. See Fig. 3 for the spatial map of domains on the columns. (**B**), The expression pattern of marker genes in CS12 clusters resulted from clustering all cells in limb bud. Right panel, p values of Kolmogorov-Smirnov test of mean UMIs of marker genes across clusters between two sets of clustering. Red line, p=0.05.



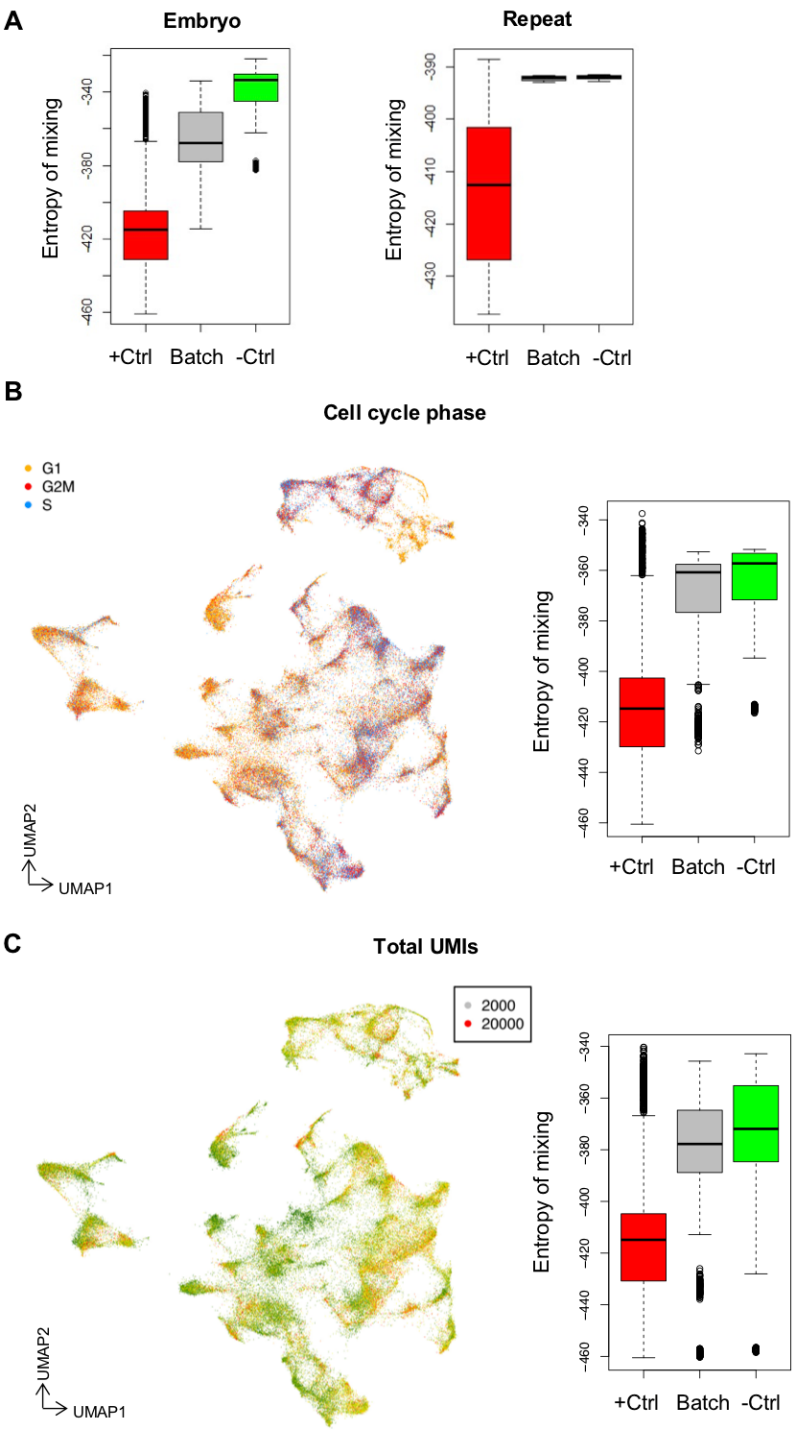
Extended Data Fig. 5. Reorganization of cell types. The map between clusters in super-clusters during clustering and cell types in developmental systems as presented in results. Red arrows denote clusters transferred to a different system according to annotation. Black arrows denote that clusters without clear belonging of developmental system were grouped as “miscellaneous”.



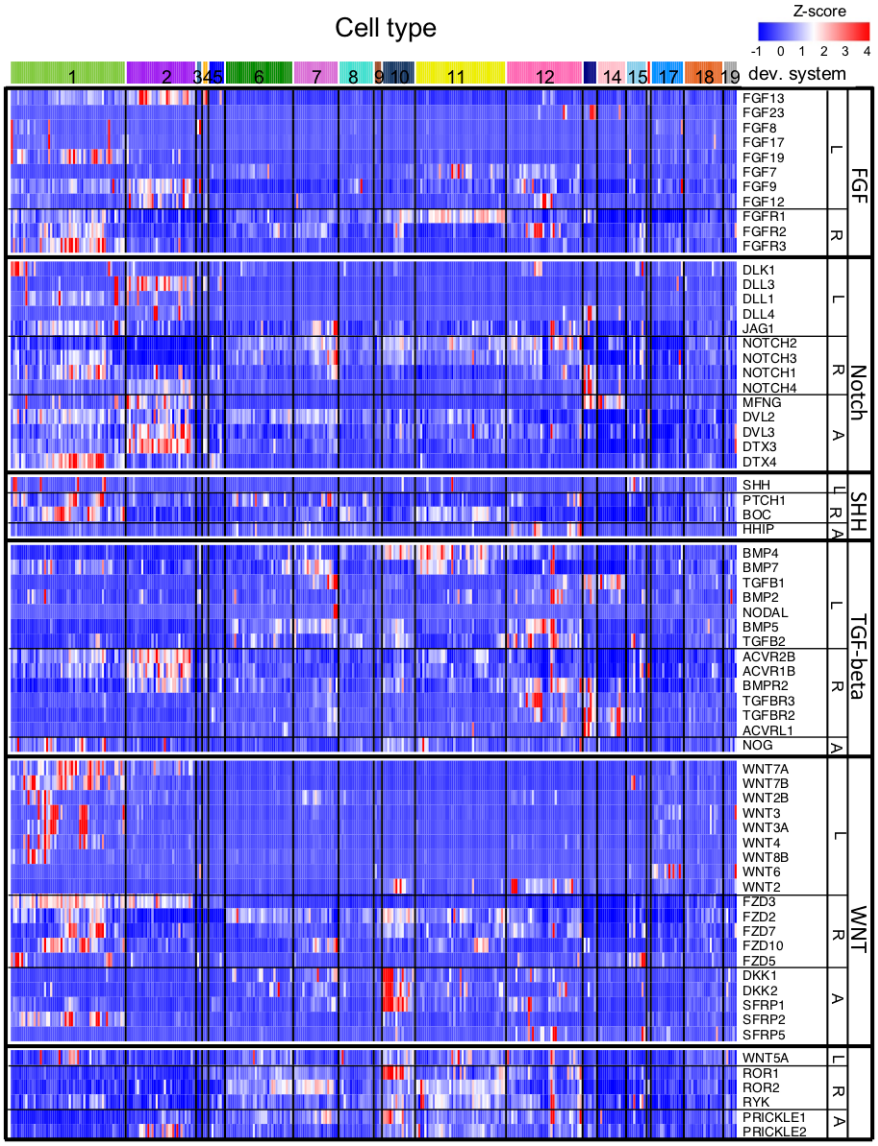
Extended Data Fig. 6. Distribution of embryonic stages in each cell type. (**A**), Stage distribution of cells in each cell type ordered by system. * denotes cell types missing cells from CS13, defined by total number of cells > 50, ratio of cells from CS12 embryo > 0.03, number of cells from CS12 embryo > 5, ratio of cells from any CS15~16 embryo > 0.03, number of cells from any CS15~16 embryo > 5, ratio of cells from CS13 embryo < 0.03, and number of cells from CS13 embryo < 5. Given this definition, the probably of cell types missing cells from CS13 is 0.001 by Poisson distribution. (**B**), The stage of cells on Umap.



Extended Data Fig. 7. Similarity between cell types. Pairwise Pearson's correlation of all 333 cell types based on the expression of DEGs. Color bars denote developmental systems (see Fig. 1 for convention). Cell types are ordered as in Supplementary Table 1C.

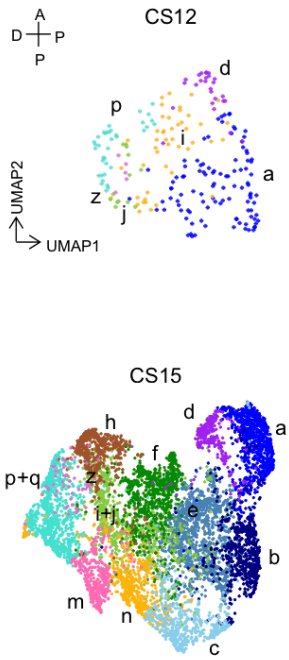


Extended Data Fig. 8. Benchmark of clustering. (**A**), Batch effects across embryos and across technical repeats as estimated by the entropy of mixing[84]. Batch effect is anti-correlated to entropy of mixing. Boxplots showing the entropy of mixing using cluster identities as batch ('+Ctrl', clustering totally driven by batch effect), entropy of mixing by this effect ('Batch'), and entropy of mixing using randomly assigned batch ('-Ctrl', no batch effect). (**B**), Left panel, the estimated cell cycle phases on Umap (Seurat, 'CellCycleScoring'). Right panel, the effect of cell cycle phases on clustering estimated by the entropy of mixing. (**C**), Left panel, total UMIs in each cell on Umap. Right panel, the effect of total UMIs on clustering estimated by the entropy of mixing, where total UMIs were stratified by 4-quantiles.



Extended Data Fig. 9. Expression pattern of signaling pathways. The expression (z-score) of ligands, receptors, and antagonists of Fgf, TGF-β, Hedgehog, Notch and Wnt pathways in all cell types. Color bars denote developmental systems (see Fig. 1 for convention). Cell types are ordered as in Supplementary Table 1C.

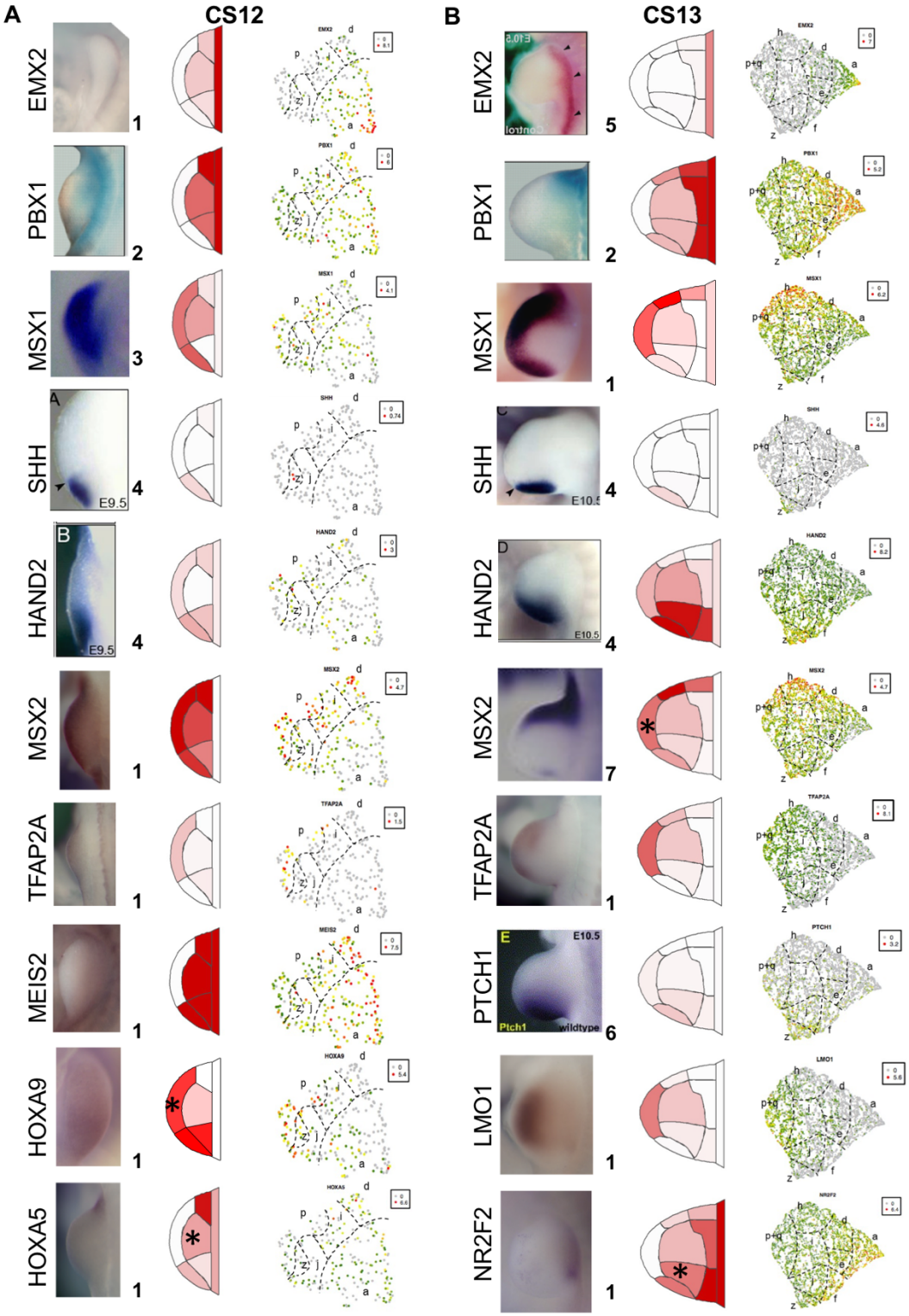
A



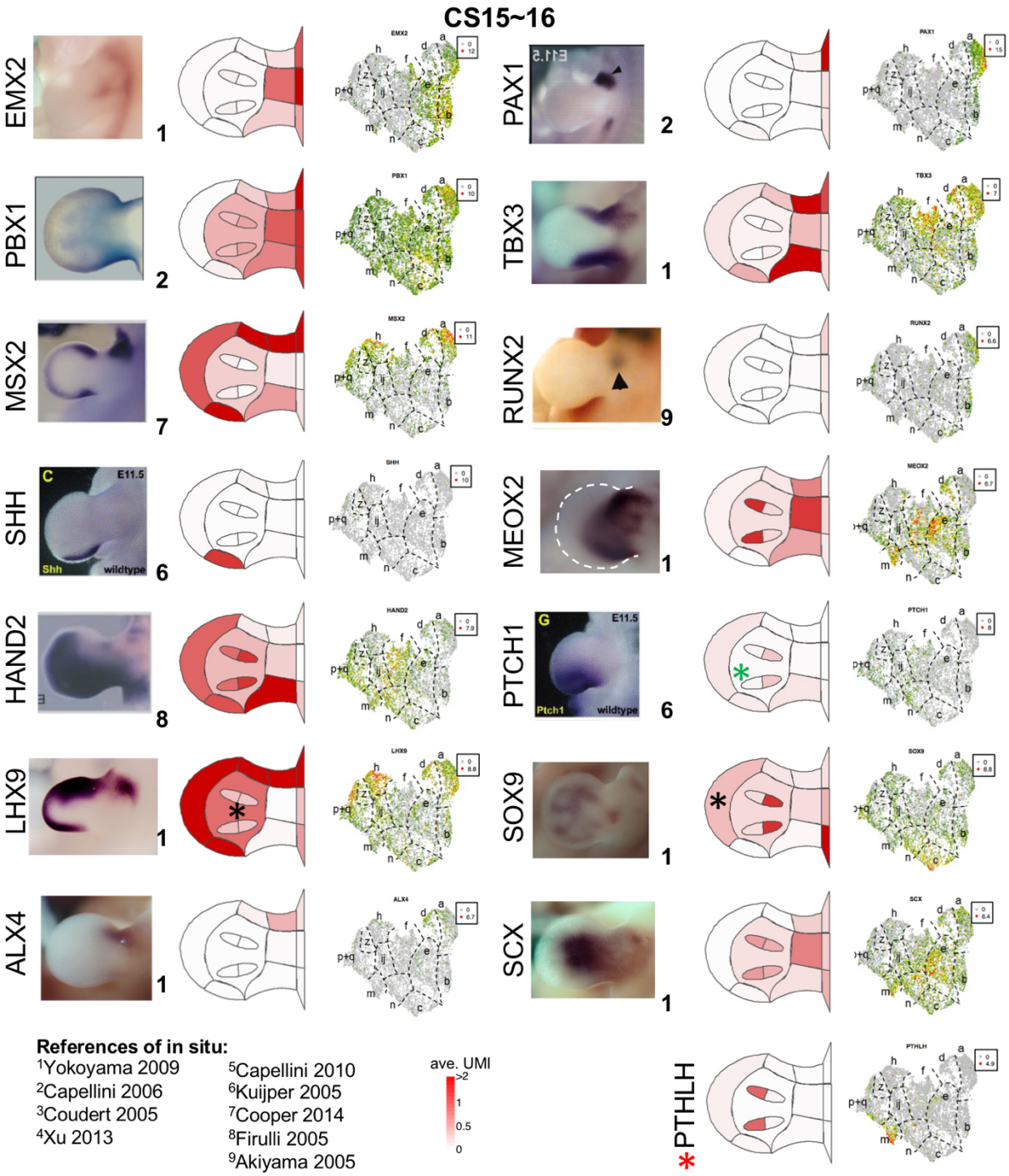
B

Domain	Markers of matched, *in situ* specific, and scRNA-seq specific
CS12-a	EMX2, PBX1, MEIS2
CS12-d	PBX1, MSX2, MEIS2, HOXA5
CS12-i	MSX1, MEIS2, HOXA5*
CS12-j	HAND2, MEIS2, HOXA9
CS12-z	MSX1, HAND2, SHH, MSX2, HOXA9
CS12-p	MSX1, MSX2, TFAP2A, HOXA9*
CS13-a	EMX2, PBX1, NR2F2
CS13-d	PBX1, MSX1, MSX2
CS13-i	-
CS13-j	PTCH1, HAND2, NR2F2*
CS13-z	SHH, MSX1, PTCH1, HAND2
CS13-e	PBX1, NR2F2
CS13-f	HAND2, PBX1, NR2F2
CS13-h	MSX1, MSX2
CS13-p	MSX1, LMO1, TFAP2A
CS13-q	MSX1, LMO1, TFAP2A, MSX2*, HAND2*
CS15-a	LHX9, PAX1, PBX1, MSX2, TBX3, EMX2, RUNX2, SOX9
CS15-d	LHX9, ALX4, PBX1, MSX2, TBX3, MEOX2
CS15-n	SOX9, SCX, HAND2
CS15-ij	HAND2, SCX, SOX9, PTCH1*, LHX9*
CS15-z	SHH, MSX2, LHX9, HAND2, PTCH1
CS15-e	SCX, PBX1, EMX2, MEOX2
CS15-f	PBX1, TBX3, HAND2, MEOX2, PTCH1
CS15-h	LHX9, MSX2
CS15-p	LHX9, MSX2, HAND2, SOX9*
CS15-q	LHX9, MSX2, HAND2, SOX9*, PTCH1*
CS15-b	EMX2, RUNX2, SOX9
CS15-c	EMX2, PAX1, PBX1, TBX3, HAND2, SOX9
CS15-m	HAND2, MEOX2, SCX, SOX9, PTHLH*

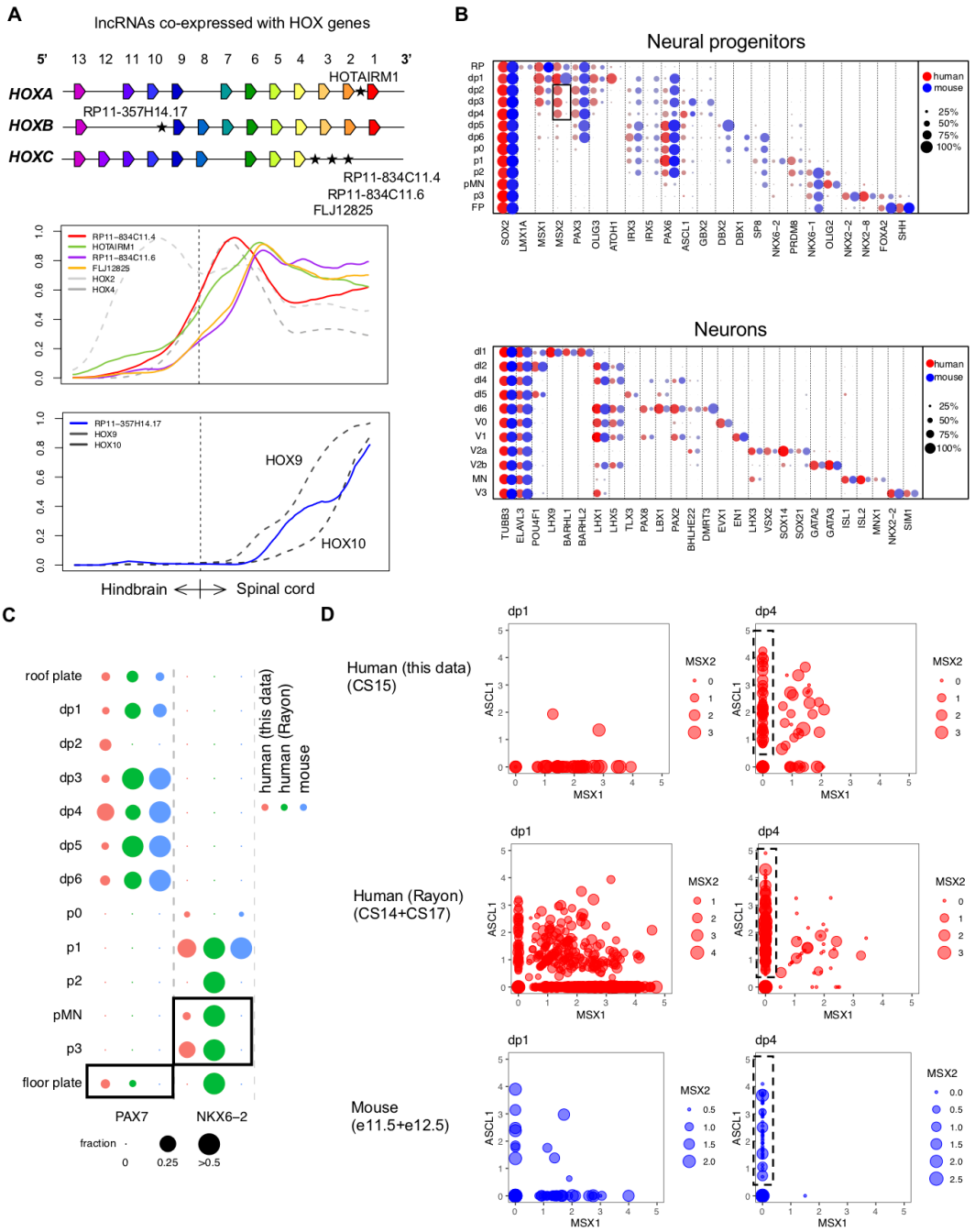
Extended Data Fig. 10. Analysis of spatial domains in limb mesenchymal cells. (**A**), UMAP visualization of forelimb cell types at CS12 and CS15~16. Cells are colored by domains as labeled in Fig. 3. (**B**), Summary of marker genes used for domain annotation. See Supplementary Note 1 for stepwise annotation. The inconsistency between *in situ* and scRNA-seq were indicated in blue (*in situ* specific) and orange (scRNA-seq specific).



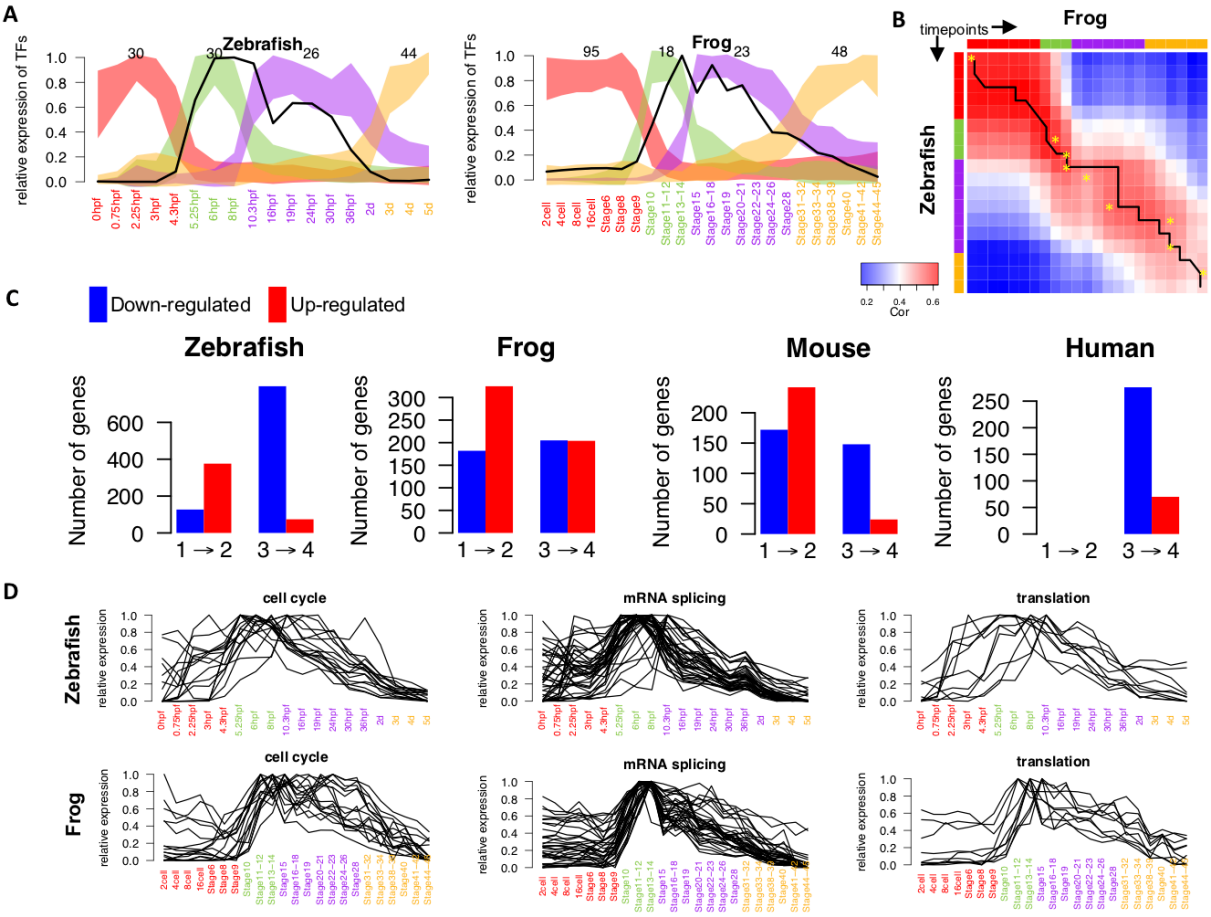
Extended Data Fig. 11. *In situ* evidence of spatial domains in limb mesenchymal cells. The comparison of marker genes in *in situ* and scRNA-seq at CS12 (**A**) and CS13 (**B**). *In situ* results were collected from mouse studies at the corresponding stage (numbers indicate references, see Extended Data Fig. 12). The gene expression in spatial domains in scRNA-seq were showed in schematic diagram colored by mean UMIs (2nd column). The gene expression in each cell in scRNA-seq were showed on Umap (3rd column). Asterisks denote domains that display inconsistent expression between *in situ* and scRNA-seq results (black asterisk, only detected in scRNA-seq; green asterisk, only detected by *in situ*).



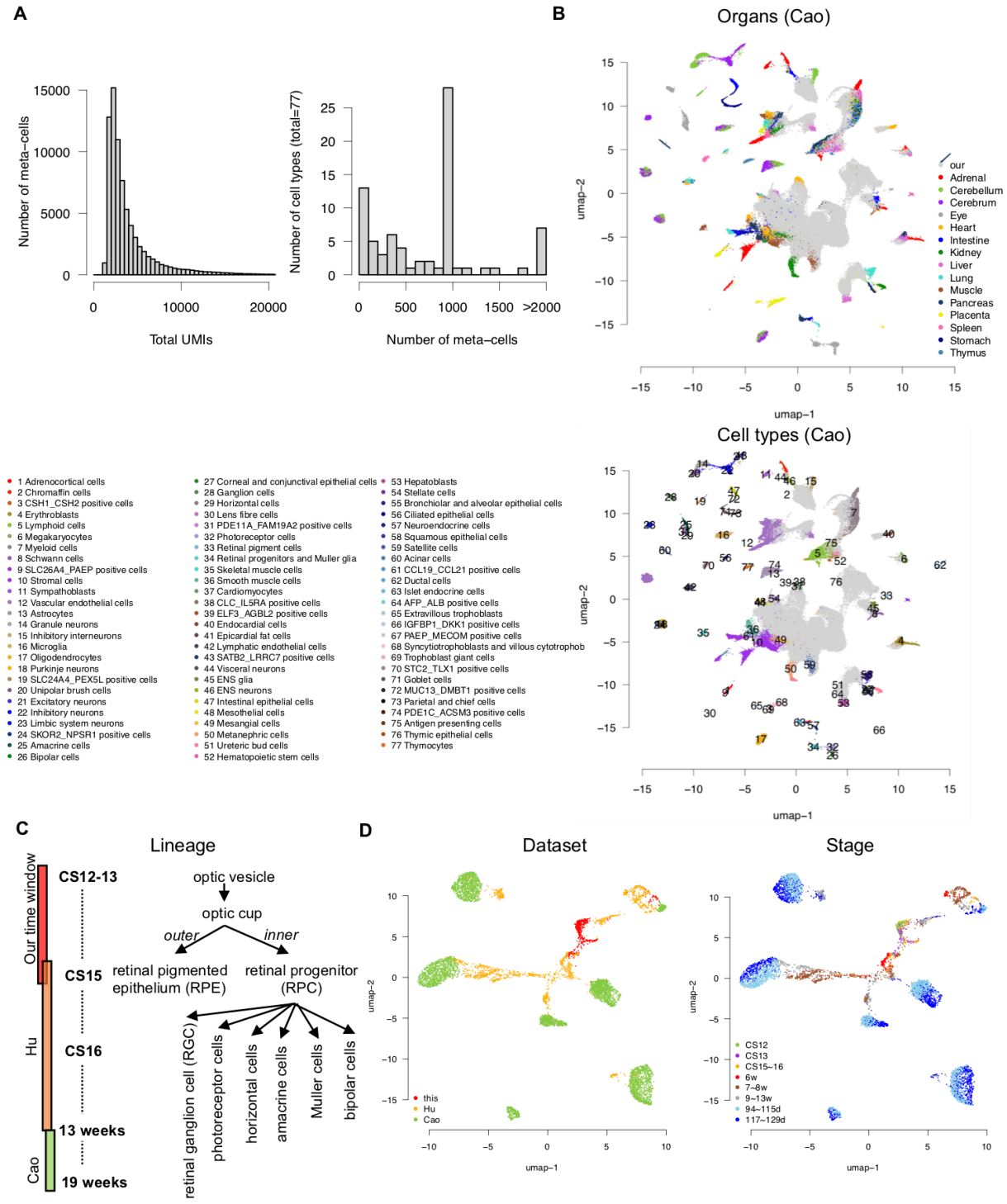
Extended Data Fig. 12. *In situ* evidence of spatial domains in limb mesenchymal cells. The comparison of marker genes in *in situ* and scRNA-seq at CS15~16. See Extended Data Fig. 11 for convention. No *in situ* data is available for *PTHLH* at this stage of limb. For the detail of references, please see reference section.



Extended Data Fig. 13. Analysis of neural tube cell types. (**A**), Upper panel, schematic diagram of the genomic location of the 5 lncRNAs identified as AP related genes in neural tube. Lower panel, expression pattern of lncRNAs and accompanying HOX genes along the anterior-posterior axis. (**B**), The comparison of expression of canonical markers for cell types in neural progenitors (upper panel) and neurons (lower panel) between human and mouse. Box indicates the difference of *MSX2* between human and mouse in neural progenitors. (**C**), The comparison of expression of *PAX7* and *NKX6-2* between human and mouse in neural progenitors. Boxes indicate human specific expression that are consistent between two datasets of human. (**D**), The expression of markers (dp1 marker *MSX1* and dp4 marker *ASCL1*, log2 scaled) that distinguish dp1 and dp4 cells in individual cells in human (our dataset and Rayon's dataset) and mouse datasets. Dot size shows *MSX2* expression. Dash boxes indicate most confident dp4 cells (*ASCL1*>0 and *MSX1*=0), which are showed in Fig. 4E.



Extended Data Fig. 14. LIN28A in vertebrate embryogenesis. (**A**), Expression pattern of 4 groups of TFs in zebrafish and frog. Each group of TFs were identified as TFs that are highly expressed in the corresponding stage. Numbers above curves indicate the number of TFs in each group. The width of each curve represents 2 standard deviations within each group of TFs. The black lines show the expression of *LIN28A* in each species. (**B**), Pairwise correlation of timepoints between zebrafish and frog by homologous TFs. Black line indicates the alignment of timepoints between zebrafish and frog by dynamic time warping. The yellow asterisks indicate the match of timepoints from a previous study[85]. (**C**), Numbers of systemically up- (red) and down-regulated (blue) genes from stage 1 to stage 2 (labeled as 1→2) and from stage 3 to stage 4 (labeled as 3→4). No transcriptome data is available for stages 1 and 2 in human. (**D**), Expression dynamics of genes in cell cycle, mRNA splicing and translation pathways that are positively correlated to *LIN28A* in zebrafish and frog.



Extended Data Fig. 15. Data integration in human embryos. (**A**), Left panel, the distribution of total UMIs in meta-cells in Cao's data. Right panel, the distribution of numbers of meta-cells in each cell type in Cao's data. See Methods for the definition of meta-cell. (**B**), The joint Umap of our dataset (grey) and Cao's dataset colored by organ source and by cell type in Cao's dataset. (**C**), Lineage development of human eye and time windows of three studies in integration. (**D**), Joint Umap of cell types of eye in three studies colored by dataset and stage.

Supplementary tables and information

Supplementary Table 1. (separate file)
Cell types and DEGs.

Supplementary Table 2. (separate file)
LIN28A in vertebrate embryogenesis.

Supplementary Table 3. (separate file)
Systemically changing genes in vertebrate embryogenesis.

Supplementary Table 4. (separate file)
Enriched pathways in systemically changing genes.

Supplementary Table 5. (separate file)
Data integration.

Supplementary Note 1. (separate file)
Annotation vignette.